8-23-2022 1:00 PM

Capturing Within Host HIV-1 Evolution Dynamics Using Simulation Methods

Emmanuel Wong, *The University of Western Ontario*

Supervisor: Poon, Art F.Y., *The University of Western Ontario*
A thesis submitted in partial fulfillment of the requirements for the Master of Science degree in Pathology and Laboratory Medicine

Recommended Citation

Wong, Emmanuel, "Capturing Within Host HIV-1 Evolution Dynamics Using Simulation Methods" (2022). *Electronic Thesis and Dissertation Repository*. 8901.
https://ir.lib.uwo.ca/etd/8901

Abstract

The persistent latent reservoir of long-lived cells carrying integrated HIV DNA is the source of reinfection upon treatment interruption, and a primary focus for cure research. The reservoir is difficult to study because these cells are relatively rare or located in tissues that are difficult to sample. Sequencing proviral DNA in the latent reservoir is an important source of information about reservoir establishment and persistence, especially from the presence of identical (clonal) sequences. I evaluated the relationship between select measures of these clonal sequences and drivers of reservoir persistence, e.g., clonal expansion, by implementing a simulation model of within-host HIV dynamics in actively and latently infected cells. I implemented a discrete event simulation in the R package treeswithintrees, with four populations of cells corresponding to active, latent, replenishment and death compartments. To simulate molecular evolution on the resulting trees, I collapsed branches representing infected cells in a latent state and ran the program INDELible with parameters calibrated to HIV-1 on a representative *env* sequence. I propose a new clonality statistic (pairwise clonality) that can capture the genetic diversity of a sample with less information loss. I then evaluated the response of two clonality statistics used in literature (the proportion of identical sequences, Gini coefficient) and my proposed clonality statistic (number of identical pairwise comparisons) to changes in simulation parameter values by fitting a General linear Model (GLM). I found that the former clonality statistics were not as robust as the proposed pairwise clonality score. In addition, there were significant associations between clonality statistics and simulation parameters. Finally, I implemented a particle filtering method to evaluate non-linear relationships between simulation parameters and the clonality scores.

Keywords: HIV (Human Immuno-deficiency Virus), Latent Reservoir, Within-Host Evolution, Viral Evolution.

Lay Abstract

Human Immunodeficiency Virus (HIV) is a disease with significant social and economic burden without a cure – it is currently manageable through combination Antiretroviral Therapy (cART). However cART is unable to remove viruses that have inserted themselves into the DNA of host cells (the latent reservoir). The latent reservoir is relatively stable and long lasting and is a source of re-infection if cART is halted, or otherwise made ineffective. The latent reservoir is difficult to study in humans for various reasons. Therefore, simulations can be helpful to study the evolution & dynamics of HIV within a host. I created a simulation framework "simclone" that can simulate within-host evolution of HIV. Events of evolutionary significance during the course of an HIV infection were simulated: acute infection, cART initiation and chronic infection. Sequence evolution is conducted on the outputs of HIV infection. Clonality statistics (i.e. how many sequences in a sample are identical) are used to understand the latent reservoir, but the commonly used statistics (proportional clonality, GINI coefficient) have significant drawbacks. I propose the "pairwise clonality" statistic as a response to this issue. I then analyze the simulation model to see if any parameter values (e.g. infection rate) are associated with increased clonality scores for each of the metrics. I then implement a particle filtering method to attempt simulation parameter estimation based on clonality score. The results shown in this work i) provide a framework for simulating within-host evolution that enables specific hypotheses testing ii) proposes a clonality statistic that has better statistical properties than existing measures iii) highlights the difficulties in using particle filtering for parameter estimation.

Acknowlegements

First, and foremost I would like to begin by thanking my parents. Thank you for always being present for me. Your unconditional encouragement, support and love was always deeply appreciated (even if I didn't show it from time to time) through my journey to this point. Without your presence and the opportunities you have given me, this work, and who I have become would not have been possible. Mum: thanks for showing me how to be resilient and how to pursue my goals with a passion.

I am grateful for the support and guidance of my supervisor Dr. Art Poon. Thank you for being patient with me from the very first day I joined the lab as an undergraduate student, when I managed to overwrite my dataset within minutes. Your guidance, insight and knowledge helped a lot in my research and your suggestions and critical comments improved my thesis greatly.

Erin, I could not have completed this thesis without your unreserved support and motivation. You were, and are, a constant source of motivation, positive energy and encouragement. Thank you for helping me believe in myself even when sometimes I would have doubts. Your presence and insights were invaluable, I am deeply grateful for your company, tea chats, and most importantly the laughs.

To my best friends: Jake B. and John T., thank you for all the support you have given me throughout the years. You both have celebrated with me in the best of times, and consoled me in the worst. Thanks for the late night chats under the stars and dance parties. I would not be where I am today without you two being there for me through thick and thin.

My lab mates, Roux-cil: thank you for the advice and guidance through this process – I will miss our weekly update zoom calls and distracting you in lab. Laura thank you for your friendship and support - I will miss our climbing sessions at Junction. And to the rest of the Poon lab, past and present: thank you for all your help!

And finally to those I didn’t mention by name, but were still supported me through this process: thank you.

Contents

List of Figures

List of Tables

List of Appendices

Chapter 1

Introduction

1.1 Natural History of HIV Infection

Human **I**mmunodeficiency **V**irus Type I (HIV) is a *retrovirus* that has a significant global burden of disease. Retroviruses are viruses in the family *Retroviridae*, of which the defining feature is obligate reverse transcription of the viral RNA genome into DNA, and insertion of the viral DNA into the host genome (Coffin, 1995; Telesnitsky, 2010). HIV is grouped with other immunodeficiency viruses (e.g. Simian Immunodeficiency Virus, SIV) in the Lentivirus virus genus, under the subfamily Orthoretrovirinae (Luciw, 1996).

An estimated 38,000,000 [95% CI: 31,600,000 - 44,500 000] individuals are living with HIV, with a estimated yearly death toll of 690,000 [95%: CI 500,000 - 970,000] as of 2020 (UNAIDS, 2021). An estimated 1.5 million [95% CI: 1.0 - 2.0 million] individuals are newly infected each year. It is a disease with severe economic and individual impact. While HIV-related deaths and new HIV infections have declined yearly since the early 2000's (UNAIDS, 2021) due to advancements in treatments and increased funding for HIV related programs, there is currently no practical cure for HIV. Three individuals have been cured of HIV, however these therapies were extreme in nature and are not practical for widespread use. The patients that have been cured of HIV had cancers that required a bone marrow transplant, a procedure that confers significant risk. The patients were transplanted with stem cells from a donor with the CCR5Δ32/Δ32 mutation (Gupta et al., 2020; Hsu et al., 2022; Hütter et al., 2009). This mutation confers protection on the patient by preventing HIV binding to the CCR5 co-receptor (Ni et al., 2018).

Uncontrolled, progression of HIV infection to **A**cquired **I**mmuno**d**eficiency **S**yndrome (AIDS) can take months to years (T. Schacker, 1996; T. W. Schacker, 1998; UNAIDS, 2021). AIDS is clinically defined as a 200 CD4+ cells per μL or by occurrence of HIV specific associated disease (Mandell et al., 1995). There is wide variation in the course of HIV infection, the variation is generally understood to be caused by a combination of host (Fauci, 2008; Goff, 2007) and viral characteristics (Li et al., 1999, Naif et al., 1999, Cavarelli and Scarlatti, 2009). Patients who progress to AIDS are at a higher risk for opportunistic infections caused by bacteria, viruses, fungi, and parasites due to the reduced CD4+ T-cell levels (Douek et al., 2003; C. B. Holmes et al., 2003).

ART

Anti-**R**etroviral **T**herapy (ART) for HIV has evolved significantly from its original form. The earliest class of treatments for managing active HIV infection were **N**ucleoside **R**everse **T**ranscriptase **I**nhibitors (NRTI). Both sequential and concurrent application of NRTI class drugs resulted in the rise of drug resistant genotypes (Schmit et al., 1996). **C**ombination **A**nti-**R**etroviral **T**herapy (cART), also known as **H**ighly **A**ctive **A**nti-**R**etroviral **T**herapy (HAART) is the modern standard for ART. It combines anti-retroviral agents of multiple drug classes to reduce the risk of drug resistance, which typically results in failure of the treatment (Collier et al., 1996; D'Aquila, 1996; Staszewski et al., 1996). cART targets at least two distinct molecular targets in the HIV replication cycle. While statistically likely that resistance to at least one agent in the cART 'cocktail' will occur, multiple drug resistance is unlikely if the therapy regime is adhered to (Coffin, 1995; Frost and McLean, 1994; Stengel, 2008). It is statistically unlikely that drug resistance mutations would occur at multiple sites concurrently. The typical cART combination anti-retroviral agents can be divided by function into six groups: nucleoside-analog reverse transcriptase inhibitors (NARTIs), non–nucleoside reverse transcriptase inhibitors (NNRTIs), integrase inhibitors, protease inhibitors (PIs), fusion inhibitors, and co-receptor antagonists (Arts and Hazuda, 2012). It is important to note ART is not a cure, as the treatment does not excise integrated HIV sequences in the host genome (Arts and Hazuda, 2012). The significance of these integrated HIV provirus sequences will be discussed further in detail.

Viral Load and HIV Viremia

Viral Load is a density measure of HIV in peripheral blood (HIV copies/mL) (Fraser et al., 2007). It can vary between patients, and also within patients over time (Mellors et al., 1996). After the initial infection, there is a phase of exponential viral replication. At this stage, individuals are likely asymptomatic, but are the most infectious. Over the next several months, the amount of circulating RNA HIV slowly decays to the set point load. The set point value is the approximate viral load level of an asymptomatic patient. It can differ significantly between patients, with up to a 1000 fold difference (de Wolf et al., 1997; Mellors et al., 1996). HIV viral loads are highly predictive of disease prognosis - higher viral loads are associated with quicker disease progression; and of transmission direction between transmission pairs - higher viral loads are typically found in the originating individual of a tranmission pair (de Wolf et al., 1997; Mellors et al., 1996).

Viremia, the presence of circulating virus in the blood stream, is present soon after primary infection. For most individuals on ART without drug failure, HIV viral loads decrease significantly, below detectable levels (20 - 40 copies/mL). These low levels provide clinical benefits to the individual (Palella et al., 1998). HIV proviruses are HIV DNA viral sequences that are integrated into host genomes. Despite controlling the replication of HIV, there exists a set of infected cells that contain provirus that can produce replication competent HIV (Finzi et al., 1997a; Siliciano et al., 2003; Wong et al., 1997,Bruner et al., 2016; Y.-C. Ho et al., 2013).

Transient viremia with plasma viral load of up to 500 copies/mL or “viral blips” can occur even when patients adhere to ART (Young et al., 2015). Viral blips are specific situations of viremia - a sharp, transient increase in virus concentration followed by a decrease to undetectable levels. Sustained plasma viral loads above this threshold are interpreted as drug failure events (Garcia-Gasco et al., 2008). There are many possibilities for the cause of these blips: poor drug adherence or absorption; or conditions that lead to increased replication (infection, or vaccinations) (Günthard et al., 2000; L. Jones and Perelson, 2005; Nettles, 2005). Ramifications of viral blips are not yet clear. Investigations into the relationship between viral blips and subsequent drug therapy failure are mixed (Günthard et al., 2000; L. Jones and Perelson, 2005; Young et al., 2015). Viral blips could be due to integrated proviral sequences in latent (memory) CD4+ cells. Reactivation is the basis for small spikes of low level viremia and "viral blips" (Chun et al., 1997; Finzi et al., 1997b; Siliciano et al., 2003). Low level viremia is not clinically significant (i.e. no significant difference in disease progression or drug efficacy) for individuals remaining on ART (Nettles et al., 2005).

Disease Progression

The rate of HIV disease progression after a transmission event can differ between individuals. Speed of progression is governed by host factors, viral factors, and stochastic processes that are not well understood (Cavarelli and Scarlatti, 2009; Fauci, 2008; Li et al., 1999; Naif et al., 1999. There are three distinct stages of HIV infection, however individuals may progress through these stages at different speeds. The acute infection is the first clinical stage of HIV infection. This stage is charactertized by a CD4+ cell count per μL ≥ 500. Individuals may present with viral symptoms in this stage such as fever and rash (Moir et al., 2011). Typical CD4+ cell counts range from 500-1200 per μL. As a patient's infection progresses, viral load increases and CD4+ cell counts further decrease due to the cytotoxic effects of HIV replication (Davey et al., 1999; Fauci, 2008; T. Schacker, 1996). This chronic stage is characterized by a CD4+ cell count between 200-499 per μL, and can last for months or years. In the second stage, the concentration of CD4+ lymphocytes declines. The third clinical stage is characterized by high viral loads, a CD4+ cell count lower than 200 per µL or opportunistic infections (AIDS) (C. B. Holmes et al., 2003). Individuals with HIV do not necessarily progress to AIDS. Individuals that have at least 12 years of asymptomatic infection, are termed Long Term Non-Progressors (LTNP). Although the end result is the same - long term asymptomatic infection - the basis for asymptomatic disease control differs. Long term control of HIV infection has been associated with strong neutralizing antibody response and cell-mediated immunity to HIV (Grabar et al., 2009; Zanussi et al., 1996). Mutations to HIV that result in a weakly replicating strain can also result in asymptomatic infection (Deacon et al., 1995; Iversen et al., 1995). Weakly replicating strains can still infect large number of cells in a host, however the viral load is much lower due to ”attenuated productive viral infection” (Iversen et al., 1995). Although the virus is under control for a period of time, LTNP’s may eventually progress to symptomatic infection. For most infected with HIV, HIV-mediated CD4+ T-cell depletion occurs early in an infection. Immunodeficiency does not present immediately, as there is regeneration of the CD4+ T-cells resulting in an equilibrium. If the virus is left uncontrolled, this equilibrium deteriorates over time (A. A. Okoye and Picker, 2013).

1.2 Infection Cycle

The HIV infection cycle can be grouped into six stages. These are: entry; DNA synthesis; integration; transcription; RNA modification, transportation and protein synthesis; and, assembly and budding. The DNA synthesis and integration stages of the infection cycle in particular are of evolutionary and clinical significance: the DNA synthesis stage is responsible for the vast majority of mutations in HIV, and thus immune evasion (Coffin, 2002; Hu and Hughes, 2012; Takeuchi et al., 1988); and the integration stage is responsible for long term HIV infection (Z. Wang et al., 2018).

1. Entry (Receptor Recognition and Binding)

Host cell entry is the requisite first step of the replication cycle. Entry is mediated by the interaction between the HIV *env* (envelope) spike protein with a host CD4 receptor (Bosque et al., 2011; Chan et al., 1997; Douek et al., 2002; Kowalski et al., 1987). The HIV *env* spike protein constituent proteins are GP120 and GP41 glycoproteins. The GP120 protein mediates binding between HIV virions and CD4+ T-cells. Conformational changes in GP120 after the binding between CD4 and GP120 allow for subsequent co-receptor (CCR5 or CXCR4) binding (Brower et al., 2009). The CD4 receptor is necessary but not sufficient for HIV entry into a cell: at least one of the aforementioned co-receptors must be present (Brower et al., 2009; Li et al., 1999). Co-receptor binding causes surface protein re-arrangement, drawing the host membrane and viral membrane in close proximity and thus allowing the formation of a membrane fusion pore (Chan et al., 1997). The viral pore provides an avenue for deposition of virion components into the cytoplasm (Didigu and Doms, 2012). Importance of *env* proteins and the CD4 host receptor to viral entry is a motivating factor for vaccine research focusing on the interactions between these two proteins (Arts and Hazuda, 2012).

2. DNA Synthesis

RNA from the HIV virion is in the positive single strand form. *Reverse Transcriptase* (RT) is responsible for converting RNA to double stranded DNA (Hu and Hughes, 2012). RT is encoded within the HIV genome. RT starts by creating an RNA-DNA hybrid double strand using the RNA virion genome as a guide (Coffin, 2002). RNase subsequently degrades the RNA genome, leaving the single negative sense DNA (Kudesia and Wreghitt, 2009; Sarafianos et al., 2009). RT then swaps strands and uses the single negative sense DNA as a template to fill out the positive DNA strand. RT jumping between strands allows for high recombination frequency (Carr et al., 1996; Hu and Hughes, 2012; Kudesia and Wreghitt, 2009). The low fidelity of RT and the lack of proof-reading mechanism is the molecular driver for the genetic diversity found within HIV. RT is an important target for antiviral compounds because of its' importance in the life cycle (Arts and Hazuda, 2012; Coffin, 1995; Frost and McLean, 1994).

3. Integration

In the integration step, proviral DNA is inserted into cellular DNA. This is an obligate step. After HIV RNA is reverse transcribed into DNA, it is bound with proteins to create a pre-integration complex (Coffin, 1995; Craigie and Bushman, 2012). This complex contains integrase, cellular DNA and other proteins necessary for successful integration. HIV integration is similar to other transposition systems found in other pathogens such as bacterium (Kulkosky et al., 1992). The integration step is permanent – to date there are no successful therapies that can excise the integrated viruses from host genomes. Current drug therapies for HIV are not curative, rather, they focus on restricting or reducing the ability for HIV to actively replicate. Long-lived T-cells with integrated HIV proviral sequences is the genetic basis for long term infection (Chavez et al., 2015; Z. Wang et al., 2018). Integrated HIV in long-lived memory T-cells can be transcribed and result in release of active virions months to years after integration (Grabar et al., 2009; Zanussi et al., 1996). Cure efforts are focused on addressing this integrated virus with differing approaches such as: (i) complete replacement of cells with integrated virus; (ii) excision using gene editing technology such as CRISPR; (iii) "shock and kill", which activates cells with integrated virus to allow targeted immune response; and (iv) "block and lock", a technique where transcription of integrated HIV provirus is blocked via epigenetic modifications (Ahlenstiel et al., 2020). The importance of integrase in the HIV life cycle and the role of integrase in creating long-lived sources of HIV re-infection, makes integrase a useful target for inhibitors (Arts and Hazuda, 2012)).

4. Transcription

The HIV genome contains sites that resemble transcription, activator and enhancer regions. For example, the long terminal repeats in the HIV genome contain enhancers that bind to transcription factors only found in active CD4+ T-cells. When these CD4+ T-cells transition into memory cells, transcription factors (i.e. NF-KB, Sp1 and the TATA-box binding protein (TBP)) are no expressed Dutilleul et al., 2020. Without these factors, the expression of HIV proviruses is halted, resulting in latent (memory) cells that still contain integrated virus. Transcription and translation of the HIV DNA hijacks normal host cellular machinery (i.e. RNA polymerase II for reverse-transcription) (Colin and Van Lint, 2009).

5. RNA modification, Transportation and Protein Synthesis

HIV RNA is necessarily modified before protein synthesis. First the virus requires a 5' cap and poly-A-tail (Freed, 2001). Next, as with other RNA viruses, the RNA needs to be spliced to synthesize proteins, however it needs to remain intact to be packaged into virions (Ferguson et al., 2002; Freed, 2001). Thus there is a phased expression pattern of HIV patterns with feedback patterns controlling the expression levels of HIV (Ferguson et al., 2002). Initial proteins govern RNA transport (R) and transactivation (tat). Other (later) proteins are silenced through splicing (Tomezsko et al., 2020; S. Wang et al., 2022). One of the first proteins to be synthesized, *tat*, increases production of other early proteins. Transport of unspliced HIV RNA out of the nucleus follows *rev* binding to *env* (Pollard and Malim, 1998). It is also a negative feedback regulator, *rev* down-regulates the production of early genes that encourage transcription of HIV mRNA and simultaneously up-regulates the transcription of late proteins necessary for the continuation of the HIV life cycle) Fisher et al., 1986; Pollard and Malim, 1998. Late proteins include cell surface proteins (*env*) and caspid proteins (Ferguson et al., 2002; Freed, 2001).

6. Assembly and Budding

The final stage of the HIV lifecycle is the assembly of all constituent parts of the virion in the cytoplasm of a host cell. Virion assembly occurs through the random simultaneous joining of all necessary components – viral proteins, membranes and RNA (Ferguson et al., 2002; Freed, 2001). Assembly of the virion is driven by protein interactions with the viral genome. The process of budding is initiated after assembly (Sundquist and Krausslich, 2012). HIV is an membrane enveloped protein, and as such incorporates host membrane into virion particles (Go et al., 2017; Zanetti et al., 2006; Zhu et al., 2006). Cellular proteins are also randomly included, based upon cell membrane proteins located at the point of budding. The high amount of host proteins incorporated into the virion, especially on the surface provides some degree of immune evasion (Aloia et al., 1993; Ott, 2008). Immune cells have difficulty recognizing virions with endogenous surface host proteins. The immature HIV virion is released after the closure of the bud. The virion matures after release when the protease protein is expressed, and cleaves the *gag* poly-protein (Ferguson et al., 2002; Freed, 2001). Only mature virions can infect other cells.

1.3 Latent Reservoir

Cytotoxic effects of active HIV replication result in short life spans for virion producing HIV cells (the average half-life is approximately 1 day) (Pollack et al., 2017). HIV infection persists even while ART completely suppresses active replication. HIV persists in individuals adhering to ART as integrated proviral sequences. While viral blips do occur, integration events from these reactivated viruses are unlikely to occur due to ART targetting multiple points in the HIV lifecycle. Therefore the division of long lived resting CD4+ T cells are likely responsible for the longevity of infection.

The HIV latent reservoir consists of long-lived resting T-cells with integrated replication competent proviral HIV DNA (Z. Wang et al., 2018). Less than 5% of cells with integrated proviral HIV DNA are replication competent. The remainder of cells have mutations or deletions that result in non-functional virus (Wong et al., 1997). Estimates for the size of the latent reservoir range from 10^4 - 10^7 (Zhang et al., 1998). The half-life of cells in the latent reservoir is approximately 44 months. This half-life estimate translates to a natural decay timeline of over 70 years (Crooks et al., 2015; Finzi et al., 1997a; Siliciano et al., 2003. Cells with integrated replication competent HIV can sporadically reactivate, releasing virions. Reactivation is a process by which cells with integrated dormant provirus are stimulated to produce virions (Corey et al., 2004; Grabar et al., 2009; Zanussi et al., 1996). Reactivation happens at a predictable rate, when latent T-cells are activated upon encountering an antigen, or through indirect activation via an inflammatory process (Corey et al., 2004). Reactivation is a potential source for reinfection for individuals that do not adhere to an ART regimen or in cases of drug failure (Arts and Hazuda, 2012). Viral rebound can happen within weeks of drug failure or cessation (Crowell et al., 2019; Davey et al., 1999). Early models of latent reservoir persistence suggested that quiescent cells with slow turnover were solely responsible for the longevity and stability of the reservoir (Y.-C. Ho et al., 2013; Siliciano et al., 2003). This understanding of the latent reservoir persistence has been challenged by findings that show cellular proliferation (especially clonal expansion) plays a role maintaining the latent reservoir (Bosque et al., 2011; Yeh et al., 2021).

Clonal expansion is an adaptive immune process whereby cells that are stimulated by an antigen proliferate into clones (Adams et al., 2020). Clonal expansion of cells with integrated pro-viral sequences was shown to contribute to the persistence of the latent reservoir (Maldarelli et al., 2014). Three mechanisms have been identified as contributors to the clonal expansion of infected cells. Firstly, integration of pro-viral sequences in proximity to highly transcribed regions may promote cellular proliferation, and therefore clonal expansion. HIV preferentially integrates close to genes that govern cell division (Schröder et al., 2002). Viral promoter genes may increase the transcription of these genes, thereby promoting cellular proliferation (Maldarelli et al., 2014). The second mechanism that might contribute to clonal expansion of infected cells is homeostatic proliferation. Memory T-cells are maintained through cellular proliferation as a normal response to cytokines (Schluns and Lefrançois, 2003). Homeostatic proliferation is a non-selective process, whereby a stable T-cell population can be maintained for normal immune function. Homeostatic proliferation of HIV+ T-cells occurs without re-activation, suggesting this process would not trigger response from the immune system or immunotherapies (Bosque et al., 2011). The third mechanism that has been identified is antigenic stimulation due to chronic exposure. Chronic or repeated antigen exposure could stimulate clonal expansion of infected CD4+ T-cells (Douek et al., 2002). In contrary to the clonal expansion mechanism of latent reservoir persistence, a controversial mechanism is the constant seeding of the latent reservoir by low levels of ongoing HIV replication in drug sanctuary sites (C. V. Fletcher et al., 2014). Tissues such as lymph nodes where there is poor drug penetration may create drug sanctuary sites where active replication of HIV is permissible. This mechanism is contested with studies reporting conflicting results, with proponents citing low latent reservoir diversity as evidence of ongoing seeding (Barton et al., 2016; Eisele and Siliciano, 2012; Kulpa and Chomont, 2015).

The genetic diversity of the latent reservoir can be quantified using multiple approaches, each with their own benefits and limitations. Integrated proviruses can be fully sequenced (albiet missing some nucleotides between the primers) using primers that target conserved regions (Salminen et al., 1995). Only full length coverage guarantees identical sequences are truly clonal. Sampling short subgenomic regions does not adequately capture the full genetic diversity of a sample, thus biasing genetic diversity estimates (Laskey et al., 2016; Z. Wang et al., 2018). Full length sequencing techniques are limited because defective provirus sequences are included in this sequencing. It is difficult to estimate the viability of integrated provirus sequences from sequences alone. While HIV preferably integrates into gene promotor regions, it is not base specific. Therefore the probability of identical integration sites occurring from two unique integration events is infinitesimally small (Maldarelli et al., 2014; Schröder et al., 2002). As such, identical integration sites in two sequences is an indicator of clonal expansion. Approaches that sequence integration sites are important as presence of identical provirus sequences is not necessarily indicative of clonal expansion. For example, it is possible that highly identical sequences seed the latent reservoir early on in an infection, as in the case of SIV (Spira et al., 1996). In addition, it is difficult to identify viability of proviral sequences from just integration sites. Most viability analyses are limited to detecting obvious issues such as frame shifting mutations (Carrillo et al., 1998; Domingo and Holland, 1997; Maldarelli et al., 2014; Schröder et al., 2002). These defective or non-viable pro-viral sequences are not part of the latent reservoir, as the latent reservoir only includes replication competent virus. There are new methods that combine full length sequencing with integration site sequencing, completeness and replication competency analysis (Patro et al., 2019). This method is a powerful technique as it allows a previously impossible task - linking integration sites to their source proviral sequences. However, this method is time and resource intensive, and is not suited for large-scale studies. Regardless of sequencing technique, sequencing errors increase the variation observed in the viral population, and thereby inflate estimates of the size or the diversity of the latent reservoir.

The gold standard method for quantifying the latent reservoir is the **V**iral **O**utgrowth **A**ssay (VOA). Sequential serial dilutions and outgrowth stages amplify low levels of infection to detectable levels. By diluting samples, each 'well' in the assay receives on average less than one provirus. In outgrowth stages, samples in wells are cultured with HIV susceptible cells. Once HIV RNA is amplified to detectable levels, HIV RNA present in outgrowth assays are sequenced, allowing for quantification of genetic diversity. VOAs allows estimation of number of infected cells in the latent reservoir, as the dilution factor is known from the serial dilution step. VOAs are limited by their expensive and time-intensive nature. However, this method is the only method that can filter cells that contain replication incompetent proviral sequences (Siliciano et al., 2003; Taswell, 1984).

Regardless of the method used for measuring the viral diversity in the latent reservoir, summary statistics are needed to report the clonality. Clonality is a measure of the occurrence of identical genetic sequences in a population (Ferreira et al., 2021; Hosmane et al., 2017; Joos et al., 2008). A variant (v) in a population is an unique sequence. In literature, there are two widely used methods of quantifying latent reservoir clonality. First is to calculate the proportion of sequences that have at least one other identical sequence in the population, shown in Eq.1.1 (von Stockenstrom et al., 2015; Wagner et al., 2014). For a population with S sequences, where each variant has n_i individuals, and I(x) is an indicator function that returns 1 if true (i.e. when there is more than one individual in a variant), otherwise 0.

$$p_1 = \frac{\Sigma_{v_i \epsilon P} n_i I(n_i > 1)}{||S||} \tag{1.1}$$

Second is to calculate the proportion of variants in a population that are non unique (Maldarelli et al., 2014). For a population partitioned by variant, with P partitions.

$$p_2 = \frac{\Sigma_{v_i \epsilon P} I(n_i > 1)}{||P||} \tag{1.2}$$

In the second method, the clonality score does not report the membership size of a variant group - it only reports the comparisons between membership groups. To illustrate the difference between these measures, consider a population with 3 individuals with sequence 1 (variant 1), 2 individuals with sequence 2 (variant 2); and 1 individual with sequence 3 (variant 3) (S: {v_1: 3, v_2: 2, v_3: 1 }). Depending on which proportional clonality score is used (1.1 or 1.2), the clonality score is effected as the denominators are widely different. p1 (1.1) would give a clonality score of 5/6, or 0.83; and p2 (1.2)) would give a clonality score of 2/3, or 0.66. The current methods for describing clonality are not up to the task of accurately describing the underlying dynamics of the latent reservoir. Among these issues are: existing methods do not evaluate statistical significance of group membership (Ferreira et al., 2021); and under-estimate extent of clonality due to under-sampling (Reeves et al., 2017).

1.4 Viral Evolution Within Hosts

Phylogenetics is a wide ranging field of study that attempts to draw biological insights by studying the evolutionary relationships between individuals and populations – a subset of phylogenetics is phylodynamics (Yang and Rannala, 2012). Phylodynamics is the study of epidemiological, immunological, and evolutionary processes that shape viral phylogenies (E. Holmes et al., 1995). Genetic sequence data is used to infer the underlying process at play – especially in processes that are hard to study. Transmission dynamics, both within-hosts and between-hosts can be studied using these methods.

For example, phylogenetics is used to study the discordance between among-host (between individuals) and within-host (within a single individual) rates of HIV evolution. An order of magnitude separates the the two types classes of HIV evolution (Lemey, 2005). In a chronically infected individual with uncontrolled HIV, the HIV population is genetically diverse. The molecular driver for this diversity is Reverse Transcriptase (RT). RT lacks a proofreading mechanism, and therefore introduces relatively large numbers of mutations, as compared to DNA/RNA replication (Carr et al., 1996; Hu and Hughes, 2012; Kudesia and Wreghitt, 2009). A small number or even just an individual virus establishes an infection in an individual. The sharp decrease in viral diversity between individuals in a transmission pair has been termed the ”transmission bottleneck” (Bergstrom et al., 1999). To date the causes of this genetic bottleneck have not been identified. Studies have suggested that the immunological and physiological environment of viral entry points are part of the selection process (Bergstrom et al., 1999; Boeras et al., 2011). There is a strong selective filter during the transmission bottleneck, reflected in the high number of exposure events that are required for successful transmission (E. Holmes et al., 1995; Keele et al., 2008; Rieder et al., 2011; Salazar-Gonzalez et al., 2009; Shaw and Hunter, 2012). Transmitted/Founder studies have found that shortly after infection the viral population of a recipient is genetically homogenous regardless of the viral diversity of the founder individual (Boeras et al., 2011; Gnanakaran et al., 2011).

Within-host evolutionary dynamics are typically thought to be governed by selective forces and competitive fitness. Whereas between-host evolution is typically thought to be governed by neutral forces (evolution occurs due to random genetic drift), as multiple types of HIV (i.e. A, B, D) co-exist on population level. In the context of within-host evolution of viruses, neutral mutations are ones that do not effect an the viruses' potential to transmit. The neutral theory of molecular evolution posits that most mutations are either neutral or slightly deleterious (Kimura, 1968). Neutral forces are thought to govern within-host evolutionary dynamics due to lack of evidence of selection (Frost et al., 2018; Frost and Volz, 2013). There is a trade-off between within-host fitness and population-level (epidemiological) fitness - for example, traits that confer selective advantage to the virus within a host (e.g. immune adaptation) may reduce transmission potential (Alizon and Fraser, 2013; Lemey, 2005; Pybus and Rambaut, 2009; Vrancken et al., 2014). This trade-off has been shown in transmission studies - successful variants are most often not the predominant variant within the host donor (Joseph et al., 2015; Shaw and Hunter, 2012). In contrast with other viruses, such as hepatitis C, within-host and among-host HIV evolution does not differ significantly across the genome (Alizon and Fraser, 2013). For example, hepatitis C has higher rates of evolution in the the envelope genome (Gray et al., 2011). There are several hypotheses to why there is discordance between the two rates of HIV evolution. One such hypothesis is that selection pressures differs between stages of infection, with selection pressures and therefore evolution rate reduced in early infection. In this model, HIV evolves at much quicker rate during the later chronic infection stage. Slower evolution in the initial stages of the infection has been suggested to result from insufficient immune response (resulting in reduced selective pressures) (Berry et al., 2007). Transmission tends to occur early in an infection, when viral loads are typically higher. Transmissions that occur earlier in an infection would tend towards carrying fewer mutations, as each round of viral replication results in a high number of mutations. Within-host evolution tends to be measured over a long time horizon in individuals with unsuppressed HIV infections, during the chronic infection stage. An increased rate of evolution during the chronic infection stage, would result in higher observed evolution rates within-host as compared to between hosts (Berry et al., 2007; Lythgoe and Fraser, 2012).

The second model hypothesizes that the discordance between rates of evolution results from HIV becoming adapted to the host's immune response. However, these mutations only increase fitness in a particular host, and quickly revert after transmission. This model suggests that a high proportion of mutations revert once the virus is transmitted to a new host. Once again this would result in a mismatch of evolution rates, as only a small number of mutations would accumulate at the population level (Matthews et al., 2008).

The final model for HIV evolution is the “store and retrieve” hypothesis. In this model, HIV variants seed the latent reservoir early in the infection. Over time the latent reservoir is maintained through clonal expansion. Upon reactivation, possibly months to years later, early archived viral strains are able to circulate once again. It was suggested that these early variants have higher transmission fitness or remain in anatomical compartments close to routes of transmission - i.e. in the genital tract (Lythgoe and Fraser, 2012; Sagar et al., 2009). As such, these early variants would be more likely to be transmitted than variants that seed the latent reservoir later in infection (Lythgoe and Fraser, 2012). Evidence for this model of HIV evolution is mixed. In support of this model, transmission pair studies have shown that ancestral sequences are more closely related to actively reproducing/circulating virus in a newly infected individual, as opposed to actively reproducing/circulating virus in the originating individual (Herbeck et al., 2011; Sagar et al., 2009).

Much about the latent reservoir remains unknown. The mechanism underlying the establishment and persistence of the latent reservoir is still not fully known. ART is able to control active infection and taken proactively it is able to prevent the establishment of an HIV infection (A. Okoye et al., 2007). However, current therapies do not address the latent reservoir. Therefore, understanding mechanisms for persistence to prevent initial establishment of the latent reservoir are important towards the pursuit of a HIV cure.

1.5 Scope of project

The following body of work is divided into three chapters. In Chapter 2, I describe the simclone simulation pipeline I created. In this chapter I show how the HIV-1 within-host dynamics are captured using the simclone framework in an existing host-pathogen simulation program. The methods used to explore the simulation parameter space are detailed. Further, I describe the nucleotide evolution portion of the pipeline that follows population dynamics simulation. I present sample inputs and outputs for each simulation step. The mechanism of HIV- 1 latent reservoir establishment and persistence has previously been modelled using simulation methods (Perelson and Ribeiro, 2013). Modelling has provided insights into immune escape, viral evolution, impact of treatment, and other topics of research in HIV research. In this chapter I show that the host-pathogen simualtion program *twt* can be be used to simulate HIV-1 within-host dynamics.

In Chapter 3, I use the simclone simulation pipeline described in Chapter 2 to investigate methods of calculating clonality. In biological systems, including the HIV-1 latent reservoir, genetic identicality implies a particular biological mechanism. In the case of the latent reservoir, identicality implies clonal expansion contribrutes to latent reservoir persistence. The degree to which clonal expansion is responsible for persitence is contested. There are many ways of calculating the degree of identicality in a population (clonality). The means by which clonality is calculated can affect downstream correlative analyses. In chapter 3 I attempt to draw associations between clonaltiy scores and simulation parameters using three calculations for clonality, two methods used in literature: proportional clonality; and the Gini Index, and a method I propose ”pairwise clonality”.

In Chapter 4, I attempt to find non-linear associative trends between simclone parameters and clonality scores. I use particle filtering, a statistical noise reduction technique, to estimate parameter values from clonality scores. Particle filtering is a method whereby simulation samples (particles) that are poor representations of real values are systematically removed. In my implementation of particle filtering, particles are given an RMSE score against target simulations. I show that particle filtering cannot be used to predict parameter values for many of the simclone parameters.

Finally in Chapter 5, I discuss limitations and implications of my research, in particular the findings of the clonality score chapter. I also propose future directions based on the work completed in this thesis.

Chapter 2

Simulation Model (Simclone)

2.1 Introduction

Within host HIV evolution is difficult to study in-vivo for practical and economic reasons. A non-exhaustive list of practical reasons include: it is difficult to sample tissues where the latent reservoir resides (i.e. Lymph Nodes and Gut Associated Lymph Tissues (Ahlenstiel et al., 2020); it is difficult to isolate productive latent reservoir cells due to lack of biomarker, although recently advances have been made in detection (Darcis et al., 2019); and the relatively low proportion of cells that constitute the latent reservoir in proportion to circulating non latent reservoir cells (Siliciano et al., 2003). Expensive and invasive sampling techniques are required to isolate cells in the latent reservoir due to the aformentioned issues. Simulations of the dynamics of the latent reservoir has been used to address questions about immune escape, CD4+ depletion, and viral evolution (D. D. Ho et al., 1995; Perelson and Ribeiro, 2013). Previous work does not allow individual lineages to be tracked in an infection back to an index case. Having the simulated transmission, latency, and reactivation events histories can provide novel information that may motivate experimental or clinical studies. I created the simclone simulation pipeline to address this gap in simulation programs. Simclone can simulate HIV within host evolution while also allowing lineages to be tracked.

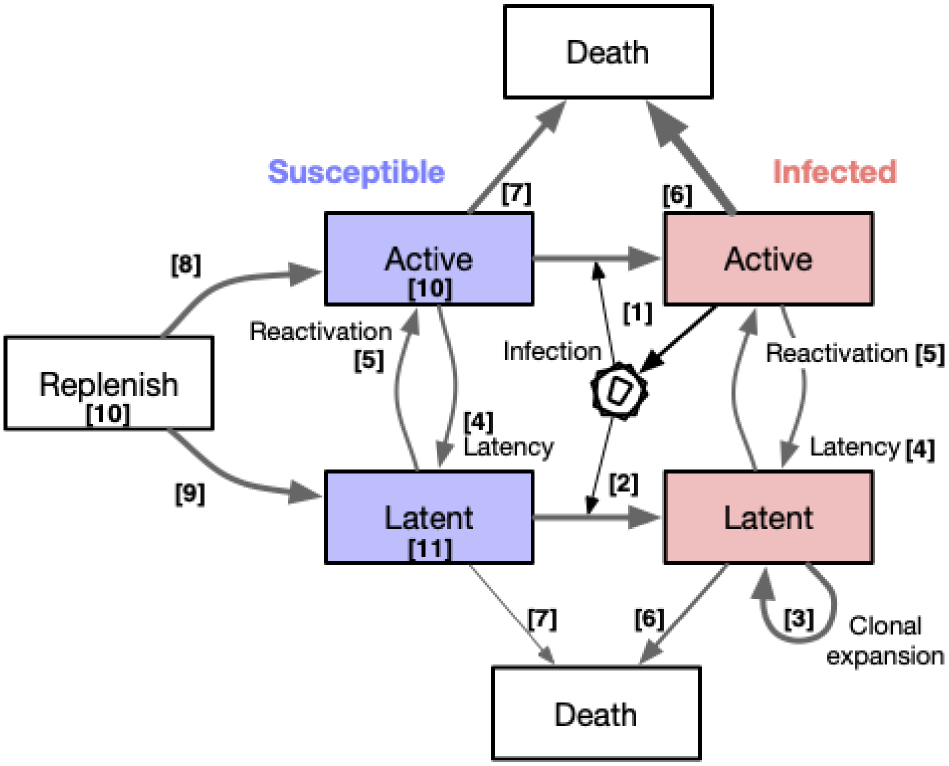


Figure 2.1: *simclone* Framework. Parameters are labelled with [n], with labels found in 2.1. Some labels are used more than once (e.g. [4,5,6,7]). In the *simclone* framework transition parameters are independent of infection status. Encapsulated virus under [1] represents an HIV Virion.

2.2 Simulation Pipeline and Performance

2.2.1 Simulation Pipeline Overview

The *simlone* simulation pipeline consists of 3 distinct overarching steps: (1) Creating specification files; (2) Within-host simulation and phylogenetic tree generation; (3) Simulation of nucleotide evolution. There are several bash and Python scripts that were used to manage the *simclone* pipeline. The first step of the *simclone* pipeline is to create sets for the simulation model. Parameter combinations were set by establishing bounds that resulted in within-host dynamics that consistent with modelling in literature (D. D. Ho et al., 1995; Perelson and Ribeiro, 2013; Wei et al., 1995). Parameter estimates from the peer-reviewed literature were used as a starting point, and adapted for the timescale of the simulation model set by computational constraints in step (2). A Python script generates yaml specification files for the simulation pipeline. The command-line interface can be used to specify options including: the number of samples generated; and the number of replicates per samples. Optionally, a sample name can be specified by the user to label an simulation run as desired. This Python script generates all necessary specification files and directory structures for step (2). An unique specification file is generated for each sample, however replicates share a common specification file. Next, using an R script, within-host interactions (e.g. transmission between cells) are simulated and transmission trees are generated for each replicate. The main output of interest from the within host simulations are phylogentic trees that represent the evolutionary history of viral sequences sampled from simulated HIV-1 infections. Following phylogenetic tree creation, guide trees generated in step (2) are used to simulate nucleotide evolution on a reference HIV-1 *env* sequence.

2.2.2 trees within trees (twt)

twt is an R package jointly developed and maintained by members of the Poon lab (www.github.com/poonlab/twt). The program combines forward-time and reverse-time methods to simulate host-pathogen trees. Forward-time simulations are used to simulate population dynamics, starting from an index case. In forward-time simulations, a Markov state process is used to simulate the population dynamics of the infection. Markov state processes are random processes where a future state is independent of past states, and only dependent on the present state. The probability of an event (transmission and transitions) occuring is drawn from a joint distribution of rates defined in specification files. The reverse-time simulation method used in twt is the coalescent, it is used to resolve the transmission and transition history of viral lineages in sampled cells back to the index case. The coalescent is a model that resolves the ancestral relationship between sampled individuals in a population. Coalescent theory estimates the time between period between lineages and their ancestral states. Reverse-time methods are used for resolving transmission and transition history, as ‘unimportant’ events can be ignored, thus saving computational effort. The basic unit of the twt framework is a cell. A cell can be susceptible or infected. Each cell belongs to a class of cells and all cells in a class share a set of user defined rates that govern its interactions with other cells in the framework, (i.e. infection rate). In twt, compartments are containers for cells of a singular class, and can contain susceptible and infected cells. It is important to note that in twt a ‘compartment’ is not the same as a compartment in a physiological sense – a compartment does not represent a spatially or physically segregated group of cells. Rather, a ‘compartment’ in twt is termed as a group of cells that share characteristics (to avoid confusion I term a compartment in twt as “cell types”). There could be compartments in twt in the physiological sense (groups of physically or spatially segregated cells), however these are not implemented in the *simclone* framework. There are two type of events that occur at a given rate: transmission – infection of a susceptible cell by another infected cell; and transition – the transition of a cell from one state to another, (e.g. from active to latent). An infected cell is always involved in a transmission event. However transition events do not require a cell to be infected. When a cell transitions between types, it maintains its infected status. Transmission rates between cells of different classes can be varied depending on the compartment of the cells in question. When a cell transitions between compartments, it assumes the dynamic characteristics of its new compartment. There is no theoretical limit to the number of compartments that can be defined, only computational (memory) limits. As the number of compartments, and the number of individuals within them increase, computation costs increase as well. Within this program, a number of parameters can be defined such as the number of compartments, compartment size, transmission rates between and within compartments, and transition rates between compartments.

The first stage of the simulation process is the forward simulation of the population dynamics of the infection. This forward simulation process is a discrete Markov process. In a discrete Markov process, the probability of events occuring are dependent on the current state of a system – independent of previous events. Starting from an index case (i.e. infection starts with one cell), the random spread of infection and transition of cells between compartments is simulated. An assumption of twt is that there can only ever be one cell infected at the start. If seeded with multiple infected cells, there would not be a single lineage tree that could be created relating all infected cells in the simulation. Each cell type is initialized with some number of susceptible cells. For every step in the Markov Chain, transmission and transition rates govern the movement of cells between cell types, and the infection of susceptible cells by infected cells.

At user defined sampling times, n_i cells are sampled from the i^{th} compartment. It is important to note that $\Sigma n_i \leq \Sigma N_i$; where N_i is the number of cells in i^{th} compartment and can vary over time. Using the transmission and transition events in the first step, a transmission tree from the index case (source) is generated for the sampled cells. A transmission tree is an hierarchal graph structure (where each node in the tree has only one parent, but may have one or more children, except the root node which has no parents) that represents the infection events between cells between the index case and all sampled cells. In this graph, internal nodes (non tip and non root nodes) are not necessarily sampled, but can are inferred. This transmission tree relates the ancestry of sampled cells from the simulation, and may include unsampled individuals where necessary. The transmission tree is also called the outer tree in the twt.

From these sampled cells, the Gillespie method (Gillespie, 1976) is used to simulate a transmission tree backwards to the index case. The Gillespie algorithm is a process by which the occurence of stochastic events is determined by drawing values from an exponentially distribution, scaled to probability rates. Reverse simulation makes it computationally feasible to model many interactions in complex systems. Reverse simulation increases the computational feasibility because the number of sampled cells is much lower than the number of total cells. Thus the simulation algorithm does not need to consider as many individuals. Finally a pathogen tree is simulated by coalescing the viral lineages in the sampled cells to resolve an ancestral history of the viral lineages to the index case.

In summation, the existing twt program first simulates the population dynamics (how viral lineages spread through cell types); resolves the ancestral relationship between sampled cells; and finally resolves the ancestral lineage of the viruses in the sampled cells.

Further details of how the population dynamics, transmission tree and pathogen tree are simulated can be found on the twt github repository (see above for repository link).

twt was designed to simulate the dynamics of virus populations being transmitted between hosts. I extended the twt framework for simulating the dynamics of cells within a single host, which I call the *simclone* framework. In the *simclone* framework (Fig. 2.1) there are four unique compartments; Active, Latent, Replenishment, and Death. To avoid confusion with "compartment" in the physiological sense, I will now refer to these compartments as cell types. This nomenclature will also make discussion regarding expansion of the *simclone* framework easier to understand. The Active cell type contains cells that can infect other Active or Latent type cells when infected. Active cell types can transition to the Latent cell type (modelling latency) or Death cell type (modelling death). Latent cell types can be infected, but cannot infect other cells. Latent cells can undergo clonal expansion, transition to the Active cell type (modelling re-activation), or transition to a Death type cell (modelling death). Clonal expansion in the *simclone* framework is modelled as a transmission event. I assume that "daughter" cells replace an uninfected cell of the same cell type. The Replenishment type cells cannot be infected, but can transition to become active or latent type cells. These transition processes mimic cell regeneration/ proliferation. Lastly death type cells can neither be infected nor transition to another cell type. The death cell type is important for modelling the death of viral lineages in the twt framework. Without the death cell type, viral lineages would continue to propagate.

Besides the parameters that define the interaction between cells, there are parameters that govern the sampling time, and simulation length that can be defined as well. The simulation length is the set period of time that the simulation runs for until sampling events, it is dimensionless and not scaled to real-world equivalents. Epochs are a distinct period of time, within an epoch transmission and transition rates and fixed. Multiple epochs can exist for a given simulation. The goal of multiple epochs is to allow for different viral transmission environments. For example, in the *simclone* framework, there are two epochs: an epoch where there is uncontrolled active transmission; then a second epoch where there is no active transmission. This is how the *simclone* framework mimics the initiation of successful ART. To model ART in the *simclone* framework, I assume that the Active cells can no longer infect other cells (i.e. $AA_{br} = 0$); in the *simclone* framework transmission events originating in Active cell types are no longer allowed. To simplify the model, the *simclone* framework assumes that all active infection is halted at the initiation point of ART.

Definable parameters within the *simclone* framework are described in Table (2.1). As discussed previously, the two epochs represent two phases of a typical HIV-1 patient that initiates ART. In the first epoch, the 12 parameters in Table (2.1) are fixed to values generated in step (1) of the *simclone* pipeline, based on models of the within-host dynamics system (D. D. Ho et al., 1995; Perelson and Ribeiro, 2013; Wei et al., 1995). The generation of parameter combinations from the above distributions will be discussed further later. At the second epoch, the transmission rates of Active type cells are set to zero (AA_br and AL_br). As cells 'die' by transitioning to the death type, active and latent type cells take their virus lineages with them. The end result is that there is a complete or near-complete removal of all viral lineages from the Active compartment, which mimics the real-world suppression of HIV-1 replication through ART. After initial parameter setting testing, some definable twt parameters were fixed to allow for transmission characteristics parameter response testing. These parameters were: sampling time (t=10); ART initiation time (t=5), and number of cells sampled per cell type (n=100).

Specification files for twt use the YAML markup language (Ingerson et al., n.d.). The YAML markup language is a structured language much like XML or HTML, but with a much simpler syntax. Data can be nested and labelled with the YAML format. The *simclone* framework is saved as a template YAML file, with string markers (i.e. *"[AA_br]"*) for modifiable parameter settings. The templates enable automated specification file generation. An example YAML specification file can be found in the supplemental section A.1.

2.2.3 Latin Hypercube Sampling

There are 11 parameters in the *simclone* framework that need to be varied in order to measure the responsiveness of clonality statistics to changes in parameter settings. Incremental sampling of parameter setting combinations (i.e. drawing n samples for each of the parameters) would be an inefficient approach to exploring the *simclone* parameter space. There are simply too many parameter combinations ($N = n^{11}$) for robust parameter space testing in a computationally feasible manner. Hence, parameter selection was preformed through Latin Hypercube Sampling (LHS).

LHS is a method for sampling a multi-dimensional space in a pseudo-random fashion. Latin Squares in statistics are much like the Sudoku square, in that it is a descrete n by n square in which a value is not repeated on a vertical nor horizontal axis. The "Hypercube" portion of the LHS moniker refers to the multiple dimensions that the sampling method is applied to. There is no theoretical limit to the number of dimensions that the LHS method can be applied to - however there may be practical limits to this method. The LHS method was described first by McKay et al., 1979 and expanded further by Iman and Conover, 1982. The LHS method works by dividing each parameter into k intervals, with each interval of size x/k; where $(x_{max} - x_{min})$ is the range of the parameter and k is also the number of samples. One value is sampled from each of the k intervals. This is repeated n times, drawing k values for each parameter value. The LHS algorithm then randomly combines the samples across parameters such that there is a sample in each interval k for each parameter value. The benefits of the LHS method is that it ensures uniform coverage across the parameter space. Unlike true random sampling, a set of parameter combinations can be generated such that the entire parameter space is covered. True random sampling could by chance miss regions of the parameter space. This method was implemented in the *simclone* framework by creating a matrix of k number of samples by 12 parameters with values (*'lhs_i'*) normally distributed between 0 and 1. The pyDOE (https://pythonhosted.org/pyDOE/) implementation of the LHS algorithm was used to generate the matrix using the *"maximin"* option. This option maximizes the minimum distance between any two sampled points. This option was chosen to increase the coverage and evenness of the parameter space. The matrix was transformed by linearly scaling the LHS generated values by the parameter ranges shown in equation 2.1.

$$x_i = lhs_i\ (\ x_{max} - x_{min}\ )\ +\ x_{min} \tag{2.1}$$

For example if a LHS generated value (lhs_i) was 0.4 for a parameter range of $x_{min} = 4$, $x_{max} = 10$, the parameter value used in the simulations would be $x_i = 6.4$.

Parameters can also be considered along a logarithmic using the LHS method as well:

$$p_i = 10^{-lhs_i} \tag{2.2}$$

Parameter values for the logarithmic scaled values are bound by (10^{-8}, 1). Values below 10^{-8} were found to be uninformative, in preliminary testing with values below 10^{-8}, simulations were unsuccesful due to lack of infected individuals to sample. The LHS values for logarithmic scaled values are bound differently (0, 8) instead of (0, 1).

2.2.4 Clonal expansion in the *simclone* framework

As discussed previously, clonal expansion is the proliferation of cells, usually triggered as a response to an antigen. As the name suggests, daughter cells from these divisions tend to be clonal. There are some mutations introduced during cell division, however cellular proliferation has a negligible rate of mutation when mutations introduced by HIV-1 Reverse Transcriptase. In the *simclone* framework we assume that clonal expansion does not introduce any mutations. Clonal expansion is a parameter that can be controlled in the *simclone* framework (ll_br). In the *simclone* framework, it is coded as an transmission event between two latent-infected cells.

A phylogenetic tree is a graph structure that represents the evolutionary relationships between sequences or individuals. Fig. 2.2 is an example of a rooted tree. In this structured graph, the nodes at the end of a tree (also called leaves) represent the observed sequences. These are labeled A, B, C, D and E in Fig. 2.2. In a rooted tree, all the tips are eventually joined until the branches collapse to a singular node called the root (R). In between the tips and the root nodes, there are internal nodes and edges. Internal nodes can represent sampled or unsampled individuals. The structure of the internal nodes and edges (i.e. how nodes and edges are arranged) can vary as it is dependent upon the relationships between the tips and preceding internal nodes. The length of the edge between two nodes can represent the genetic similarity between the two nodes; this is a weighted edge in graph theory. Two nodes connected by a shorter arc are more genetically similar than two nodes connected by a longer arc. Arcs are drawn proportional to their weight. In the tree depicted in Fig. 2.2, A and D are the same distance from the root (R), thus these two samples are equally divergent from the root sequence. It is important to note that the A and D are not necessarily identical, rather they differ from the root (R) in the same magnitude. For example if the root had the sequence “AAAA”, tip A could have the sequence “AAGG” whereas tip D could have the sequence “AATT”. Tip E is closer to the root than any other sequence, and is therefore the least different from the root. Two identical nodes will be connected by an edge with a length of zero. These are also known as polytomies.

Weighted edges representing evolutionary distance can be exploited to approximate the “store and retrieve” hypothesis in the *simclone* framework. Under the “store and retrieve” hypothesis of latent reservoir maintenance, ancestral HIV sequences are “stored” through latency then reactivated months, or possibly years later (Lythgoe and Fraser, 2012). To capture clonal expansion in this framework, evolution needs to be halted as soon as sequences enter the latent reservoir. Since twt provides transmission history, edge lengths can be replaced with zero as soon as the viral lineage is recognized in a latent type cell (through transmission or transition). As edge lengths represent evolutionary distance, setting edge lengths with 0 when sequences enter the latent reservoir in effect halts evolution. The following tree pruning algorithm (Alg. 1) was implemented in R. It recursively considers all pairs of nodes in the tree in order from

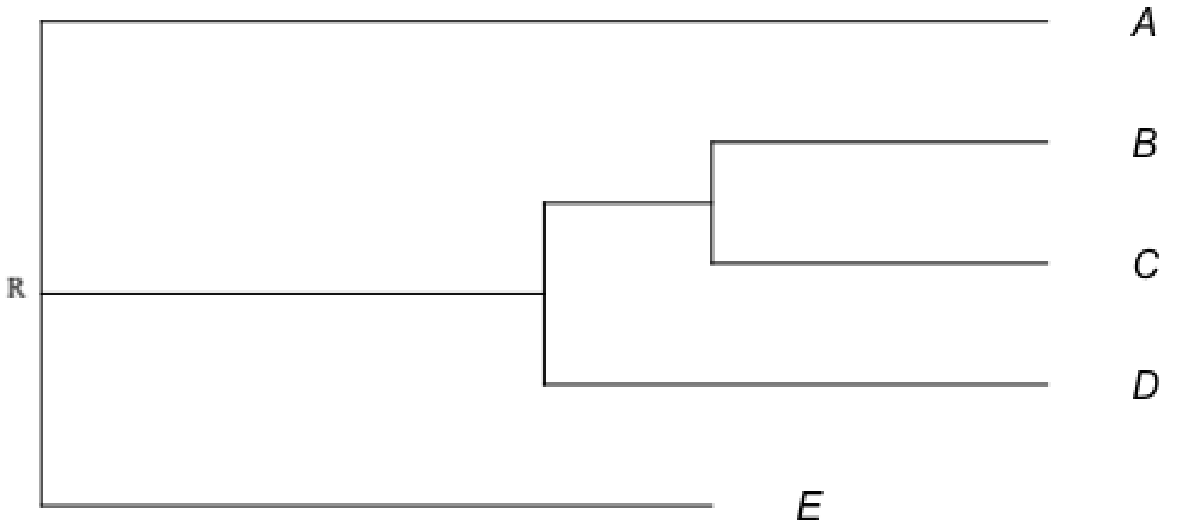


Figure 2.2: Simple Phylogenetic Tree. R represents the root node of the tree. Tips are labeled A - E, and represent observed sequences. In a phylogenetic tree, horizontal distance represents genetic similarity (distance) between two points. Internal nodes (non labelled nodes) represent unobserved, inferred ancestors between the tips and the root.

the root node to N_i; where as i is the total number of nodes in the tree. If the node represents an active type cell, the edge is tagged as 'Active'. Otherwise, if the parent node is a latent type cell, the length of the edge is set to 0, and the edge is tagged as 'Latent'.

Algorithm 1 Tree Branch Shortening Pseudo-code

```
while Remaining Nodes do
    if Parent Node is Active then
        Edge Type ← Active
    else if Parent Node is Latent then
        EdgeLength ← 0
        Edge Type ← Latent
    Next Node
```

The effect of shortening branches in the phylogenetic tree is scaled back evolution over parts of the original phylogenetic tree. As an example, consider the tree in Fig. 2.2. Sequence B and sequence C enter the latent reservoir at the parent node of the two sequences. Fig. 2.3 shows the tree that results from the branch shortening process. Sequence B and C are now a polytomy. These scaled trees are now ready for simulating nucleotide evolution.

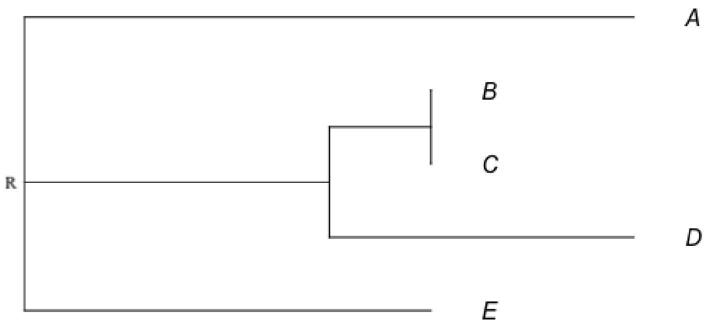


Figure 2.3: Pruned phylogenetic tree created by modifying tree found in Fig. 2.2. There is no horizontal distance between tips B and C. These two sequences are a polytomy; and would be interpreted as a clonal expansion event in the *simclone* framework.

Parameter Types	Parameter Description	Parameter Name	Label
Transmission Rates	Active - Active Transmission Rate	AA_br	1
	Active - Latent Transmission Rate	AL_br	2
	Clonal Expansion Rate	LL_br	3
Transition Rates	Active - Latent Transition (Latency)	AL_tr	4
	Latent - Active Transition (Reactivation)	LA_tr	5
Death Rates	Active Death Rate	A_dr	6
	Latent Death Rate	L_dr	7
Replenishment Rates	Active Replenishment Rate	RA_tr	8
	Latent Replenishment Rate	RL_tr	9
Initial Compartment Size	Active Compartment Size	Active_size	10
	Latent Compartment Size	Latent_size	11
	Replenishment Compartment Size	Replenish_size	12

Table 2.1: *simclone* parameters. Parameter suffixes denote type of event or cell type parameter. Prefixes denote involved compartments. First letter is source/origin, second letter (if applicable) is recipient. Label column refers to label in 2.1

2.2.5 INDELible

The software INDELible (W. Fletcher and Yang, 2009) is used for simulating nucleotide evolution. INDELible is a software that can simulate insertions, deletions, and substitutions over a variety of evolutionary models. A key feature of INDELible is that it allows for heterogeneity of substitution rates among sites, as well as along different parts of the phylogenetic tree. INDELible utilizes a Markov Chain process to simulate evolution along the entirety of the sequence. Null-order Markov Chain are processes whereby the probability of an event occurring is only related to the current state of a system (i.e. it is independent of states preceding the current state) (Gagniuc, 2017). To illustrate the Markov Chain process, consider the game Snakes and Ladders. It is a game where each player takes turns rolling die in sequence. The number rolled on the die is the number of spaces an individual can move forward on a numbered grid. The outcome (i.e. the new position of an individual) of each die roll is only dependent on the state of the player when the die are rolled. Gillespie's algorithm is used to implement the simulation process. In brief: there is a probability distribution for substitution (S), deletions (D), and insertion events (I); where as the total rates for all event types is lambda (λ) (λ= S + D + I). Depending on the model of evolution selected, the substitution rate can vary per site. At a given time point s_i, sampled from an exponential distribution ($s \sim \lambda \exp(-\lambda t)$), the type of event is randomly selected based on the model of evolution. If the chosen event is a substitution event, the site of evolution is randomly selected, but the evolution model defines the probability of the substitution type. The next sampling time for a branch is given derived through λ. Further discussion can be found in the publication. INDELible simulations are managed through control files. There are three items of note in the control files generated for INDELible: the evolution model; the phylogenetic tree; and the template for the evolution. In the *simclone* framework, the simplest model of nucleotide simulation is the default. The Jukes Cantor model of nucleotide substitution (JC69) assumes equal base swapping frequencies, as well as equal base frequencies. Other models of nucleotide evolution can be chosen as well. Thus, any model that INDELible supports can be implemented within the *simclone* framework.

A Python script I wrote inserts the modified phylogenetic trees (with pruned branches) from the twt simulation step into the control file templates. This Python script has command line options that grants flexibility when creating INDELible control files. The phylogenetic trees are in the Newick format. A representative HIV-1 *env* sequence retrieved from NCBI (Reference Sequence: NC_001802.1) was used as a the template sequence for evolution. NC_001802.1 is a 9181 basepair ssRNA (single strand RNA). The *env* gene was chosen as a candidate simulation gene because of its widespread study in literature. This gene is under strong selective pressure due to its role in immune evasion, consequently there is high variability in this sequence (Aloia et al., 1993; Zanetti et al., 2006; Zhu et al., 2006). Finally, a seed is included in the control file to ensure the experiment can be replicated, seeds are randomly generated so they differ between replicates. A sample control file is attached in the supplementary section. The input and output genetic sequence for INDELible are in the FASTA format. The FASTA format is a text-based format for storing nucleotide or amino acid sequences. In the FASTA format, sequence headers contain information about the following sequence. Headers are denoted by the '>' symbol. The line that immediately follows the header sequence contains the sequence. Each letter in the sequence line represents a nucleotide (A, C, G, or T) or amino acid (A, R, N, ...).

2.3 Parallel Computing

Relatively long simulation times due to high replicate and sample numbers necessitated the implementation of parallel computing. The longest and most computationally intensive step is typically the twt simulation stage. While each twt simulation was not computationally expensive, the high number of replicates per experiment resulted in a high total cost. Since each twt simulation were independent from each other, simulations could be run in parallel. Parallel computing was implemented through the mpi4py library in Python (https://github.com/mpi4py/mpi4py). Parallel computing is not strictly necessary to run this pipeline. However, it can greatly increase the overall speed of a set of simulations by removing the need for each simulation to wait for the completion of the previous simulation. Due to the low compute cost of the nucleotide evolution simulation step, only the population dynamics simulation step was refactored for parallel computing. It is possible to run the nucleotide evolution step on cluster computing, however.

MPI and Parallel Computing

Message Passing Interface (MPI) is a protocol for passing commands between distributed computing resources. It is a structured protocol that allows processes to be aware of the status of other ongoing processes. The mpi4py library provides methods for making use of MPI classes such as communicators. To enable MPI methods in a Python script, the mpi4py module in Python and the MPI module on the cluster must be enabled – this is specific to the RHEL distributions of Linux. Generally, there are two ways of using MPI for parallel computing. First is to distribute tasks equally over the available nodes and processing threads on initialization. This is the simplest method to implement, and does not require an active listener to assign new tasks to idle processes. For each process, a queue is initialized with an equal number of tasks to run. If these tasks do not require the same amount of time, or there is excessive variation in the time to complete different tasks, this method of task distribution can result in idle nodes. twt simulations do not necessarily take the same amount of time to complete. There are scenarios where the simulation will stochastically fail on the extreme ends of parameter combinations. For example, at the extreme end of parameter settings the number of infected cells can fall below the number of cells needed to be sampled, therefore resulting in a failed simulation. twt will re-attempt the simulation n times (default n= 3) before marking a simulation as unfeasible. The task distribution method can be modified for this behaviour. On initialization, a listening process is created with a task queue. Upon receiving a 'completed' signal from the worker thread, the listening process assigns the next task in the queue to the idle worker thread. This continues until all the tasks are completed.

2.4 Performance

There were two computing resources used in the performing simulation and analysis. The "workstation" is a custom built desktop computer used for bioinformatics work. Most work preformed on the workstation was conducted through (**s**ecure **sh**ell) ssh/command-line access due to the restrictions imposed by the COVID-19 pandemic. The computing cluster "BEVi – **B**io-informatics and **E**volution of **Vi**ruses" is a shared computing resource used among members of the lab group. BEVi in the Western University Data Centre, and was also accessed through ssh/command-line. Technical specifications for the computing resources can be found in Table 2.2. Performance estimates were based on default settings of the simulation pipeline: 1000 samples taken from some parameter distribution, with 10 replicates per sample.

Computing Resource		Specifications			
Name	Node Name	CPU	Cores	RAM	OS
Workstation (Langley)	N/A	Intel Xeon E5-1650v4	6	16 GB	Ubuntu 18.04.5
Computing Cluster (BEVi)	Head Node	2 x Intel Xeon E5-2667v4	16	64 GB	CentOS 7.9
	Compute Node 0	2 x Intel Xeon E5-2667v4	28	56 GB	
	Compute Node 1	2 x Intel Xeon E5-2667v4	28	112 GB	
	Compute Node 2	2 x Intel Xeon E5-2667v4	28	112 GB	
	Compute Node 3	2 x Intel Xeon Gold 6248	40	192 GB	

Table 2.2: Technical specifications of computing resources

The three distinct sections of the simulation pipeline are: parameter combination generation; population dynamics simulation and tree generation (twt); and nucleotide evolution simulation. On BEVi, the computation time estimate for a simulation experiment of 1000 samples, of 10 replicates each was approximately 2 days. On the workstation, the same simulation experiment would take over 7 days. The parameter and control file generation, and nucleotide evolution simulation stages were relatively quick (2 minutes, and 30 minutes respectively)- thus implementing these two stages into the cluster computing pipeline was not done. In addition, the parameter and control file generation stage is not a good fit for cluster computing, as LHS generation and control file generation is not easily broken down into smaller tasks.

2.5 Simulation outputs

Simulations are an important tool for understanding processes that are not easily measured *in vivo* for practical or for financial reasons. Simulations can reveal underlying mechanisms behind processes that can then be applied to real-world analyses. For example, *simclone* pipeline captures cell level transmission and transition, data that could never be captured when studying the latent reservoir *in vivo*. The parameters can also be tweaked to study hypotheses that would be unpractical or unethical *in vivo*. Samples of the outputs mentioned in this section can be found in supplementary materials.

2.5.1 Parameter Combination Generation

The output of the parameter combination generation is a directory that contains all the necessary files and sub-directories for the entire pipeline. Two key outputs are the comma-separated values file (*csv*) that contains all parameter combinations, and the control files for the next step: population dynamics simulation and tree generation. In this stage the Latin Hypercube Sampling method from the pyDOE Python library is used to generated a 12 by n matrix. This matrix is scaled to the parameter distributions defined by the user. The naming scheme for each sample/replicate is “{n}_{sample/rep}”. The parameters for each replicate is stored as a *CSV* for further downstream analyses.

2.5.2 Population dynamics simulation and tree generation

As discussed, the first step of the twt simulation stage is modelling the growth of the infection in the cell types. The output of the population dynamics simulation is an R6 object that contains all transmission and transition events that occurred during the simulation. Fig. 2.4 shows a single representative outcome of a stochastic population dynamics simulation process. It is a non deterministic process (i.e. there is a different outcome for the same parameter values). This graph shows the changes in population sizes for a representative simulation. The index case always starts in an active cell. Depending on the parameter combinations being tested, this population dynamic graph will appear differently. Blue lines in this figure represent the susceptible population for a cell type. Red lines represent the infected population. The graph is plotted on reverse simulation time axis, where the x-axis starts at the maximum simulation time (10 in Fig. 2.4) and ends at zero. This is the effect of cells moving from the replenishment cell type. As time passes, the number of infected active type cells rapidly increases exponentially. Latent type cells are infected much slower, but still at a steady rate (9.92×10^{-6} transition events / Unit Time). ART initialization happens at t = 5. At this point, infection between active type cells no longer occurs, the infected active type population starts to decrease. In addition, the number of susceptible active type cells start to slowly reappear. Outputs of the population dynamics simulation are saved to various folders in the experiment directory. The transmission and transition events are extracted and saved as a CSV file. A line plot displaying the trajectory of population changes (e.g. Fig. 2.4) is saved for visual inspection. In addition the outer plot showing the transmission tree for sampled individuals is saved for the same reason.

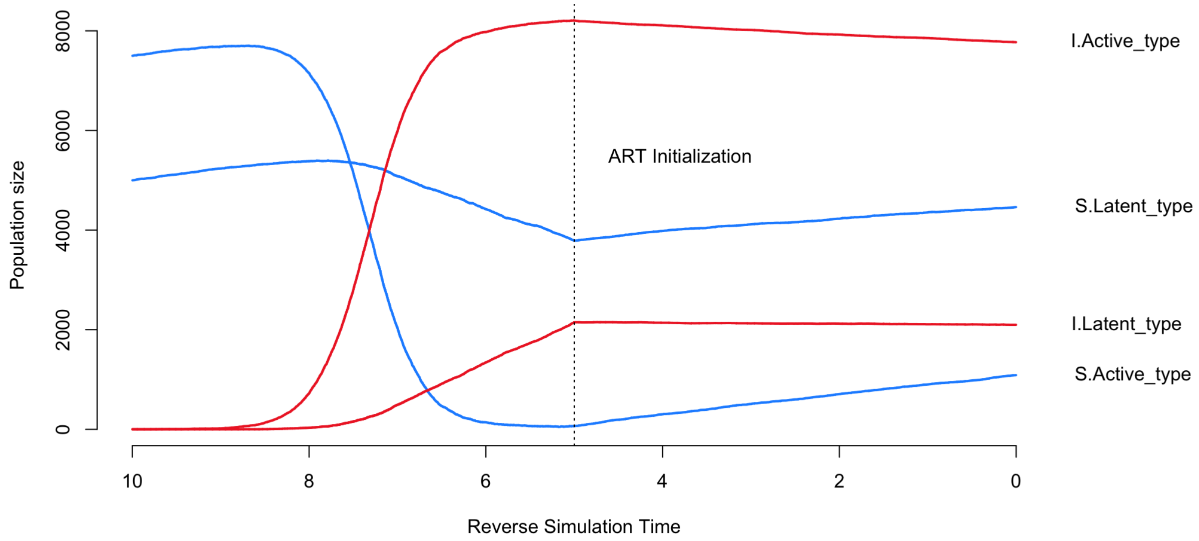
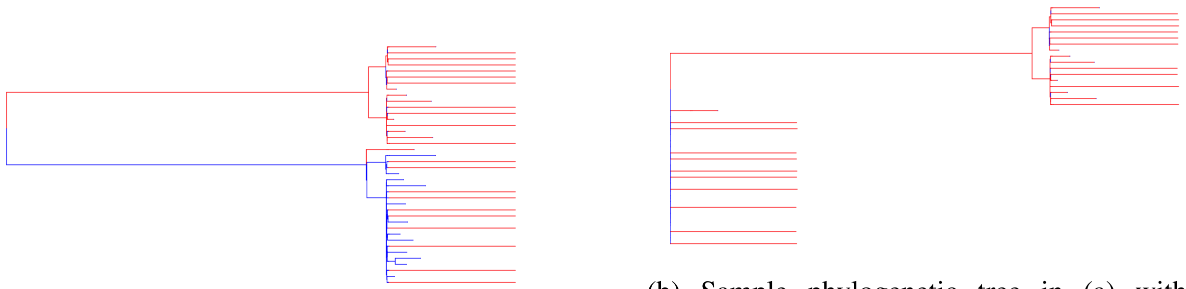


Figure 2.4: Line graph showing changes in susceptible (labelled with ”S” prefix) and infected (labelled with ”I” prefix) populations in all cell types during initial population dynamics simulation.

As discussed previously, the transmission ‘outer’ tree is used to simulate the phylogenetic



(a) Sample phylogenetic tree without branch shortening

(b) Sample phylogenetic tree in (a) with branches shortened when lineages enter latent type cells

Figure 2.5: Phylogenetic trees generated by twt simulation. Branches are coloured based on the cell type. Branches that are red are when viral lineages are in active type cell. Branches that are blue are when viral lineages are in latent type cell

‘inner’ tree. The are modified using the branch shortening method described previously. The phylogenetic tree is then saved as a Newick file for the nucleotide evolution simulation. A coloured version of the phylogenetic tree is saved as well (see Fig. 2.5b). In addition, the Newick file is modified to remove artifacts that INDELible cannot process. The tree structure does not change, only the interior node names are removed.

2.5.3 INDELible

The modified phylogenetic trees generated by the previous stage are inserted into a control file template for INDELible. See supplementary section for an example INDELible input. The rest of the control file is constructed as per the specified nucleotide evolution model. These control files are saved to tree specific directories. INDELible is then directed to these directories on a replicate-by-replicate basis. INDELible outputs two FASTA files that contain a number of evolved HIV *env* sequences. The number of sequences in the INDELible output FASTA file depends on the number of samples specified within the twt simulation. This number is the number of cells that are sampled from the active and latent cell types respectively.

Chapter 3

Summary statistics

3.1 Background

3.1.1 Measures of Clonality

Replication of the HIV genome is highly error prone, due to the lack of a proof reading mechanism in RT (Boeras et al., 2011; Cavarelli and Scarlatti, 2009; Hu and Hughes, 2012; Telesnitsky, 2010). As such, the presence of identical integrated HIV-1 DNA sequences, is taken as an indication that clonal expansion is the underlying cause of HIV persistence in the latent reservoir (Chavez et al., 2015; Crooks et al., 2015; Siliciano et al., 2003; Yeh et al., 2021). However, it is difficult to quantify the size and diversity of the latent reservoir, reasons include: relative abundance of the latent reservoir is low; assays are insensitive; and different quantification techniques systematically over or underestimate the diversity and size of the latent reservoir (Ferreira et al., 2021; Massanella and Richman, 2016).

Clonality is a measure of the occurrence of identical genetic sequences in a population (Ferreira et al., 2021; Hosmane et al., 2017; Joos et al., 2008). Clonality of the latent reservoir is used to infer the mechanism of latent reservoir maintenance. Higher levels of clonality are taken as evidence that clonal proliferation is the method for latent reservoir maintenance. (Hosmane et al., 2017). Tracking the genetic drift of a population, and the speed at which mutations become fixed in the latent reservoir can address conflicting establishment and maintenance hypotheses. The latent reservoir has been shown to become increasingly clonal over time, suggesting cellular proliferation to be the mechanism of maintenance (Cohn et al., 2015). In the “store and retrieve” hypothesis, HIV-1 variants seed the latent reservoir early in the infection, and is maintained through cellular proliferation. In support of the “store and retrieve” hypothesis, it has been reported that the vast majority of integration events occurs near the start of ART (Chavez et al., 2015, Abrahams et al., 2019; Brodin et al., 2016). Another possibility is that the latent reservoir could be seeded continuously during active infection (B. R. Jones et al., 2018). The latent reservoir could also be maintained through low level HIV activity in drug sanctuaries or environments otherwise permissive of viral replication. HIV activity in these sites results in ongoing infection and then latency of the reservoir (C. V. Fletcher et al., 2014). However it should be noted that apparent clonality can occur for reasons other than clonal proliferation. For example, ineffective ART can provide a strong selective force for a particular set of drug-resistant mutations. A ‘hard sweep’ event can occur, limiting the genetic diversity. Hard sweeps are selection events that result in fixation of mutations in a population (e.g. selection of antibiotic drug resistance genes in a population of bacteria after exposure to antibiotics). After this sweep, the population would appear to be less heterogeneous, even without clonal expansion (Feder et al., 2016).

There are clonality metrics that report clonality by quantifying the genetic diversity of the latent reservoir from observed quantities of genetic variation in the sequences.

3.1.2 Proportional Clonality

A proportional clonality score is the most frequently used method of calculating identicality in latent reservoir research (Wagner et al., 2014, von Stockenstrom et al., 2015). It can be calculated as the proportion of sequences that are identical to one other sequence in a sample (Eq. 3.1). In Figure 3.1., there are 5 individuals that share at least one identical sequence: a clonality score of 5/6 = 0.83 (von Stockenstrom et al., 2015; Wagner et al., 2014).

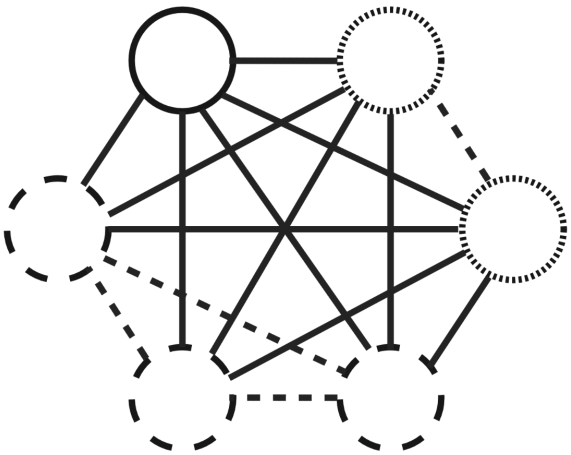


Figure 3.1: Circles that share the same border pattern are identical. Edges that connect two identical circles are dashed. Solid edges are non-identical comparisons.

The proportional clonality metric has limitations. In particular, proportional clonality statistics are biased as a result of incomplete sampling (Reeves et al., 2017), resulting in systematic underestimates of the size of the latent reservoir. This is because latent reservoir has a skewed distribution, where a small number of unique variants constitute a large proportion of the individuals in the reservoir (Hoehn et al., 2015). Therefore, there is an equal probability of sampling an unique sequence as sampling proportion increases. In particular, as sampling proportion increases, more is revealed about the genetic structure. Lower sampling proportion However, with the proportional clonality score, information about the underlying structure, such as number and size of polytomies are lost. Notably, in the proportional clonality metric, information about sequences are collapsed into a binary “singleton” vs. “not-singleton” count (Ferreira et al., 2021).

$$Proportional\ Clonality = \frac{\Sigma_i n_i I(n_i > 1)}{\Sigma_i n_i} \tag{3.1}$$

3.1.3 Gini coefficient

There are clear limitations to the proportional clonality scores. Therefore a score that takes into account the genetic diversity and structure of a sample is necessary. The Gini coefficient (G) is a method for measuring the distribution of some resource over of groups (Gini, 1936). It is typically used in economics, to estimate the distribution of wealth over the population. It has also been used in ecology and immunology research to measure the distribution of genetic variability over some distribution of individuals (Thapa et al., 2015, Sadras and Bongiovanni, 2004). The Gini co-efficient has been use to varied effects by (Bashford-Rogers et al., 2013; Hoehn et al., 2015) to estimate within host HIV genomic diversity. The Gini co-efficient ranges from 0 to 1 where populations with Gini coefficients closer to 0 have the most homogeneous distribution (most evenly dispersed genetic variation among individuals); populations with values closer to 1 have the most skewed distribution. G is calculated by taking the proportion of the space between the two curves in Fig. 3.2 ($\frac{i}{(i+ii)}$). In 3.2, the area under the dotted line (i + ii) represents perfect equality - where each sequence contributes an equal amount of genetic diversity in a population. The solid line curve separating areas i and ii is the Lorenz curve. The Lorenz curve describes inequality of genetic diversity in a population. For calculating genetic variablity, the cumulative share of individuals is plotted on the x axis, and the cumliative share of diversity is plotted on the y axis. Unlike the proportional clonality score, G is able to differentiate between a completely identical population, and a population that has two groups of identical sequences. For example, G for a population with the distribution [1,2,3] (G = 0.222) is lower than G for a population with the distribution [1,5] (G = 0.333). For both of these populations, the proportional clonality score would have been 0.833.

An approximation for the G can be found in Eq. 3.2 for population with n groups of y individuals, with values y_i to y_n, that are indexed in monotonically increasing order ($y_i \leq y_{i+1}$). This formula approximates the area of ii) in Fig. 3.2 instead of fitting a curve. G does not behave optimally as the number of unique sequences decreases, especially as it approaches 1. This is a result of the G approximation for discrete values. For a population that is completely identical,

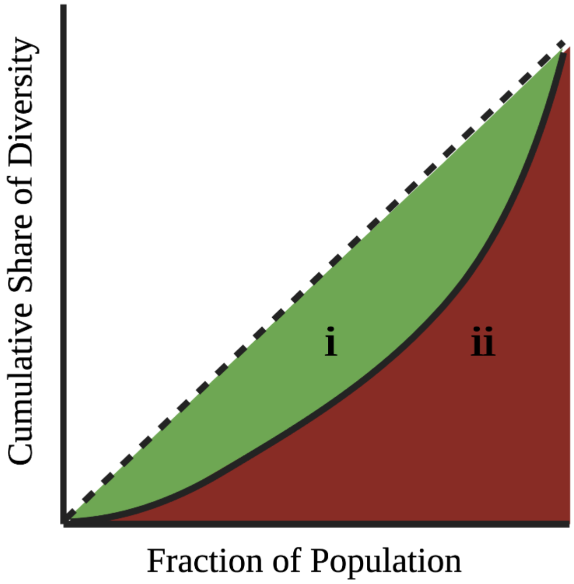


Figure 3.2: Graphical representation of the Gini co-efficient calculation. The Gini co-efficient is calculated by dividing area *i* by areas *i* + *ii*. For the approximation of the Gini co-efficient, *ii* is approximated by step wise discrete values.

G = 0. In general, as the number of unique sequences decreases the Gini coefficient starts to fluctuate wildly. G is dependent on the number of groups in the sample. For example consider two populations; one with the distribution [0,100] (G = 0.500); and the other [0,0,0,100] (G = 0.750). There is higher in-equality in the second population because there are a greater number of groups with 0, as compared to the one group with 100. This property of G is one reason that may not be ideal for calculating genetic diversity for a population that can be highly clonal. It is impossible for there to be a group with '0' members when using the G to measure sequence diversity.

$$G = \frac{1}{n}(n + 1 - 2\frac{\sum_{i=1}^{n}(n + 1 - i)y_i}{\sum_{i=1}^{n} y_i}) \tag{3.2}$$

3.1.4 Pairwise Clonality

Due to the drawbacks of proportional clonality and the Gini coefficient, I propose the pairwise clonality measure. The pairwise clonality calculation is shown in Eq. 3.3.

$$Pairwise = \frac{\sum_{i=1}^{n} \sum_{j>i} I(i, j)}{\binom{n}{2}} \tag{3.3}$$

where I(i,j) = 1 if sequence i == sequence j, otherwise 0

Plainly, the pairwise clonality statistic is the proportion of pairwise comparisons between sequences that are identical. It addresses some of the shortcomings identified with the previous clonality scores. The pairwise clonality score is a linear score. Unlike the Gini co-efficient, as a population becomes more homogeneous, the pairwise clonality score increases in value. The inverse is true - as a population becomes more heterogeneous, the pairwise clonality score decreases in value. The rationale behind this measure is that identity is not a attribute of a single individual. The comparison between two observations can be identical. A single observation cannot be identical in-itself. In addition, the pairwise clonality statistic has nice statistical characteristic. Each identity comparison - that is each comparison between two individuals – is an independent event from other comparisons. This provides a more reasonable statistical framework for evaluating clonality results.

3.2 Methods

3.2.1 Calculating Clonality

The FASTA files from the INDELible simulation stage were processed to calculate each aforementioned clonality metric (proportional clonality, Gini coefficient and pairwise clonality). For each FASTA file, the number of each unique sequence was counted. To automate sequence processing, FASTA files were processed in Python. Hashing is a deterministic algorithmic method that transforms data into a set length value. The Python implementation of the hash function (https://docs.python.org/3/library/ hashlib.html) transforms any input into a fixed length integer. The hash function implemented in Python ensures there is low chance of collision (two objects generating the same hash). Thus, hashing is a vastly superior alternative than an exhaustive linear comparison of two sequences. Each simulated sequence was hashed once, and a dictionary was created with the name of the first sequence processed for each unique sequence; the hashed value keyed by the unique sequence; and a count of how many times the hash was observed. The dictionary was used to calculate the proportional clonality score (Eq. 3.1) and the Gini co-efficient (Eq. 3.2). For the pairwise clonality score, a recursive function was written that considered every comparison possible between all sequences. For a population with n individuals, there are $\binom{n}{2}$ possible unique comparisons. After the pairwise comparison between individuals i_0 and i_1 has been considered, the comparison between individuals i_1 and i_0 does not need to be considered. The recursive function I wrote in Python considers all comparisons for an individual i_x indexed on x (where 0 < x < n), then moves on to consider i_{x+1}, removing individual i_x from consideration. The function recursively considers all individuals until none are left. As shown in Eq. 3.3, the pairwise clonality score is the proportion of all pairwise comparisons that are identical.

3.2.2 Regression Analysis

At its core, the purpose of linear regression is to predict a value y on the basis of another value x. The variable x is the '*predictor*' or '*independent*' variable whereas the variable y is the '*outcome*' or '*dependent*' variable. Linear regressions can rely on one or more independent variables to predict the value of the dependent variable. A linear regression is fit to the data by estimating coefficients of the independent variable(s). Regression coefficients are estimates of parameters, they quantify the relationship between the independent variable and the response variable. These regressions can be compared to determine what coefficients result in the best fit for the data. Eq. 3.4 shows the general formula for a multi-variable linear regression. To estimate "*y*", the coefficient "*b*" is calculated for *n* independent variables (*x*) ("Simple linear regression", 2005).

$$y = b_0 + \Sigma_i b_i x_i \tag{3.4}$$

Generalized Linear Models (GLM) are an expansion of linear regression. A key component of the GLM is the 'link function' that allows response variables to vary in a non-linear fashion (Nelder and Wedderburn, 1972). Another component of the GLM is allowing the distribution of response variable to vary (i.e. binary response to a linear regression (0,1)). For example using simple linear regression to predict a binary categorical response (i.e. clonal vs. singleton) would return nonsensical results, as linear regression makes the assumption that y is a linear function of x. Instead a binomial distribution link function can be used in a GLM. A variety of link functions can be used in a GLM; Poison, Binomial, Bernoulli to name a few.

3.3 Results

3.3.1 Analyzing Clonality Scores

A benefit of simulations is the ability to control the input parameters. Unlike real-life experiments, simulations can be tuned to test specific hypotheses. As discussed previously, transmission and transition rates can be controlled in the simclone framework. Identifying the strength of relationship between transmission/transition rates and clonality scores is possible through GLM analyses. Identifiability of parameter values from clonality scores (e.g. estimating the levels of clonal expansion from the clonality score alone) can provide validation for analyses that use clonality scores to infer the method of latent reservoir persistence.

Identifiability of parameter values using clonality metrics (proportional, Gini co-efficient, pairwise coefficient) was tested by fitting GLM on simulation results. Increasing clonal expansion the rate of clonal expansion should be correlated with increased number of identical sequences (clones). Analysis of simulated sequences reveal that this is not the case for all clonality measures.

A GLM was fit with the three clonality scores as the response variable with all simulation parameter values as independent variables. The link function used to fit the GLM was the binomial family. Proportional Clonality was not significantly associated with increased clonal expansion when all parameters were varied (p = 0.324) nor when only a subset of parameters were varied [Active to Active infection rate, Active to Latent infection rate, Clonal Expansion] (p = 0.846). The Gini coefficient was significantly associated with increased clonal expansion when all parameters were varied ($p = 1.0\times10^{-4}$), but not when only the same subset of parameters as above were varied (p = 0.242). The pairwise clonality score was significantly associated with increased clonal expansion when all parameters were varied (p < 2e-16) and also when only the same subset of the parameters were varied (p < 2e-16). Parameters that introduce or propagate viral lineages to the latent reservoir were chosen for the subset. The subset was used to reduce simulation noise to test associations. Other associations were tested using the GLM method as well.

Figures 3.3, 3.4, and 3.5 show the three clonality scores (proportional, Gini and pairwise) calculated for the same simulation run. All simulation parameters were varied. The number of simulated sequences for each replicate is 100. Thus the possible range of clonality scores are: proportional clonality (0 - 100); Gini co-efficient (0 - 1); and pairwise clonality (0 - 1). The Gini co-efficient was significantly associated with Active to Active infection rate (p = 0.006), and Latent to Latent infection rate (p = 0.0001) when all parameter values were varied. When only a subset of simulation parameters were varied, only the Active to Active infection rate was significant at the $\alpha = 0.05$ level (p = 0.007).

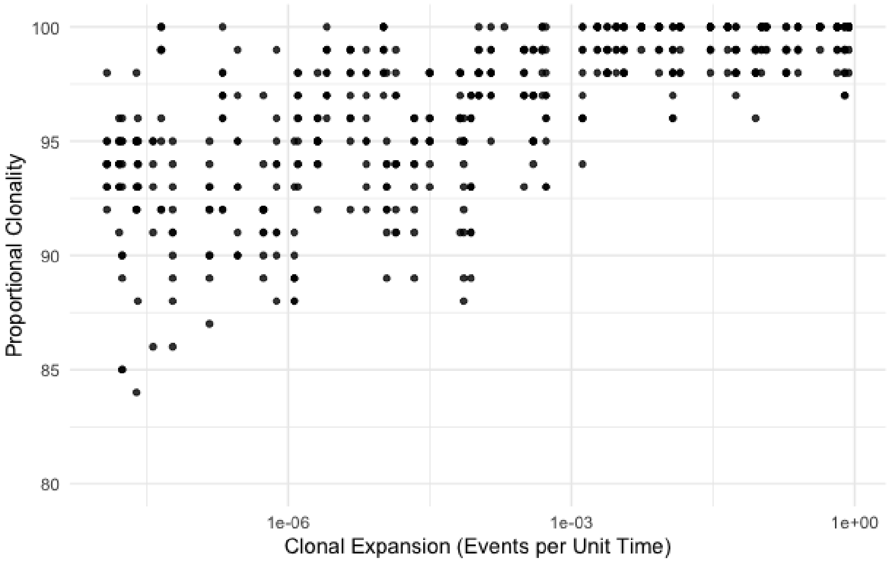


Figure 3.3: Proportional clonality score drifts upwards as clonal expansion rate was increased. Clonal expansion parameter is on the x-axis, log-normal scaled between 0 and 1. X values represent clonal expansion events per simulation unit time. Y-axis is proportional clonality score, possible values range between 0 and 100. Each point represents one simulation replicate. All simulation parameter values were varied for this simulation run. Proportional clonality was not significantly associated with clonal expansion p = 0.324.

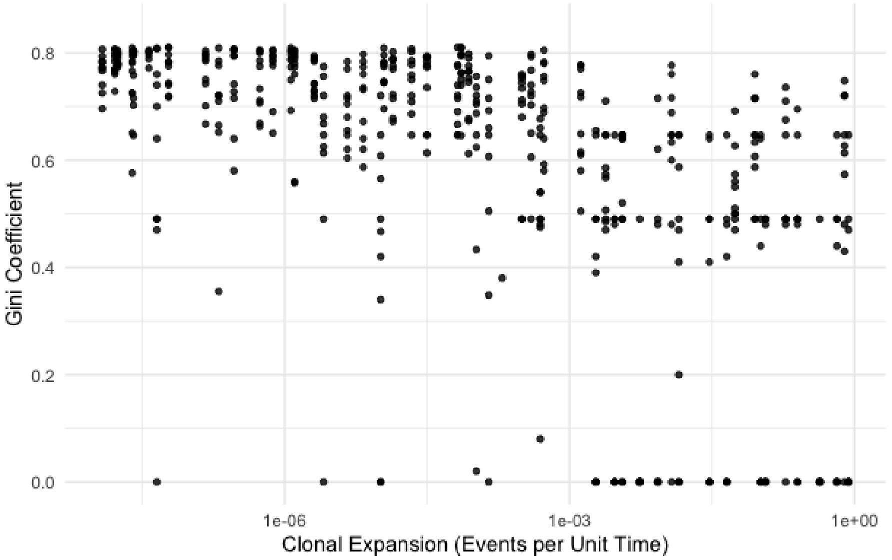


Figure 3.4: Gini coefficient is inversely correlated with clonal expansion. Clonal expansion parameter is on the x-axis, log-normal scaled between 0 and 1. Y-axis is Gini coefficient score, possible values range between 0 and 1. All simulation parameter values were varied for this simulation run. Gini coefficient was significantly associated with clonal expansion when all parameter values were varied p = 0.0001.

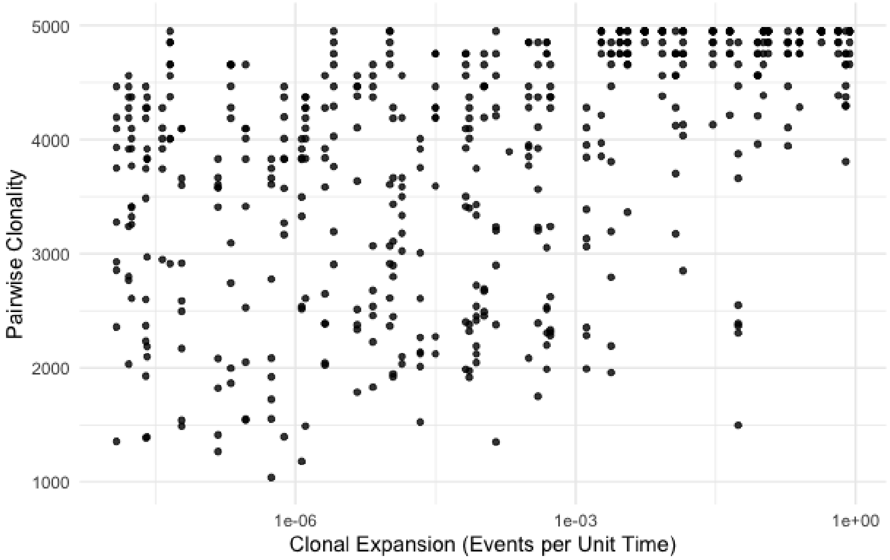


Figure 3.5: Pairwise clonality is correlated with clonal expansion. Clonal expansion parameter is on the x-axis, log-normal scaled between 0 and 1. Y-axis is pairwise clonality score, possible values range between 0 and 4950. All simulation parameter values were varied for this simulation run. Pairwise clonality was significantly associated with clonal expansion when all parameter values were varied p ¡ 2e-16

	Adjusted Odds Ratio (95% CI)		
Simulation Parameter	Proportional Clonality	Gini Coefficient	Pairwise Clonality
(Intercept)	1.53e+01 (9.72e-02, 3.23+03)	1.95 (2.88e-01, 1.34e+01)	**1.19e+01 (1.15e+01, 1.23e+01)**
Initial Active-type Population	1.0 (1.0, 1.0)	1.0 (1.0, 1.0)	1.0 (1.0, 1.0)
Initial Latent-type Population	1.0 (1.0, 1.0)	1.0 (1.0, 1.0)	1.0 (1.0, 1.0)
Initial Replenishment Population	1.0 (1.0, 1.0)	1.0 (1.0, 1.0)	1.0 (1.0, 1.0)
Active-to-Active Transmission	Inf. (0, Inf.)	0 (0, 0)	Inf. (Inf., Inf.)
Active-to-Latent Transmission	0 (0, Inf.)	Inf. (0, Inf.)	0 (0, 0)
Clonal Expansion	61.42 (4.39e-01, 1.543e+08)	**9.62e-02 (2.72e-02, 3.08e-01)**	**134.20 (127.36, 141.08)**
Active-to-Latent Transition	1.91e-01 (0 , Inf.)	1.44e+221 (4.36e-113, Inf.)	0 (0, 0)
Latent-to-Active Transition	Inf. (0, Inf.)	Inf. (0, Inf.)	0 (0, 0)
Active Death Rate	3.256e+05 (2.59e-85, 1.71e+94)	1.60e-21 (1.61e-54, 8.89e+11)	**1.68e+19 (4.71e+18, 6.01e+19)**
Latent Death Rate	3.12e+11 (4.58e-79, 1.09e+102)	1.21e-06 (6.32e-40, 1.96e+27)	0 (0, 0)
Active Replenishment	Inf. (0, Inf.)	Inf. (0, Inf.)	0 (0, 0)
Latent Replenishment	0 (0, 0)	0 (0, 0)	0 (0, 0)

Table 3.1: No parameters were significantly associated with clonality across all three clonality measures. Bolded adjusted Odds Ratios indicate point estimates with 95% CI intervals that do not overlap OR = 1. Adjusted odds ratios were calculated with GLM (binomial link function) fitted to simulation data with all parameters varying with clonality measures as the response variable.

	Adjusted Odds Ratio (95% CI)		
Simulation Parameter	Proportional Clonality	Gini Coefficient	Pairwise Clonality
(Intercept)	**2.08e+02 (1.02e+01, 1.10e+04)**	**4.75-01 (3.01e-01, 7.90e-01)**	**3.89 (3.85, 3.92)**
Active-to-Active Transmission	7.48+03 (7.54e-02, NA)	**2.21e-1 (6.85-02, 6.25e-01)**	**1.62 (1.60, 1.65**)
Active-to-Latent Transmission	0 (0, Inf.)	Inf. (1.57e-87, Inf.)	0 (0, 0)
Clonal Expansion	Inf. (0, Inf.)	0 (0, 7.13e+154)	Inf. (Inf., Inf.)

Table 3.2: When only a subset of parameters were varied, none of the parameters avoided overlapping AOR = 1 for all three clonality measures. Bolded adjusted Odds Ratios indicate point estimates with 95% CI intervals that do not overlap OR = 1. Adjusted odds ratios were calculated with GLM (binomial link function).

Adjusted odds ratios (AOR) were calculated for each simulation parameter value that was varied for both of the above GLMs. Adjusting the odds ratios was necessary to adjust for confounding variables. AOR 95% Confidence Intervals (95% CI) that do not span AOR = 1 indicate significant predictor variables. When all parameters were varied, no simulation parameter was significant for all three clonality measures (Table 3.1). No parameter Confidence Interval spanned the AOR = 1 for the proportional clonality score. For G, only the clonal expansion parameter was significant. The clonal expansion parameter was negatively associated with the Gini coefficient. For the pairwise clonality score, both the clonal expansion parameter and the Active Death Rate parameters were significantly associated with increases in pairwise clonality.

The GLM analyses show clonality scores that take into account the population structure (i.e. Gini coefficient and pairwise clonality) perform better at predicting changes in clonal expansion than clonality scores that do not (proportional clonality). The discrepancy maybe due to the inability for the proportional clonality score to take into account the difference between a completely identical sample and a sample where there are several groups of non-unique sequences. In Fig. 3.1, the proportional clonality score would give a clonality score of 0.83; the Gini coefficient would give a clonality score of 0.22; and the pairwise clonality score is 0.27. The proportional clonality score, or any other proposed clonality measure that does not consider the difference between groups of non-unique sequences, would lose vital information about the genetic population structure. The empirical distribution of proportional clonality scores is extremely narrow. These clonality values range from 84 to 100, with a mean of 96.5 and an interquartile range of 94 - 99 in Fig. 3.3. Since proportional clonality is an integer count value, only 5% of possible values contain 50% of the observations. The reduced granularity and inability to differentiate between common clonality scenarios severely limits the effectiveness of the proportional clonality score.

3.4 Discussion

The Gini coefficient could differentiate between the ambiguous clonality scenario that the proportional clonality score fails to address. Gini coefficient values ranged from 0 to 0.81, with a mean of 0.57 and a standard deviation of 0.27. The IQR range of Gini coefficient values was 0.49 to 0.773. As previously demonstrated, the Gini coefficient was able to differentiate between the scenarios that the proportional clonality score failed to do. The GLM results show that the Gini coefficient was significantly associated with the clonal expansion parameter. The main flaw with the Gini coefficient stems from a property of identity: the number of unique individuals in a population decreases as the population becomes more homogeneous. These results mirror the findings of (Ferreira et al., 2021; Hoehn et al., 2015), who found noted that as sampling fraction increased, the number of singletons likewise increased, resulting in higher Gini coefficients. However, unlike Hoehn et al., 2015, there were significant associations between clonality scores and simulation parameter values - with the caveat that Hoehn et al. were using attempting to fit clinical variables, not simulation parameters. The simclone parameters may be more identifiable then clinical variables because of system complexity. It can be reasonably supposed that the simclone framework has less complexity, and opportunity for variability than real-world data due to simplifying assumptions. The number of unique sequences has a disproportionate effect on the G, especially as the number of unique individuals approaches 1. For example consider the following 3 populations: population a with the distribution [98,1,1; G = 0.647]; population b with the distribution [99,1; G = 0.490]; and population c with the distribution [79,16,5; G = 0.491]. The Gini coefficient would suggest that population *a* is more unequal than populations *b* and *c*; and population *b* and *c* are approximately unequal to the same extent. The Gini coefficient loses specificity as the number of unique sequences drops. The Gini coefficient is a measure of inequality across unique groups, as the number of unique groups decreases, the level of inequality decreases for remaining groups. In this way, it falls into a similar trap as the proportional clonality score. There exists multiple population distributions with very different population structures that can be represented by the same Gini coefficient. Beyond the collision of Gini coefficients for distinct distributions, as the number of unique sequences decreases, the Gini score does not decrease in a linear fashion. The Gini coefficient for a completely clonal population is 0; whereas as the population approaches complete clonality, the Gini coefficient increases. The population is becomes more unequal as the membership to unique sequence groups increases. These two issues significantly limit the usefulness of the Gini coefficient in situations where the a population starts at, or nears complete clonality.

The pairwise clonality measure considers the genetic structure of a population, while maintaining a stable denominator that does not vary with the homogeneity/heterogeneity of a population. The denominator of the pairwise clonality measure is set by the number of possible pairwise comparisons between all individuals in a population $\binom{n}{2}$, where n is the size of the population. The pairwise clonality measure can capture the changes as homogeneity increases, and many small unique groups merge into larger unique groups. Considering the same 3 populations as before: population a with the distribution [98,1,1; pairwise = 0.96]; population b with the distribution [99,1; pairwise = 0.98]; and population c with the distribution [88,15,7]; pairwise = 0.80]. Population a and b have a much more similar population structure, and are much closer to homogeneity than population c. As shown in this example, the pairwise clonality score is able to capture the not-so-subtle differences in population structures that the Gini coefficient and the proportional clonality scores cannot differentiate between. The pairwise clonality score also has a nice statistical property. Identity is not a trait of one object – one object cannot be identical. Identity is a trait of the relationship between two objects. As such, the pairwise clonality score is able to comment on identity in a internally-valid way. It has the added bonus of ensuring each observation is a quasi independent measure. This provides statistical robustness to following statistical analyses. The pairwise clonality measure also addresses some of the short comings of the Gini coefficient identified by (Bashford-Rogers et al., 2013; Ferreira et al., 2021; Hoehn et al., 2015). For the Gini coefficient, as the read depth increases, the number of clones in the large clones increases, while the number of unique singletons also increases. The skew of the clonal distribution increases; thus increasing the Gini coefficient (Hoehn et al., 2015). However with the pairwise clonality statistic, the addition of individuals to large clonal groups, or as singletons does not cause as large variations in the clonality statistic. These analyses suggest that clonality measures that consider the population structure would be more robust in capturing changes in the rates of clonal expansion. Whereas the Gini coefficient may still be a flawed measure, it may still provide usefulness when analysing heterogeneous in populations, as in ecology or immunology research (Sadras and Bongiovanni, 2004; Thapa et al., 2015). There may be situations where the Gini coefficient or the proportional clonality score are a better fit for identifying trends in clonality. Therefore, the goal of the analysis of the latent reservoir data (i.e. sequence data) should be considered when choosing the best method of analyses.

Chapter 4

Particle Filtering

4.1 Background

General Linear Models are limited, because they require prior knowledge of the type of distribution and relationship between independent and dependent variables to establish associations. Due to complexity of the *simclone* model, it is difficult to establish these relationships, therefore I employed the particle filtering method. The particle filtering method is a simulation based method of fitting a model to observed data using summary statistics. In the domain of phylogenetics, this particle filtering method has been used to fit complex population dynamic models to genealogies; and to infer phylodynamic parameters (e.g. reproductive number) from genetic sequence data without fitting a phylogenetic tree (Park et al., 2021; Rasmussen et al., 2011). The approach that I took to particle filtering mirrors the approach of the two aforementioned papers. However, calculating likelihood for this complex model is too difficult, as there are too many parameters to consider. The particle filtering method approximates the posterior distribution of a model given a noisy observation.

My implementation of the particle filtering method uses a truth simulation run and an experiment dataset simulated using samples drawn from the same distribution that the truth sample was drawn from (see 4.2 for the experimental design). The clonality scores for the truth sample is taken as the true clonality values for each of the three clonality scores - [Proportional, Gini, and pairwise]. These random simulation runs are taken as the 'truth', as we don't have actual outcome data with the ground truth. The truth samples are used as the target for the particle filtering method. The experimental dataset consists of *x* replicates for each sample with each of the three clonality values calculated for each replicate. Each replicate is considered a particle. The Root Mean Squared Error (RMSE) is calculated using Eq. 4.1 for each of the particles; where *Exp* are experimental clonality scores; and *Targ* is the target clonality score.

$$RMSE = \sqrt{\frac{1}{n}\Sigma_{i=1}^{n}\left(\frac{Exp_i - Targ}{\sigma_i}\right)^2} \tag{4.1}$$

The resulting RMSE scores can be interpreted as an objective (cost) function, where a particle with a RMSE score of 0 against a target represents an identical comparison. Lower RMSE scores represent a clonality score closer to the target simulation. Fig. 4.1 shows the ideal scenario. In the ideal scenario, the RMSE is a quadratic curve, centered around the true parameter value. Values that are further away from the target are penalized exponentially. The exponential penalty should be beneficial to particle filtering. RMSE was also chosen for proof-of-concept testing since the measure has a for its low computational cost, due to the high number of comparisons involved. However, the cost function for particle filtering I have implemented can be changed to a Bayesian framework as in Rasmussen et al., 2011 and Park et al., 2021.

I treated each replicate in the experimental dataset as a particle. After a RMSE score is calculated for each particle, the particles are filtered, removing the upper 50% of particles. In the filtering step, the median RMSE score for all replicates in the experiment dataset is calculated, and any value above the median RMSE score is removed. The particle filtering method is outlined in Fig 4.2.

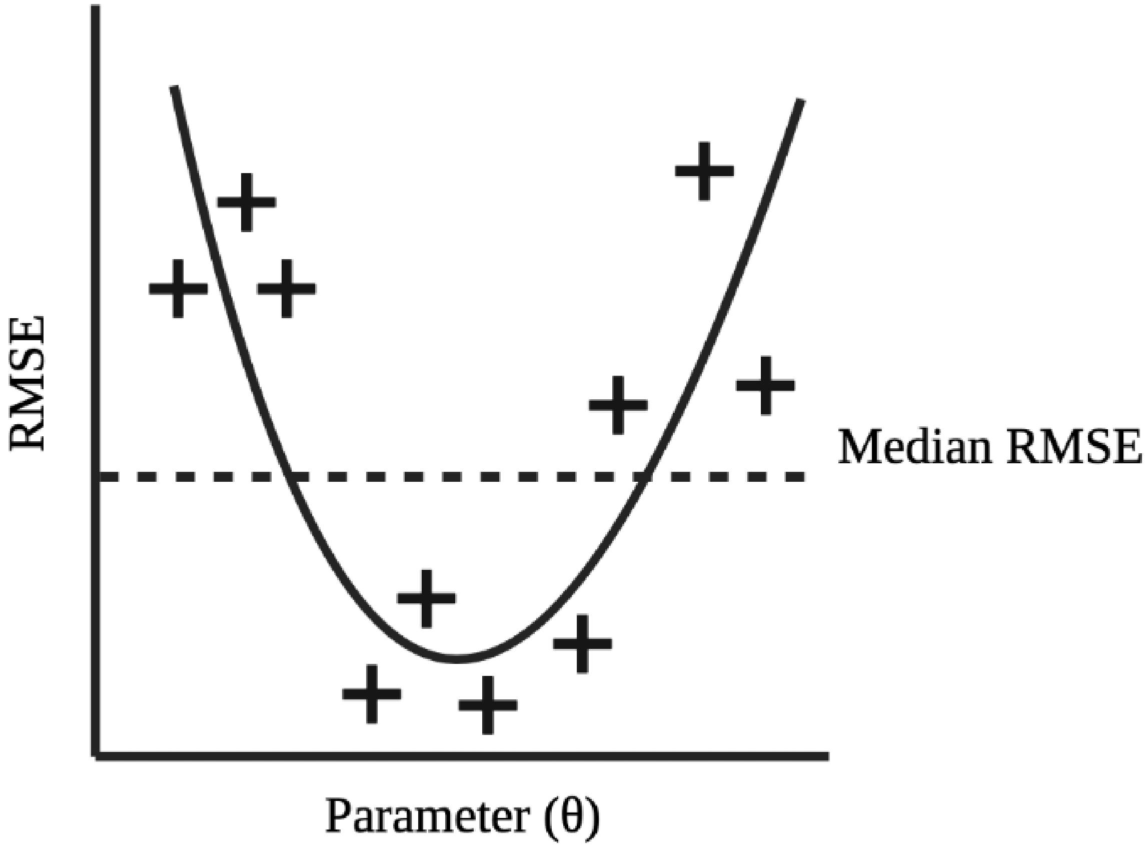


Figure 4.1: Visualization of the RMSE particle filtering method. Each (+) represents a particle. In the ideal scenario, the RMSE values decrease to a minima for particles close to the true parameter value. The dotted line shows the median RMSE score; particles above the median are filtered (removed) when estimating parameter scores.

4.2 Method

The target samples and the experiment samples are drawn from the same parameter distribution using the LHS method discussed in chapter 2. The parameter values are normally distributed across intervals initially drawn from literature (but adapted to the simclone framework). Target and experiment samples were sampled using the LHS method to ensure that the chosen targets were representative of the overall parameter space. As described in chapter 2, the LHS algorithm ensures sampling from the entire parameter range. This is important for the particle filtering method, because of the possibility that discrete regions of parameter space could be in-

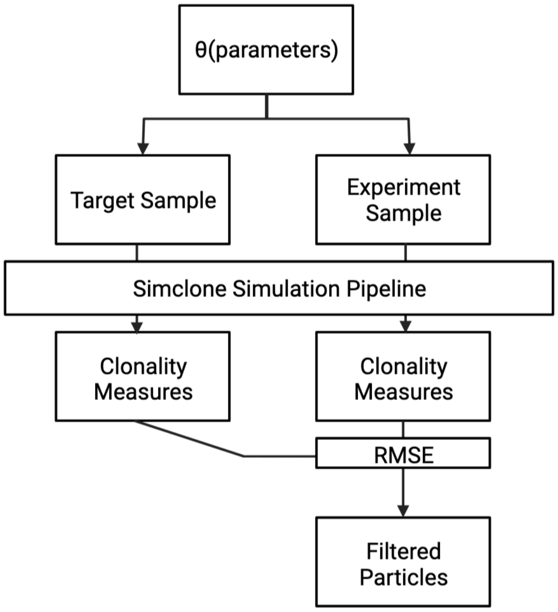


Figure 4.2: The particle filtering method is an experimental design that sits on top of the simclone simulation pipeline. The filtered particles are samples in the lower 50% of RMSE values. Ten target samples were drawn (1 replicate per sample simulated), and 1000 experiment samples were drawn (10 replicates per sample)

formative; multiple regions of the parameter space could have low RMSE values. Using varied parameters is important for the target dataset, as portions of the parameter space could conceivably have unique interactions. The experiment simulations were also drawn under varying model parameter values to calculate a pseudo-likelihood surface based on the RMSE of clonality measures between target and experimental samples. The RMSE surface approximates a posterior probability, without needing to calculate the likelihood of the models. This surface was considered for each parameter - i.e. were there regions of parameter space that had lower RMSE values when considering RMSE against the parameter? The target sample and experiment samples were simulated through the simclone simulation pipeline normally without any additional modification: i) twt simulation specification files were generated; ii) twt simulations were conducted on the computing cluster; iii) twt outputs (i.e. lineage tree) were modified to shorten branch lengths in the latent reservoir; iv) INDELible was used to simulate nucleotide evolution on twt outputs; and v) clonality scores (proportional clonality, Gini coefficient, pairwise clonality) were calculated using the method discussed in chapter 2. The RMSE values were calculated for each of the targets using Eq. 4.1. The filtering step did not necessarily filter the same particles (replicate) for each target considered, nor for each clonality score. The RMSE values were appended to the same dataframe with the parameter settings and the clonality scores.

4.3 Results

There were ten targets considered with this method. These targets were drawn from the same distribution as the experiment samples using the LHS method. The targets were equally distributed across all parameters. The RMSE values for each particle was calculated using the described methods above. Each particle had 10 RMSE values for each clonality score: one for each simulation against a target. Fig. 4.3 shows example histograms for the distribution of clonality RMSE values for particles against the target.

Histograms show multi-modal distributions for RMSE values across all clonality measures. However these distributions are highly varied, and are dependent on the target. Histograms are included to highlight the difficulty of choosing a RMSE threshold. Static thresholds were difficult to parameterize, as the range and distribution of RMSE values could greatly vary between targets. The median RMSE value was chosen as the cutoff, to maintain the same number of particles for each comparison across targets and clonality statistics.

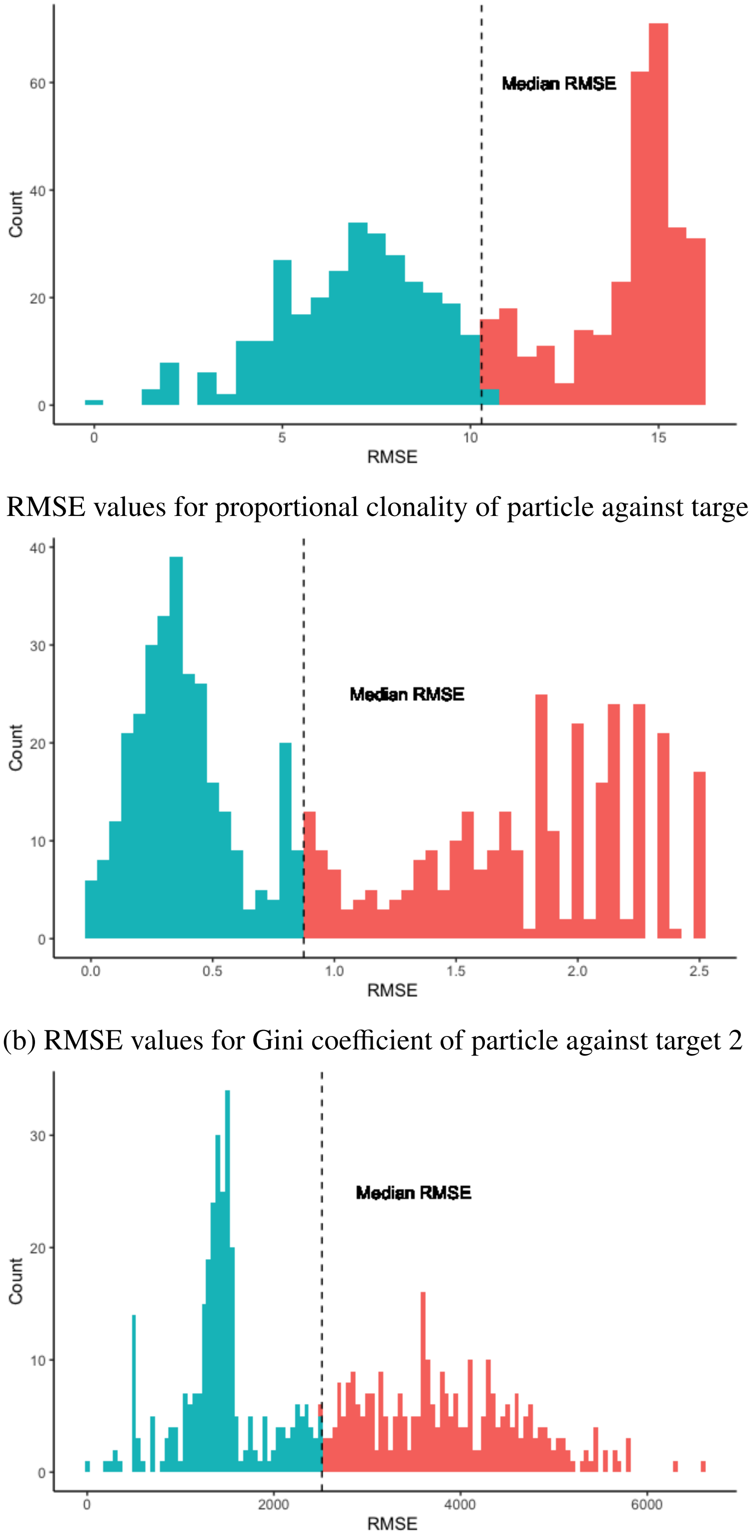
As described previously, particles with an RMSE score above the median RMSE value are filtered out. Table 4.1 shows the median particle RMSE value for each target, given its target clonality score. Median value represents the maximum deviation of the filtered particles from target clonality, as any particle with an RMSE below that value is filtered out. Proportional clonality trends towards lower median particle RMSE when target proportional clonality score is high (Fig. 4.4a). Median particle RMSE follows a quadratic trend as target pairwise clonality and Gini coefficient increase (Fig. 4.4c and 4.4b, respectively).

Proportional clonality is an integer count value that is bounded by 0 - 100 for this ex-

Target	Proportional Clonality		Gini Coefficient		Pairwise Clonality	
	Median Particle RMSE	Target Clonality	Median Particle RMSE	Target Clonality	Median Particle RMSE	Target Clonality
1	13.12	94	0.901	0.797	2444	4371
2	4.95	100	1.544	0.000	3334	4950
3	4.95	100	0.900	0.440	2449	4386
4	4.95	100	1.900	0.440	4278	2931
5	15.92	93	0.903	0.798	3182	3756
6	4.95	100	1.494	0.020	5479	2454
7	5.66	98	0.754	0.706	4308	2917
8	10.30	95	0.845	0.768	3716	3229
9	4.95	100	1.348	0.080	5322	2514
10	4.95	100	1.398	0.060	7776	1647

Table 4.1: Median RMSE values of particles against their target are presented for targets (n = 10) for proportional clonality, Gini coefficient and pairwise clonality. The same target is used across the row, but differ between rows. The same particles are considered across all rows.

periment. There are fewer potential outcomes for proportional clonality scores than for Gini coefficient and pairwise clonality. Proportional clonality collapses many population distributions into a single score (i.e. [(99, 1); (54, 45, 1); and (60, 30, 4, 3, 2, 1)] are all assigned the proportional clonality score of 99). The reduction in variability, and thus lower RMSE score may reflect this process. Therefore, comparing RMSE values between targets given a type of clonality calculation (proportional, Gini, pairwise) is possible, but comparing RMSE values between clonality measures for a single measure is not. RMSE values for different clonality statistics cannot be compared because RMSE is a scaled measure; it is dependent on the magnitude of the target and particle clonality values (Exp_i and $Targ$). The clonality statistics operate on different scales (proportional clonality from 0 - 100; Gini coefficient 0 - 1; and pairwise clonality 0 - 4950), and therefore RMSE values are not comparable. However, comparing particle RMSE scores for various targets can be informative. Given that clonality metrics respond in some fashion to simulation parameter(s), comparing RMSE values for particles can allow estimation of the target parameter values from the particles. The median RMSE values does not necessarily represent how identifiable parameters are, it simply represents the distribution of the clonality score for the 50% of particles closest to a given target.

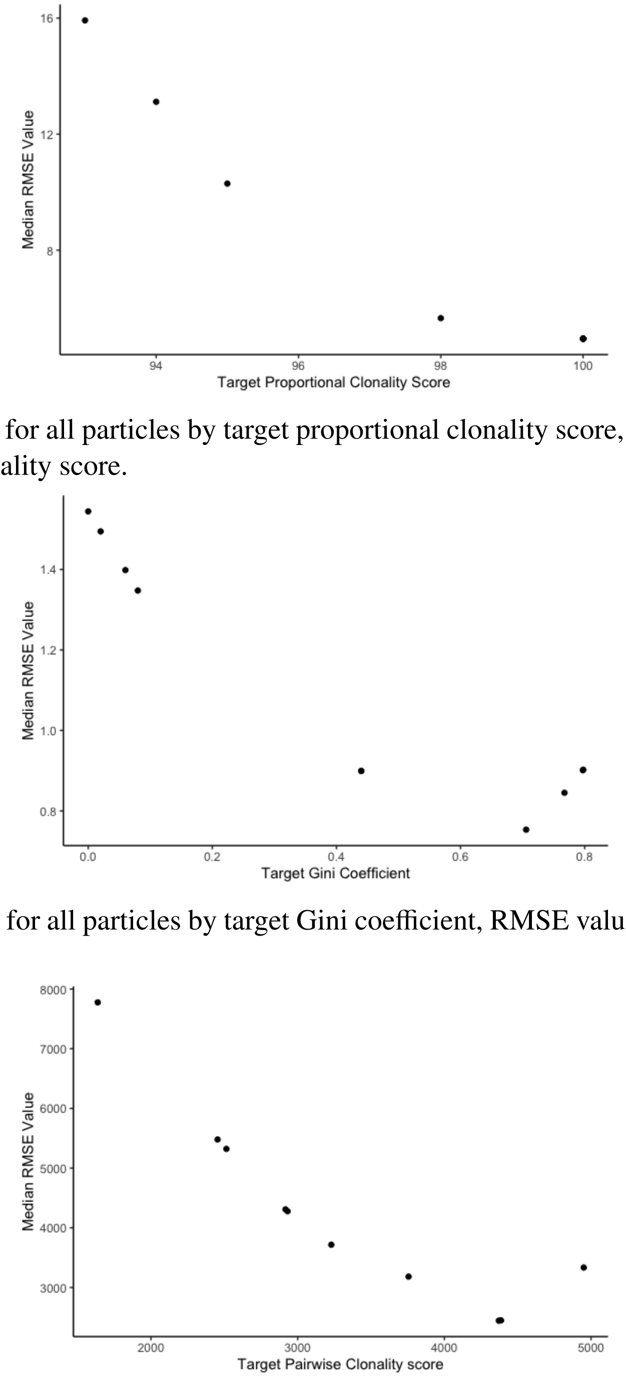


(a) RMSE values for proportional clonality of particle against target 2

(b) RMSE values for Gini coefficient of particle against target 2

(c) RMSE values for pairwise coefficient of particle against target 2

Figure 4.3: Histogram of RMSE values for clonality scores of particles against the same target. A,B, and C show the same particles measured against the same target. Dashed horizontal line shows the median RMSE value of all particles. Red are particles with an RMSE value above the median for this target. Blue are particles with an RMSE value below the median for this target



(a) Median RMSE value for all particles by target proportional clonality score, RMSE values calculated for the proportional clonality score.

(b) Median RMSE value for all particles by target Gini coefficient, RMSE values calculated for the Gini coefficient.

(c) Median RMSE value for all particles by target pairwise clonality score, RMSE values calculated for the pairwise clonality score

Figure 4.4: Median RMSE decreases as the target clonality increases for proportional clonality (a). Median RMSE generally decreases to an inflection point for (a) Gini coefficient and (b) pairwise clonality. Scatters show Median RMSE score for each target particle, per clonality statistic (a) Proportional Clonality (b) Gini coefficient (c) Pairwise clonality. Points are the median RMSE value for a target, plotted against the target's actual clonality value.

To test parameter identifiability for a given target, I calculate the mean of the parameter of interest for only filtered particles. This mean value is the estimate of the true parameter value. The median RMSE value is calculated for each combination of target and clonality metric, and a rank value is appended for each RMSE value (1 for below, 0 for above median). For the filtered particles, the mean of each parameter is taken as the parameter estimate for the filtered particles. Fig. 4.5 is a visualization for the process. Red points indicate a particle that have been excluded, blue points indicate particles remaining. The vertical red line on a) and b) is the actual parameter value. Fig. 4.5 shows the ideal scenario, where the true parameter value is close to the parameter estimate (mean parameter value of filtered values). Particle filtering visualizations for all targets are grouped by clonality statistic in the supplementary files.

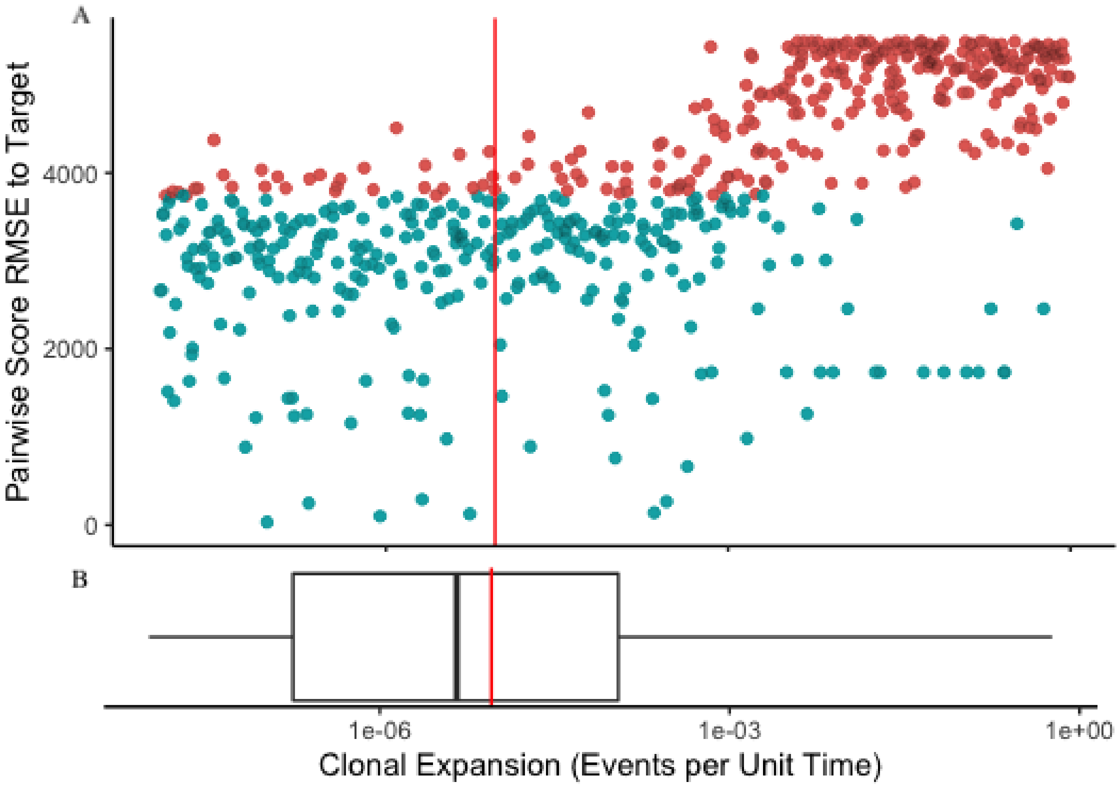
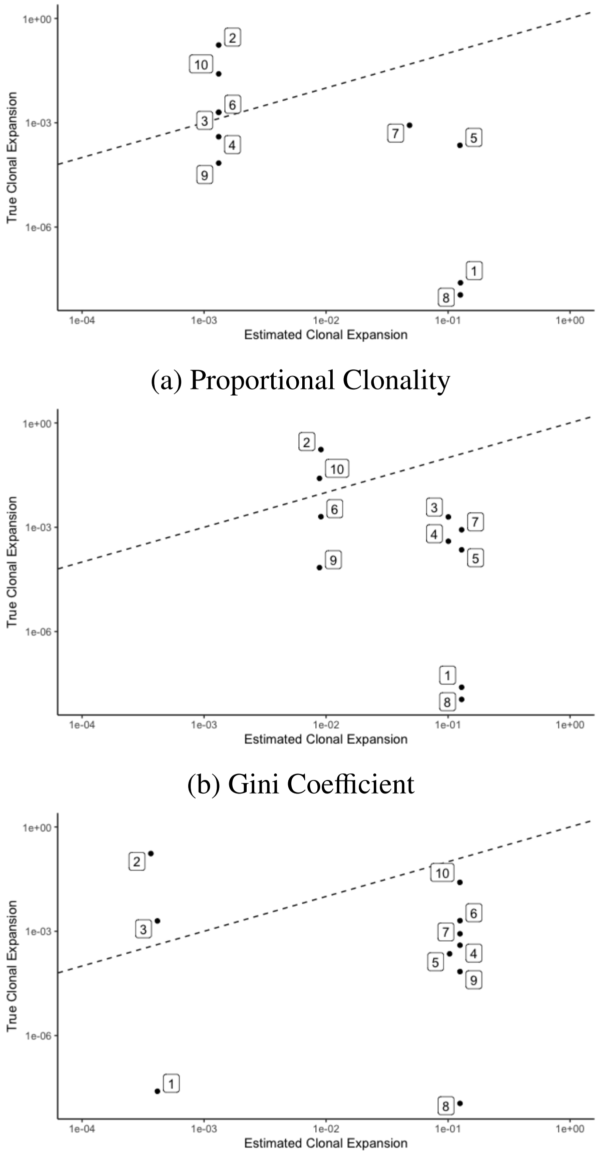


Figure 4.5: **A)** RMSE values for all particles plotted against the Clonal Expansion (ll_br) parameter. Particle colour indicates whether the particle has been filtered. Blue coloured particles are the remaining particles, with RMSE values in the lower 50%. **B)** Distribution of the Clonal Expansion parameter for the filtered particles in A. The red vertical lines show the target Clonal Expansion parameter.

Fig. 4.5 was the exception, rather than the rule for this method. The validity of the particle filtering method was investigated by visual inspection. In most combinations of the target and clonality measure, the parameter estimate was uninformative. Either the true parameter value for the target simulation lay outside of even the IQR (Fig. 4.7a), or the distribution of high RMSE and low RMSE values did not differ with respect to parameter of interest (Fig. 4.7b). Identifiability of the Clonal Expansion parameter was most thoroughly investigated. Out of 10 targets particle filtering method was able to capture the true parameter value for: 2 of the targets using proportional clonality; 3 of the targets using the Gini coefficient; and 2 of the targets using pairwise clonality (See Table. 4.2 for estimates). Box-and-whisker plots showing parameter estimates using each clonality metric, grouped by target are shown in the supplementary section (Fig. A.31). Scatter plots showing point estimates for Clonal Expansion against true parameter value for Clonal Expansion are shown in Figure (4.6. Dotted line on the figures (with slope y = x) represent the line of perfect prediction. Points closer to this line are targets where predicted Clonal Expansion values were closer to the true (target) Clonal Expansion value.

Target	True Target Clonal Expansion	Particle Parameter Estimate (IQR)		
		Proportional Clonality	Gini Coefficient	Pairwise Clonality
1	2.45e-08	1.26e-01 (1.74e-03, 1.48e-01)	1.29e-01 (2.06e-03, 1.56e-01)	4.14e-04 (2.00e-07, 1.10e-04)
2	1.72e-01	1.32e-03 (1.40e-07, 4.06e-05)	9.08e-03 (2.00e-07, 7.73e-05)	3.66e-04 (1.80e-07, 9.99e-05)
3	1.98e-03	1.32e-03 (1.40e-07, 4.06e-05)	1.01e-01 (9.70e-06, 6.87e-02)	4.14e-04 (2.00e-07, 1.10e-04)
4	3.96e-04	1.32e-03 (1.40e-07, 4.06e-05)	**1.01e-01 (2.06e-03, 6.87e-02)**	1.25e-01 (1.52e-03, 1.46e-01)
5	2.24e-04	1.25e-01 (1.48e-03, 1.46e-01)	1.29e-01 (2.06e-03, 1.56e-01)	1.03e-01 (5.03e-05, 8.01e-02)
6	2.01e-03	1.32e-03 (1.40e-07, 4.06e-05)	**9.08e-03 (2.00e-07, 7.73e-05)**	1.25e-01 (1.48e-03, 1.46e-01)
7	8.45e-04	**4.83e-02 (3.00e-07, 1.42e-02)**	1.29e-01 (2.06e-03, 1.56e-01)	**1.25e-01 (1.52e-03, 1.46e-01)**
8	1.12e-08	1.26e-01 (1.75e-03, 1.47e-01)	1.29e-01 (2.06e-03, 1.56e-01)	1.25e-01 (1.53e-03, 1.46e-01)
9	6.91e-05	1.32e-03 (1.40e-07, 4.06e-05)	8.83e-03 (2.00e-07, 7.12e-05)	1.25e-01 (1.48e-03, 1.46e-01)
10	2.54e-02	**1.32e-03 (1.40e-07, 4.06e-05)**	**8.83e-03 (2.00e-07, 7.12e-05)**	**1.25e-01 (1.03e-03, 1.46e-01)**

Table 4.2: Clonal Expansion estimate for each each target is reported for each clonality statistic. The true target Clonal Expansion parameter is reported as well. Bolded estimates represent estimates where the true parameter fell within the IQR

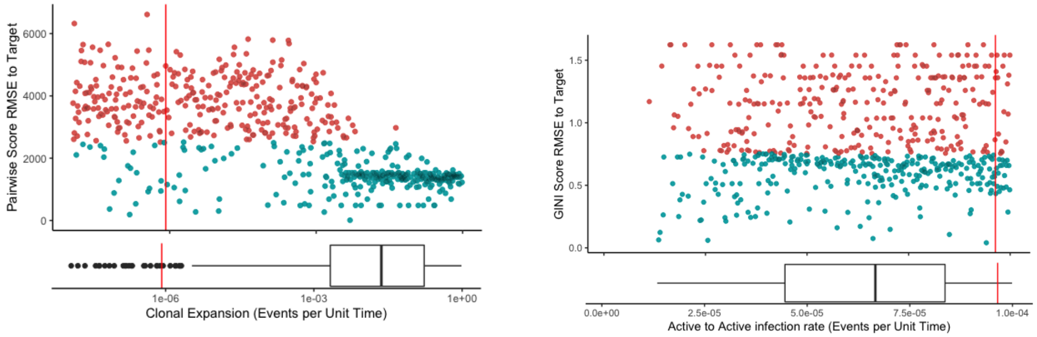


(a) Proportional Clonality

(b) Gini Coefficient

(c) Pairwise Clonality

Figure 4.6: True (Target) Clonal Expansion values are shown on the Y axis, with point estimates (mean Clonal Expansion value) for Clonal Expansion using particle filtering method on the X axis. Dotted line is a reference 'perfect prediction' line (slope y = x). Each point represents the prediction for a given target (labeled with target number), points closer to the prediction line are better estimates



(a) True Clonal Expansion falls squarely outside of the IQR range.

(b) No difference in particle distribution (besides RMSE values)

Figure 4.7: Examples of uninformative particle filtering. Particle Filtering visualization using same technique as 4.5. The target value for Clonal Expansion is denoted by the vertical red line.

4.4 Discussion

Particle filtering yielded poor results for identifying parameter settings of target simulations when using RMSE as a scoring measure and the median RMSE value as the threshold for simulation parameters other than Clonal Expansion. Uninformative particle filtering scenarios could be classified into two types of scenarios. Targets where the parameter e Targets in the former scenario could simply represent targets for which the method would not estimate well (e.g. the targets were outside the reasonable range for parameter estimation). A lack of difference with respect to the parameter of interest (i.e. AA_br in Figure 4.7b) would suggest that the particle filtering step did not add value to the parameter estimation. Two possible reasons for this are: i) there was no association between the parameter of interest and the clonality measure; or ii) the RMSE cost function was ineffective at separating particles

However, I believe the RMSE threshold value was not the issue behind these poor results. For many of the uninformative particle filtering comparisons (where the true parameter value fell outside of the ITR range), the parameter estimate was within the outlier region of particles (Fig. 4.7a). Thus, changing the cutoff threshold for RMSE values (i.e. reducing to lowest 25% RMSE values instead of 50%) would not increase the precision of the estimate. Additionally, for other uninformative particle filtering comparisons, there was no clear separation in the distribution of particles that are filtered and unfiltered. Fig. 4.7b shows an example of this. In figure 4.7b, there is no distinction between the distribution of red (removed) and blue particles (retained) along the active to active infection rate in this figure other than the RMSE values. It is unlikely that another error metric would fix either of these inherent imperfections in the particle filtering method.

For the former issue (poor estimation), the results would suggest that: (i) the simulation model parameter is not suitable for parameter estimation using any of the clonality measures; (ii) that particle filtering is not an effective method for estimating parameters; or (iii) that the RMSE cost function is not an appropriate cost function for this use case. However, further investigation is required into which, if any, or if all of these three are the issue. Using a different error metric that is reliant on the same comparison between particle clonality metric and target clonality metric would most likely yield the same result. Possible workarounds for fixing the latter issue (uniform RMSE scores across parameter values) may be limited as well.

Particle filtering is a method that may be able to identify the Clonal Expansion parameter in the simclone framework under certain conditions and for certain clonality measures. It may be possible to use the particle filtering method to identify sets of within host data that exhibit high Clonal Expansion. Classifying the type of within host dynamic that drives the proliferation and maintenance of the latent reservoir for individual hosts may lead to personalized therapy responses. However, caution must be advised for applying the particle filtering method to estimating rates of infection and transition, especially in real world data. The particle filtering method shown here is applied to a simplistic simulation framework with many assumptions. There is a high amount of 'noise' in the data due to the complexity of the simulation pipeline. There are many moving parts in the simulation pipeline, with the added difficulty that parameter identification is a moving target. The particle filtering results here would suggest that the goal of the particle filtering method should drive the choice for error metric and clonality metric. For situations where there is a non-uniform distribution of RMSE values across a parameter range, the RMSE error metric with the Gini clonality metric could be a good starting point for tuning a specific particle filtering method.

Chapter 5

Concluding remarks

The simclone framework can be used to simulate various scenarios of within host pathogen dynamics. It is a flexible pipeline that can be modified to test more complex scenarios. These complex scenarios are difficult to study *in vivo* for ethical and practical reasons. Retaining a study population for the timescale at which the latent reservoir changes at is difficult. Anatomic compartments that provide location for low level HIV-1 replication, and/or proliferation of the latent reservoir remain hard to sample. On the practical side, it is remains expensive to extensively sample the latent reservoir. The half-life of cells in the latent reservoir is estimated to be approximately 44 months, translating to a natural decay timeline of over 70 years (Crooks et al., 2015; Finzi et al., 1997a; Siliciano et al., 2003). In addition, due to the relatively low number circulating latent reservoir cells to uninfected cells, large blood draws are required for participants. Combination of long blood draws and lengthy timelines, make longitudinal studies impractical. Longitudinal studies that track the latent reservoir *in vivo* or *in vitro* are able to capture informative stop-restart ART scenarios, but these analyses are typically done retrospectively and have many caveats. Some examples of limitations to these longitudinal studies include: distinct clinical experiments between studies used in meta-analyses (Borges et al., 2018; Gwadz et al., 2021); recency bias for drug adherence or behavioural studies (Gwadz et al., 2021); and lack of verifiable data, for example prescription data used as a stand-in for drug adherence (Reed et al., 2021). Therefore, simulations would remain a suitable alternative method, or at least a useful supplement to study within host HIV-1 dynamics.

Simulations provide granular data, and in particular, unlimited snapshots in time. For example, simclone pipeline captures cell level transmission and transition, data that could never be captured when studying the latent reservoir *in vivo*. Simulations allow us to control rates of transmission and transition, allowing us to understand the role each rate plays in creating genetic diversity in the latent reservoir. An example hypothesis that could be tested is: what effect does increasing rates of active to active transmission have on the clonality of latent reservoir; and what would happen if ART was started, then halted in this system. These can be useful for modelling and understanding complex systems that are poorly understood, like the HIV-1 latent reservoir. *Simclone* is a robust framework, however as with any other simulations, there are limitations that can be addressed in future work. For one, the lack of anatomical compartments (i.e. circulating CD4+ cells in peripheral blood vs. CD4+ cells in lymph nodes); and different types of cells with unique rates of transmission and transition (i.e. Gut-Associated Lymphoid Tissues that have low level HIV-1 replication (Thompson et al., 2017).

The simclone framework as described here is a starting point, and can already be used for downstream analyses. Increasing the complexity of the system will only add to the usefulness of this tool. It can be expanded to include multiple strain competition, anatomical compartments with differing within host dynamic interactions, and different models of nucleotide evolution can be selected. Increasing the complexity of the simclone framework would allow novel hypotheses to be tested, as described previously. Some possible scenarios for future evaluation using the simclone framework include: effect of ART cessation and restart after a viral rebound on the latent reservoir (Borges et al., 2018); effect of low level HIV-1 replication and infection in drug sanctuary sites (Thompson et al., 2017); and effect of HIV-1 ART initiation timing on the latent reservoir (Brodin et al., 2016; Chavez et al., 2015). On the other hand, users should be cautious about increasing the complexity of the tool. The simclone framework already has a high number of parameters. Parameterizing the model to produce rational outcomes was already difficult for the simclone framework. Creating additional definable parameters to increase model complexity will make this process more difficult. The trade-off for simulations that could reveal novel insights may be a narrow range of simulation parameter values. In summary, the simclone framework provides a robust simulation pipeline for measuring the within host dynamics of the HIV-1, and can be used as the foundational pieces for further simulation work.

The analysis of clonality measures in Chapter 3 spoke to the shortcomings of the widely used proportional clonality measure (von Stockenstrom et al., 2015; Wagner et al., 2014), and the proposed Gini coefficient (Sadras and Bongiovanni, 2004; Thapa et al., 2015). As discussed in Chapter 3, there were frequent collisions in clonality scores for differing genetic population structures for the proportional clonality measure and the Gini coefficient. Collisions occur when two population structures are given the same clonality score. This happened more frequently when using the Gini Coefficient (due to the effect of diminishing unique groups), and the proportional clonality score (due to low number of possible scores). Collisions are especially problematic when the distribution of genetic sequences are vastly different (i.e. proportional clonality score for population *a* [3,2,1] = 0.83; and for population *b* [5,1] = 0.83. Collisions are unideal when using clonality measures because it conflates scenarios in further analyses. This compounds the loss of information caused by having low numbers of possible clonality outcomes (as in the case of proportional clonality). The Gini coefficient in particular was not effective for measuring clonality in diminishing groups. Gini coefficient did not respond to increasing heterogeneity in a directional fashion - especially in situations where the number of unique sequences are low. This limits the usefulness of the Gini Coefficient for HIV-1 latent reservoir research. There are situations where there is high homogeneity in samples (Feder et al., 2016; Wolinsky et al., 1992; Yeh et al., 2021), especially when the sampling fraction is low. Gini coefficient was flawed, and caution is advised when using Gini coefficient to describe genetic diversity in populations with high homogeneity.

The presence of identical sequences has been interpreted in previous studies as an indication of clonal expansion (Bosque et al., 2011, Douek et al., 2002, Maldarelli et al., 2014). However, my simulations showed high measures of clonality (regardless of clonality measure) are not necessarily the result of clonal expansions. Two identical sequences that shared ancestry in the simulation were not necessarily identical due to clonal expansion. Instead, these simulated sequences could be identical if by chance if there was no mutations accumulated between the two sequences since their descent from a common ancestor (prior to their respective integration events). This mirrors research by various others (Lorenzi et al., 2016, Maldarelli et al., 2014) who have shown that sequencing integration sites are better indicators of clonal expansion. The findings shown here re-iterate this paradigm and support the use of sequencing integration sites to measure clonal expansion in the latent reservoir.

Clonality has been used in other domains to characterize biological processes. For example, in cancer research, clonality is used to identify driver mutations (Martincorena, 2019, Biermann et al., 2018, McGranahan et al., 2015). Driver mutations are mutations that drive abnormal cellular proliferation, and thus tumourgenesis. These mutations are usually identified by observing large numbers of identical cells with excessive mutations clustered around or in a particular region. In McGranahan et al., 2015, the authors attempted to identify actionable driver mutations using clonality metrics. In their research, the authors used the proportional clonality metric to describe the distribution of mutations a pan-cancer dataset. The findings here suggest that information about the presence and diversity of sub-clonal groups may be lost. The pairwise clonality metric could be applied in research such as the above to provide more information about the true genetic distribution of the population than the often using proportional clonality score. Identifying driver mutations in cancers is just one avenue of research where reductive clonality metrics are used. There are other research topics beyond driver mutations in cancer and HIV-1 where clonality cited as an indicator for an underlying event. Others include: immunology research - T-cell antigen response dynamics (Chiffelle et al., 2020);and ecology research - organism genetic diversity (Hughes and Stachowicz, 2009), and the relationship between carbon cycling and clonality (Cornelissen et al., 2014). The results found here indicate more care needs to be taken when selecting the correct clonality measure for a particular analysis.

There is potential for the particle filtering method described in this work to be applied with a narrow scope. The particle filtering method for parameter estimation can only be applied to parameter-target combinations that have a non-uniform distribution with regard to RMSE values. Estimating the true value of the clonal expansion parameter may be possible with this technique. The RMSE cost function may not have been an appropriate cost function for this technique. The RMSE cost function was only able to predict clonal expansion in a small subset of cases for clonal expansion (3/10 targets). When visually inspecting the filtered particles using the median RMSE score as a threshold, there did not appear to be a significance difference for most parameters based on filtered status. This suggests two possibilities; i) the parameters (excluding clonal expansion) were unresponsive to changes in clonality metric - unlikely, as multivariate GLM results showed significance; or ii) the RMSE cost function did not accurately penalize uninformative particles. A different approach to establishing likelihood functions can be pursued, including incorporating the Bayesian likelihood framework to the particle filtering (Park et al., 2021; Rasmussen et al., 2011). In future work, this parameter estimation technique could be used to estimate the degree to which clonal expansion is responsible for an individual's latent reservoir. Clonality metrics would be extracted from an individual's latent reservoir sequence data. Simclone simulations would be created, using the real-world clonality metrics as the target. Parameter estimates would be calculated using parameters that were most similar to the target. These parameter estimates could provide information about about an individual's HIV-1 infection, and would add to our understanding of the latent reservoir as a whole. This knowledge could help in many active areas of HIV-1 research including therapy and prevention. In addition, there is potential to use this particle filtering method to identify parameters that are non-linearly associated with clonality measures. Visualizing these trends, while reducing the amount of variation by filtering the particles could provide insights into responsive variables. This would be particularly useful if the simclone framework is expanded to test more complex scenarios. Particle filtering as described here is not ineffective, however opens the door to many more possibilities.

Using the simclone framework I have created, I have shown pitfalls of the proportional clonality measure that is widely used; the Gini coefficient that has been suggested for describing genetic population diversity; and have proposed a new pairwise clonality statistic with more desirable statistical properties. Analysis of the clonality measures could apply to other research domains, including cancer research. Further, I have described a particle filtering technique that could be used to visually identify non-linear trends, and further used under stringent conditions to estimate parameter values.

Appendix A

Supplementary

	Proportional Clonality		Gini Coefficient		Pairwise Clonality	
Simulation Parameter	Coefficient Estimate	p	Coefficient Estimate	p	Coefficient Estimate	p
(Intercept)	2.73	0.299	6.70e-01	0.494	**2.48**	**<2e-16**
Initial Active-type Population	6.46e-06	0.612	4.45e-07	0.917	**-2.70e-06**	**<2e-16**
Initial Latent-type Population	1.64e-06	0.986	-5.35e-05	0.169	**-6.22e-06**	**<2e-16**
Initial Replenishment Population	-8.63e-05	0.685	5.28e-05	0.539	**-1.77e-04**	**<2e-16**
Active-to-Active Transmission	1.55e+04	0.262	**-1.27e+04**	**0.006**	**9.53e+03**	**<2e-16**
Active-to-Latent Transmission	-6.57e+04	0.537	2.40e+04	0.518	**-5.87e+04**	**<2e-16**
Clonal Expansion	4.12	0.324	**-2.34**	**0.0001**	**4.89**	**<2e-16**
Active-to-Latent Transition	-1.66	0.999	5.09e+02	0.195	**-7.98e+02**	**<2e-16**
Latent-to-Active Transition	8.25e+03	0.930	2.11e+04	0.544	**-1.32e+03**	**<2e-16**
Active Death Rate	1.27	0.903	-4.79e+01	0.214	**4.27e+01**	**<2e-16**
Latent Death Rate	2.65e+01	0.800	-1.36e+01	0.727	**-1.31e+01**	**<2e-16**
Active Replenishment	-2.22e+04	0.857	2.84e+04	0.548	**-8.40e+03**	**<2e-16**
Latent Replenishment	3.60e+04	0.690	2.42e+03	0.945	**-3.39e+04**	**<2e-16**

Table A.1: Coefficient estimates and p scores are reported for GLM results. All parameters were varied. The 3 clonality statistics were calculated from the same set of genetic sequences. Bolded Coefficient values denote $p \leq 0.05$

```
InitialConditions:
  originTime: 10
  size:
    Active_type: 7500
    Latent_type: 5000
    Replenish_type: 20000
    Death_type: 0
  indexType: Active_type

CompartmentTypes:
  'Active_type':
    branching.rates:
      10: (Active_type=0.00043, Latent_type=2.57e-05, Replenish_type= 0.0, Death_type= 0.0)
      # No infection of Death_type compartment
      5:   (Active_type = 0.0, Latent_type= 0.0, Replenish_type= 0.0, Death_type= 0.0)
      #No more active infection
    transition.rates: (Active_type=0.0, Latent_type= 5e-05, Replenish_type= 0.0, Death_type = 0.01)
    coalescent.rate: 1.5
  'Latent_type':
    branching.rates: (Active_type=0, Latent_type=1.66e-06, Replenish_type= 0.0, Death_type = 0.0)
    #constant re-infection
    transition.rates: (Active_type=5e-05, Latent_type= 0.0, Replenish_type= 0.0, Death_type = 0.01)
    coalescent.rate: 1.5
  'Replenish_type':
    branching.rates: (Active_type=0, Latent_type=0, Replenish_type= 0.0, Death_type= 0.0)
    transition.rates:
      10: (Active_type=0, Latent_type=0, Replenish_type= 0.0, Death_type= 0.0)
      5: (Active_type=1.25e-02, Latent_type= 1.25e-02, Replenish_type= 0.0, Death_type=0.0)
  'Death_type':
    branching.rates: (Active_type=0, Latent_type=0, Replenish_type= 0.0, Death_type= 0.0)
    transition.rates: (Active_type=0, Latent_type=0, Replenish_type= 0.0, Death_type= 0.0)
Compartments:
  'Activecomp':
    type: Active_type
    replicates: 20
  'Latentcomp':
    type: Latent_type
    replicates: 20
Lineages:
  'A':
    sampling.time: 0
    location: Activecomp
    replicates: 1
  'B':
    sampling.time: 0
    location: Latentcomp
    replicates: 1
```

Listing A.1: Sample YAML simclone specification file

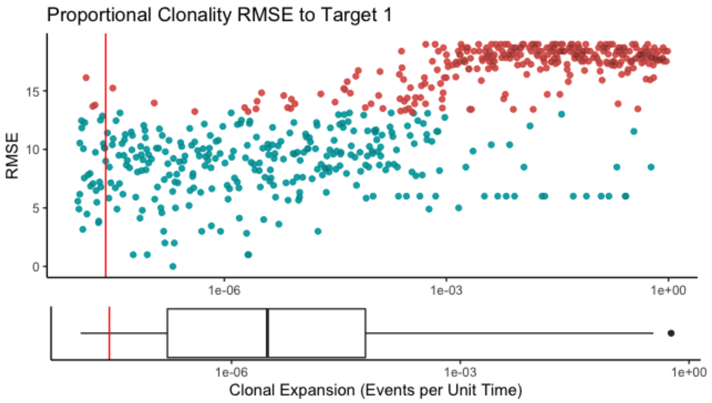


Figure A.1: Particle filtering parameter estimation for clonal expansion for target 1 using proportional clonality

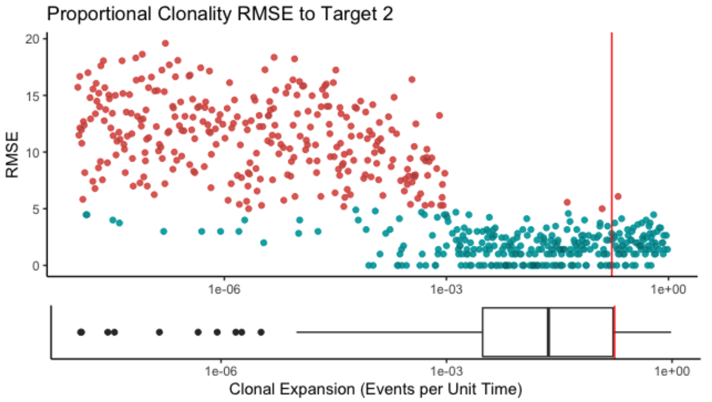


Figure A.2: Particle filtering parameter estimation for clonal expansion for target 2 using proportional clonality

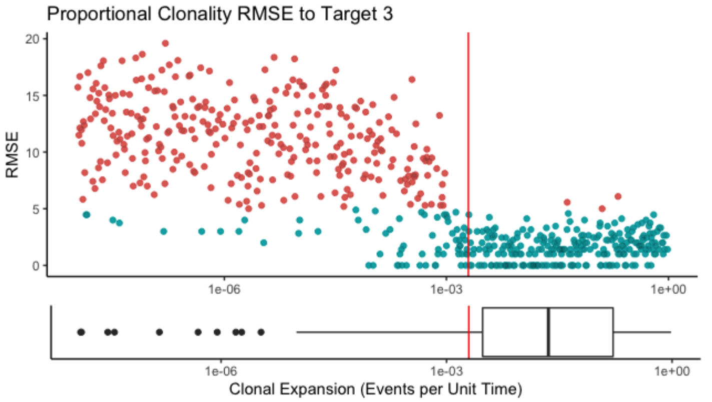


Figure A.3: Particle filtering parameter estimation for clonal expansion for target 3 using proportional clonality

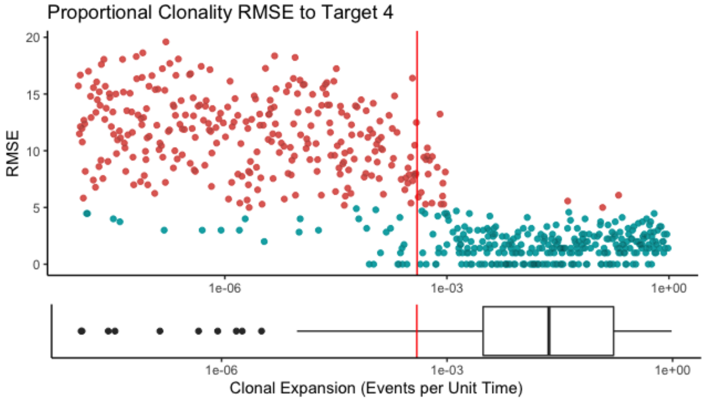


Figure A.4: Particle filtering parameter estimation for clonal expansion for target 4 using proportional clonality

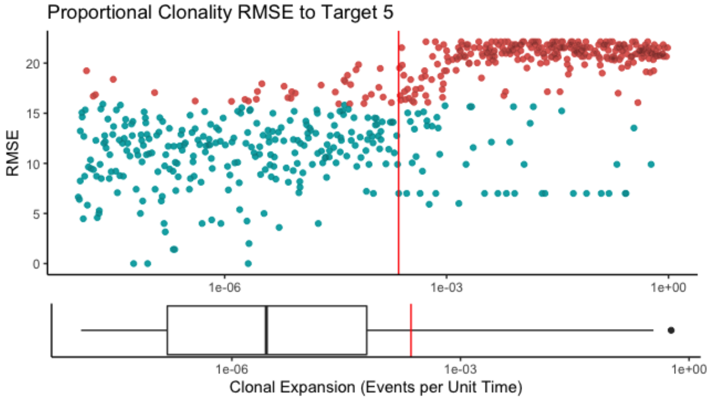


Figure A.5: Particle filtering parameter estimation for clonal expansion for target 5 using proportional clonality

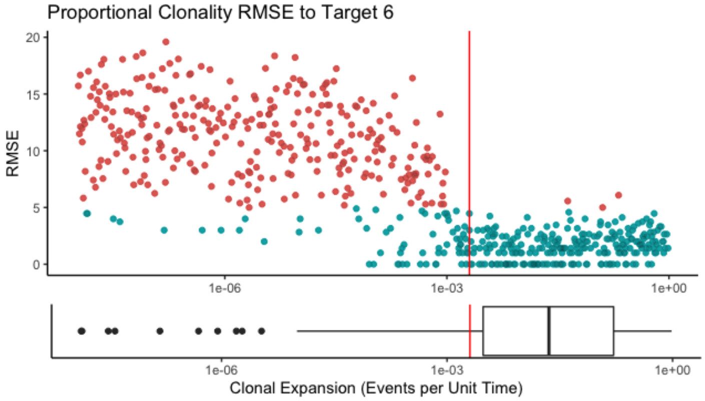


Figure A.6: Particle filtering parameter estimation for clonal expansion for target 6 using proportional clonality

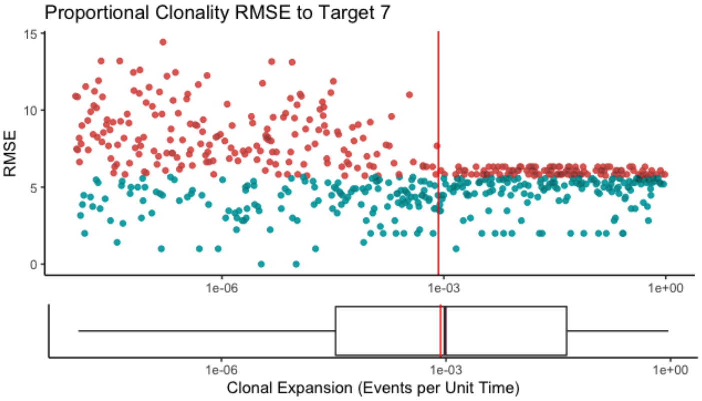


Figure A.7: Particle filtering parameter estimation for clonal expansion for target 7 using proportional clonality

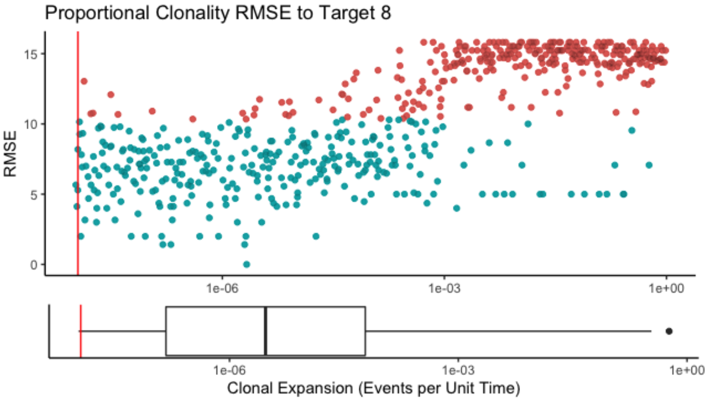


Figure A.8: Particle filtering parameter estimation for clonal expansion for target 8 using proportional clonality

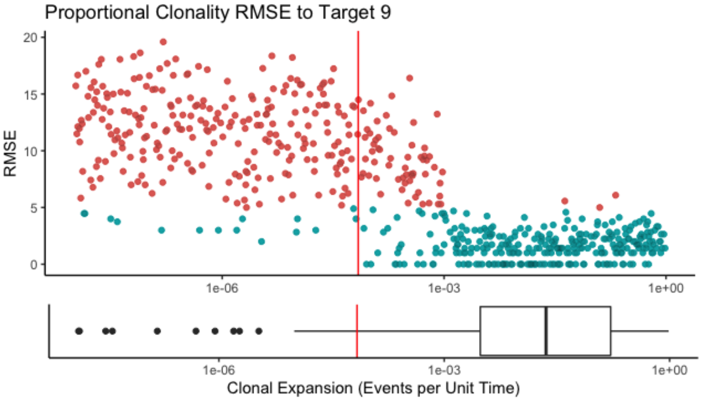


Figure A.9: Particle filtering parameter estimation for clonal expansion for target 9 using proportional clonality

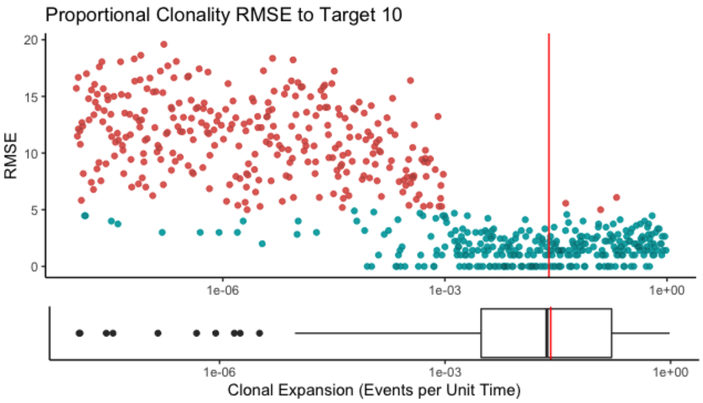


Figure A.10: Particle filtering parameter estimation for clonal expansion for target 10 using proportional clonality

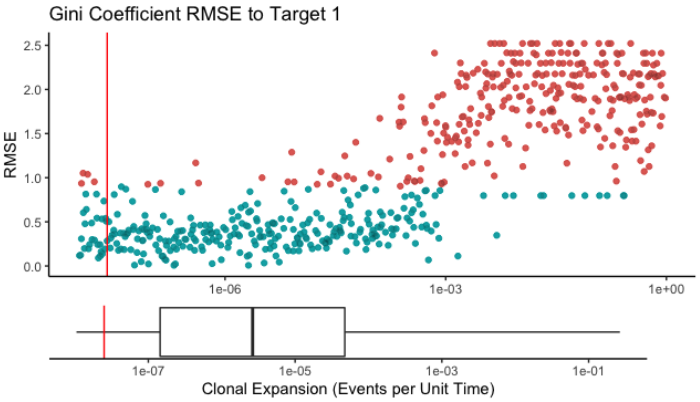


Figure A.11: Particle filtering parameter estimation for clonal expansion for target 1 using Gini Coefficient

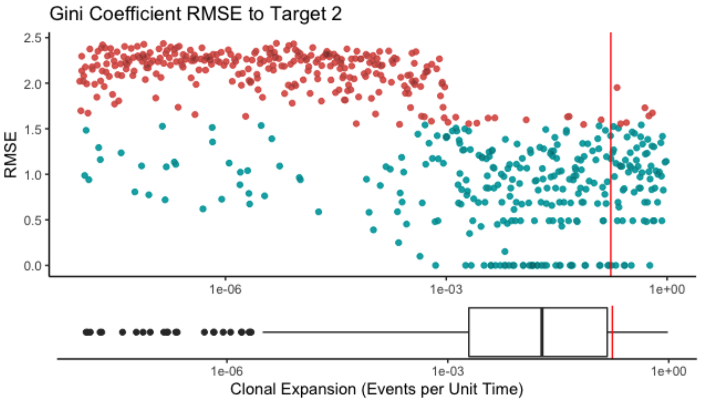


Figure A.12: Particle filtering parameter estimation for clonal expansion for target 2 using Gini Coefficient

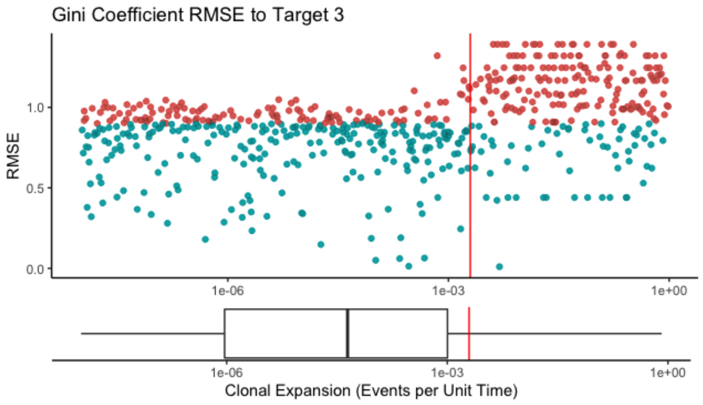


Figure A.13: Particle filtering parameter estimation for clonal expansion for target 3 using Gini Coefficient

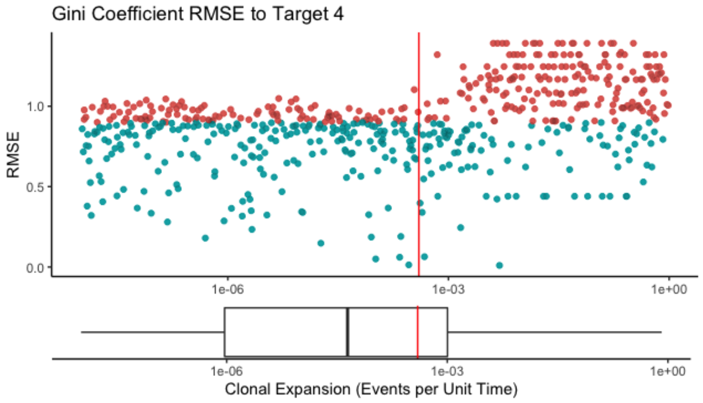


Figure A.14: Particle filtering parameter estimation for clonal expansion for target 4 using Gini Coefficient

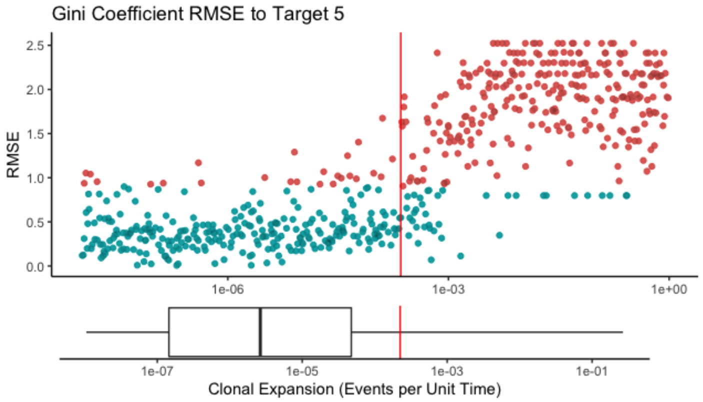


Figure A.15: Particle filtering parameter estimation for clonal expansion for target 5 using Gini Coefficient

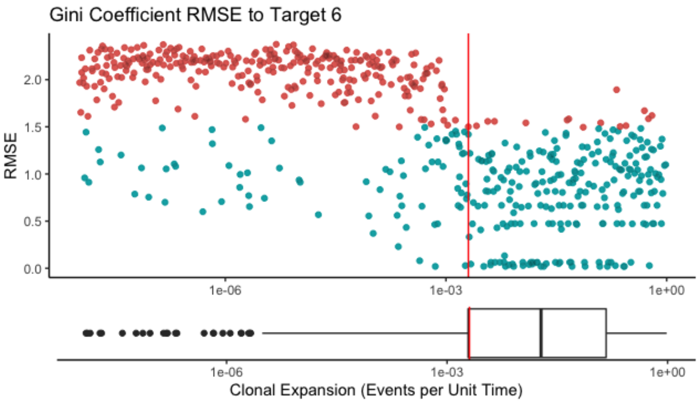


Figure A.16: Particle filtering parameter estimation for clonal expansion for target 6 using Gini Coefficient

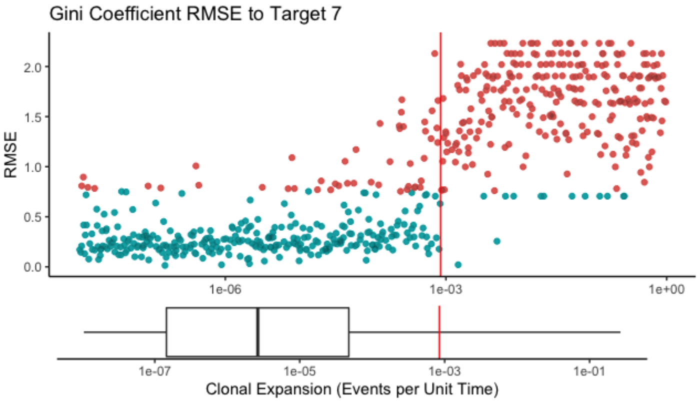


Figure A.17: Particle filtering parameter estimation for clonal expansion for target 7 using Gini Coefficient

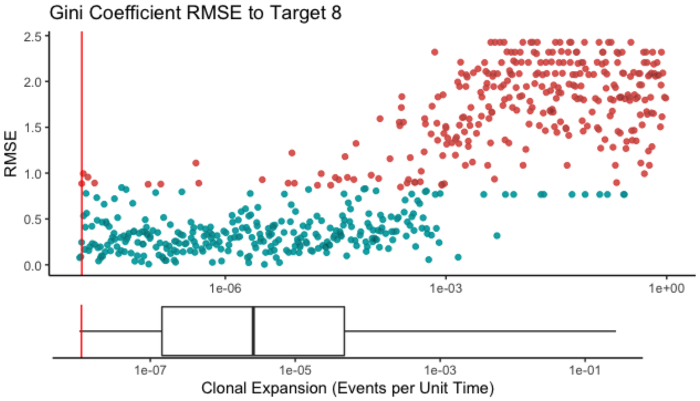


Figure A.18: Particle filtering parameter estimation for clonal expansion for target 8 using Gini Coefficient

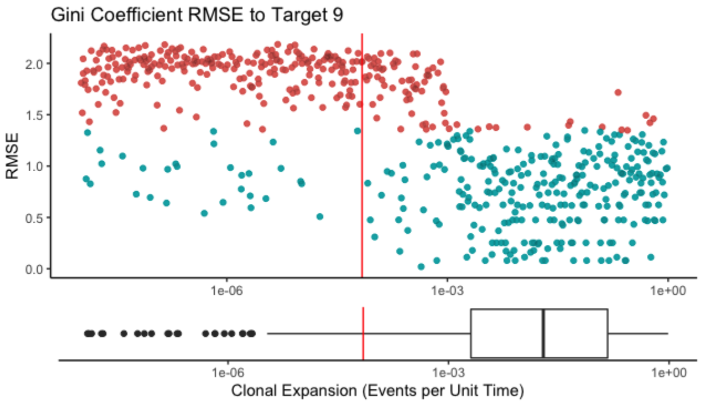


Figure A.19: Particle filtering parameter estimation for clonal expansion for target 9 using Gini Coefficient

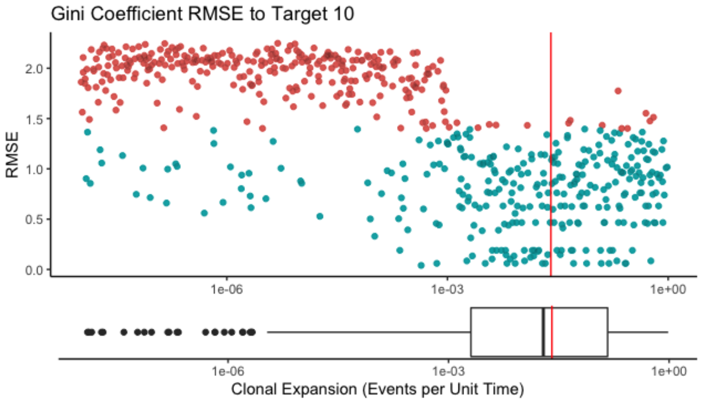


Figure A.20: Particle filtering parameter estimation for clonal expansion for target 10 using Gini Coefficient

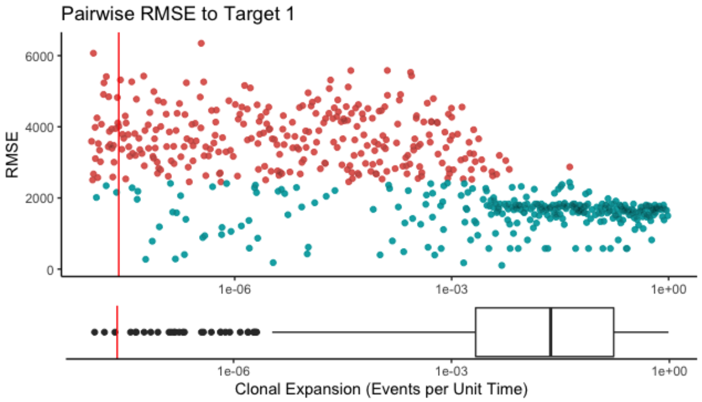


Figure A.21: Particle filtering parameter estimation for clonal expansion for target 1 using pairwise clonality

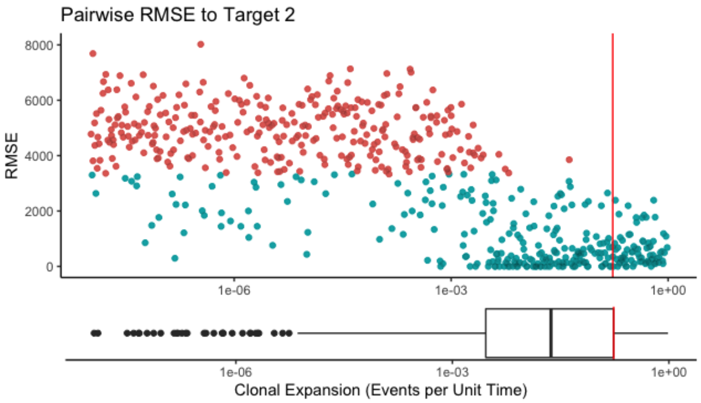


Figure A.22: Particle filtering parameter estimation for clonal expansion for target 2 using pairwise clonality

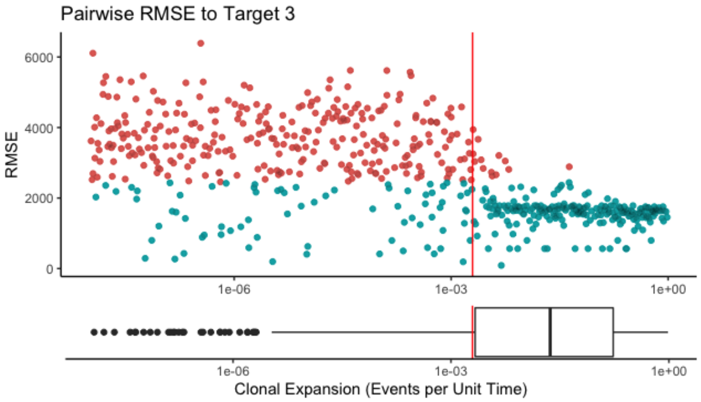


Figure A.23: Particle filtering parameter estimation for clonal expansion for target 3 using pairwise clonality

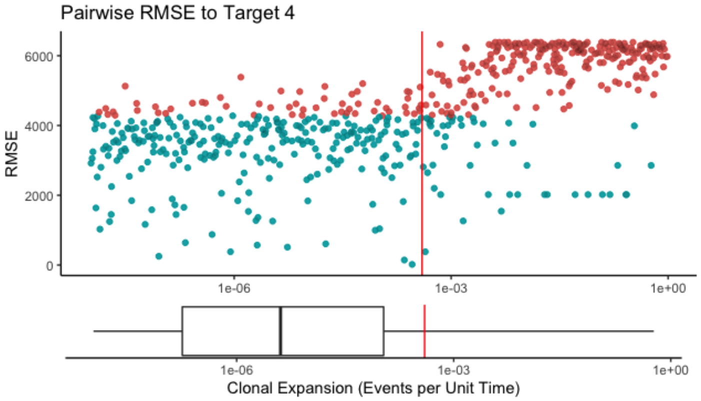


Figure A.24: Particle filtering parameter estimation for clonal expansion for target 4 using pairwise clonality

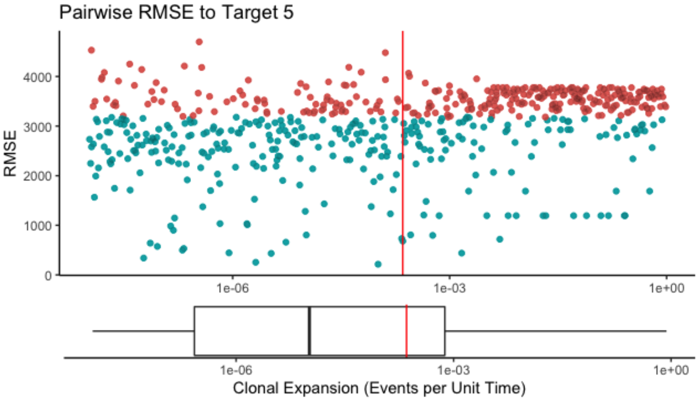


Figure A.25: Particle filtering parameter estimation for clonal expansion for target 5 using pairwise clonality

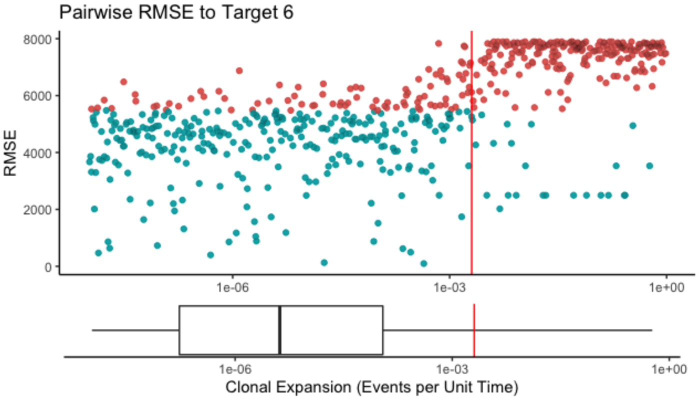


Figure A.26: Particle filtering parameter estimation for clonal expansion for target 6 using pairwise clonality

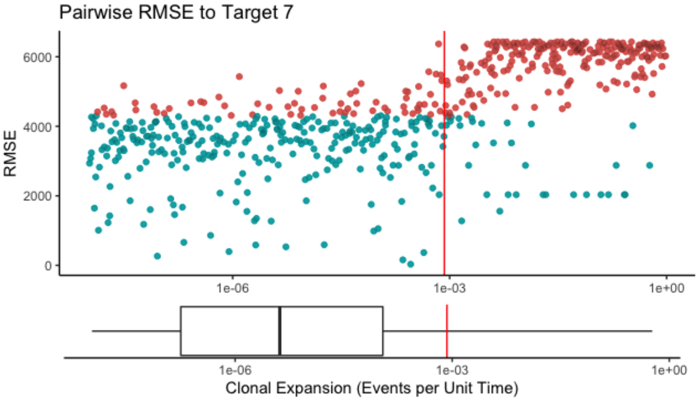


Figure A.27: Particle filtering parameter estimation for clonal expansion for target 7 using pairwise clonality

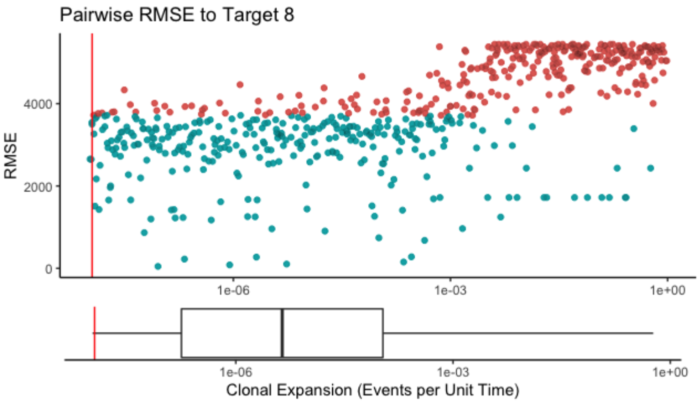


Figure A.28: Particle filtering parameter estimation for clonal expansion for target 8 using pairwise clonality

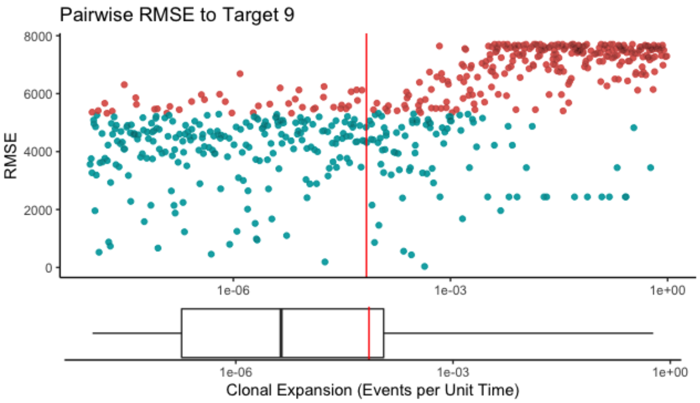


Figure A.29: Particle filtering parameter estimation for clonal expansion for target 9 using pairwise clonality

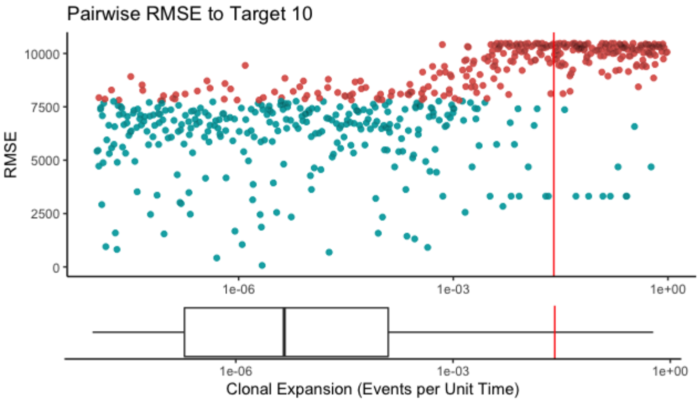
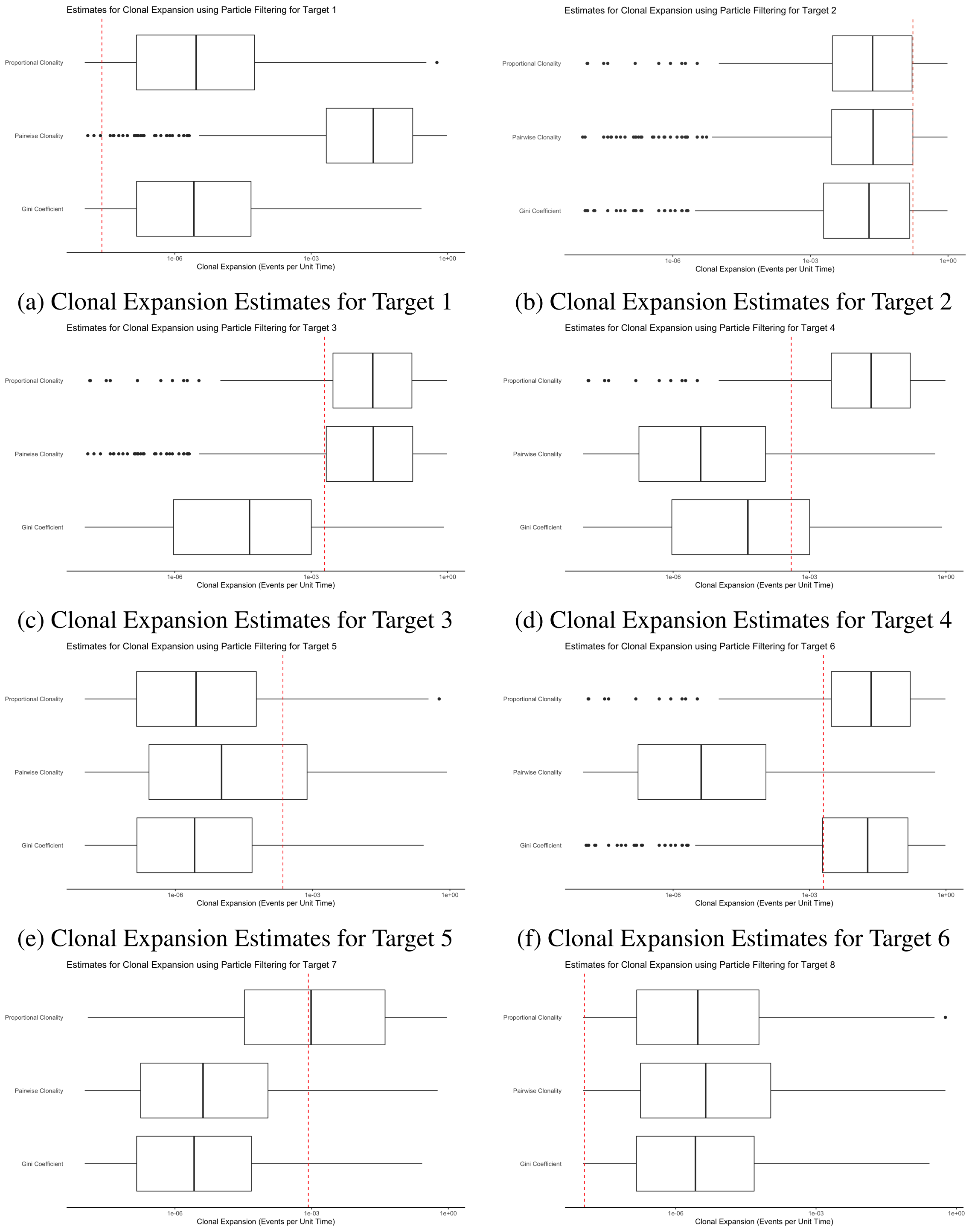


Figure A.30: Particle filtering parameter estimation for clonal expansion for target 10 using pairwise clonality



(a) Clonal Expansion Estimates for Target 1

(b) Clonal Expansion Estimates for Target 2

(c) Clonal Expansion Estimates for Target 3

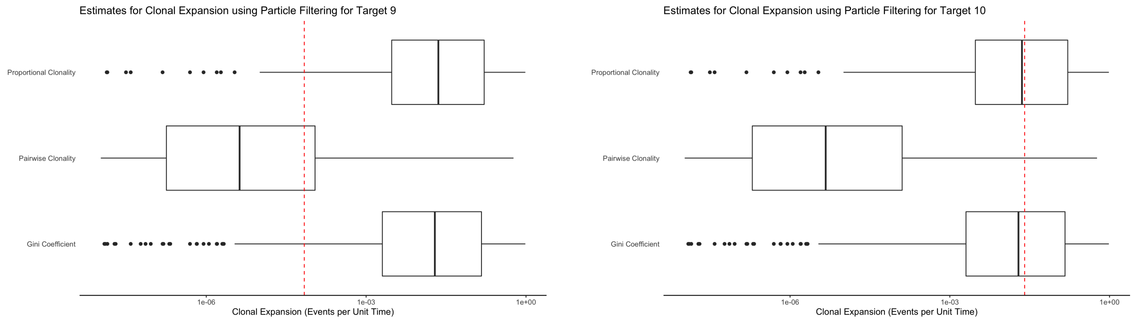
(d) Clonal Expansion Estimates for Target 4

(e) Clonal Expansion Estimates for Target 5

(f) Clonal Expansion Estimates for Target 6

(g) Clonal Expansion Estimates for Target 7

(h) Clonal Expansion Estimates for Target 8



(i) Clonal Expansion Estimates for Target 9

(j) Clonal Expansion Estimates for Target 10

Figure A.31: Clonal expansion estimates for all targets by clonality statistic used. Dashed vertical line represents true clonal expansion value. Parameter estimate lands within IQR using the Gini Coefficient for Targets 4, 6, and 10 (d, f, and j respectively). Parameter lands within IQR using proportional and pairwise clonality for Targets 7, and 10 (g, and j respectively) (cont.)

```
[TYPE] NUCLEOTIDE 1
[MODEL]      modelname
[submodel] JC
[TREE] treename ((((((Latentcomp_83__B_1:0.0):0.0):4.9e10-6):1.1e10-6):1.26e10-8,
((((((((Latentcomp_91__B_1:0.0):0.0):3.2e10-5):1.4e10-9):e10-3):2.5e10-6,
((((((((((((Latentcomp_11__B_1:0.0):0.0):7.2e10-8):4.7e10-6):8.3e10-6,
(((Latentcomp_92__B_1:0.0):0.0):4.5e10-7):4.4e10-9):0.0):2.0e10-8)
:1.0e10-9):7.8e10-6,
(((Activecomp_99__A_1:1.9e10-3):4.2e10-3):4.4e10-8):4.8e10-8):0.0):7.8e10-6):4.7e10-4)
:3.6e10-9):0.0):3.8e10-6,(((((Latentcomp_81__B_1:0.0):0.0):3.6e10-5)
:5.8e10-10):7.2e10-6)))));
[PARTITIONS] partitionname
  [treename modelname /simclone/NC_001802.1.fa]
[EVOLVE] partitionname 1 outputname
```

Listing A.2: Sample INDELible control file

Bibliography

Abrahams, M.-R., Joseph, S. B., Garrett, N., Tyers, L., Moeser, M., Archin, N., Council, O. D., Matten, D., Zhou, S., Doolabh, D., & et al. (2019). The replication-competent hiv-1 latent reservoir is primarily established near the time of therapy initiation. https://doi.org/10.1101/512475

Adams, N. M., Grassmann, S., & Sun, J. C. (2020). Clonal expansion of innate and adaptive lymphocytes. *Nature Reviews Immunology*, *20*(11), 694–707. https://doi.org/10.1038/s41577-020-0307-4

Ahlenstiel, C. L., Symonds, G., Kent, S. J., & Kelleher, A. D. (2020). Block and lock hiv cure strategies to control the latent reservoir. *Frontiers in Cellular and Infection Microbiology*, *10*. https://doi.org/10.3389/fcimb.2020.00424

Alizon, S., & Fraser, C. (2013). Within-host and between-host evolutionary rates across the hiv-1 genome. *Retrovirology*, *10*(1). https://doi.org/10.1186/1742-4690-10-49

Aloia, R. C., Tian, H., & Jensen, F. C. (1993). Lipid composition and fluidity of the human immunodeficiency virus envelope and host cell plasma membranes. *Proceedings of the National Academy of Sciences*, *90*(11), 5181–5185. https://doi.org/10.1073/pnas.90.11.5181

Arts, E. J., & Hazuda, D. J. (2012). Hiv-1 antiretroviral drug therapy. *Cold Spring Harbor Perspectives in Medicine*, *2*(4). https://doi.org/10.1101/cshperspect.a007161

Barton, K., Winckelmann, A., & Palmer, S. (2016). Hiv-1 reservoirs during suppressive therapy. *Trends in Microbiology*, *24*(5), 345–355. https://doi.org/10.1016/j.tim.2016.01.006

Bashford-Rogers, R. J., Palser, A. L., Huntly, B. J., Rance, R., Vassiliou, G. S., Follows, G. A., & Kellam, P. (2013). Network properties derived from deep sequencing of human b-cell receptor repertoires delineate b-cell populations. *Genome Research*, *23*(11), 1874–1884. https://doi.org/10.1101/gr.154815.113

Bergstrom, C. T., McElhany, P., & Real, L. A. (1999). Transmission bottlenecks as determinants of virulence in rapidly evolving pathogens. *Proceedings of the National Academy of Sciences*, *96*(9), 5095–5100. https://doi.org/10.1073/pnas.96.9.5095

Berry, I. M., Ribeiro, R., Kothari, M., Athreya, G., Daniels, M., Lee, H. Y., Bruno, W., & Leitner, T. (2007). Unequal evolutionary rates in the human immunodeficiency virus type 1 (hiv-1) pandemic: The evolutionary rate of hiv-1 slows down when the epidemic rate increases. *Journal of Virology*, *81*(19), 10625–10635. https://doi.org/10.1128/jvi.00985-07

Biermann, J., Parris, T. Z., Nemes, S., Danielsson, A., Engqvist, H., Werner Rönnerman, E., Forssell-Aronsson, E., Kovács, A., Karlsson, P., Helou, K., & et al. (2018). Clonal relatedness in tumour pairs of breast cancer patients. *Breast Cancer Research*, *20*(1). https://doi.org/10.1186/s13058-018-1022-y

Boeras, D. I., Hraber, P. T., Hurlston, M., Evans-Strickfaden, T., Bhattacharya, T., Giorgi, E. E., Mulenga, J., Karita, E., Korber, B. T., Allen, S., & et al. (2011). Role of donor genital tract hiv-1 diversity in the transmission bottleneck. *Proceedings of the National Academy of Sciences*, *108*(46). https://doi.org/10.1073/pnas.1103764108

Borges, Á. H., Neuhaus, J., Sharma, S., Neaton, J. D., Henry, K., Anagnostou, O., Staub, T., Emery, S., & Lundgren, J. D. (2018). The effect of interrupted/deferred antiretroviral therapy on disease risk: A smart and start combined analysis. *The Journal of Infectious Diseases*, *219*(2), 254–263. https://doi.org/10.1093/infdis/jiy442

Bosque, A., Famiglietti, M., Weyrich, A. S., Goulston, C., & Planelles, V. (2011). Homeostatic proliferation fails to efficiently reactivate hiv-1 latently infected central memory cd4+ t cells. *PLoS Pathogens*, *7*(10). https://doi.org/10.1371/journal.ppat.1002288

Brodin, J., Zanini, F., Thebo, L., Lanz, C., Bratt, G., Neher, R. A., & Albert, J. (2016). Establishment and stability of the latent hiv-1 dna reservoir. *eLife*, *5*. https://doi.org/10.7554/elife.18889

Brower, E. T., Schön, A., Klein, J. C., & Freire, E. (2009). Binding thermodynamics of the n-terminal peptide of the ccr5 coreceptor to hiv-1 envelope glycoprotein gp120. *Biochemistry*, *48*(4), 779–785. https://doi.org/10.1021/bi8021476

Bruner, K. M., Murray, A. J., Pollack, R. A., Soliman, M. G., Laskey, S. B., Capoferri, A. A., Lai, J., Strain, M. C., Lada, S. M., Hoh, R., & et al. (2016). Defective proviruses rapidly accumulate during acute hiv-1 infection. *Nature Medicine*, *22*(9), 1043–1049. https://doi.org/10.1038/nm.4156

Carr, J. K., Salminen, M. O., Koch, C., Gotte, D., Artenstein, A. W., Hegerich, P. A., Louis, D. S., Burke, D. S., & Mccutchan, F. E. (1996). Full-length sequence and mosaic structure of a human immunodeficiency virus type 1 isolate from thailand. *Journal of Virology*, *70*(9), 5935–5943. https://doi.org/10.1128/jvi.70.9.5935-5943.1996

Carrillo, C., Moore, D. M., Sobrino, F., Borca, M., & Morgan, D. O. (1998). In vivo analysis of the stability and fitness of variants recovered from foot-and-mouth disease virus quasispecies. *Journal of General Virology*, *79*(7), 1699–1706. https://doi.org/10.1099/0022-1317-79-7-1699

Cavarelli, M., & Scarlatti, G. (2009). Phenotype variation in human immunodeficiency virus type 1 transmission and disease progression. *Disease Markers*, *27*(3-4), 121–136. https://doi.org/10.1155/2009/685608

Chan, D. C., Fass, D., Berger, J. M., & Kim, P. S. (1997). Core structure of gp41 from the hiv envelope glycoprotein. *Cell*, *89*(2), 263–273. https://doi.org/10.1016/s0092-8674(00)80205-6

Chavez, L., Calvanese, V., & Verdin, E. (2015). Hiv latency is established directly and early in both resting and activated primary cd4 t cells. *PLOS Pathogens*, *11*(6). https://doi.org/10.1371/journal.ppat.1004955

Chiffelle, J., Genolet, R., Perez, M. A., Coukos, G., Zoete, V., & Harari, A. (2020). T-cell repertoire analysis and metrics of diversity and clonality. *Current Opinion in Biotechnology*, *65*, 284–295. https://doi.org/10.1016/j.copbio.2020.07.010

Chun, T.-W., Stuyver, L., Mizell, S. B., Ehler, L. A., Mican, J. A., Baseler, M., Lloyd, A. L., Nowak, M. A., & Fauci, A. S. (1997). Presence of an inducible hiv-1 latent reservoir during highly active antiretroviral therapy. *Proceedings of the National Academy of Sciences*, *94*(24), 13193–13197. https://doi.org/10.1073/pnas.94.24.13193

Coffin, J. M. (1995). Hiv population dynamics in vivo: Implications for genetic variation, pathogenesis, and therapy. *Science*, *267*(5197), 483–489. https://doi.org/10.1126/science.7824947

Coffin, J. M. (2002). *Retroviruses*. Cold Spring Harbor Laboratory Press.

Cohn, L. B., Silva, I. T., Oliveira, T. Y., Rosales, R. A., Parrish, E. H., Learn, G. H., Hahn, B. H., Czartoski, J. L., McElrath, M. J., Lehmann, C., & et al. (2015). Hiv-1 integration landscape during latent and active infection. *Cell*, *160*(3), 420–432. https://doi.org/10.1016/j.cell.2015.01.020

Colin, L., & Van Lint, C. (2009). Molecular control of hiv-1 postintegration latency: Implications for the development of new therapeutic strategies. *Retrovirology*, *6*(1). https://doi.org/10.1186/1742-4690-6-111

Collier, A. C., Coombs, R. W., Schoenfeld, D. A., Bassett, R. L., Timpone, J., Baruch, A., Jones, M., Facey, K., Whitacre, C., McAuliffe, V. J., & et al. (1996). Treatment of human immunodeficiency virus infection with saquinavir, zidovudine, and zalcitabine. *New England Journal of Medicine*, *334*(16), 1011–1018. https://doi.org/10.1056/nejm199604183341602

Corey, L., Wald, A., Celum, C. L., & Quinn, T. C. (2004). The effects of herpes simplex virus-2 on hiv-1 acquisition and transmission: A review of two overlapping epidemics. *JAIDS Journal of Acquired Immune Deficiency Syndromes*, *35*(5), 435–445. https://doi.org/10.1097/00126334-200404150-00001

Cornelissen, J. H., Song, Y.-B., Yu, F.-H., & Dong, M. (2014). Plant traits and ecosystem effects of clonality: A new research agenda. *Annals of Botany*, *114*(2), 369–376. https://doi.org/10.1093/aob/mcu113

Craigie, R., & Bushman, F. D. (2012). Hiv dna integration. *Cold Spring Harbor Perspectives in Medicine*, *2*(7). https://doi.org/10.1101/cshperspect.a006890

Crooks, A. M., Bateson, R., Cope, A. B., Dahl, N. P., Griggs, M. K., Kuruc, J. D., Gay, C. L., Eron, J. J., Margolis, D. M., Bosch, R. J., & et al. (2015). Precise quantitation of the latent hiv-1 reservoir: Implications for eradication strategies. *Journal of Infectious Diseases*, *212*(9), 1361–1365. https://doi.org/10.1093/infdis/jiv218

Crowell, T. A., Colby, D. J., Pinyakorn, S., Sacdalan, C., Pagliuzza, A., Intasan, J., Benjapornpong, K., Tangnaree, K., Chomchey, N., Kroon, E., & et al. (2019). Safety and efficacy of vrc01 broadly neutralising antibodies in adults with acutely treated hiv (rv397): A phase 2, randomised, double-blind, placebo-controlled trial. *The Lancet HIV*, *6*(5). https://doi.org/10.1016/s2352-3018(19)30053-0

D'Aquila, R. T. (1996). Nevirapine, zidovudine, and didanosine compared with zidovudine and didanosine in patients with hiv-1 infection. *Annals of Internal Medicine*, *124*(12), 1019. https://doi.org/10.7326/0003-4819-124-12-199606150-00001

Darcis, G., Berkhout, B., & Pasternak, A. O. (2019). The quest for cellular markers of hiv reservoirs: Any color you like. *Frontiers in Immunology*, *10*. https://doi.org/10.3389/fimmu.2019.02251

Davey, R. T., Bhat, N., Yoder, C., Chun, T.-W., Metcalf, J. A., Dewar, R., Natarajan, V., Lempicki, R. A., Adelsberger, J. W., Miller, K. D., & et al. (1999). Hiv-1 and t cell dynamics after interruption of highly active antiretroviral therapy (haart) in patients with a history of sustained viral suppression. *Proceedings of the National Academy of Sciences*, *96*(26), 15109–15114. https://doi.org/10.1073/pnas.96.26.15109

Deacon, N. J., Tsykin, A., Solomon, A., Smith, K., Ludford-Menting, M., Hooker, D. J., McPhee, D. A., Greenway, A. L., Ellett, A., Chatfield, C., & et al. (1995). Genomic structure of an attenuated quasi species of hiv-1 from a blood transfusion donor and recipients. *Science*, *270*(5238), 988–991. https://doi.org/10.1126/science.270.5238.988

de Wolf, F., Spijkerman, I., Schellekens, P. T., Langendam, M., Kuiken, C., Bakker, M., Roos, M., Coutinho, R., Miedema, F., Goudsmit, J., & et al. (1997). Aids prognosis based on hiv-1 rna, cd4+ t-cell count and function. *AIDS*, *11*(15), 1799–1806. https://doi.org/10.1097/00002030-199715000-00003

Didigu, C. A., & Doms, R. W. (2012). Novel approaches to inhibit hiv entry. *Viruses*, *4*(2), 309–324. https://doi.org/10.3390/v4020309

Domingo, E., & Holland, J. J. (1997). Rna virus mutations and fitness for survival. *Annual Review of Microbiology*, *51*(1), 151–178. https://doi.org/10.1146/annurev.micro.51.1.151

Douek, D. C., Betts, M. R., Brenchley, J. M., Hill, B. J., Ambrozak, D. R., Ngai, K.-L., Karandikar, N. J., Casazza, J. P., & Koup, R. A. (2002). A novel approach to the analysis of specificity, clonality, and frequency of hiv-specific t cell responses reveals a potential mechanism for control of viral escape. *The Journal of Immunology*, *168*(6), 3099–3104. https://doi.org/10.4049/jimmunol.168.6.3099

Douek, D. C., Picker, L. J., & Koup, R. A. (2003). T cell dynamics in hiv-1 infection. *Annual Review of Immunology*, *21*(1), 265–304. https://doi.org/10.1146/annurev.immunol.21.120601.141053

Dutilleul, A., Rodari, A., & Van Lint, C. (2020). Depicting hiv-1 transcriptional mechanisms: A summary of what we know. *Viruses*, *12*(12), 1385. https://doi.org/10.3390/v12121385

Eisele, E., & Siliciano, R. (2012). Redefining the viral reservoirs that prevent hiv-1 eradication. *Immunity*, *37*(3), 377–388. https://doi.org/10.1016/j.immuni.2012.08.010

Fauci, A. S. (2008). 25 years of hiv. *Nature*, *453*(7193), 289–290. https://doi.org/10.1038/453289a

Feder, A. F., Rhee, S.-Y., Holmes, S. P., Shafer, R. W., Petrov, D. A., & Pennings, P. S. (2016). More effective drugs lead to harder selective sweeps in the evolution of drug resistance in hiv-1. *eLife*, *5*. https://doi.org/10.7554/elife.10670

Ferguson, M. R., Rojo, D. R., von Lindern, J. J., & O'Brien, W. A. (2002). Hiv-1 replication cycle. *Clinics in Laboratory Medicine*, *22*(3), 611–635. https://doi.org/10.1016/s0272-2712(02)00015-x

Ferreira, R.-C., Prodger, J. L., Redd, A. D., & Poon, A. F. (2021). Quantifying the clonality and dynamics of the within-host hiv-1 latent reservoir. *Virus Evolution*, *7*(1). https://doi.org/10.1093/ve/veaa104

Finzi, D., Hermankova, M., Pierson, T., Carruth, L. M., Buck, C., Chaisson, R. E., Quinn, T. C., Chadwick, K., Margolick, J., Brookmeyer, R., & et al. (1997a). Identification of a reservoir for hiv-1 in patients on highly active antiretroviral therapy. *Science*, *278*(5341), 1295–1300. https://doi.org/10.1126/science.278.5341.1295

Finzi, D., Hermankova, M., Pierson, T., Carruth, L. M., Buck, C., Chaisson, R. E., Quinn, T. C., Chadwick, K., Margolick, J., Brookmeyer, R., & et al. (1997b). Identification of a reservoir for hiv-1 in patients on highly active antiretroviral therapy. *Science*, *278*(5341), 1295–1300. https://doi.org/10.1126/science.278.5341.1295

Fisher, A. G., Feinberg, M. B., Josephs, S. F., Harper, M. E., Marselle, L. M., Reyes, G., Gonda, M. A., Aldovini, A., Debouk, C., Gallo, R. C., & et al. (1986). The trans-activator gene of htlv-iii is essential for virus replication. *Nature*, *320*(6060), 367–371. https://doi.org/10.1038/320367a0

Fletcher, C. V., Staskus, K., Wietgrefe, S. W., Rothenberger, M., Reilly, C., Chipman, J. G., Beilman, G. J., Khoruts, A., Thorkelson, A., Schmidt, T. E., & et al. (2014). Persistent hiv-1 replication is associated with lower antiretroviral drug concentrations in lymphatic tissues. *Proceedings of the National Academy of Sciences*, *111*(6), 2307–2312. https://doi.org/10.1073/pnas.1318249111

Fletcher, W., & Yang, Z. (2009). Indelible: A flexible simulator of biological sequence evolution. *Molecular Biology and Evolution*, *26*(8), 1879–1888. https://doi.org/10.1093/molbev/msp098

Fraser, C., Hollingsworth, T. D., Chapman, R., de Wolf, F., & Hanage, W. P. (2007). Variation in hiv-1 set-point viral load: Epidemiological analysis and an evolutionary hypothesis. *Proceedings of the National Academy of Sciences*, *104*(44), 17441–17446. https://doi.org/10.1073/pnas.0708559104

Freed, E. O. (2001). *Somatic Cell and Molecular Genetics*, *26*(1/6), 13–33. https://doi.org/10.1023/a:1021070512287

Frost, S. D., Magalis, B. R., & Kosakovsky Pond, S. L. (2018). Neutral theory and rapidly evolving viral pathogens. *Molecular Biology and Evolution*, *35*(6), 1348–1354. https://doi.org/10.1093/molbev/msy088

Frost, S. D., & Volz, E. M. (2013). Modelling tree shape and structure in viral phylodynamics. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *368*(1614), 20120208. https://doi.org/10.1098/rstb.2012.0208

Frost, S. D., & McLean, A. R. (1994). Quasispecies dynamics and the emergence of drug resistance during zidovudine therapy of hiv infection. *AIDS*, *8*(3), 323–332. https://doi.org/10.1097/00002030-199403000-00005

Gagniuc, P. A. (2017). *Markov chains: From theory to implementation and experimentation*. Wiley Blackwell.

Garcia-Gasco, P., Maida, I., Blanco, F., Barreiro, P., Martin-Carbonero, L., Vispo, E., Gonzalez-Lahoz, J., & Soriano, V. (2008). Episodes of low-level viral rebound in hiv-infected patients on antiretroviral therapy: Frequency, predictors and outcome. *Journal of Antimicrobial Chemotherapy*, *61*(3), 699–704. https://doi.org/10.1093/jac/dkm516

Gillespie, D. T. (1976). A general method for numerically simulating the stochastic time evolution of coupled chemical reactions. *Journal of Computational Physics*, *22*(4), 403–434. https://doi.org/10.1016/0021-9991(76)90041-3

Gini, C. (1936). On the measure of concentration with special reference to income and statistics. *Colorado College Publication*, *General Series*(208), 73–79.

Gnanakaran, S., Bhattacharya, T., Daniels, M., Keele, B. F., Hraber, P. T., Lapedes, A. S., Shen, T., Gaschen, B., Krishnamoorthy, M., Li, H., & et al. (2011). Recurrent signature patterns in hiv-1 b clade envelope glycoproteins associated with either early or chronic infections. *PLoS Pathogens*, *7*(9). https://doi.org/10.1371/journal.ppat.1002209

Go, E. P., Ding, H., Zhang, S., Ringe, R. P., Nicely, N., Hua, D., Steinbock, R. T., Golabek, M., Alin, J., Alam, S. M., & et al. (2017). Glycosylation benchmark profile for hiv-1 envelope glycoprotein production based on eleven env trimers. *Journal of Virology*, *91*(9). https://doi.org/10.1128/jvi.02428-16

Goff, S. P. (2007). Host factors exploited by retroviruses. *Nature Reviews Microbiology*, *5*(4), 253–263. https://doi.org/10.1038/nrmicro1541

Grabar, S., Selinger-Leneman, H., Abgrall, S., Pialoux, G., Weiss, L., & Costagliola, D. (2009). Prevalence and comparative characteristics of long-term nonprogressors and hiv controller patients in the french hospital database on hiv. *AIDS*, *23*(9), 1163–1169. https://doi.org/10.1097/qad.0b013e32832b44c8

Gray, R. R., Parker, J., Lemey, P., Salemi, M., Katzourakis, A., & Pybus, O. G. (2011). The mode and tempo of hepatitis c virus evolution within and among hosts. *BMC Evolutionary Biology*, *11*(1). https://doi.org/10.1186/1471-2148-11-131

Günthard, H., Wong, J., Spina, C., Ignacio, C., Kwok, S., Christopherson, C., Hwang, J., Haubrich, R., Havlir, D., Richman, D., & et al. (2000). Effect of influenza vaccination on viral replication and immune response in persons infected with human immunodeficiency virus receiving potent antiretroviral therapy. *The Journal of Infectious Diseases*, *181*(2), 522–531. https://doi.org/10.1086/315260

Gupta, R. K., Peppa, D., Hill, A. L., Gálvez, C., Salgado, M., Pace, M., McCoy, L. E., Griffith, S. A., Thornhill, J., Alrubayyi, A., & et al. (2020). Evidence for hiv-1 cure after ccr5/32 allogeneic haemopoietic stem-cell transplantation 30 months post analytical treatment interruption: A case report. *The Lancet HIV*, *7*(5). https://doi.org/10.1016/s2352-3018(20)30069-2

Gwadz, M., Cleland, C. M., Freeman, R., Wilton, L., Collins, L. M., L. Hawkins, R., Ritchie, A. S., Leonard, N. R., Jonas, D. F., Korman, A., & et al. (2021). Stopping, starting, and sustaining hiv antiretroviral therapy: A mixed-methods exploration among african american/black and latino long-term survivors of hiv in an urban context. *BMC Public Health*, *21*(1). https://doi.org/10.1186/s12889-021-10464-x

Herbeck, J. T., Rolland, M., Liu, Y., McLaughlin, S., McNevin, J., Zhao, H., Wong, K., Stoddard, J. N., Raugi, D., Sorensen, S., & et al. (2011). Demographic processes affect hiv-1 evolution in primary infection before the onset of selective processes. *Journal of Virology*, *85*(15), 7523–7534. https://doi.org/10.1128/jvi.02697-10

Ho, D. D., Neumann, A. U., Perelson, A. S., Chen, W., Leonard, J. M., & Markowitz, M. (1995). Rapid turnover of plasma virions and cd4 lymphocytes in hiv-1 infection. *Nature*, *373*(6510), 123–126. https://doi.org/10.1038/373123a0

Ho, Y.-C., Shan, L., Hosmane, N., Wang, J., Laskey, S., Rosenbloom, D., Lai, J., Blankson, J., Siliciano, J., Siliciano, R., & et al. (2013). Replication-competent noninduced proviruses in the latent reservoir increase barrier to hiv-1 cure. *Cell*, *155*(3), 540–551. https://doi.org/10.1016/j.cell.2013.09.020

Hoehn, K. B., Gall, A., Bashford-Rogers, R., Fidler, S. J., Kaye, S., Weber, J. N., McClure, M. O., Kellam, P., & Pybus, O. G. (2015). Dynamics of immunoglobulin sequence diversity in hiv-1 infected individuals. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *370*(1676), 20140241. https://doi.org/10.1098/rstb.2014.0241

Holmes, C. B., Losina, E., Walensky, R. P., Yazdanpanah, Y., & Freedberg, K. A. (2003). Review of human immunodeficiency virus type 1–related opportunistic infections in sub-saharan africa. *Clinical Infectious Diseases*, *36*(5), 652–662. https://doi.org/10.1086/367655

Holmes, E., Nee, S., Rambaut, A., Garnett, G., & Harvey, P. (1995). Revealing the history of infectious disease epidemics through phylogenetic trees. *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences*, *349*(1327), 33–40. https://doi.org/10.1098/rstb.1995.0088

Hosmane, N. N., Kwon, K. J., Bruner, K. M., Capoferri, A. A., Beg, S., Rosenbloom, D. I., Keele, B. F., Ho, Y.-C., Siliciano, J. D., Siliciano, R. F., & et al. (2017). Proliferation of latently infected cd4+ t cells carrying replication-competent hiv-1: Potential role in latent reservoir dynamics. *Journal of Experimental Medicine*, *214*(4), 959–972. https://doi.org/10.1084/jem.20170193

Hsu, J., Bryson, Y., Persaud, D., Browning, R., Mellors, J., Riches, M., Tobin, N., Bone, F., Golner, A., Warshaw, M., & et al. (2022). Conference on retroviruses and opportunistic infection. In *Hiv-1 remission with ccr5delta32/delta32 haplo-cord transplant in a us woman: Impaact p1107*.

Hu, W.-S., & Hughes, S. H. (2012). Hiv-1 reverse transcription. *Cold Spring Harbor Perspectives in Medicine*, *2*(10). https://doi.org/10.1101/cshperspect.a006882

Hughes, A. R., & Stachowicz, J. J. (2009). Ecological impacts of genotypic diversity in the clonal seagrasszostera marina. *Ecology*, *90*(5), 1412–1419. https://doi.org/10.1890/07-2030.1

Hütter, G., Nowak, D., Mossner, M., Ganepola, S., Müßig, A., Allers, K., Schneider, T., Hofmann, J., Kücherer, C., Blau, O., & et al. (2009). Long-term control of hiv byccr5delta32/delta32 stem-cell transplantation. *New England Journal of Medicine*, *360*(7), 692–698. https://doi.org/10.1056/nejmoa0802905

Iman, R. L., & Conover, W. J. (1982). A distribution-free approach to inducing rank correlation among input variables. *Communications in Statistics - Simulation and Computation*, *11*(3), 311–334. https://doi.org/10.1080/03610918208812265

Ingerson, B., Evans, C., & Ben-Kiki, O. (n.d.). Yet another markup language (yaml) 1.0. https://yaml.org/spec/history/2001-08-01.html

Iversen, A. K., Shpaer, E. G., Rodrigo, A. G., Hirsch, M. S., Walker, B. D., Sheppard, H. W., Merigan, T. C., & Mullins, J. I. (1995). Persistence of attenuated rev genes in a human immunodeficiency virus type 1-infected asymptomatic individual. *Journal of Virology*, *69*(9), 5743–5753. https://doi.org/10.1128/jvi.69.9.5743-5753.1995

Jones, B. R., Kinloch, N. N., Horacsek, J., Ganase, B., Harris, M., Harrigan, P. R., Jones, R. B., Brockman, M. A., Joy, J. B., Poon, A. F., & et al. (2018). Phylogenetic approach to recover integration dates of latent hiv sequences within-host. *Proceedings of the National Academy of Sciences*, *115*(38). https://doi.org/10.1073/pnas.1802028115

Jones, L., & Perelson, A. (2005). Opportunistic infection as a cause of transient viremia in chronically infected hiv patients under treatment with haart. *Bulletin of Mathematical Biology*, *67*(6), 1227–1251. https://doi.org/10.1016/j.bulm.2005.01.006

Joos, B., Fischer, M., Kuster, H., Pillai, S. K., Wong, J. K., Böni, J., Hirschel, B., Weber, R., Trkola, A., Günthard, H. F., & et al. (2008). Hiv rebounds from latently infected cells, rather than from continuing low-level replication. *Proceedings of the National Academy of Sciences*, *105*(43), 16725–16730. https://doi.org/10.1073/pnas.0804192105

Joseph, S. B., Swanstrom, R., Kashuba, A. D., & Cohen, M. S. (2015). Bottlenecks in hiv-1 transmission: Insights from the study of founder viruses. *Nature Reviews Microbiology*, *13*(7), 414–425. https://doi.org/10.1038/nrmicro3471

Keele, B. F., Giorgi, E. E., Salazar-Gonzalez, J. F., Decker, J. M., Pham, K. T., Salazar, M. G., Sun, C., Grayson, T., Wang, S., Li, H., & et al. (2008). Identification and characterization of transmitted and early founder virus envelopes in primary hiv-1 infection. *Proceedings of the National Academy of Sciences*, *105*(21), 7552–7557. https://doi.org/10.1073/pnas.0802203105

Kimura, M. (1968). Evolutionary rate at the molecular level. *Nature*, *217*(5129), 624–626. https://doi.org/10.1038/217624a0

Kowalski, M., Potz, J., Basiripour, L., Dorfman, T., Goh, W. C., Terwilliger, E., Dayton, A., Rosen, C., Haseltine, W., Sodroski, J., & et al. (1987). Functional regions of the envelope glycoprotein of human immunodeficiency virus type 1. *Science*, *237*(4820), 1351–1355. https://doi.org/10.1126/science.3629244

Kudesia, G., & Wreghitt, T. (2009). Human immunodeficiency virus (hiv) and acquired immunodeficiency syndrome (aids). *Clinical and Diagnostic Virology*, 54–61. https://doi.org/10.1017/cbo9780511575778.013

Kulkosky, J., Jones, K. S., Katz, R. A., Mack, J. P., & Skalka, A. M. (1992). Residues critical for retroviral integrative recombination in a region that is highly conserved among retroviral/retrotransposon integrases and bacterial insertion sequence transposases. *Molecular and Cellular Biology*, *12*(5), 2331–2338. https://doi.org/10.1128/mcb.12.5.2331-2338.1992

Kulpa, D. A., & Chomont, N. (2015). Hiv persistence in the setting of antiretroviral therapy: When, where and how does hiv hide? *Journal of Virus Eradication*, *1*(2), 59–66. https://doi.org/10.1016/s2055-6640(20)30490-8

Laskey, S. B., Pohlmeyer, C. W., Bruner, K. M., & Siliciano, R. F. (2016). Evaluating clonal expansion of hiv-infected cells: Optimization of pcr strategies to predict clonality. *PLOS Pathogens*, *12*(8). https://doi.org/10.1371/journal.ppat.1005689

Lemey, P. (2005). Evolutionary dynamics of human retroviruses investigated through full-genome scanning. *Molecular Biology and Evolution*, *22*(4), 942–951. https://doi.org/10.1093/molbev/msi078

Li, S., Juarez, J., Alali, M., Dwyer, D., Collman, R., Cunningham, A., & Naif, H. M. (1999). Persistent ccr5 utilization and enhanced macrophage tropism by primary blood human immunodeficiency virus type 1 isolates from advanced stages of disease and comparison to tissue-derived isolates. *Journal of Virology*, *73*(12), 9741–9755. https://doi.org/10.1128/jvi.73.12.9741-9755.1999

Lorenzi, J. C., Cohen, Y. Z., Cohn, L. B., Kreider, E. F., Barton, J. P., Learn, G. H., Oliveira, T., Lavine, C. L., Horwitz, J. A., Settler, A., & et al. (2016). Paired quantitative and qualitative assessment of the replication-competent hiv-1 reservoir and comparison with integrated proviral dna. *Proceedings of the National Academy of Sciences*, *113*(49). https://doi.org/10.1073/pnas.1617789113

Luciw, P. A. (1996). Human immunodeficiency viruses and their replication. in: Fields, b.n., knipe, d.m., howley, p.m., chanock, r.m., melnick, j.l., monath, t.p., roizman, b. & straus, s.e., eds, fields virology, vol. 2, 3rd ed., philadelphia, lippincott-raven, pp. 1881–1952. *Fields Virology*, *2*(3), 1881–1952.

Lythgoe, K. A., & Fraser, C. (2012). New insights into the evolutionary rate of hiv-1 at the within-host and epidemiological levels. *Proceedings of the Royal Society B: Biological Sciences*, *279*(1741), 3367–3375. https://doi.org/10.1098/rspb.2012.0595

Maldarelli, F., Wu, X., Su, L., Simonetti, F. R., Shao, W., Hill, S., Spindler, J., Ferris, A. L., Mellors, J. W., Kearney, M. F., & et al. (2014). Specific hiv integration sites are linked to clonal expansion and persistence of infected cells. *Science*, *345*(6193), 179–183. https://doi.org/10.1126/science.1254194

Mandell, G. L., Douglas, R. G., & Bennett, J. E. (1995). *Mandell, douglas and bennett's principles and practice of infectious diseases*. Churchill Livingstone.

Martincorena, I. (2019). Somatic mutation and clonal expansions in human tissues. *Genome Medicine*, *11*(1). https://doi.org/10.1186/s13073-019-0648-4

Massanella, M., & Richman, D. D. (2016). Measuring the latent reservoir in vivo. *Journal of Clinical Investigation*, *126*(2), 464–472. https://doi.org/10.1172/jci80567

Matthews, P. C., Prendergast, A., Leslie, A., Crawford, H., Payne, R., Rousseau, C., Rolland, M., Honeyborne, I., Carlson, J., Kadie, C., & et al. (2008). Central role of reverting mutations in hla associations with human immunodeficiency virus set point. *Journal of Virology*, *82*(17), 8548–8559. https://doi.org/10.1128/jvi.00580-08

McGranahan, N., Favero, F., de Bruin, E. C., Birkbak, N. J., Szallasi, Z., & Swanton, C. (2015). Clonal status of actionable driver events and the timing of mutational processes in cancer evolution. *Science Translational Medicine*, *7*(283). https://doi.org/10.1126/scitranslmed.aaa1408

McKay, M. D., Beckman, R. J., & Conover, W. J. (1979). A comparison of three methods for selecting values of input variables in the analysis of output from a computer code. *Technometrics*, *21*(2), 239. https://doi.org/10.2307/1268522

Mellors, J. W., Rinaldo, C. R., Gupta, P., White, R. M., Todd, J. A., & Kingsley, L. A. (1996). Prognosis in hiv-1 infection predicted by the quantity of virus in plasma. *Science*, *272*(5265), 1167–1170. https://doi.org/10.1126/science.272.5265.1167

Moir, S., Chun, T.-W., & Fauci, A. S. (2011). Pathogenic mechanisms of hiv disease. *Annual Review of Pathology: Mechanisms of Disease*, *6*(1), 223–248. https://doi.org/10.1146/annurev-pathol-011110-130254

Naif, H. M., Li, S., Alali, M., Chang, J., Mayne, C., Sullivan, J., & Cunningham, A. L. (1999). Definition of the stage of host cell genetic restriction of replication of human immunodeficiency virus type 1 in monocytes and monocyte-derived macrophages by using twins. *Journal of Virology*, *73*(6), 4866–4881. https://doi.org/10.1128/jvi.73.6.4866-4881.1999

Nelder, J. A., & Wedderburn, R. W. (1972). Generalized linear models. *Journal of the Royal Statistical Society. Series A (General)*, *135*(3), 370. https://doi.org/10.2307/2344614

Nettles, R. E. (2005). Intermittent hiv-1 viremia (blips) and drug resistance in patients receiving haart. *JAMA*, *293*(7), 817. https://doi.org/10.1001/jama.293.7.817

Nettles, R. E., Kieffer, T. L., Kwon, P., Monie, D., Han, Y., Parsons, T., Cofrancesco, J., Gallant, J. E., Quinn, T. C., Jackson, B., Flexner, C., Carson, K., Ray, S., Persaud, D., & Siliciano, R. F. (2005). Intermittent HIV-1 Viremia (Blips) and Drug Resistance in Patients Receiving HAART. *JAMA*, *293*(7), 817–829. https://doi.org/10.1001/jama.293.7.817

Ni, J., Wang, D., & Wang, S. (2018). The ccr5-delta32 genetic polymorphism and hiv-1 infection susceptibility: A meta-analysis. *Open Medicine*, *13*(1), 467–474. https://doi.org/10.1515/med-2018-0062

Okoye, A., Meier-Schellersheim, M., Brenchley, J. M., Hagen, S. I., Walker, J. M., Rohankhedkar, M., Lum, R., Edgar, J. B., Planer, S. L., Legasse, A., & et al. (2007). Progressive cd4+ central–memory t cell decline results in cd4+ effector–memory insufficiency and overt disease in chronic siv infection. *Journal of Experimental Medicine*, *204*(9), 2171–2185. https://doi.org/10.1084/jem.20070567

Okoye, A. A., & Picker, L. J. (2013). Cd4(+)t-cell depletion in hiv infection: Mechanisms of immunological failure. *Immunological Reviews*, *254*(1), 54–64. https://doi.org/10.1111/imr.12066

Ott, D. E. (2008). Cellular proteins detected in hiv-1. *Reviews in Medical Virology*, *18*(3), 159–175. https://doi.org/10.1002/rmv.570

Palella, F. J., Delaney, K. M., Moorman, A. C., Loveless, M. O., Fuhrer, J., Satten, G. A., Aschman, D. J., & Holmberg, S. D. (1998). Declining morbidity and mortality among patients with advanced human immunodeficiency virus infection. *New England Journal of Medicine*, *338*(13), 853–860. https://doi.org/10.1056/nejm199803263381301

Park, Y., Martin, M., & Koelle, K. (2021). Epidemiological inference for emerging viruses using segregating sites. https://doi.org/10.1101/2021.07.07.451508

Patro, S. C., Brandt, L. D., Bale, M. J., Halvas, E. K., Joseph, K. W., Shao, W., Wu, X., Guo, S., Murrell, B., Wiegand, A., & et al. (2019). Combined hiv-1 sequence and integration site analysis informs viral dynamics and allows reconstruction of replicating viral ancestors. *Proceedings of the National Academy of Sciences*, *116*(51), 25891–25899. https://doi.org/10.1073/pnas.1910334116

Perelson, A. S., & Ribeiro, R. M. (2013). Modeling the within-host dynamics of hiv infection. *BMC Biology*, *11*(1). https://doi.org/10.1186/1741-7007-11-96

Pollack, R. A., Jones, R. B., Pertea, M., Bruner, K. M., Martin, A. R., Thomas, A. S., Capoferri, A. A., Beg, S. A., Huang, S.-H., Karandish, S., & et al. (2017). Defective hiv-1 proviruses are expressed and can be recognized by cytotoxic t lymphocytes, which shape the proviral landscape. *Cell Host & Microbe*, *21*(4). https://doi.org/10.1016/j.chom.2017.03.008

Pollard, V. W., & Malim, M. H. (1998). The hiv-1 rev protein. *Annual Review of Microbiology*, *52*(1), 491–532. https://doi.org/10.1146/annurev.micro.52.1.491

Pybus, O. G., & Rambaut, A. (2009). Evolutionary analysis of the dynamics of viral infectious disease. *Nature Reviews Genetics*, *10*(8), 540–550. https://doi.org/10.1038/nrg2583

Rasmussen, D. A., Ratmann, O., & Koelle, K. (2011). Inference for nonlinear epidemiological models using genealogies and time series. *PLoS Computational Biology*, *7*(8). https://doi.org/10.1371/journal.pcbi.1002136

Reed, J. B., Shrestha, P., Were, D., Chakare, T., Mutegi, J., Wakhutu, B., Musau, A., Nonyana, N. M., Christensen, A., Patel, R., & et al. (2021). Hiv prep is more than art-lite: Longitudinal study of real-world prep services data identifies missing measures meaningful to hiv prevention programming. *Journal of the International AIDS Society*, *24*(10). https://doi.org/10.1002/jia2.25827

Reeves, D. B., Duke, E. R., Wagner, T. A., Palmer, S. E., Spivak, A. M., & Schiffer, J. T. (2017). A majority of hiv persistence during antiretroviral therapy is due to infected cell proliferation. https://doi.org/10.1101/146977

Rieder, P., Joos, B., Scherrer, A. U., Kuster, H., Braun, D., Grube, C., Niederost, B., Leemann, C., Gianella, S., Metzner, K. J., & et al. (2011). Characterization of human immunodeficiency virus type 1 (hiv-1) diversity and tropism in 145 patients with primary hiv-1 infection. *Clinical Infectious Diseases*, *53*(12), 1271–1279. https://doi.org/10.1093/cid/cir725

Sadras, V., & Bongiovanni, R. (2004). Use of lorenz curves and gini coefficients to assess yield inequality within paddocks. *Field Crops Research*, *90*(2-3), 303–310. https://doi.org/10.1016/j.fcr.2004.04.003

Sagar, M., Laeyendecker, O., Lee, S., Gamiel, J., Wawer, M., Gray, R., Serwadda, D., Sewankambo, N., Shepherd, J., Toma, J., & et al. (2009). Selection of hiv variants with signature genotypic characteristics during heterosexual transmission. *The Journal of Infectious Diseases*, *199*(4), 580–589. https://doi.org/10.1086/596557

Salazar-Gonzalez, J. F., Salazar, M. G., Keele, B. F., Learn, G. H., Giorgi, E. E., Li, H., Decker, J. M., Wang, S., Baalwa, J., Kraus, M. H., & et al. (2009). Genetic identity, biological phenotype, and evolutionary pathways of transmitted/founder viruses in acute and early hiv-1 infection. *Journal of Experimental Medicine*, *206*(6), 1273–1289. https://doi.org/10.1084/jem.20090378

Salminen, M. O., Koch, C., Sanders-Buell, E., Ehrenberg, P. K., Michael, N. L., Carr, J. K., Burke, D. S., & McCutchan, F. E. (1995). Recovery of virtually full-length hiv-1 provirus of diverse subtypes from primary virus cultures using the polymerase chain reaction. *Virology*, *213*(1), 80–86. https://doi.org/10.1006/viro.1995.1548

Sarafianos, S. G., Marchand, B., Das, K., Himmel, D. M., Parniak, M. A., Hughes, S. H., & Arnold, E. (2009). Structure and function of hiv-1 reverse transcriptase: Molecular mechanisms of polymerization and inhibition. *Journal of Molecular Biology*, *385*(3), 693–713. https://doi.org/10.1016/j.jmb.2008.10.071

Schacker, T. (1996). Clinical and epidemiologic features of primary hiv infection. *Annals of Internal Medicine*, *125*(4), 257. https://doi.org/10.7326/0003-4819-125-4-199608150-00001

Schacker, T. W. (1998). Biological and virologic characteristics of primary hiv infection. *Annals of Internal Medicine*, *128*(8), 613. https://doi.org/10.7326/0003-4819-128-8-199804150-00001

Schluns, K. S., & Lefrançois, L. (2003). Cytokine control of memory t-cell development and survival. *Nature Reviews Immunology*, *3*(4), 269–279. https://doi.org/10.1038/nri1052

Schmit, J.-C., Cogniaux, J., Hermans, P., Van Vaeck, C., Sprecher, S., Van Remoortel, B., Witvrouw, M., Balzarini, J., Desmyter, J., De Clercq, E., & et al. (1996). Multiple drug resistance to nucleoside analogues and nonnucleoside reverse transcriptase inhibitors in an efficiently replicating human immunodeficiency virus type 1 patient strain. *Journal of Infectious Diseases*, *174*(5), 962–968. https://doi.org/10.1093/infdis/174.5.962

Schröder, A. R., Shinn, P., Chen, H., Berry, C., Ecker, J. R., & Bushman, F. (2002). Hiv-1 integration in the human genome favors active genes and local hotspots. *Cell*, *110*(4), 521–529. https://doi.org/10.1016/s0092-8674(02)00864-4

Shaw, G. M., & Hunter, E. (2012). Hiv transmission. *Cold Spring Harbor Perspectives in Medicine*, *2*(11). https://doi.org/10.1101/cshperspect.a006965

Siliciano, J. D., Kajdas, J., Finzi, D., Quinn, T. C., Chadwick, K., Margolick, J. B., Kovacs, C., Gange, S. J., & Siliciano, R. F. (2003). Long-term follow-up studies confirm the stability of the latent reservoir for hiv-1 in resting cd4+ t cells. *Nature Medicine*, *9*(6), 727–728. https://doi.org/10.1038/nm880

Simple linear regression. (2005). *Applied Linear Regression*, 19–46. https://doi.org/10.1002/0471704091.ch2

Spira, A. I., Marx, P. A., Patterson, B. K., Mahoney, J., Koup, R. A., Wolinsky, S. M., & Ho, D. D. (1996). Cellular targets of infection and route of viral dissemination after an intravaginal inoculation of simian immunodeficiency virus into rhesus macaques. *Journal of Experimental Medicine*, *183*(1), 215–225. https://doi.org/10.1084/jem.183.1.215

Staszewski, S., Miller, V., Rehmet, S., Stark, T., De Creé, J., De Brabander, M., Peeters, M., Andries, K., Moeremans, M., De Raeymaeker, M., & et al. (1996). Virological and immunological analysis of a triple combination pilot study with loviride, lamivudine and zidovudine in hiv-1-infected patients. *AIDS*, *10*(5). https://doi.org/10.1097/00002030-199605000-00001

Stengel, R. F. (2008). Mutation and control of the human immunodeficiency virus. *Mathematical Biosciences*, *213*(2), 93–102. https://doi.org/10.1016/j.mbs.2008.03.002

Sundquist, W. I., & Krausslich, H.-G. (2012). Hiv-1 assembly, budding, and maturation. *Cold Spring Harbor Perspectives in Medicine*, *2*(7). https://doi.org/10.1101/cshperspect.a006924

Takeuchi, Y., Nagumo, T., & Hoshino, H. (1988). Low fidelity of cell-free dna synthesis by reverse transcriptase of human immunodeficiency virus. *Journal of Virology*, *62*(10), 3900–3902. https://doi.org/10.1128/jvi.62.10.3900-3902.1988

Taswell, C. (1984). Limiting dilution assays for the determination of immunocompetent cell frequencies. iii. validity tests for the single-hit poisson model. *Journal of Immunological Methods*, *72*(1), 29–40. https://doi.org/10.1016/0022-1759(84)90430-7

Telesnitsky, A. (2010). Retroviruses: Molecular biology, genomics and pathogenesis. *Future Virology*, *5*(5), 539–543. https://doi.org/10.2217/fvl.10.43

Thapa, D. R., Tonikian, R., Sun, C., Liu, M., Dearth, A., Petri, M., Pepin, F., Emerson, R. O., & Reanger, A. (2015). Longitudinal analysis of peripheral blood t cell receptor diversity in patients with systemic lupus erythematosus by next-generation sequencing. *Arthritis Research & Therapy*, *17*(1). https://doi.org/10.1186/s13075-015-0655-9

Thompson, C. G., Gay, C. L., & Kashuba, A. D. (2017). Hiv persistence in gut-associated lymphoid tissues: Pharmacological challenges and opportunities. *AIDS Research and Human Retroviruses*, *33*(6), 513–523. https://doi.org/10.1089/aid.2016.0253

Tomezsko, P. J., Corbin, V. D., Gupta, P., Swaminathan, H., Glasgow, M., Persad, S., Edwards, M. D., Mcintosh, L., Papenfuss, A. T., Emery, A., & et al. (2020). Determination of rna structural diversity and its role in hiv-1 rna splicing. *Nature*, *582*(7812), 438–442. https://doi.org/10.1038/s41586-020-2253-5

UNAIDS. (2021). *Global hiv & aids statistics - fact sheet*. https://www.unaids.org/en/resources/fact-sheet

von Stockenstrom, S., Odevall, L., Lee, E., Sinclair, E., Bacchetti, P., Killian, M., Epling, L., Shao, W., Hoh, R., Ho, T., & et al. (2015). Longitudinal genetic characterization reveals that cell proliferation maintains a persistent hiv type 1 dna pool during effective hiv therapy. *Journal of Infectious Diseases*, *212*(4), 596–607. https://doi.org/10.1093/infdis/jiv092

Vrancken, B., Rambaut, A., Suchard, M. A., Drummond, A., Baele, G., Derdelinckx, I., Van Wijngaerden, E., Vandamme, A.-M., Van Laethem, K., Lemey, P., & et al. (2014). The genealogical population dynamics of hiv-1 in a large transmission chain: Bridging within and among host evolutionary rates. *PLoS Computational Biology*, *10*(4). https://doi.org/10.1371/journal.pcbi.1003505

Wagner, T. A., McLaughlin, S., Garg, K., Cheung, C. Y., Larsen, B. B., Styrchak, S., Huang, H. C., Edlefsen, P. T., Mullins, J. I., Frenkel, L. M., & et al. (2014). Proliferation of cells with hiv integrated into cancer genes contributes to persistent infection. *Science*, *345*(6196), 570–573. https://doi.org/10.1126/science.1256304

Wang, S., Li, H., Lian, Z., & Deng, S. (2022). The role of rna modification in hiv-1 infection. *International Journal of Molecular Sciences*, *23*(14), 7571. https://doi.org/10.3390/ijms23147571

Wang, Z., Simonetti, F. R., Siliciano, R. F., & Laird, G. M. (2018). Measuring replication competent hiv-1: Advances and challenges in defining the latent reservoir. *Retrovirology*, *15*(1). https://doi.org/10.1186/s12977-018-0404-7

Wei, X., Ghosh, S. K., Taylor, M. E., Johnson, V. A., Emini, E. A., Deutsch, P., Lifson, J. D., Bonhoeffer, S., Nowak, M. A., Hahn, B. H., & et al. (1995). Viral dynamics in human immunodeficiency virus type 1 infection. *Nature*, *373*(6510), 117–122. https://doi.org/10.1038/373117a0

Wolinsky, S. M., Wike, C. M., Korber, B. T., Hutto, C., Parks, W. P., Rosenblum, L. L., Kunstman, K. J., Furtado, M. R., & Muñoz, J. L. (1992). Selective transmission of human immunodeficiency virus type-1 variants from mothers to infants. *Science*, *255*(5048), 1134–1137. https://doi.org/10.1126/science.1546316

Wong, J. K., Hezareh, M., F., G. H., Havlir, D. V., Ignacio, C. C., Spina, C. A., & Richman, D. D. (1997). Recovery of replication-competent hiv despite prolonged suppression of plasma viremia. *Science*, *278*(5341), 1291–1295. https://doi.org/10.1126/science.278.5341.1291

Yang, Z., & Rannala, B. (2012). Molecular phylogenetics: Principles and practice. *Nature Reviews Genetics*, *13*(5), 303–314. https://doi.org/10.1038/nrg3186

Yeh, Y.-H. J., Yang, K., Razmi, A., & Ho, Y.-C. (2021). The clonal expansion dynamics of the hiv-1 reservoir: Mechanisms of integration site-dependent proliferation and hiv-1 persistence. *Viruses*, *13*(9), 1858. https://doi.org/10.3390/v13091858

Young, J., Rickenbach, M., Calmy, A., Bernasconi, E., Staehelin, C., Schmid, P., Cavassini, M., Battegay, M., Günthard, H. F., Bucher, H. C., & et al. (2015). Transient detectable viremia and the risk of viral rebound in patients from the swiss hiv cohort study. *BMC Infectious Diseases*, *15*(1). https://doi.org/10.1186/s12879-015-1120-8

Zanetti, G., Briggs, J. A., Grünewald, K., Sattentau, Q. J., & Fuller, S. D. (2006). Cryo-electron tomographic structure of an immunodeficiency virus envelope complex in situ. *PLoS Pathogens*, *2*(8). https://doi.org/10.1371/journal.ppat.0020083

Zanussi, S., Simonelli, C., D'Andrea, M., Caffau, C., Clerici, M., Tirelli, U., & De Paoli, P. (1996). Cd8+ lymphocyte phenotype and cytokine production in long-term nonprogressor and in progressor patients with hiv-1 infection. *Clinical and Experimental Immunology*, *105*(2), 220–224. https://doi.org/10.1046/j.1365-2249.1996.d01-746.x

Zhang, Z.-Q., Notermans, D. W., Sedgewick, G., Cavert, W., Wietgrefe, S., Zupancic, M., Gebhard, K., Henry, K., Boies, L., Chen, Z., & et al. (1998). Kinetics of cd4+ t cell repopulation of lymphoid tissues after treatment of hiv-1 infection. *Proceedings of the National Academy of Sciences*, *95*(3), 1154–1159. https://doi.org/10.1073/pnas.95.3.1154

Zhu, P., Liu, J., Bess, J., Chertova, E., Lifson, J. D., Grisé, H., Ofek, G. A., Taylor, K. A., & Roux, K. H. (2006). Distribution and three-dimensional structure of aids virus envelope spikes. *Nature*, *441*(7095), 847–852. https://doi.org/10.1038/nature04817