2018

Single Cell Analysis of the HIV-1 Latent Reservoir

Lillian Brumer Cohn

SINGLE CELL ANALYSIS OF THE HIV-1 LATENT RESERVOIR

A Thesis Presented to the Faculty of

The Rockefeller University

in Partial Fulfillment of the Requirements for

the degree of Doctor of Philosophy

by

Lillian Brumer Cohn

June 2018

SINGLE CELL ANALYSIS OF THE HIV-1 LATENT RESERVOIR

Lillian Brumer Cohn, Ph.D.

The Rockefeller University 2018

Human immunodeficiency virus type 1 (HIV-1), the virus that causes acquired immune deficiency syndrome (AIDS), is one of the world's most serious health and development challenges. Worldwide there are approximately 36.7 million people living with HIV, and tens of millions have died of AIDS-related causes since the beginning of the epidemic. Treatment of HIV-1 infection with combinations of antiretroviral drugs has significantly reduced the death rate and improved the quality of life of HIV-1 infected individuals. Despite over thirty years of HIV-1 research, however, both a cure and a vaccine remain elusive. Complete eradication of HIV-1 by antiretroviral drugs is prevented by the persistence of rare, long-lived, latently infected cells. These cells, called the latent reservoir, are thought to resist immune clearance and viral cytopathic effects by harboring a transcriptionally quiescent integrated HIV-1 provirus. As a result, interruption of suppressive therapy almost inevitably results in rapid viral rebound, which originates from these latently infected cells and prevents HIV-1 cure. It is thought that establishing the reservoir requires intact retroviral integration into the host cell genome and subsequent transcriptional silencing of the integrated provirus. These are rare events and these cells have no known distinguishing surface markers, which has made it difficult to define the precise cellular and molecular nature of the reservoir. The long half-life of the latent reservoir has been attributed to a stable pool of long-lived latently infected CD4+ T cells.

An alternative explanation, consistent with the frequent occurrence of monotypic viral sequences, is that infected latent cells are maintained in part by cell proliferation. T cell division and productive HIV-1 transcription are mediated by shared metabolic and transcriptional pathways, and productive HIV-1 infection typically leads to CD4+ T cell death. Thus, how infected cells survive while dividing is unknown. I focused my thesis on characterizing this latent reservoir in virally suppressed, HIV-1 infected individuals and examining the mechanisms of HIV-1 latency.

In the first part of this thesis, using a novel single-cell, high throughput integration site sequencing method, I demonstrate that HIV-1 infected cells are capable of cell division, but that the great majority of the largest expanded clones contain defective proviruses which cannot contribute to the replication competent rebound virus. In the second part of this thesis, using an assay to qualitatively and quantitatively characterize the latent reservoir, I suggest that the replication competent latent reservoir may, in fact, be maintained in part by rare cell division events. And finally, I developed a novel isolation strategy which allowed single cell characterization of recently reactivated latent cells. I was able to obtain reactivated latent T cells that produced intact, replication competent HIV-1. By sequencing the T cell receptors, I prove that these isolated latent cells are expanded T cell clones. Single cell gene expression analysis revealed that latent cells share a specific gene profile that prominently includes genes implicated in silencing the virus, T cell exhaustion markers, and genes that may aid in identification of specific CD4+ T cell subsets prone to latent infection. Together, the data supports a model for latency whereby infected T cells turn on a gene expression program that suppresses viral

replication during cell division thereby preventing activation of the cell death pathways that are normally triggered by HIV-1 infection.

ACKNOWLEDGEMENTS

Completing this PhD would not have been possible without immense support:

First and foremost, I would like to thank Michel. Being a part of the Nussenzweig laboratory has been a deeply formative experience for me scientifically and personally. I cannot imagine a better mentor, a better scientific environment to pursue important work, nor can I imagine a better group of scientists to have as teammates. Michel, thank you for your generosity, your commitment, and for the example you've set. Thank you for advocating for me. Thank you for the stories of my grandfather and of Ralph, for carrying on their legacies, and for teaching me to continue that legacy – it has been a privilege to work with you.

Since my first day, and every day since, Mila Jankovic has been my tireless champion, scientifically and otherwise. From morning brainstorming, to baking birthday cakes, no one person has impacted my Rockefeller experience more. Mila, I am forever grateful for everything. You have become family.

None of my experiments would be possible without the tireless work of our human machine, Israel Tojal da Silva. Israel is a rare bioinformatician whose thoughtfulness has made every single experiment we've done better and more meaningful. And importantly, he never panics when I start a conversation with "Israel, we have a major problem," which somehow happens often, yet is almost never true.

There's only one person whose daily presence in the lab is arguably more necessary than even Michel's. Zoran Jankovic keeps everything going. Without him, experiments would be slower, less efficient and we'd be less happy.

I am grateful to Daniel Mucida for his friendship and guidance. I am inspired by his consistent fight to make the world a better, more inclusive, sustainable place and the uncompromising way he manages to do science with those values at the forefront.

I have had the joy of mentoring two students during my PhD: Lion Uhl and Amy Huang. Thank you for believing in the work, for trusting me, and for your contributions to the data in this thesis. I would also like to thank all Nussenzweig lab members, past and present, for making this group such a scientific force, and having fun while doing it. In particular, Kristie Gordon, Neena Thomas and Kal Chhosphel for sorting; Julio Lorenzi and Yehuda Cohen for performing Q^2VOA experiments and discussion; Marina Caskey, Allison Butler, Katrina Millard, and Maggi Pack for incredible participant coordination and recruitment; Shiraz Belblidia, Juan Dizon, Roshni Patel, Cecile Unson-O'Brien, Irina Shimeliovich for processing samples; Thiago Oliviera and Joy Pai for bioinformatics support; Qiao Wang, Davide Robbiani, Ervin Kara, and Josh Horwitz for discussion; Anna Gazumyan for antibody production; Pia Dosenovic and Pilar Mendoza Daroca for their friendship; Gaelle Breton for FACS advice; Johannes Scheid for bequeathing me the best bench in the lab; and Jennifer McQuillian and Adriana Barillas-Batarse for administrative support.

I was privileged to work with incredible scientists before my PhD, and I would like to thank Julie McElrath, Lelia Delamarre, and Ira Mellman for preparing me, and their

constant encouragement and advice during this PhD. I would like to acknowledge David Drubin for his support during my decision to leave UC Berkeley. Finally, from middle school through college, I had dedicated science teachers who ignited my curiosity – in particular Dan Bloedel, Walter P. Spangenberg, and Ken Miller.

The best part of doing science is collaboration and conversation – and there is perhaps no greater example of that than the Oaxaca crew. Thank you to Gustavo Reyes-Teran for including me in your yearly meeting and curating a group of such wonderful and kind scientists. This group models what lifelong scientific friendship should look like.

To my thesis committee, Charlie Rice and Paul Bieniasz: thank you for your insight and expertise. The data in this thesis was made better with your continuous input and rigorous critique of our science. I am grateful to my external thesis committee member Rafi Ahmed, whose work I have long admired, for his time and participation during my thesis defense.

Rockefeller is a special place to be a graduate student thanks to the tireless work of the Dean's Office. Thank you to Sid Strickland, Emily Harms, Cris Rosario, Stephanie Fernandez, Kristen Cullen, Marta Delgado and Andrea Morris for making the Rockefeller Graduate Program science and student focused.

I would like to thank Svetlana Mosjov, for championing women and reminding me that sometimes too much thought doesn't help. And to the women who are heads-of-lab at Rockefeller, thank you for your hard work to make the path a little easier for the women in my generation – in particular Leslie Vosshall and Vanessa Ruta have offered guidance, support and a model to follow.

A number of people have made my life in New York more fun and more full: first, Elie Hirschfeld and Sally Schlesinger, for your immense generosity, for warmly welcoming me into your family, and including me for holidays. My apartment has been a sanctuary from the city and living here for the last years has kept me sane. Sally, thank you for dressing me and feeding me since 2008. Sharon Lee, who brought Seattle to New York, regularly. Gigi, for coming to be family when no one else could. To my friends here – all of you – thank you. And to my grandmother, Fern, for her love, support, and dinners.

Everyone should be fortunate enough to have an inspiring, humbling group of girlfriends, and I certainly won the lottery with mine. The strength that lies within a group of formidable women is unrivaled and I have relied daily on that power to get me through this experience.

To Daliso Leslie, thank you for giving me your light and your darkness. I am a better person because of you.

We don't choose our family, but to my cousins, distant and near, my aunts, uncle, and my two nonagenarian grandmothers: I would choose you if I could.

Finally, I dedicate this thesis to my mom Rachel, my dad David (who also generously proofread sections of this thesis), my brother Noah, and my sister Eliza. Your steadfast love has carried me through this PhD and will continue to carry me onwards. I look at you four and am awed by how you each, in your own way, approach the world with quiet confidence and compassion. I feel like the luckiest person to go through life with you. I love you.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

CHAPTER 1:

INTRODUCTION

History of Global HIV Epidemic

Acquired Immune Deficiency Syndrome (AIDS) was first identified in 1981 when young homosexual men began dying of opportunistic infections and rare malignancies. A retrovirus named Human Immunodeficiency Virus type 1 (HIV-1) was determined to be the causative agent of AIDS, now one of the worst infectious disease epidemics in recent human history. HIV-1 is primarily a sexually transmitted disease, but can be transmitted also through percutaneous, perinatal and intravenous routes (Maartens et al., 2014). There are an estimated 36.7 million people currently infected worldwide, and 25 million HIV/AIDS related deaths thus far (UNAIDS, 2017). The greatest disease burden is found in developing countries, with young adults in sub-Saharan Africa representing the highest disease prevalence worldwide (Figure 1.1) (Collaborators, 2016).

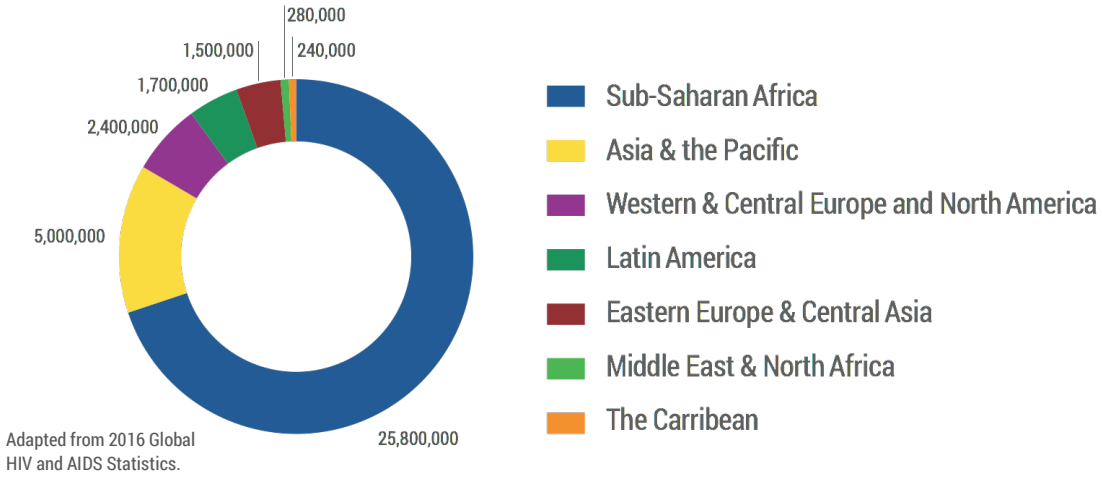


Figure 1.1 Worldwide prevalence of HIV.

The spread of infection and illness caused by HIV-1 infection have been mitigated by the administration of antiretroviral drug cocktails, either for prevention or treatment (Arts and Hazuda, 2012). Therapy, although not universally available, has improved survival rates and lengthened disease-free state. These virologically suppressed HIV-1 infected individuals, however, still have shorter life expectancy and slightly worse health outcomes than their uninfected counterparts (Hunt, 2014). Additionally, efforts to discover a cure or an effective vaccine have been largely unsuccessful. As a result, HIV infection and AIDS related illnesses are positioned to continue their attack on the global population, especially under-resourced communities, for the coming decades.

HIV-1 life cycle

HIV-1 is a human-tropic, single-stranded, positive sense, enveloped, RNA virus in the *retroviridae* family. The virus envelope protein interacts with cell surface receptors CD4 (Maddon et al., 1986; McDougal et al., 1986) and CXCR4 or CCR5 (Berger et al., 1998) to enter its target cells – primarily CD4+ T cells of the immune system. Viral entry occurs upon membrane fusion (Wilen et al., 2012), followed by reverse transcription of the RNA genome into double stranded DNA by the virally encoded Reverse Transcriptase enzyme (Hu and Hughes, 2012). HIV-1 reverse transcription is error-prone, leading to an extremely high mutation rate which results in the impressive diversity of HIV-1 genomes, even within a single individual. These mutations may confer resistance, allowing HIV-1 to evade individual treatment modalities and vaccines (Abram et al., 2010; Keele et al., 2008). The resulting double stranded DNA is imported into the nuclease where the viral

protein integrase and host co-factors mediate host genome cutting, viral DNA insertion and DNA repair around the inserted provirus (Craigie and Bushman, 2012). After successful integration, the cell is permanently infected. The fate of this provirus can take two forms: first, the virus can be targeted by host transcription machinery for the production of HIV-1 RNA. These RNAs are either translated into viral proteins to assemble new virions, or full length viral RNAs are packaged into new virus particles, which release from the host cell membrane to begin the replication cycle again. Second, and much less frequently, the virus can become latent and is not transcribed, allowing the cell to survive and avoid detection by the immune system (Finzi et al., 1997). This pool of latently infected cells, termed the latent reservoir, is long lived (Finzi et al., 1999), insensitive to antiretroviral therapy (Chun et al., 1997b; Dinoso et al., 2009), and is the source of viral rebound after therapy cessation (Joos et al., 2008). Thus, the persistence of latently infected cells is the major barrier to HIV-1 cure.

Immune response to HIV-1 and broadly neutralizing antibodies

Most HIV-1 infections occur via the mucosal route during sexual contact, though it is unclear whether HIV-1 is transmitted as a free or a cell-associated virus (Pope and Haase, 2003). The mechanism used by HIV-1 to cross the mucosal epithelium during transmission is unknown, but it's thought that virus reaches the mucosal epithelium by either vesicular transport through epithelial cells or by making contact with dendritic processes of intraepithelial dendritic cells. Transmission frequency increases if the genital mucosa is damaged by physical trauma or co-existing infection, which allows

transmission via the epithelium to occur more easily (Galvin and Cohen, 2004). Viral transmission across the mucosal barrier seems to be a rare event, however, as studies of the first detectable viremia suggest that infection is initiated by a single infectious virus called the transmitted-founder (TF) virus (Keele et al., 2008; Parrish et al., 2013). Homogeneity of the TF virus is indicative of early infection events originating from a single virus and focused to a close group of mucosal resident CD4+ cells, with early viremia enhanced by recruitment of non-tissue resident susceptible CD4+ T cells to the initial site of infection. Peak viremia occurs around 28 days post exposure and correlates with a dramatic HIV-1 induced loss of CD4+ T cells from tissues and the periphery (Brenchley et al., 2004; Keele et al., 2008). Following peak viremia, plasma levels of HIV-1 RNA gradually decrease for a period of up to 20 weeks until viral set-point is reached. Without therapeutic intervention, this viral set-point represents the homeostasis between the anti-viral activity of the immune system and the virus' ability to evade immune response. In part, the ability of the immune system to manage viral infection lies in CD4+ T cell function. Actively infected CD4+ T cells have a measured half-life of approximately one day, demonstrating the vulnerability of these cells in the context of HIV-1 infection (Markowitz et al., 2003). HIV-1-specific CD4+ T cells are required for helping B cells in the lymphoid germinal centers make high affinity antibodies. While the immune system mounts a complex and multi-tiered response against HIV-1, for the purposes of this thesis introduction, I will now focus only on the development of antibody responses.

Antibodies develop early in infection and are directed to HIV-1 Envelope (Env), a sparsely expressed protein on the surface of virions and cells actively infected by the

virus. This protein serves as the target for neutralizing humoral immune responses since it is the only viral protein expressed on the surface of the virus. Gp120 and gp41 comprise one HIV Env monomer, which come together to form a functional Env protein trimer (Ward and Wilson, 2015). The trimer is highly unstable and antibodies arise against gp41 as early as 2 weeks after infection and against gp120 3-4 weeks later (Binley et al., 1996; Tomaras et al., 2008). Despite their abundance, these antibodies do not detectably control viremia or exert evolutionary pressure on the diversifying Envelope protein (Keele et al., 2008). Months later, neutralizing antibodies arise that target autologous virus. Neutralizing antibodies have 2 main functions: 1) they disrupt viral replication by binding cell-free virus and prevent the virion from infecting new target cells and 2) they bind to Env expressed on the cell surface and via the FC region, mediate antibody-dependent cellular cytotoxicity. These autologous neutralizing antibodies drive virus neutralization escape (Escolano et al., 2017). This was revealed during studies of antibody and virus co-evolution which demonstrated that earlier circulating viruses are more sensitive to current serum antibodies than concurrent circulating viruses (Doria-Rose et al., 2014; Liao et al., 2013). Thus, it seems that autologous neutralizing antibodies drive evolution of the virus Envelope, which in turn drives evolution of the antibody response in the germinal center reaction (Figure 1.2). Viral escape occurs when favorable mutations arise during the random reverse transcriptase-mediated mutagenesis of HIV-1. These mutations take the form of amino acid deletions, substitutions or insertions, particularly via the shielding of functionally constrained regions by the addition or subtraction of glycans (Doores, 2015; Wei et al., 2003).

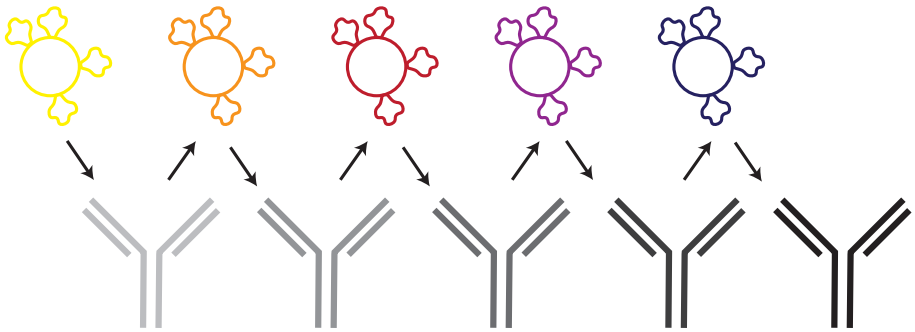


Figure 1.2 The evolution of the HIV-1 envelope on the virus drives the diversification of the antibody response.

Rarely, neutralizing antibodies in a single individual evolve to increase in breadth and potency and develop the ability to neutralize heterologous viruses. The mechanism for the development of broadly neutralizing antibodies (bNAbs) is widely studied, and while there is not a consensus, it seems that persistent antigen exposure is important such that 1) length of untreated infection, 2) viral load and 3) viral diversity, all contribute to broadly neutralizing antibody development *in vivo* (Gray et al., 2011; Rusert et al., 2016). However, the number of people who develop broad and potent antibodies is much fewer than the number of people who experience persistent antigen exposure, which suggests that there are additional, currently undiscovered factors (perhaps host-specific) which contribute to bNAb development.

Through techniques pioneered by the Nussenzweig laboratory, a new, extremely potent and broad generation of bNAbs have been cloned from single B cells obtained from individuals with highly neutralizing serum in the last 10 years (Klein et al., 2013). The study of these antibodies has led to the discovery of vulnerable epitopes on the Envelope

spike protein, novel therapeutic and vaccine strategies, and important developments in understanding cellular mediated control of HIV-1 infection in non-human primates (Escolano et al., 2017).

CD4+ T cell dynamics in HIV-1 infection

CD4+ T cell responses play a necessary role in effective cellular and humoral antiviral immune responses. In response to cognate antigen presented by dendritic cells, naïve CD4+ T cells proliferate and differentiate into effector cells. Depending on the context of the stimulus, naïve CD4+ T cells may differentiate into one of several lineages of T helper (Th) cells, including Th1, Th2, Th17, and induced Treg, as defined by their pattern of cytokine production and function (Sallusto, 2016; Zhu et al., 2010). Discovery of these subsets has revealed phenotypic hallmarks of each lineage, including the definition of molecular mechanisms (usually one or multiple transcription factors) that dictate differentiation and function. However, due to the complexity of immune responses *in vivo*, the relevant function of these subsets has yet to be fully understood in the context of the human immune system (Swain et al., 2012). Upon resolution of the immune response, the majority of the effector cells die, leaving behind only a small fraction of the clone in a pool of memory cells with diverse phenotypic and functional properties and gene expression profiles which are poised to mount a faster secondary immune response when the host encounters the same antigen (Mahnke et al., 2013).

Since HIV primarily infects and kills CD4+ T cells, efforts have focused on understanding HIV-1-specific CD4+ T cell responses. Progress has been hindered by a

number of considerable challenges inherent in studying *ex vivo* isolated CD4+ T cells and their role in HIV-1 infection. These include: 1) the lack of necessary tools (MHC Class II tetramers are much laborious to produce than MHC Class I tetramers, for example), 2) the short lifespan of HIV-specific CD4+ T cells, as they are thought to be preferential targets of HIV (Douek et al., 2002), and 3) perhaps most importantly for *ex vivo* human studies, the indirect experimental readouts of CD4+ T cell function: CD4+ T cells act primarily by helping other immune cell subsets within complex tissue architecture.

Chronic infection like HIV-1 can lead to dysfunction of responding immune cells, including CD4+ T cells. During early, acute HIV-1 infection, CD4+ T cells produce the hallmark Th1 cytokine IFNγ and T cell clones specific to several HIV-1 peptides can be readily identified (Oxenius et al., 2001). These early robust Th1 responses decrease as infection progresses, and within a few weeks, the detectable HIV-1 specific CD4 response is dramatically decreased. The HIV-1 specific response by CD4+ T cells does not return to initial levels, even after sustained antiretroviral therapy. This loss of specific CD4+ T cells could be explained by favored infection of HIV-1 specific CD4+ T cells (Douek et al., 2002), bystander cell death induced by HIV infection (Doitsh and Greene, 2016) or persistent CD4+ T cell fatigue. There is also evidence that early initiation of therapy may mitigate HIV-specific CD4+ T cell dysfunction (Cellerai et al., 2011), which could imply that the initial insult of chronic HIV infection affects lasting change to the immune system's overall function.

Understanding CD4+ T cell responses during HIV infection is tightly linked to understanding the establishment of the latent reservoir. The latent reservoir is established

very early during infection (Archin et al., 2012; Chun et al., 1998), and because of its long half-life of 44 months (Crooks et al., 2015; Finzi et al., 1999) it is the main impediment to curing HIV-1 infection (Siliciano and Greene, 2011). Resting CD4+ T cells have a longer half-life *in vivo* than effector cells but resting CD4+ T cells are relatively resistant to infection compared to activated effector cells. Since activated CD4+ T cells are prime targets for HIV-1 infection for a number of reasons (CCR5, the co-receptor for HIV-1 entry, is upregulated upon CD4+ T cell activation, NF-κB and other transcription factors required for HIV-1 RNA transcription are present in activated cells, and finally SAMHD1, a protein that disrupts reverse transcriptase, is highly expressed in resting cells (Murray et al., 2016)) but die quickly, how is the latent reservoir established and maintained in the overwhelming majority of HIV-1 infected individuals? The predominating hypothesis posits that a few activated CD4+ T cells become infected during the cellular transition to a long-lived resting memory state that does not support high level viral gene transcription. These events occurring in concert to support latency is a rare occurrence, which corresponds with the relative rarity of latent cells *in vivo* (~1 per million CD4+ T cells) (Crooks et al., 2015; Finzi et al., 1999).

It is thought that these latent cells do not transcribe viral RNAs or make viral proteins, and thus avoid detection by the immune system and are not eliminated by antiretroviral therapy. If it's assumed that latent cells are simply long-lived memory CD4+ T cells harboring a copy of HIV-1 DNA integrated somewhere in their genome, it's not unexpected that these cells might have a long half-life *in vivo*. This model attempts to explain HIV-1 latency in the framework of the normal physiology of immunologic CD4+ T

cell memory. It is consistent with the idea that antiretroviral therapy may merely reveal latency, rather than a hypothesis which requires the virus to harbor some special mechanism for latent infection.

Establishment of the latent reservoir and HIV integration

For the purposes of this thesis, I will define HIV latency as the reversibly nonproductive state of infection of individual host cells with a stable, intact integrated form of the viral genome. Although the latent reservoir remains to be completely described, it is thought that establishing the reservoir requires intact retroviral integration into the genome and subsequent transcriptional silencing (Siliciano and Greene, 2011). Latency may be enforced by silencing epigenetic modifications of the integrated provirus (Wang et al., 2007) and the virus persists essentially as genetic information, being protected with the host genome in the nucleus of the cell as the cell carries out its normal function.

The first essential step in establishment of the latent reservoir is the integration of the provirus into the genome. After reverse transcription, the double stranded viral DNA forms a complex with host and viral proteins called the preintegration complex (PIC). The PIC is a large complex which is imported into the nucleus via a largely unknown mechanism, but since HIV-1 can infect non-dividing cells, import of the PIC into the nucleus is likely an active process. Once in the nucleus, the PIC accesses the host chromosomes. Viral protein integrase is an integral part of the PIC and is responsible for cutting the DNA and joining the host and viral genomes. Many different studies provide evidence to show that HIV-1 preferentially integrates into a subset of transcriptionally

active genes of the host cell genome (Craigie and Bushman, 2012). More recently, it was demonstrated that this preference may be due to nuclear topology (Marini et al., 2015). HIV-1 integration seems to occur in the outer shell of the nucleus in close proximity to the nuclear pore. This region contains a series of number of host genes, which are preferentially targeted by the virus, and characterized by the presence of active transcription chromatin marks before viral infection. Additional evidence suggests that HIV integration into the genome is known to favor the introns of expressed genes (Han et al., 2004), some of which, like *BACH2* and *MKL2* carry multiple independent HIV-1 integrations in different individuals and are considered hotspots for integration (Maldarelli et al., 2014; Wagner et al., 2014). Functional viral integrase, and the presence of the cellular Nup153 and LEDGF/p75 integration cofactors are indispensable for the perinuclear integration site selection of the virus (Marini et al., 2015). Thus, virally encoded and host proteins coordinate to direct HIV-1 to specific regions of accessible chromatin.

Whether or not the genomic location of the integration impacts on latency is debated (Jordan et al., 2003; Jordan et al., 2001). Several mechanisms could silence HIV gene expression and replication. These include but are not limited to: transcriptional interference (Lenasi et al., 2008), problems with or limitations to RNA processing and transport (Lassen et al., 2006), epigenetic silencing such as changes in DNA methylation (Kauder et al., 2009), and the presence of repressive transcription factors, or the absence of necessary positive transcription factors (Van Lint et al., 2013). These mechanisms

could contribute individually and/or in concert to the transcriptional silencing of HIV and permit infected host cell survival.

Measuring the latent reservoir

The HIV-1 latent reservoir has been difficult to define for two major reasons: first, the majority of integrated proviruses are not replication competent and do not contribute to viral rebound after treatment cessation but do confound direct measurements of viral nucleic acids by overestimating the number of latent cells. Second, a complete understanding of viral reactivation is lacking and thus induction from latency is incomplete, which prevents accurate measurement of intact latent viruses.

The gold standard for latent reservoir quantitation is the Quantitative Viral Outgrowth Assay (QVOA) (Laird et al., 2013). Blood or leukopheresis product is obtained from HIV-1-infected individuals, and CD4+ T cells are isolated. The CD4+ T cells are plated in limiting dilution to allow quantification and are stimulated with PHA or other latency reversing agents (LRAs) and irradiated peripheral blood mononuclear cells (PBMC) from a non-infected donor to stimulate viral outgrowth. The next day, target cells (either CD8 depleted lymphoblasts or an HIV-1 permissive cell line) are added as a source for amplification of the released virus. After several weeks, wells that contain replication competent HIV-1 are detected by ELISA measuring supernatant HIV-1 p24 gag protein. QVOA is the only assay which detects solely replication competent virus. However, the drawbacks of this assay are significant – primarily that the assay does not detect all replication competent virus because reactivation of the latent reservoir is incomplete with

current modalities. It is also time-consuming, resource-intensive, and requires large blood draws from the study participants.

Efforts to more accurately and more quickly quantitate the latent reservoir are underway. The tat/rev induced limiting dilution assay (TILDA) measures inducible multiply spliced HIV-1 RNA (Procopio et al., 2015). Multiply spliced HIV-1 RNA is only produced in actively transcribing cells who harbor an intact proviral LTR and is therefore a more accurate predictor of the active reservoir than quantitation of genomic proviral DNA. Furthermore, the PCR primers used in TILDA are specific to the tat/rev region, which is the most commonly deleted region in defective proviruses. As one example from a treated individual, TILDA estimated a latent reservoir that was 48 times higher than that measured by QVOA and approximately 6–27 times lower than that predicted by PCR-based assays. This demonstrates its ability to measure a larger proportion of the latent reservoir than QVOA while being more selective in measuring likely intact virus than traditional DNA PCR-based approaches (Procopio et al., 2015).

Although QVOA is the gold standard, the assay tends to underestimate the size of the reservoir because some proviruses are relatively resistant to reactivation *in vitro* as revealed by near full-length genome sequencing (Ho et al., 2013). QVOA estimates the reservoir size in treated individuals to be, on average, 1 latent cell per million CD4+ T cells. Since QVOA requires proviral reactivation, recently three studies performed characterization of the replication competent reservoir by PCR to avoid any bias introduced by reactivation (Bruner et al., 2016; Hiener et al., 2017; Ho et al., 2013). These studies employ an unbiased single genome amplification approach to define the

sequence landscape of persistent proviruses. Using a near full-length viral genome PCR, proviruses in patient samples are amplified at limiting dilution to avoid any PCR artifacts. Following outer PCR amplification, inner segments of the virus are amplified and directly sequenced, allowing the reconstruction of the full viral genome while limiting the potential for PCR generated mutations. This method has defined a new subset of reservoir viruses which are non-induced, but genetically intact. The size of the reservoir measured by these full-genome sequencing methods is estimated to be 10 to 100-fold greater than that measured by QVOA. These experiments are time intensive and laborious but require many fewer cells than QVOA. Additionally, though these viruses have an intact genome, whether these viruses could reactivate *in vivo* is still undetermined.

These results may suggest that 1) the barrier to cure may be much larger than originally thought, 2) our current methods are insufficient to reactivate all intact viruses, and 3) more efficient methods to measure the size of the reservoir are needed.

Evidence for clones of infected cells

CD4+ T cell death is the primary result of productive HIV-1 infection (Doitsh and Greene, 2016). The mechanism of CD4+ T cell depletion *in vivo* is not entirely understood, but it was thought that infected cells would die before they were able to divide because cell activation triggers both cell division and HIV-1 transcription (Nabel and Baltimore, 1987; West et al., 2001; Williams and Greene, 2007). However, increasing evidence has begun to suggest that infected cells are able to divide *in vivo*.

Successful antiretroviral therapy reduces HIV-1 plasma viremia to below the detection limit of ultrasensitive clinical assays (20 copies of HIV-1 RNA/ml plasma). Nevertheless, very low levels of free virus can be found in the plasma. Sequencing the residual plasma viremia revealed persistent virus released from infected cells for months to years without evident sequence change (Bailey et al., 2006). Comparison of the plasma virus with integrated, cell associated viruses showed that occasionally some sequences in the plasma were identical to viruses in resting CD4+ T cells (Anderson et al., 2011). Despite the large diversity of HIV-1 sequenced from resting CD4+ T cells, the residual viremia was dominated by a homogeneous population of viruses with identical sequences (Bailey et al., 2006). Thus, in individuals on antiretroviral therapy, a mechanism for residual viremia involves persistent production of a small number of viral clones without evident evolution. Since new viral replication results in mutagenesis of viral sequence, one plausible explanation for this observation is the proliferation of infected cells carrying a single integrated provirus. An alternative, but less likely, explanation is that identical sequences could result from a burst of infectious virions which each infected unrelated CD4+ T cells.

Upon antiretroviral therapy cessation, viremia rebounds from the latent reservoir. When it does, it appears to involve an increasing proportion of monotypic, archived HIV-1 sequences, further suggesting the proliferation of latently infected cells (Joos et al., 2008; Wagner et al., 2013). Based on this observation and the finding that a subset of cells bearing integrated HIV-1 undergoes clonal expansion in individuals receiving suppressive antiretroviral therapy (Maldarelli et al., 2014; Wagner et al., 2014), it has

been proposed that the clonally expanded cells play a critical role in maintaining the latent reservoir.

This thesis describes three methods developed to study the latent reservoir from treated individuals at the single cell level. First, integration site sequencing was used to reveal clonally expanded infected cells which harbored primarily defective proviruses. Second, qualitative analysis of virus from single latent cells suggested that rarely, cells harboring replication competent virus could divide. And finally, the inability to isolate latently infected cells has limited the study of the HIV-1 latent reservoir. As a first step in this direction, I developed a method to identify and characterize single recently reactivated latently infected cells.

CHAPTER 2:

HIV-1 INTEGRATION IN PRODUCTIVE AND LATENT INFECTION

Integration Library Construction

To obtain additional insights into role of clonal expansion in maintaining the reservoir and the regions of the genome that are favored by HIV-1 for integration, we developed a single cell method to identify a large number of HIV-1 integration sites from treated and untreated individuals, including “viremic controllers” who spontaneously maintain viral loads of <2000 RNA copies/ml and “typical progressors” who display viral loads >2000 RNA copies/ml.

Twenty-four integration libraries were constructed from CD4+ T cells from 13 individuals: 3 provided longitudinal samples before and after (0.1-7.2 years) initiation of therapy; 4 were untreated; 2 were treated; and 4 were viremic controllers (Table 2.1). Individuals were grouped into three categories based on viral loads and therapy: 1. viremic progressors were untreated individuals with viral loads higher than 2000 viral RNA copies/mL of plasma; 2. progressors were treated individuals whose initial viral loads were higher than 2000 viral RNA copies/mL before therapy; 3. controllers were individuals who maintain low viral loads spontaneously in the absence of therapy (less than 2000 viral RNA copies/mL).

Table 2.1. Clinical profile of human subjects.

	Gender	Year of diag-nosis	cART Start Date	cART Regimen	Date of Sample Collection	Viral load (Copies/ mL)	CD4 Count	Group
1	Male	2000	6/17/09					
				naïve	11/20/07	63500	291	Viremic
				naïve	11/17/08	8350	165	Viremic
				Truvada, Atazanavir, Ritonivir	11/16/09	<40	298	Treated
				Truvada, Atazanavir, Ritonivir	5/17/10	<40	344	Treated
				Truvada, Atazanavir, Ritonivir	2/11/14	<40	486	Treated
2	Female	1990	1/12/07					
				naïve	2/8/05	15200	359	Viremic
				naïve	8/9/05	60000	295	Viremic
				Atripla	3/10/09	<30	560	Treated
				Atripla	4/6/10	<30	757	Treated
				Atripla	3/11/14	<40	965	Treated
3	Male	2006	12/17/09					
				naïve	8/5/09	4734	433	Viremic
				Truvada, Atazanavir, Ritonivir	11/18/10	0	616	Treated
				Truvada, Atazanavir, Ritonivir	12/23/13	<40	869	Treated
4	Male	2011		NA	9/20/13	71857	530	Viremic
5	Male	2003		NA	3/28/14	3210	674	Viremic
6	Female	2000		NA	4/7/14	43650	520	Viremic
7	Male	2006		NA	4/11/14	5340	607	Viremic
8	Male	1996	4/22/97	Tenofovir/ Emtriva, Nevirapine	4/22/13	<40	280	Treated
9	Male	2011	9/19/11	Truvada, Efavirenz	5/2/13	<40	440	Treated
10	Male	2003		NA	5/27/10	49	1070	Controller
11	Male	2002						Controller
				NA	5/08	410	518	Controller
				NA	7/10	880	565	Controller
12	Male	1997		NA	7/23/10	505	430	Controller
13	Male	1989		NA	7/23/10	<50	580	Controller

Libraries were produced from genomic DNA by a modification of the translocation-capture sequencing method that we refer to in this paper as integration sequencing (Figure 2.1) (Janovitz et al., 2013; Klein et al., 2011).

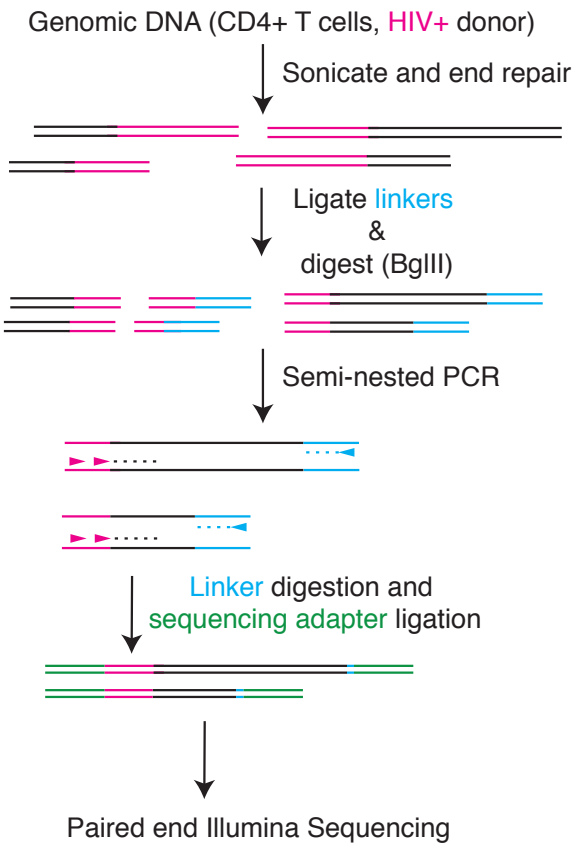


Figure 2.1 Diagrammatic representation of integration library construction.

Virus integration sites were recovered by semi-nested ligation-mediated PCR from fragmented DNA using primers specific to the HIV-1 3' LTR. PCR products were subjected to high-throughput paired-end sequencing, and reads were aligned to the human genome. Since sonication is random, it produces unique linker ligation points that identify the specific integration events in each infected CD4+ T cell, which allows both single cell resolution and identification of expanded clones of cells with identical integrations ((Berry et al., 2012) Figure 2.1). Thus, integration sequencing can enumerate both the number of integration sites and the number of infected cells.

A total of 6719 unique virus integration sites were determined: 873 unique integrations in viremic controllers; 987 integrations in untreated progressors; and 4859 integrations in treated progressors.

Integrations enriched in highly expressed genes

We analyzed the genomic location of the integration sites obtained from viremic controllers, untreated and treated progressors and compared our results to published data obtained from HIV-1 infected individuals (Brady et al., 2009; Han et al., 2004; Ho et al., 2013; Ikeda et al., 2007; Schroder et al., 2002; Sherrill-Mix et al., 2013; Wang et al., 2007). In agreement with the work of others, the majority of integration sites in each group are genic (Figure 2.2a). Moreover, integrations are found more frequently in the introns of highly expressed genes, and there is a slight bias for viral orientation that leads to convergent transcription (Figures 2.2b-d) (Mitchell et al., 2004). Thus, the general features of integrations defined by our integration sequencing assay are similar to those obtained by others.

Although the differences between groups were small in magnitude, they were significant in that treated progressors had a smaller proportion of integrations in genic regions ($p<0.0001$ and $p<0.0001$, respectively) and in highly expressed genes ($p<0.0001$ and $p<0.0001$, respectively) when compared to viremic controllers and untreated progressors (Figure 2.2c). Conversely, the proportion of viral integrations in genes expressed at lower levels was increased in treated progressors compared to viremic controllers and untreated progressors ($p= 0.002$ and $p<0.0001$, respectively). Viremic

controllers and treated progressors were not significantly different from each other in terms of the level of expression of the genes at the sites of integration (Figure 2.2c). Thus, therapy is associated with a relative decrease in the number of cells with viral integrations in highly expressed genes.

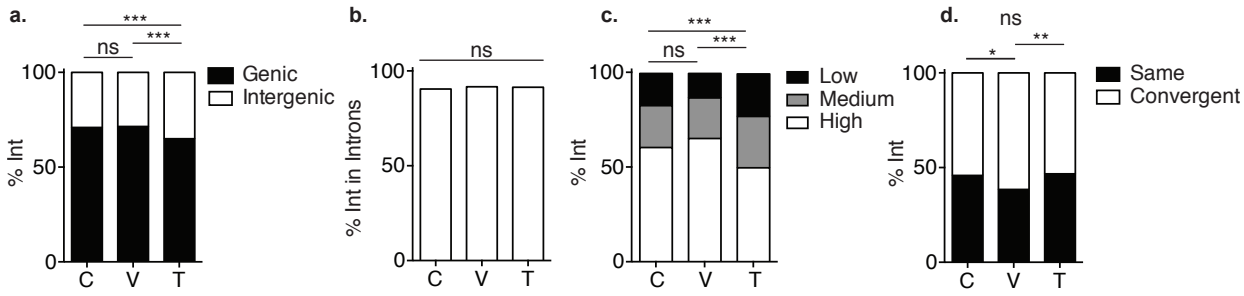


Figure 2.2. HIV Integration Libraries. a) Proportion of integrations that are in genic or intergenic regions in controllers (C), viremic (V) or treated progressors (T). **b)** Proportion of genic integrations located in introns in controllers, viremic or treated progressors **c)** Proportion of integrations in genes with high, medium or low expression. Integrations in genes with silent and trace expression contribute a minor proportion and were not included in this analysis. P-values refer to proportion of integrations in highly expressed genes. **d)** Transcriptional orientation of integrated HIV-1 relative to host gene in controllers, viremic or treated progressors. ns: not significant *P<0.05**P< 0.01 ***P<0.0001 using two-proportion z-test.

Identification of clonally expanded infected cells

Since we shear DNA ends randomly to produce our libraries, and by paired end sequencing can determine the precise site of both the integration and sheared end, we infer that identical integrations with unique sheared ends arise from clones of expanded cells (Figure 2.1). Since HIV-1 integration is semi-random, it is extremely unlikely that any

2 *de novo* integration sites would be identical, and thus we infer that identical integration sites arise from cell division of a single infected cell. Integrations can therefore be classified as clonally expanded (i.e. identical integrations with distinct sheared ends, deriving from the clonal expansion of an original unique, single integration event) or single integrations (i.e. unique integration site with a single sheared end).

Clonally expanded viral integrations were present in all individuals irrespective of therapy or viremia. However, the proportion of clonally expanded viral integrations is significantly lower in viremic controllers (30%) and viremic progressors (27%) than in treated progressors (40%) ($p<0.0001$ and $p<0.0001$, Figure 2.3a and b). Although the size of individual clones varied from 2-295 cells (Figure 2.3c), the relative increase in clonally expanded integrations during therapy consistently translated into an increase in the number of infected cells that derive from expanded clones (Figure 2.3d and e). The percentage of cells containing clonally expanded HIV-1 integrations was similar in untreated progressors (78%) and controllers (79%), but it was significantly increased in treated progressors (90%) ($p<0.0001$ and $p<0.0001$, Figure 2.3d and e). Thus, therapy is associated with an increase in the frequency of clonal HIV-1 integrations and infected clonally expanded cells.

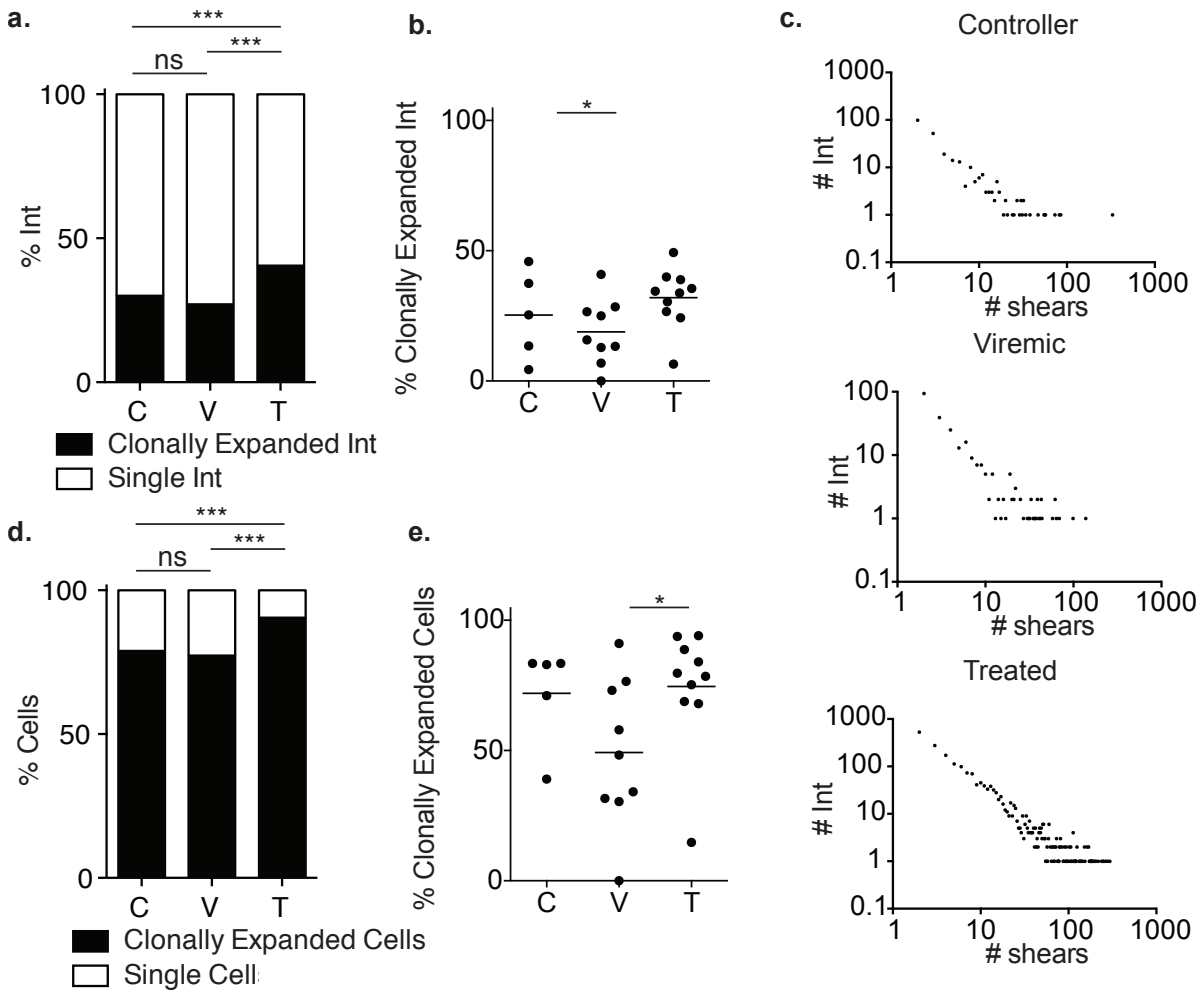


Figure 2.3 Identification of clonally expanded cells bearing integrated HIV-1. See also Figure S2. **a)** Proportion of viral integrations (Int) that are clonally expanded, as identified by the same integration site with multiple shears in controllers, viremic or treated progressors. **b)** Proportion of integrations (Int) that are clonally expanded in controllers, viremic or treated progressors; each dot represents data from an individual integration library. **c)** Graph shows the size of proliferating clones of infected cells. Plotted is the number of shears (X-axis) by the number of integrations (Y-axis). **d)** Proportion of infected cells deriving from clonal expansion in controllers, viremic or treated progressors. **e)** Proportion of infected cells deriving from clonal expansion in controllers, viremic or treated progressors; each dot represents data from an individual integration library.

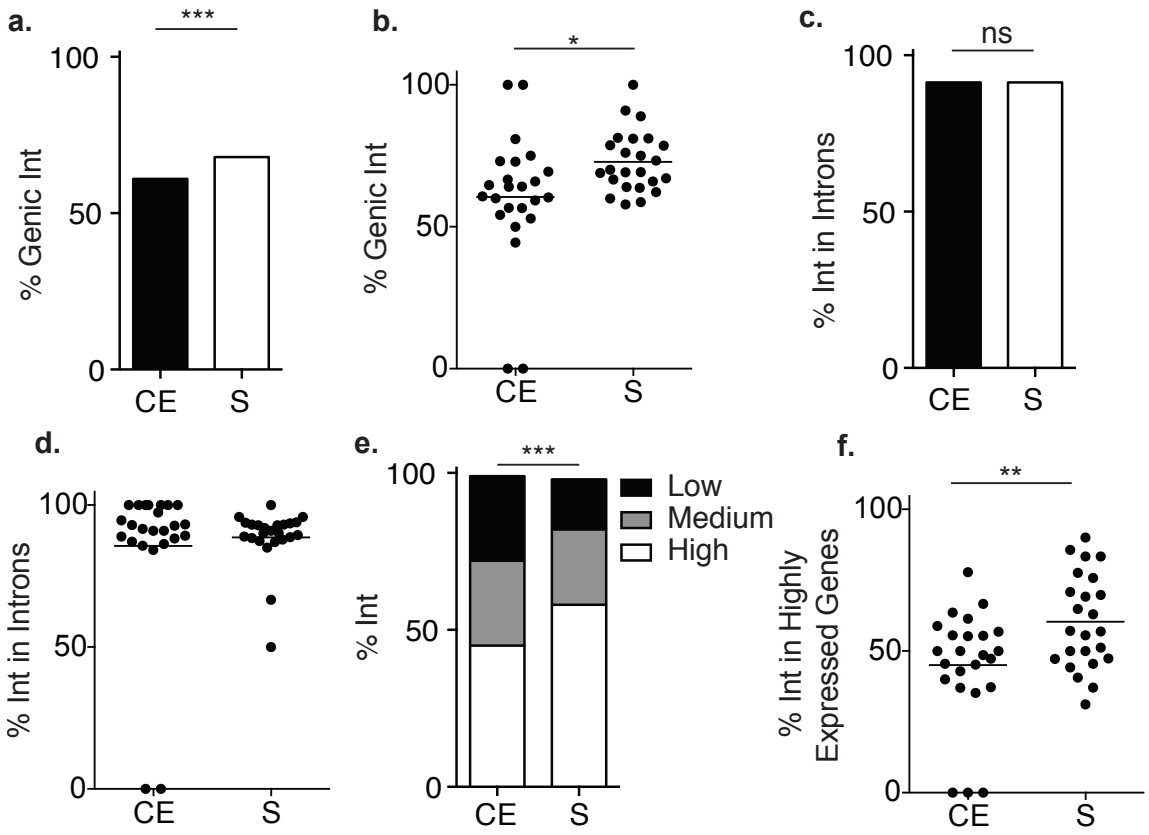


Figure 2.4 Comparing clonally expanded and single integrations. a) Proportion of clonally expanded (CE) and single (S) viral integrations in genic or intergenic regions. b) Proportion of clonally expanded (CE) and single (S) genic integrations; each dot represents data from an individual integration library. c) Proportion of clonally expanded and single viral integrations in introns. d) Proportion of clonally expanded (CE) and single (S) integrations in introns. e) Proportion of clonally expanded or single viral integrations in genes with high, medium or low expression. Integrations in genes with silent and trace expression contribute a minor proportion and were not included in this analysis. P values refer to proportion of integrations in highly expressed genes. f) Proportion of clonally expanded (CE) and single (S) integrations in highly expressed genes.

To determine whether the position of viral integration in the genome correlates with clonal expansion we compared the genomic clonally expanded to single integrations.

Both types of integrations favored genes and their introns (Figures 2.4a-d). However, the proportion of clonally expanded integrations in intergenic regions was greater than that of single integrations (Figure 2.4a and b, $p<0.0001$). Moreover, of the integrations in genes, single integrations were more likely to be found in highly expressed genes than clonal integrations (Figure 2.4e and f, $p<0.0001$). Thus, cells harboring viral integrations in intergenic regions and genes that are expressed at lower levels are more likely to be clonally expanded.

Hotspots for virus integration

Overlap between integrations in the genes of different individuals suggests the existence of hotspots for HIV-1 integration. A number of individual genes have been identified as preferential sites for HIV-1 integration including *BACH2, MKL2, DMNT1, MDC1 and STAT5B* (Ikeda et al., 2007; Maldarelli et al., 2014; Wagner et al., 2014). To identify hotspots for HIV-1 integration genome-wide, we subjected our data set to hot_scan analysis (Silva et al., 2014), which defines hotspots by identifying regions of local enrichment using scan statistics. This analysis identified 55, 85, and 247 hotspots for controllers, viremic and treated progressors, respectively. For example, the intron between exons 5 and 6 in *MKL2* is a hotspot for integration in participant 11, contains an expanded clonal family in participant 10 and was also identified as a site of enrichment for integration by others (Maldarelli et al., 2014) (Figure 2.5a).

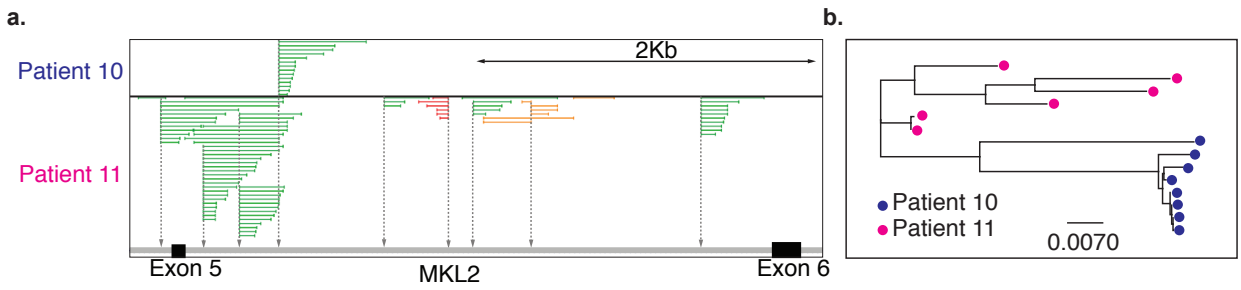


Figure 2.5 MKL2 hotspot characterization. a) Integrations in *MKL2* from participants 10 and 11. Gray vertical arrows indicate site of integrations. Colored horizontal lines show fragments of DNA spanning the point of integration through sheared end. Green: viruses integrated in the same orientation as gene. Red: convergent orientation. Orange: viruses integrated at same site with both orientations. **b)** HIV-1 *gag* was amplified from integrated proviruses in *MKL2* from participant 10 and 11. PCR was performed using nested integration site-specific primers and HIV-1 *gag* primers. Sequences were clustered by the Tamura-Nei model to assess DNA sequence similarity. The scale bar represents 0.007 substitutions per site.

To validate our *in-silico* analysis and to further characterize the *MKL2* hotspot, we sequenced the *gag* gene from proviruses integrated into *MKL2* by amplification with nested genomic primers specific for *MKL2* and HIV-1 *gag*. Sequences obtained from participant 10, who showed only one expanded clone are very closely related to each other, which is consistent with a single clonally expanded integration (Figure 2.5b). In contrast, sequences obtained from participant 11 are far more diverse suggesting that there were several different viral integrations in the *MKL2* hotspot (Figure 2.5b). We conclude that the hotspots defined by hot_scan represent multiple distinct integration events in close proximity.

Viremic progressors had the highest proportion of integration events in hotspots, indicating that in the case of high-level viremia there are specific genomic locations that

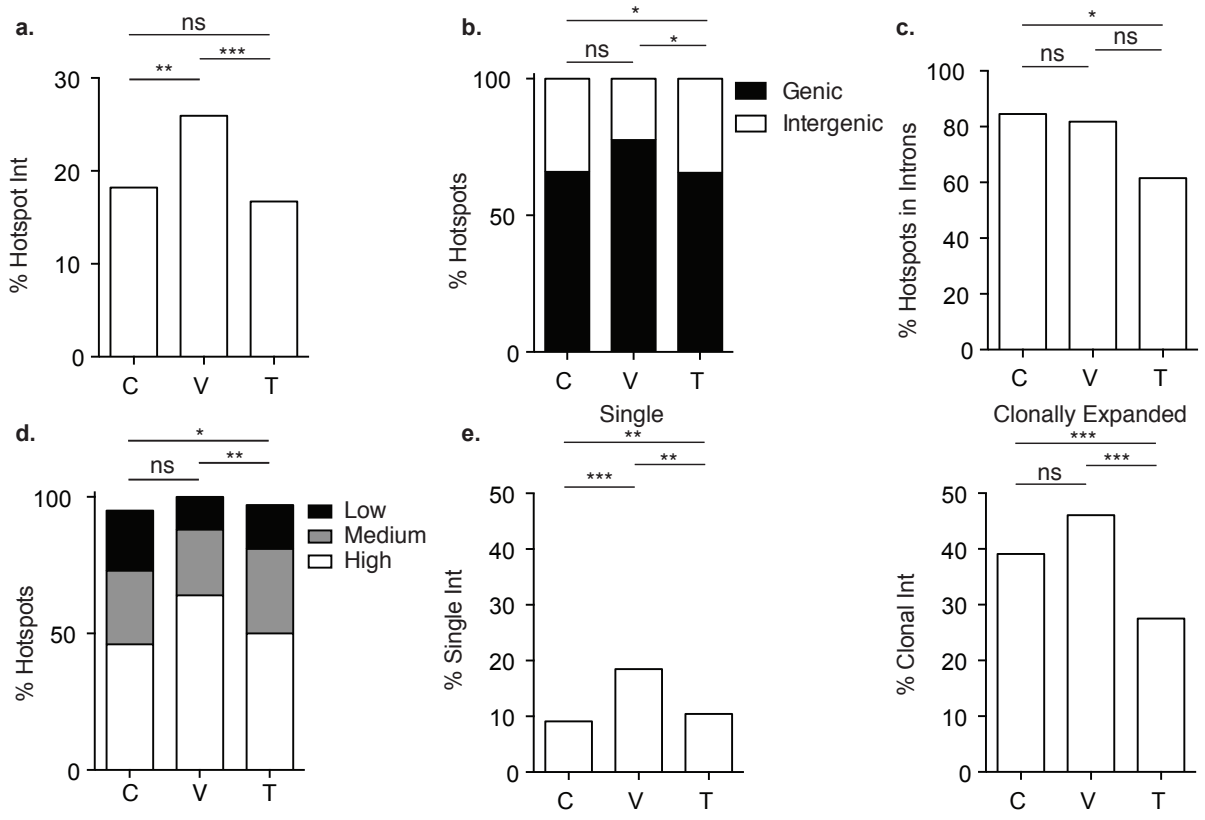


Figure 2.6 Hotspots for HIV-1 integration. a) Proportion of virus integrations inside hotspots in controllers, viremic and treated progressors. b) Proportion of hotspots in genic and intergenic regions in controllers, viremic and treated progressors. c) Proportion of hotspots in introns in controllers, viremic and treated progressors. d) Proportion of hotspots in genes with high, medium or low expression. Hotspots in genes with silent and trace expression contribute a minor proportion and were not included in this analysis. P values refer to proportion of integrations in highly expressed genes. e) Percentage of total single and clonally expanded viral integrations inside hotspots. Enrichment of clonally expanded viral integrations compared to single integrations is significant, $p < 0.0001$. ns: not significant *$P<0.05$ **$P< 0.01$ ***$P<0.0001$ using proportion test

favor integration (Figure 2.6a). Although the majority of all integrations fall outside of hotspots (Figure 2.6a), hotspot integrations resemble others in that they are preferentially found within genes with a preponderance of these in introns (Figure 2.6b and c). In all cases, hotspots are enriched in highly expressed genes, and consistent with the overall decrease in viral integrations in highly expressed genes during therapy, the proportion of hotspots in these genes also decreases (Figure 2.6d and 2.2c). Thus, the general characteristics of hotspots are similar to features of all integrations.

To determine whether there is a relationship between hotspots and clonally expanded viral integrations we enumerated single and clonally expanded integrations in hotspots (Figure 2.6e). Only a small fraction (11-18%) of all single integrations were found in hotspots with untreated viremic progressors showing the highest level (Figure 2.6e). In contrast, there was a much higher proportion of clonal integrations in hotspots (30-46%) with the lowest proportion in treated progressors (Figure 2.6e).

Integrations enriched near *Alu* repeats

We next wondered whether a specific genomic feature could partially explain the enrichment of integrations in hotspots. We observed a significant enrichment of integrations inside *Alu* repeats (Figure 2.7a), and in close proximity to *Alu* repeats, irrespective of whether the integration is inside genes or in intergenic regions (Figure 2.7b). Thus, a preference for *Alu* is independent of a preference for integration in genes.

Previous studies have suggested that a preference for *Alu* repeats, at least in part reflects a preference for highly expressed genes (Schroder et al., 2002). To examine the

relationship between *Alu* repeats and transcription, we determined the distance between *Alu* repeats and the center of all genes. There was no positive correlation between the position of *Alu* and the level of transcription (Figure 2.7c). To determine whether the distance between integration and *Alu* repeats correlates with transcription, we measured the distance between the sites of integration and *Alu* repeats in all genes (Figure 2.7d). There was no significant difference between integration distance to *Alu* repeats in highly expressed, silent or trace level expressed genes. Therefore, the rate of transcription does not impact integration distance to *Alu* repeats and integration at these sites must be independent of transcription.

Finally, the number of *Alu* repeats in a hotspot is directly correlated with the number of integration events in that hotspot (Figure 2.7e, ρ=0.86). The data suggests that HIV-1 has a preference for integration in close proximity to sites in the genome that are enriched in *Alu* repeats and that this preference is independent of the level of transcription.

Increase in clonally expanded infected cells during antiretroviral treatment

The proportion of clonally expanded viral integrations is increased in treated progressors (Figure 2.3a (Wagner et al., 2014)). To further examine the effect of therapy on clonal expansion we analyzed longitudinal samples from three typical progressors before and during therapy (Table 2.1). We found an increase in the number of clonally expanded integrations throughout the treatment period of up to 7.2 years in two of the three individuals (Figure 2.8a, p=0.017) as well as an increase in the number of cells that contained clonally expanded viral integrations (Figure 2.8b).

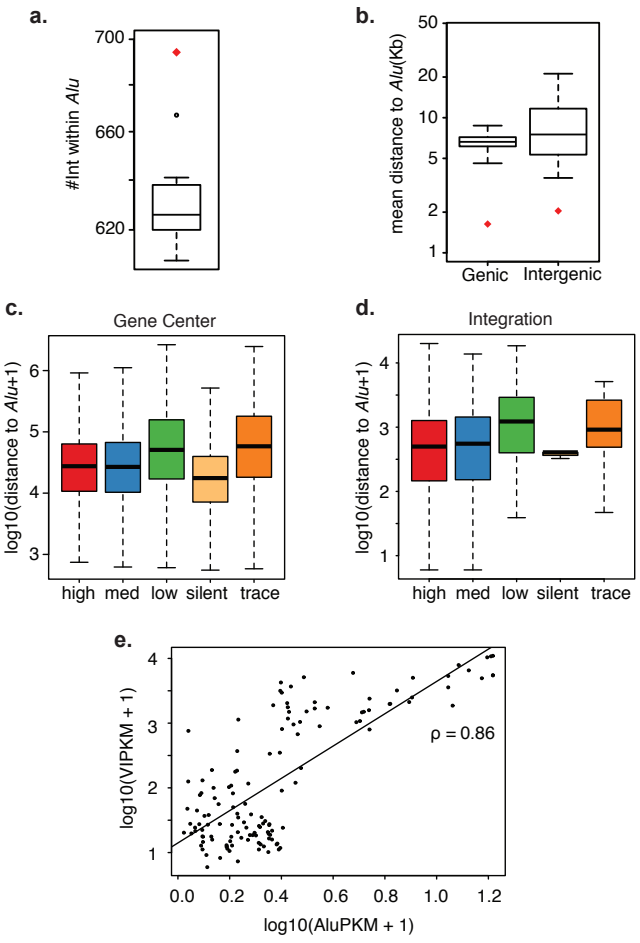


Figure 2.7 Enrichment of integrations in Alu repeats.

a) Integrations are enriched inside *Alu* repeats. Total integrations identified inside *Alu* repeats were enumerated (red diamond) and compared to the expected value as defined by Monte Carlo simulation. The boxplot displays the variation of the number of random integrations identified inside *Alu* repeats in each iteration of the simulation. **b)** Integrations are near *Alu* repeats in genes and intergenic regions. Average distance to the nearest *Alu* repeat for all integrations inside genes or intergenic regions was calculated (red diamond) and compared to the expected distance as defined by Monte Carlo simulation. The boxplot displays the variation of the distance of random integrations from *Alu* repeats in genes or intergenic regions in each iteration of the simulation. **c)** Distance to *Alu* repeats from the center of highly, medium, low, trace or silently expressed genes. **d)** Distance to *Alu* repeats in highly, medium, low, trace or silently expressed genes. **e)** Positive correlation between *Alu* repeats and integrations inside hotspots. Graph shows number of *Alu* repeats (X axis) vs. integrations in hotspots (Y axis). Hotspots not containing *Alu* repeats were removed from this analysis. The scatter plot shows the linear relationship between the number of INT-motifs and integrations inside hotspots (Pearson's correlation, $\rho = 0.86$).

Correspondingly, there was also an overall decrease over time in single integrations (p=0.017), with a half-life of 127 months assuming a non-linear regression model for one-phase decay (Figure 2.8c). Thus, our data suggests that the numbers of single integrations decay very slowly over time, while clonally expanded integrations increase with time on cART.

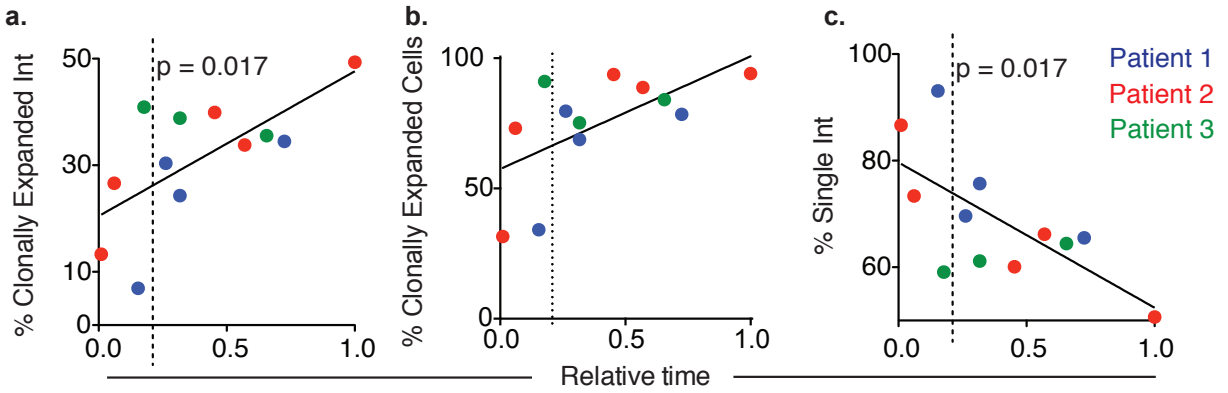


Figure 2.8 Clonally expanded viral integrations increase and single integrations decrease during therapy. Graphs show data from participant 1 (blue), 2 (red) and 3 (green) from longitudinal time points. Time was normalized from 0 to 1 (727 days pre therapy to 2617 days post therapy). Dotted line at t = 0.21 marks therapy initiation. Trendline was determined by linear regression model. Solid lines indicate significant change in proportion of events. **a)** Proportion of clonally expanded viral integrations (Int). **b)** Proportion of clonally expanded cells. **c)** Proportion of single viral integrations.

The increase in the number of clonal integrations during cART did not favor genic or intergenic regions (p=0.65), indicating that this effect is independent of the location of the integration in the genome (Figure 2.9a). In contrast, single integrations decrease significantly in genic regions and increase proportionally in intergenic regions (Figure

2.9b, p=0.036). Thus, the fate of cells harboring single viral integrations in ART treated progressors differs from clonal integration.

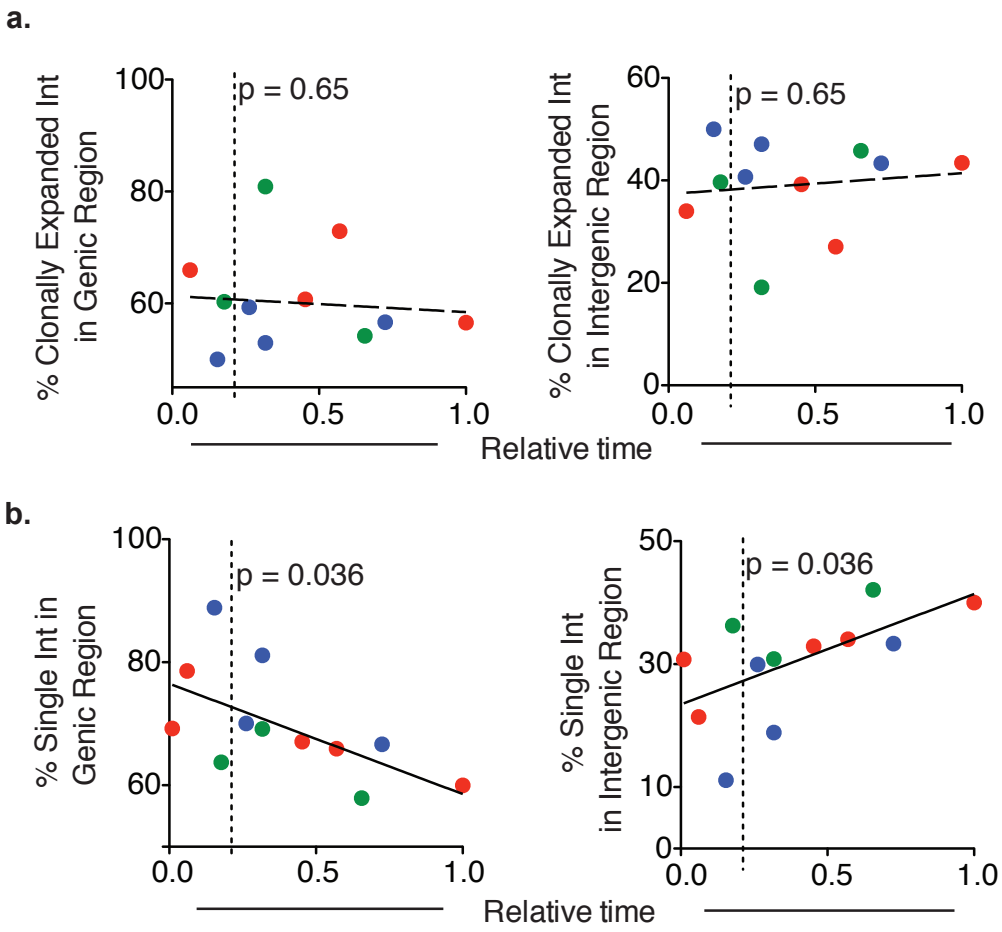


Figure 2.9 Single integrations decrease preferentially from genic regions during time on antiretroviral therapy. Graphs show data from participant 1 (blue), 2 (red) and 3 (green) from longitudinal time points (Table S1). Time was normalized from 0 to 1 (727 days pre therapy to 2617 days post therapy). Dotted line at t = 0.21 marks therapy initiation. Trendline was determined by linear regression model. Solid lines indicate significant change in proportion of events; dashed lines indicate insignificant change in proportion of events. **a)** Proportion of genic and intergenic clonally expanded viral integrations. **b)** Proportion of genic and intergenic single viral integrations.

Moreover, the fate of single integrations is dependent on their location in the genome whereas the clonal integrations are not. These results suggest that cells bearing genic single integrations are selected against during therapy and that clonal expansion is not.

Integrations in cancer related genes decrease over time on therapy

In the 3 progressors who provided longitudinal samples, approximately 5% of the clonal integrations persisted through successive time points without selection for genic or intergenic regions compared to all clonal integrations (Figure 2.10a and b). Furthermore, of the genic integrations that persisted, there was also no selection for or against those in highly expressed genes (Figure 2.10c). Thus, the persistent clonal integrations are indistinguishable from the larger pool of clonally expanded viral integrations in terms of their position in the genome.

Since clonal integrations have been associated with genes involved in malignant transformation (Wagner et al., 2014) we examined our entire data set for enrichment of integrations in cancer-associated genes (n = 743 cancer associated genes (Vogelstein et al., 2013; Zhao et al., 2013)). Although there was an overall enrichment among for integrations in cancer genes (329/4410 = 7.5%) compared to all genes in the human genome (743/25,660 =2.8%) ($p<0.0001$), this preference does not seem to be significant because it is similar to the overall preference for integration into highly expressed genes (Figure 2.10d). Furthermore, we observed no overrepresentation of single, clonal or persistent integrations in cancer genes (Figure 2.10e). Importantly, a significant decrease in integrations in cancer related genes was observed in longitudinal samples (Figure 2.9f) suggesting that these are selected against with therapy.

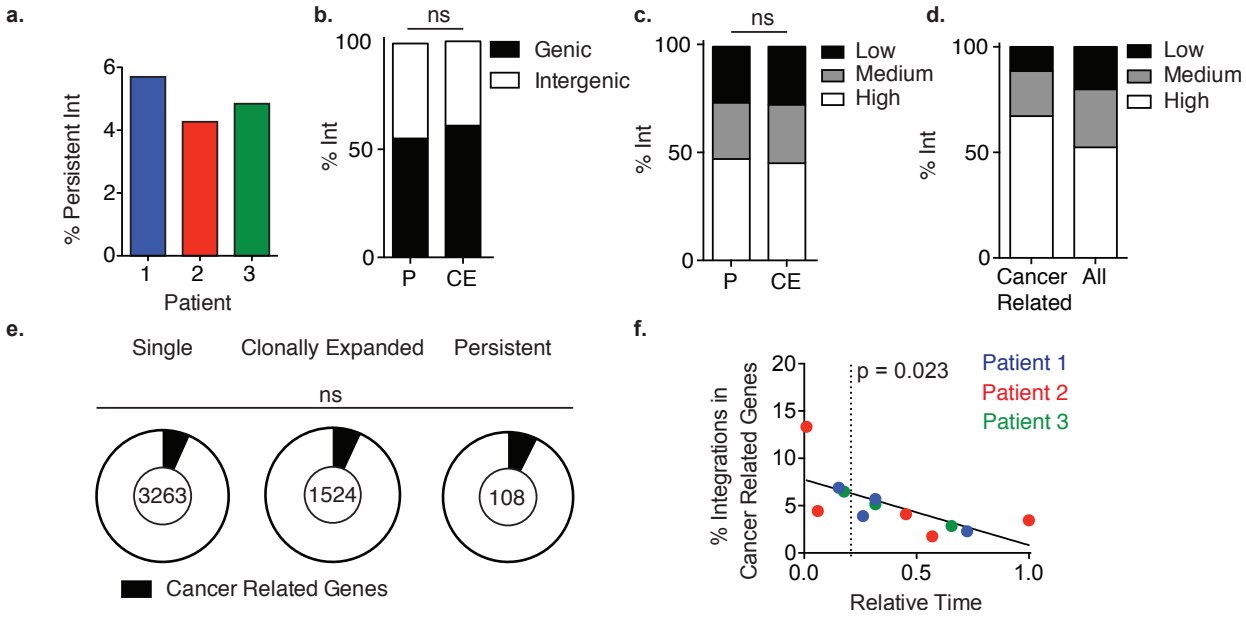


Figure 2.10. Integrations in cancer related genes decrease over time on therapy

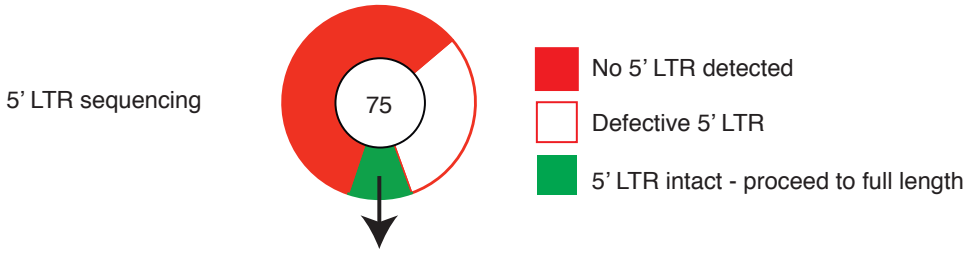
a) Percent viral integrations present in more than one time point (persistent integrations) in participants 1, 2 and 3 (Table S1). **b)** Comparison of persistent (P) and clonally expanded (CE) viral integrations in genic or intergenic region. **c)** Proportion of persistent and clonally expanded viral integrations in genes with high, medium or low expression. Integrations in genes with silent and trace expression contribute a minor proportion and were not included in this analysis. P values refer to proportion of integrations in highly expressed genes. **d)** Proportion of cancer related (left) or total (right) integrations in genes with high, medium or low expression. **e)** Genes with integrations were analyzed for their association with cancer. Proportions of cancer-associated genes are shown for single, clonally expanded and persistent viral integrations. The number indicates the total number of genes from each category. **f)** Graph shows proportion of integrations in cancer-related genes from participants 1 (blue), 2 (red) and 3 (green) from longitudinal time points (Table S1). Time was normalized from 0 to 1 (727 days pre therapy to 2617 days post therapy). Dotted line at t = 0.21 marks therapy initiation. Trendline was determined by linear regression model. Solid line indicates significant change in proportion of events, p=0.023. ns: not significant *P<0.05**P< 0.01 ***P<0.0001 using two-proportion z-test.

Expanded clones contain defective virus

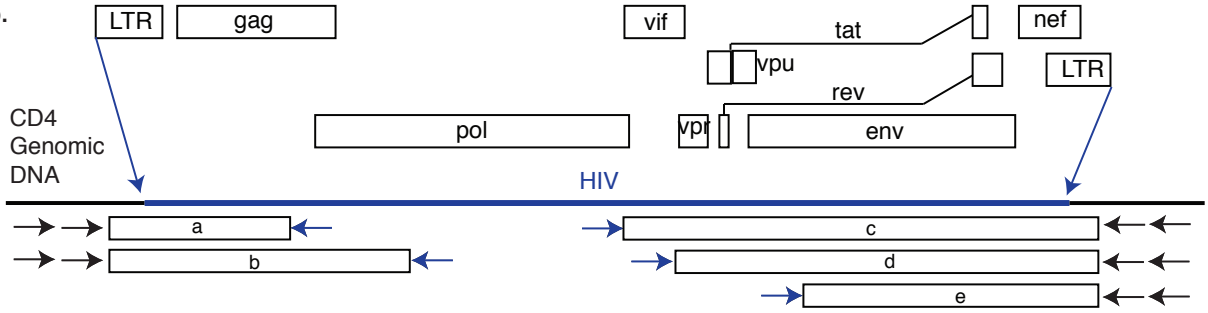
Our method of integration sequencing captures the end of the 3' LTR and identifies the genomic site of viral integration. To determine whether the viruses found in expanded clones are intact, we used nested integration site-specific PCR primers that were anchored in the host genome to amplify the 5'LTRs of 75 expanded clones from 8 individuals. The clones selected for PCR verification varied in size from 5-200 out of 0.3-2x10^6 CD4+ T cells. Of the 75 sequences obtained, 24 showed fragmented 5'LTRs flanked by the correct genomic site, and an additional 44 of the proviruses did not have a recoverable 5' end (Figure 2.11a). The remaining 8 proviruses with intact 5' LTRs were amplified in limiting dilution conditions using integration site-specific primers and HIV-1 primers (Figure 2.11b). Three of the 8 proviruses could not be amplified; 4 had large deletions in Env, 1 had a frameshift mutation in *pol* and 1 had undergone APOBEC3G mediated hypermutation to produce a premature stop codon in *env* (Figure 2.11c). Thus, we were unable to find a single intact integrated provirus among 75 expanded clones.

Figure 2.11. Large expanded clones are defective. **a)** Sequence analysis of 5'LTRs in clonally expanded integrations. Of 75 different clonally expanded integrations from 8 individuals, 24 showed fragmented 5' LTRs, 44 didn't have a recoverable 5' LTR, and 8 contained intact 5'LTRs. **b)** Strategy for HIV-1 sequencing. 8 proviruses were analyzed for intact viral sequence. Nested genomic primers and internal HIV primers were used in a PCR walking strategy to amplify fragments a-e from specific clonally expanded integrations. PCR products were sequenced directly. **c)** Summary of HIV-1 sequencing from large expanded clones. Sequences were aligned to HXB2 and examined for presence of large internal deletions. Intact sequences were analyzed for G → A hypermutation by Los Alamos Hypermut algorithm. Non hypermutated products were analyzed for intact reading frames and frameshift mutations by Los Alamos HIVQC. Green dot: intact, non hypermutated sequence. Red dot: no PCR product recovered. Red triangle: sequence with internal deletion. –: not done.

a.



b.



c.

Patient	a	b	c	d	e
3	no viral sequence detected by PCR				
8	no viral sequence detected by PCR				
9	no viral sequence detected by PCR				
10	●	●	△	●	●
11	—	frameshift	—	—	—
11	—	—	△	—	—
11	—	—	—	—	hypermut/stop
13	●	—	△	—	—

● Intact
● No PCR product
△ Deletion
— Not done

Using a novel integration sequencing method, we studied the integration profile of HIV-1 in viremic progressors, individuals receiving antiretroviral therapy and viremic controllers. We identified infected clonally expanded T cells which represented the majority of all integrations and increased during therapy. However, we could not recover intact virus from any of the 75 clones we assayed. This data indicates that the HIV-1 reservoir likely resides in CD4+ T cells that have not undergone extensive clonal expansion.

CHAPTER 3:

QUANTITATIVE AND QUALITATIVE VIRUS CHARACTERIZATION

Quantitative and Qualitative Viral Outgrowth Assay

After the identification of clonally expanded infected cells, and the observation that the majority of the largest clones harbor defective viruses, I sought to understand the composition of the replication competent reservoir. To investigate the genetic and phenotypic complexity of the replication-competent reservoir, the quantitative viral outgrowth assay (QVOA) protocol was modified to increase the number of unique outgrowth cultures.

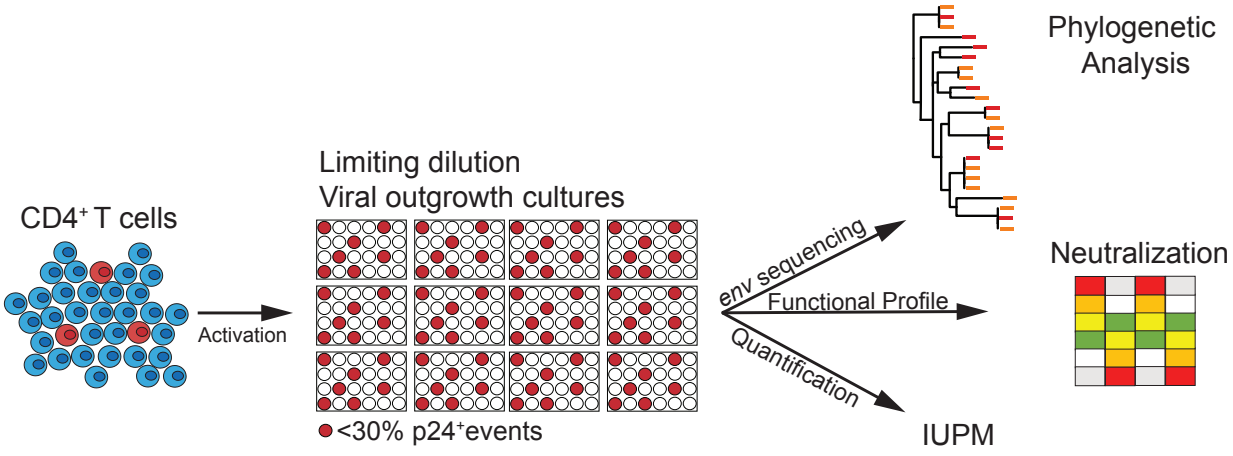


Figure 3.1 Quantitative and qualitative analysis of the replication-competent reservoir. Diagrammatic representation of the assay. CD4+ T cells are cultured at a limiting dilution under conditions whereby a single virus emerges from the latent reservoir in each positive well (red). The number of infectious units per million (IUPM) is determined directly from the number of p24-positive wells. Virus-containing supernatants from positive cultures are harvested for env sequencing and neutralization assays.

Unlike QVOA, where multiple dilutions are assayed to determine the frequency of infected cells, Q^2VOA is performed using a single predetermined dilution that produces less than 30% positive wells. This maximizes the total number of individual viruses that can be quantified and then subsequently phenotypically assayed by sequencing or antibody neutralization. Based on Poisson distribution, this technique produces individual viral outgrowth cultures that are likely to contain single replication-competent proviruses (Figure 3.1).

CD4+ T lymphocytes were isolated from each of four chronically infected individuals who had been virologically suppressed by combination ART for 4–22 years, at two time-points 4–6 months apart (Table 3.1). Between 0.40–1.44 × 10^8 CD4+ T lymphocytes from each ART-treated individual were tested at each time point. On average, 13.5% of cultures were positive for p24. The number of cells yielding replication-competent viruses varied across individuals from 0.19 to 1.07 infectious units per million, which is similar to values obtained by others (Laird et al., 2013) (Table 3.2).

Table 3.1 Clinical Characteristics of Study Subjects.

Study ID	Age	Sex	Year of HIV diagnosis	CD4 nadir	Years since HIV diagnosis	Years on ART	ART regimen
B106	27	M	2008	390	7	7	TDF/FTC/RPV
B115	44	M	1993	200	22	22	DRV/r, ABC, 3TC
B155	59	M	1993	444	22	15	TDF/FTC/RPV
B199	49	M	2009	200	6	4	TDF/FTC, RAL

Table 3.2 Overall Q²VOA results and IUPM

	Months between time points	Total CD4⁺ cells tested	Wells tested	Positive wells (%)	IUPM	Total CD4⁺ cells tested	Wells tested	Positive wells (%)	IUPM
Study ID			**Time point 1**				**Time point 2**		
B106	4	39.6×10^6	132	31 (23.5)	0.89	75.6×10^6	252	28 (11.1)	0.39
B115	4	57.6×10^6	192	40 (20.8)	0.57	140×10^6	468	50 (10.7)	0.38
B155	6	75.6×10^6	252	69 (27.4)	1.07	72×10^6	240	24 (16.7)	0.35
B199	4	43.2×10^6	144	24 (16.7)	0.61	144×10^6	480	26 (5.4)	0.19

To characterize the cultured viruses molecularly, cDNA was produced from culture supernatants and sequenced the *env* gene using primers that resulted in a clonal prediction score of 94 of 100 (silicianolab.johnshopkins.edu/cps) (Laskey et al., 2016). Thus, there was a high probability that identical *env* sequences represented identical full-length genomes. 234 *env* sequences were obtained from Q²VOA, of which 13.7% were excluded from further analysis due to the presence of short reads (3.8%) or the presence of reads producing an inconclusive consensus (9.8%). The phylogenetic analysis of the remaining 202 *env* sequences showed that the four individuals were infected with epidemiologically unrelated clade B viruses (Figure 3.2). Phylogenetic analysis of individual sequences revealed the existence of a diverse viral population composed of multiple (bootstrap-supported) clusters for each of the four individuals (Figure 3.3a).

To compare the diversity of viruses obtained from a “bulk” culture with the diversity of viruses derived by Q²VOA, single genome analysis (SGA) was performed on bulk culture supernatants established at the same time from the same four individuals.

Figure 3.2 Four individuals are infected with epidemiologically unrelated clade B viruses. Maximum likelihood phylogenetic tree was constructed from viral env sequences from outgrowth culture supernatants as well as archived proviral DNA from all participants. Hypervariable (as defined in https://www.hiv.lanl.gov/content/sequence/VAR_REG_CHAR) and other poorly aligned regions were excluded from the analysis. The tree was constructed using RAxML v. 8.0.22 (55) with a GTRGAMMA substitution model, with 1,000 bootstrap replicates and midpoint rooted. Scale bar indicates diversity.

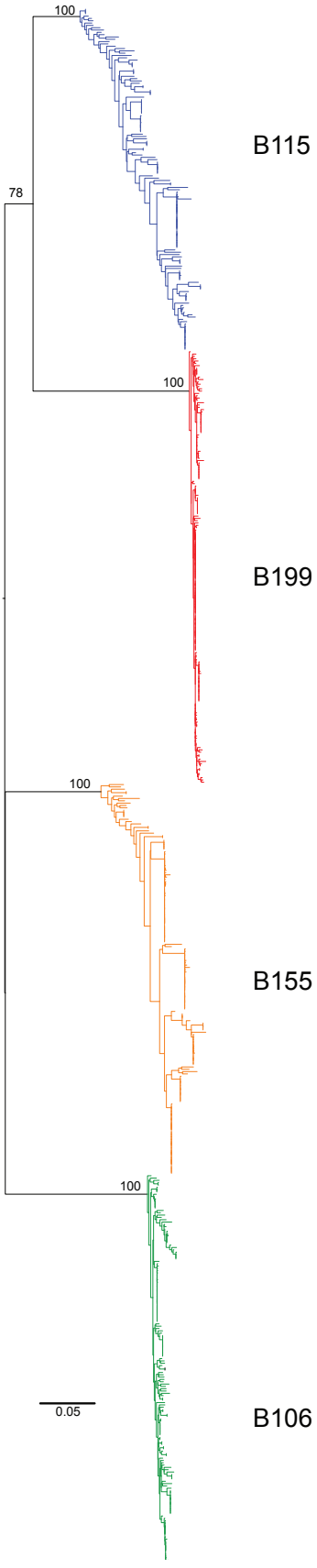

100
78
100
100
100
0.05
B115
B199
B155
B106

Figure 3.3 Env sequences from outgrowth cultures. a) Maximum likelihood phylogenetic trees of full-length env sequences of viruses from Q2VOA outgrowth cultures from four individuals. Viruses from time point 1 are green, viruses from time point 2 are red, and bulk culture SGA is gray. Asterisks indicate nodes with significant bootstrap values (bootstrap support ≥ 70%). Numbers next to sequences correspond to viruses assayed for neutralization in Fig. 6. **b)** Pie charts depict the distribution of culture-derived env sequences from the two time points. The number in the inner circle indicates the total number of env sequences analyzed. White represents sequences isolated only once across both time points, and colored areas represent identical sequences that appear more than once. The size of the pie slice is proportional to the number of sequences in the clone. Clones found at both time points are the same color and denoted by asterisks. Percentages of identical sequences are displayed at the bottom right of each pie chart. **c)** Representation of overlapping sequences between the two time points. The size of the hemisphere is proportional to the number of sequences. Light blue hemispheres represent overlapping sequences and gray hemispheres represent the total number of sequences. The percentage of overlap is indicated at the bottom of each hemisphere.

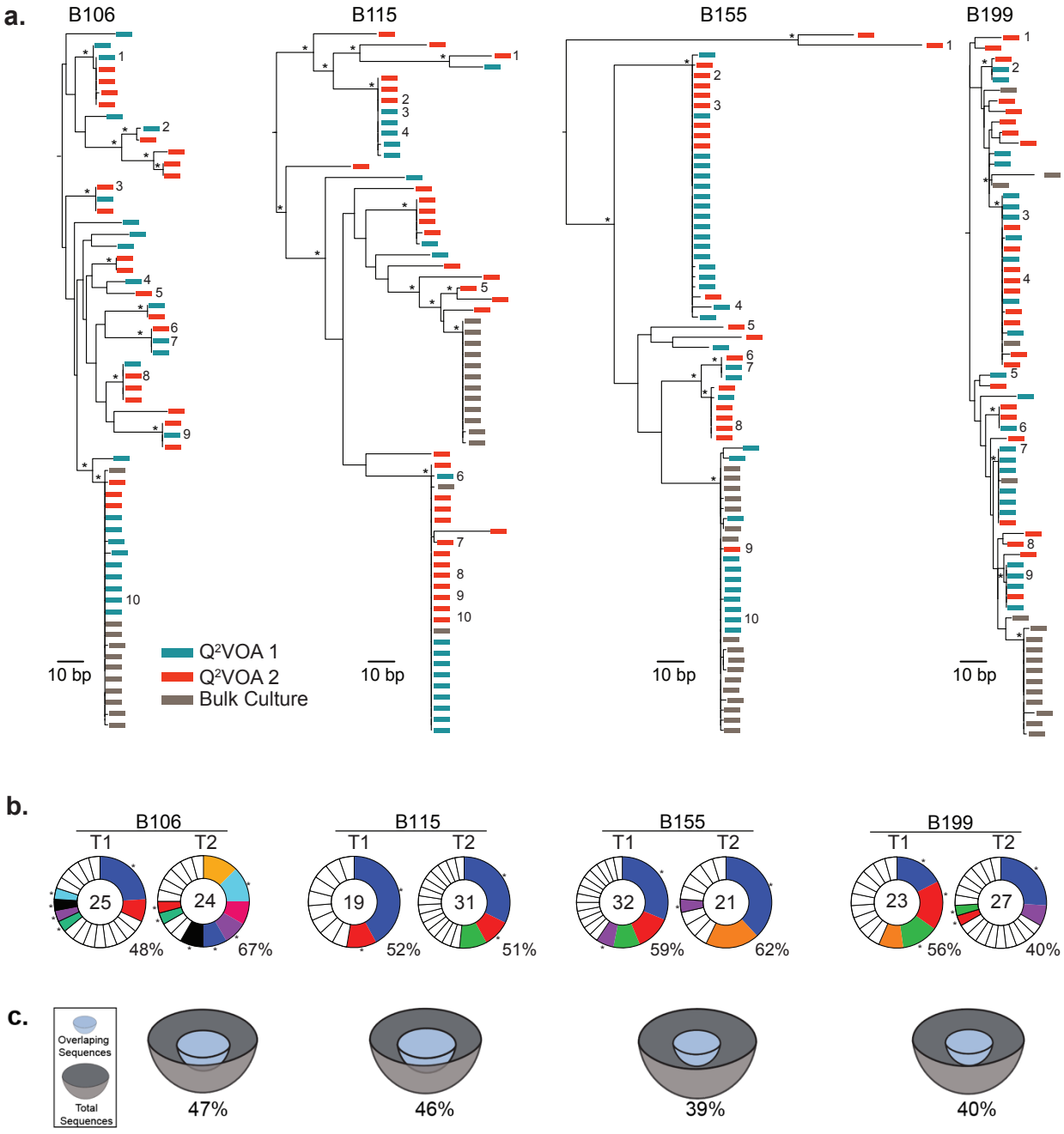

a.
B106
B115
B155
B199
10 bp
Q²VOA 1
Q²VOA 2
Bulk Culture
b.
B106
T1
T2
25
24
48%
67%
B115
19
31
52%
51%
B155
32
21
59%
62%
B199
23
27
56%
40%
c.
Overlaping Sequences
Total Sequences
47%
46%
39%
40%

Table 3.3 Distribution of observed sequences in Q²VOA.

	Non-clonal			Clonal			Overlapping sequences between
Study ID	T1	T2	Total (%)	T1	T2	Total (%)	T1 and T2 (%)
B106	13	8	21/49 (42.9)	12	16	28/49 (57.1)	23/49 (46.9)
B115	9	15	24/50 (48.0)	10	16	26/50 (52.0)	23/50 (46.0)
B155	13	8	21/53 (39.6)	19	13	32/53 (60.4)	21/53 (39.6)
B199	10	16	26/50 (52.0)	13	11	24/50 (48.0)	20/50 (40.0)

In contrast to Q²VOA, bulk culture supernatants were mainly monotypic and showed much reduced overall diversity (Figure 3.3a). Thus, when multiple infected cells are reactivated in a single culture, the strain or strains with the fastest growth kinetics, or the greatest fitness in culture, dominate.

Stability of replication competent reservoir over time

Individual replication-competent viruses obtained from different Q²VOA cultures frequently encoded identical *env* sequences. When Q²VOA-derived viruses obtained at the two time points were compared for each subject, typically less than half (40–52%) of their *env* sequences were unique (Table 3.3).

The majority of sequences were identical to at least one other independently derived replication-competent virus obtained from the same subject. For example, of a total of 49 *env* sequences isolated from the two time points from B106, only 21 were unique. The majority (22) were identical to at least one other sequence that appeared at one of the two time-points. Irrespective of whether they appear at one or both time points, these repeated sequences will be referred to as “clones” because they must originate from at

least two different CD4+ T cells. This finding does not necessarily imply that the viruses are integrated in the same location in the genome because it is possible that identical viruses can infect different CD4+ T cells. The size of these clones ranged from two to 10 members, with a mean of 3.69 when all four individuals and both time points were considered (Figure 3.3b, c and Table 3.3). Fifty-four percent of all replication-competent viruses emerging in Q^2VOA cultures were derived from expanded clones (Figure 3.3b, c and Table 3.3).

To determine whether the viral sequences obtained from the replication-competent reservoir remained stable over time, the sequences from the two time points for each individual were compared. Many branches in the phylogenetic trees contained sequences derived from both time points (Figures 3.3a-c and Table 3.3). The relationship between the sequences from both time points was formally assessed by determining their Genealogical Sorting Index (GSI), which quantitates the degree of phylogenetic association between sequences (Cummings et al., 2008). GSI values, which range between 0 (complete interspersion) and 1 (complete monophyly), showed that for each of the four individuals, the sequences obtained by Q^2VOA from both time points could not be segregated as distinct groups (Table 3.4). This finding demonstrates that the viral population emerging from the latent reservoir in four individuals was stable over the 4- to 6- month time interval analyzed.

Table 3.4 GSI and probability values for HIV env trees under the null hypothesis that Q²VOA-derived sequences from both visits are a single mixed group.

	Q²VOA			
	Visit 1		Visit 2	
Study ID	GSI	P-value	GSI	P-value
B106	0.067	0.45	0.02	0.92
B115	0.066	0.14	0	1
B155	0.059	0.26	0.012	0.82
B199	0.053	0.3	0	1

Comparison of replication competent viruses with proviral sequences

To examine the relationship between proviruses integrated into CD4+ T-cell DNA and the replication-competent viruses obtained by Q²VOA, SGA on DNA isolated from primary CD4+ T lymphocytes was performed from the same individuals at both time points. We obtained a total of 498 env sequences, of which 16.3% were excluded from further analysis due to the presence of hypermutated regions (2.4%), short reads (5.6%), or reads producing an inconclusive consensus (8.2%). The remaining 417 full-length env sequences, or 85–113 per individual, fell within the same four patient-specific clades as the Q²VOA-derived env sequences, indicating absence of sample mix-up or contamination (Figure 3.2).

As previously reported for archived proviral DNA amplified from primary CD4+ T cells, we found unique sequences, as well as large groups of identical sequences, marking presumptive expanded cell clones (Cohn et al., 2015; Maldarelli et al., 2014; von Stockenstrom et al., 2015; Wagner et al., 2013; Wagner et al., 2014). For example, in B106, 67 of 113 env sequences obtained from the two time points were unique and the remaining 46 were members of clones. In addition, 36 of the 46 identical sequences overlapped between the two time points (Figures 3.4a and b and Table 3.5). Thus, like

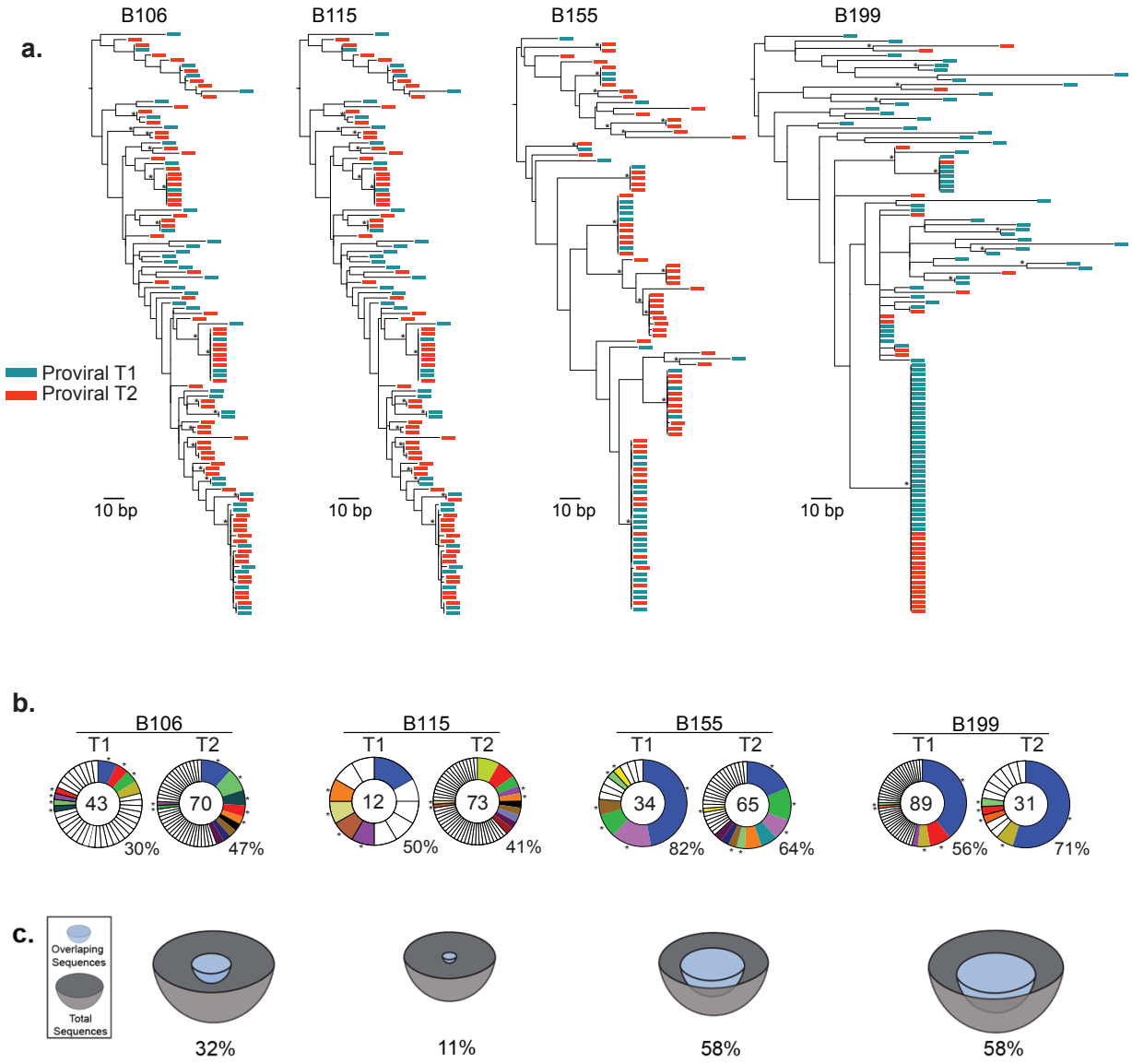
the Q²VOA-derived sequences, the archived proviral population was stable over the time interval analyzed.

Table 3.5 Distribution of observed sequences in proviral DNA

	Non-clonal			Clonal			Overlapping sequences between
Study ID	T1	T2	Total (%)	T1	T2	Total (%)	T1 and T2 (%)
B106	30	37	67/113 (59.3)	13	33	46/113 (40.7)	36/113 (31.9)
B115	6	43	49/85 (57.6)	6	30	36/85 (42.4)	10/85 (11.8)
B155	6	23	29/99 (29.3)	28	42	70/99 (70.7)	58/99 (58.6)
B199	39	9	48/120 (40.0)	50	22	72/120 (60.0)	70/120 (58.3)

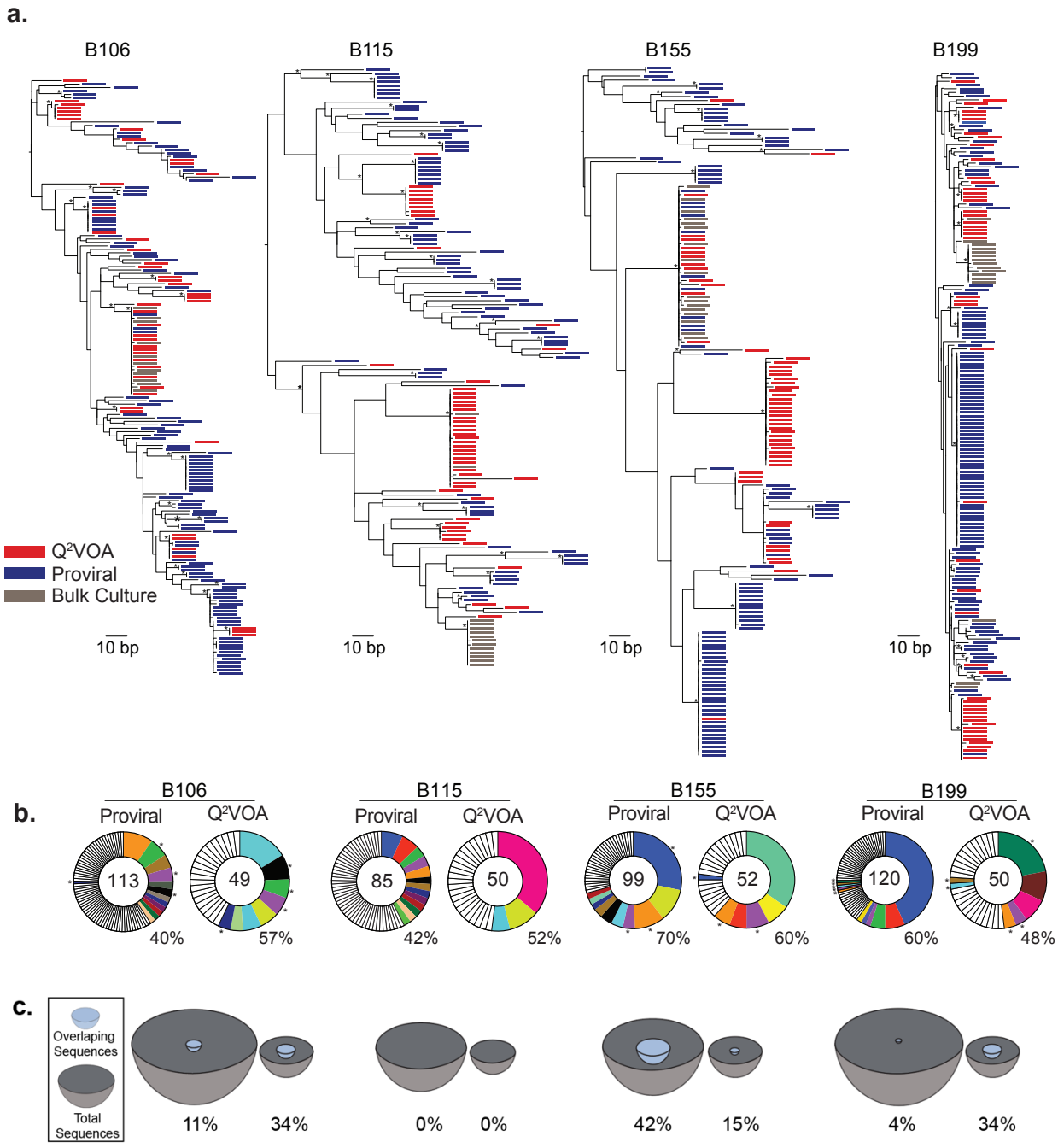
Archived proviral DNA sequences were then compared with replication-competent Q²VOA-derived sequences. Because both the proviral DNA and the replication-competent viral sequences for any given individual were overlapping at the two time points, we combined each of the two sets of sequences (Figure 3.5). As might be expected, given that the sample size is limited, we found relatively limited overlap

Figure 3.4 Env sequences from archived proviral DNA. a) Maximum likelihood phylogenetic trees of full-length env sequences derived by SGA from primary CD4+ T cells from four individuals. Viruses from time point 1 are green, and viruses from time point 2 are red. Asterisks indicate nodes with significant bootstrap values (bootstrap support ≥ 70%). **b)** Pie charts depict the distribution of archived env sequences from the two time points. The number in the inner circle indicates the total number of env sequences analyzed. White represents unique sequences isolated only once across both time points, and colored areas represent identical sequences that appear more than once. The size of the pie slice is proportional to the number of sequences in the clone. Clones found at both time points are the same color and denoted by asterisks. Percentages of identical sequences are displayed at the bottom right of each pie chart. **c)** Representation of overlapping sequences between the two time points. The size of the hemisphere is proportional to the number of sequences. Light blue hemispheres represent overlapping sequences and gray hemispheres represent the total number of sequences. The percentage of overlap is indicated at the bottom of each hemisphere.


a.
B106
B115
B155
B199
Proviral T1
Proviral T2
10 bp
10 bp
10 bp
10 bp
b.
B106
T1
T2
43
70
30%
47%
B115
T1
T2
12
73
50%
41%
B155
T1
T2
34
65
82%
64%
B199
T1
T2
89
31
56%
71%
c.
Overlaping Sequences
Total Sequences
32%
11%
58%
58%

between the archived proviral sequences and Q²VOA culture-derived sequences in phylogenetic trees, with multiple large clusters composed of sequences isolated from only a single source (Figure 3.5a). The 417 archived proviral sequences contained 50 expanded clones, 12 of which were also found in the replication-competent outgrowth cultures. The relative frequency of expanded clones from these two sources differed: Some greatly expanded clones identified in primary CD4+ T-cell DNA represented only a small fraction of clones identified in the outgrowth cultures, and rare archived clones were disproportionally abundant among the reactivated latent replication-competent viruses (Figure 3.5b). This finding is best illustrated in subject B155, where the largest archived proviral clone, which comprised 28 of 99 total sequences, was found only once among 52 replication-competent viruses (Figures 3.5b and c). In contrast, in B199, one archived proviral sequence, which appeared only once in a total of 120 sequences derived from primary CD4+ T-cell DNA, was found in 11 of 50 Q²VOA culture-derived sequences emerging from the latent reservoir. B115 provided the clearest example of the discrepancy between proviral DNA and cultured viruses, with no instances of matching sequences between 85 proviral sequences and 50 outgrowth viruses (Figures 3.5b and c). Although this discrepancy is most likely due to the prevalence of defective archived proviral sequences (Bruner et al., 2016; Cohn et al., 2015; Eriksson et al., 2013; Ho et al., 2013), differences in proviral accessibility to polymerase may also contribute.

Figure 3.5 Comparison of env sequences from archived proviruses and replication-competent viruses. Sequences from the two time points were pooled for each participant. **a)** Maximum likelihood phylogenetic trees of env sequences. Limiting dilution outgrowth viruses are red, bulk culture viruses are gray, and viral sequences amplified from primary CD4+ T cells are blue. Asterisks indicate nodes with significant bootstrap values (bootstrap support ≥ 70%). **b)** Pie charts depicting the distribution of archived proviruses and culture-derived sequences. The numbers in the inner circles indicate the total number of env sequences analyzed. White represents sequences isolated only once, and colored areas represent identical sequences. The size of the pie slice is proportional to the number of sequences in the clone. Clones found in proviral DNA and outgrowth cultures are the same color and denoted by asterisks. Percentages of identical groups of sequences are displayed at the bottom right of each pie chart. **c)** Representation of overlapping sequences between the two sources. The size of the hemisphere is proportional to the number of sequences. Light blue hemispheres represent overlapping sequences and gray hemispheres represent the total number of sequences. The percentage of overlap is indicated at the bottom of each hemisphere.


a.
B106
B115
B155
B199
Q²VOA
Proviral
Bulk Culture
10 bp
10 bp
10 bp
10 bp
b.
B106
Proviral
Q²VOA
113
49
40%
57%
B115
Proviral
Q²VOA
85
50
42%
52%
B155
Proviral
Q²VOA
99
52
70%
60%
B199
Proviral
Q²VOA
120
50
60%
48%
c.
Overlaping Sequences
Total Sequences
11%
34%
0%
0%
42%
15%
4%
34%

between the archived proviral sequences and Q²VOA culture-derived sequences in phylogenetic trees, with multiple large clusters composed of sequences isolated from only a single source (Figure 3.5a). The 417 archived proviral sequences contained 50 expanded clones, 12 of which were also found in the replication-competent outgrowth cultures. The relative frequency of expanded clones from these two sources differed: Some greatly expanded clones identified in primary CD4+ T-cell DNA represented only a small fraction of clones identified in the outgrowth cultures, and rare archived clones were disproportionally abundant among the reactivated latent replication-competent viruses (Figure 3.5b). This finding is best illustrated in subject B155, where the largest archived proviral clone, which comprised 28 of 99 total sequences, was found only once among 52 replication-competent viruses (Figures 3.5b and c). In contrast, in B199, one archived proviral sequence, which appeared only once in a total of 120 sequences derived from primary CD4+ T-cell DNA, was found in 11 of 50 Q²VOA culture-derived sequences emerging from the latent reservoir. B115 provided the clearest example of the discrepancy between proviral DNA and cultured viruses, with no instances of matching sequences between 85 proviral sequences and 50 outgrowth viruses (Figures 3.5b and c). Although this discrepancy is most likely due to the prevalence of defective archived proviral sequences (Bruner et al., 2016; Cohn et al., 2015; Eriksson et al., 2013; Ho et al., 2013), differences in proviral accessibility to polymerase may also contribute.

To determine the extent to which sequences from replication- competent viruses and archived proviruses were compartmentalized, we calculated their GSI. This analysis

demonstrated that archived proviral and Q²VOA culture-derived sequences were significantly segregated for every single individual analyzed (Table 3.6). Thus, proviral sequences derived from primary CD4+ T-cell DNA do not provide an accurate representation of viruses comprising the replication-competent reservoir.

Table 3.6 GSI and probability values for HIV env trees under the null hypothesis that Q²VOA-derived sequences and proviral-derived sequences are a single mixed group.

	Q²VOA		Proviral	
Study ID	**GSI**	**P-value**	**GSI**	**P-value**
B106	0.184	<0.01	0.114	<0.001
B115	0.61	<0.001	0.43	<0.001
B155	0.421	<0.001	0.125	<0.001
B199	0.15	<0.001	0.071	<0.05

To understand the relationship between the two groups of sequences better, we analyzed them by a mathematical model that describes every sequence by two variables. The first (p) is the frequency of a sequence in the proviral compartment, and the second (r) is its probability of reactivation in the active viral culture. These parameters were extracted from the experimental data using Bayesian inference methods. The data show that clone size is negatively correlated with the activation probability ($r = –0.94$, $p = 3.4 \times 10–34$) (Figure 3.6). These results indicate that the larger the clone of archived proviral sequences, the lower is its probability of representing a replication-competent virus in the Q²VOA culture.

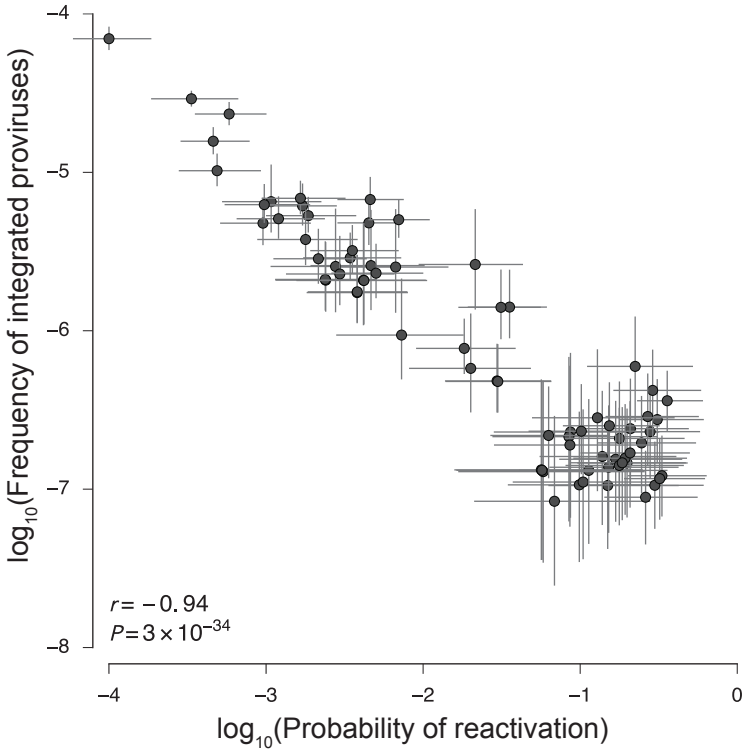


Figure 3.6 Negative correlation between proviral clone size and probability of reactivation in culture. The frequency of integrated provirus (p) is negatively correlated with its probability of reactivation in culture (r). Bars denote interquartile ranges of posterior parameter estimates. The Pearson correlation is computed using median values for large clones. Clones with diverse frequencies and reactivation probabilities are observed in each individual.

CHAPTER 4:

LATENT CELL CAPTURE AND CHARACTERIZATION

Latency capture enriches HIV-1 RNA producing cells

To investigate the cells that contribute to the latent reservoir, I developed a method to enrich and isolate reactivated latent cells by combining antibody staining, magnetic enrichment, and flow cytometry (Pape et al., 2011) (latent cell capture, or LURE). Purified CD4+ T cells from ART suppressed donors were activated with PHA, a robust *in vitro* latency reactivation agent, for 36h in the presence of 5 potent antiretroviral drugs to prevent new infection and virion maturation, and a pan-caspase inhibitor to reduce cell death associated with active infection. Reactivated latent cells expressing surface HIV-1 Envelope (Env) protein were labeled with a cocktail of biotinylated anti-Env broadly neutralizing antibodies (bNAbs), streptavidin-PE, and anti-PE magnetic beads, followed by enrichment over a magnetic column (Figure 4.1).

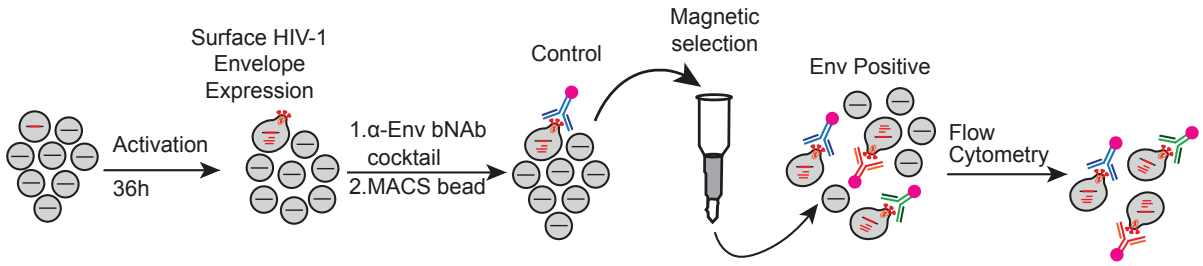


Figure 4.1 Diagrammatic representation of latency capture (LURE) protocol. CD4+ T cells from ART suppressed donors are cultured in conditioned media with PHA, IL-2, antiretroviral drug cocktail and pan-caspase inhibitor for 36h. Cells are labeled with a biotinylated bNAb cocktail, followed by Streptavidin PE and anti-PE magnetic beads, passed over a magnetic column, and FACS analysis.

Our anti-Env antibody cocktail consists of 3BNC117 (Scheid et al., 2011), 10-1074 (Mouquet et al., 2012), and PG16 (Walker et al., 2009), that together cover over 90% of all viral envelopes (Yoon et al., 2015).

Relative enrichment of the magnetically isolated, Env+ cellular fraction was measured by comparison to unfractionated control cells from the same culture by flow cytometry (Figure 4.2) and by quantitative PCR for HIV-1 *gag* mRNA (Figure 4.3a).

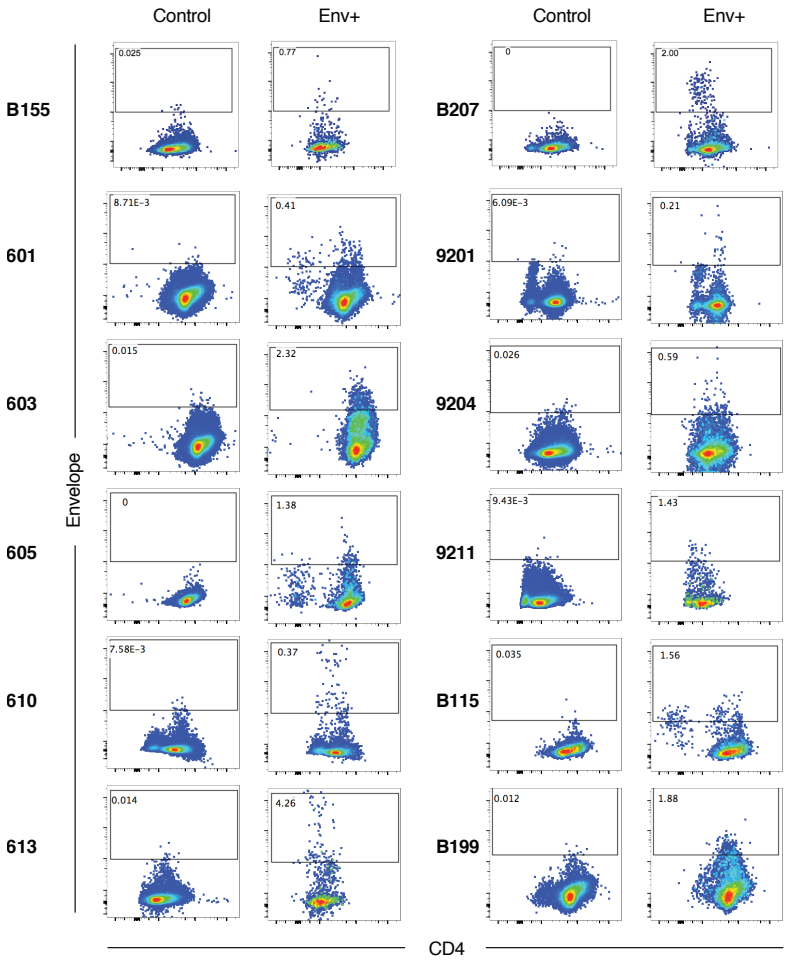


Figure 4.2 Enrichment Env expressing cells by LURE. Dot plots show Env vs. CD4 staining on pre-enrichment control and positively selected cells for all donors. Gate shows frequency of Env+ cells in each population.

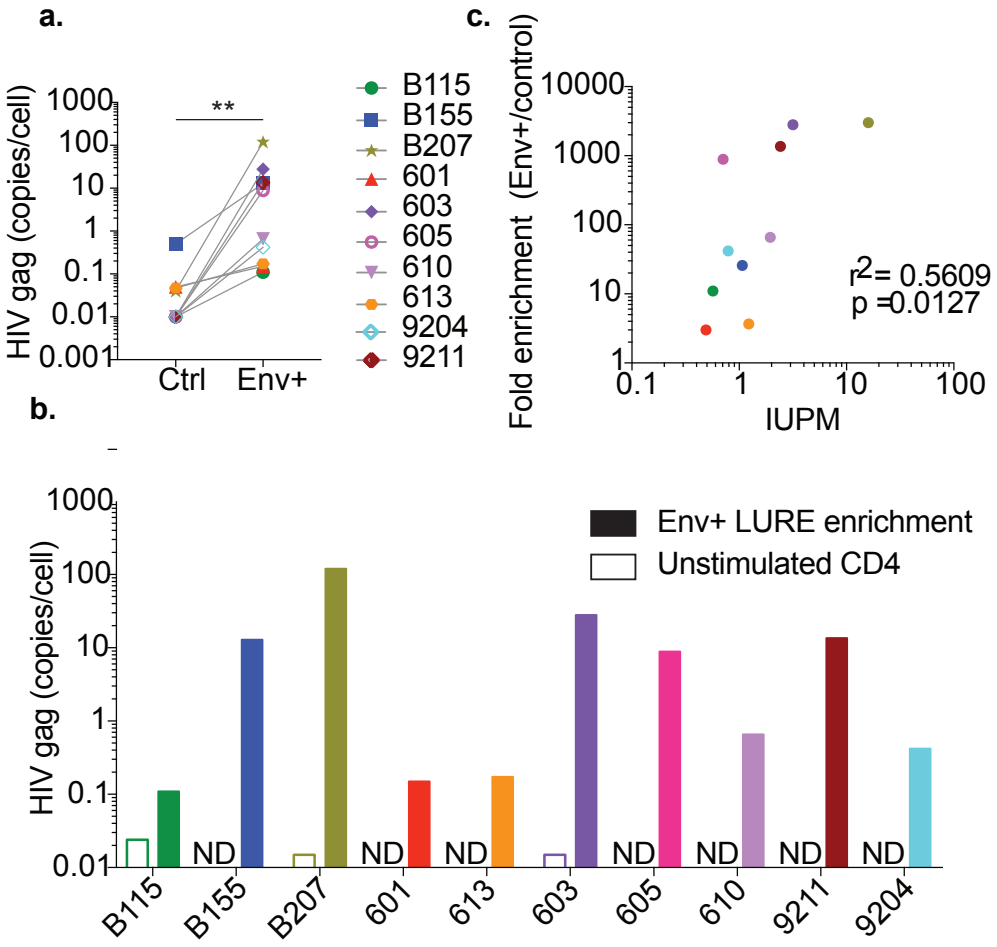


Figure 4.3 HIV-1 RNA enrichment in cells isolated by LURE. a) HIV-gag mRNA was measured in equivalent numbers of Env+ and control cells. Graph shows results of qPCR (12.8-copy limit of detection) for HIV-gag mRNA, normalized to the number of sorted cells. p = 0.002, Wilcoxon matched-pairs signed rank test. **b)** HIV-gag mRNA was measured in equivalent numbers of Env+ and unstimulated CD4+ T cells. Graph shows results of qPCR (12.8-copy limit of detection) for HIV-gag mRNA, normalized to the number of sorted cells. ND: none detected. **c)** Fold-enrichment (Env+/control) in (a) compared to IUPM.

Enrichment of cell associated HIV-1 RNA was entirely dependent on cellular activation with PHA (Figure 4.3b). The degree of enrichment achieved by magnetic columns was measured in samples from 10 individuals and was found to be dependent in part (r^2 = 0.5609, p = 0.0127) on the size of the latent reservoir as measured by viral outgrowth assays in infectious units per million (IUPM) (Figure 4.3c).

For example, participant B207 has an IUPM of 16 and a relative enrichment of HIV-1 *gag* RNA of 3000-fold, participant 610 an IUPM of 1.95 and a 66-fold enrichment, and participant 601 has an IUPM of 0.49 and only a 3-fold enrichment (Figure 4.3c). We conclude that reactivated latently infected cells can be enriched based on HIV-1 Env surface expression.

Full length virus recovered by single-cell RNA Sequencing

To further purify the reactivated latent cells, we used flow cytometry to sort single cells from the magnetically enriched fraction based on Env staining. Individual cells expressing both *env* and *gag* were identified by the combination of surface Env staining and single cell HIV-1 *gag* mRNA expression. The frequency of *gag* mRNA expressing single cells in individuals with high IUPMs ranged from 10-50% of sorted cells (603 and B207, IUPM 3.17 and 16 respectively, Table 4.1). In individuals with relatively lower IUPMs (0.49-2.43), the percent of Env+gag+ single cells isolated varied from 0-4% (12 individuals were examined: Env+gag+ single cells were isolated from 10 of the 12, Table 4.1).

Table 4.1 Patient demographics and LURE experiments. ART abbreviations, ATV: atazanavir, R: ritonavir, ABC: abacavir, 3TC: lamivudine, RPV: rilpivirine, FTC: emtricitabine, TDF: tenofovir disoproxil, RAL: raltegravir, EFV: efavirenz, LPV: lopinavir, EGV: elvitegravir, TAF: tenofovir alafenamide, cobi: cobicistat. Env+ bulk gag RNA enrichment, LURE: gag RNA enrichment performed on immunomagnetically isolated Env+ cellular fraction. YES: significant enrichment in Env+ fraction compared to controls. ND: not done.Single Cell LURE: single cell sort of Env+ enriched LURE cells. YES: gag+ cells identified by single cell qPCR. NO: no gag+ cells identified by single cell qPCR.

Patient ID	Year of birth	Gender	Race	Year Dx	Years on ART (uninterrupted)	ART regimen	IUPM	Env+ bulk gag RNA enrichment, LURE	Single Cell LURE
B115	1971	M	Black	1993	24	ATV/r/ABC/3TC	0.57	YES	NO
B155	1956	M	Black	1993	17	RPV/FTC/TDF	1.07	YES	YES
B199	1966	M	White	2009	6	RAL/FTC/TDF	0.61	ND	YES
B207	1969	M	White/Hisp	2006	11	EFV/FTC/TDF	16	YES	YES
601	1959	M	White	1994	20	LPV/r/ABC/3TC	0.49	YES	NO
603	1972	M	White/Hisp	2003	12	EFV/FTC/TDF	3.17	YES	YES
605	1979	M	White/Hisp	2001	15	RPV/FTC/TDF	0.71	YES	YES
610	1986	M	White	2011	5	RPV/FTC/TDF	1.95	YES	YES
613	1966	M	Multiple/Hisp	1997	19	ATV/r/FTC/TDF	1.23	YES	YES
9201	1973	M	White/Hisp	2013	4	EGV/cobi/FTC/TDF	0.78	ND	YES
9204	1994	M	White/Hisp	2012	5	EGV/cobi/FTC/TAF	0.79	YES	YES
9211	1977	M	Black	2011	5	EGV/cobi/FTC/TDF	2.43	YES	YES

To obtain a more comprehensive understanding of the nature of the cells captured by LURE and the viruses they harbor, we performed single cell RNA sequencing (scRNASeq). Donors 603, 605 and B207, were selected based on length of ART therapy (undetectable viremia for 11-15 years), sample availability, and the range of IUPMs (603: 3.17, 605: 0.71 and B207: 16). We profiled the transcriptome of 2 groups of cells obtained from these 3 individuals: reactivated gag+Env+ single cells captured by LURE, and control unfractionated single cells from the exact same PHA activated culture. In addition,

we performed scRNASeq on activated CD4+ T cells that were productively infected with HIV-1_{YU2} (YU2) *in vitro* and purified by cell sorting using anti-Env antibodies (Figure 4.4).

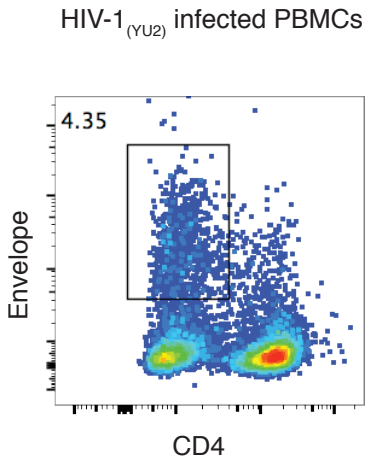


Figure 4.4 Gating strategy for HIV-1YU2 infected cells. Cells infected in vitro with HIV-1_{YU2} for 2 days were FACSorted by gating on Env+CD4lo cells.

Overall 249 cells were characterized, of which 22 cells (8.8%) were removed by quality metrics (Gaublomme et al., 2015). Of the 227 cells retained, 33 were YU2 infected cells, 85 were cells captured by LURE, and 109 were unfractionated control cells from the same cultures. On average, we obtained ~1500 expressed genes per cell (Figure 4.5).

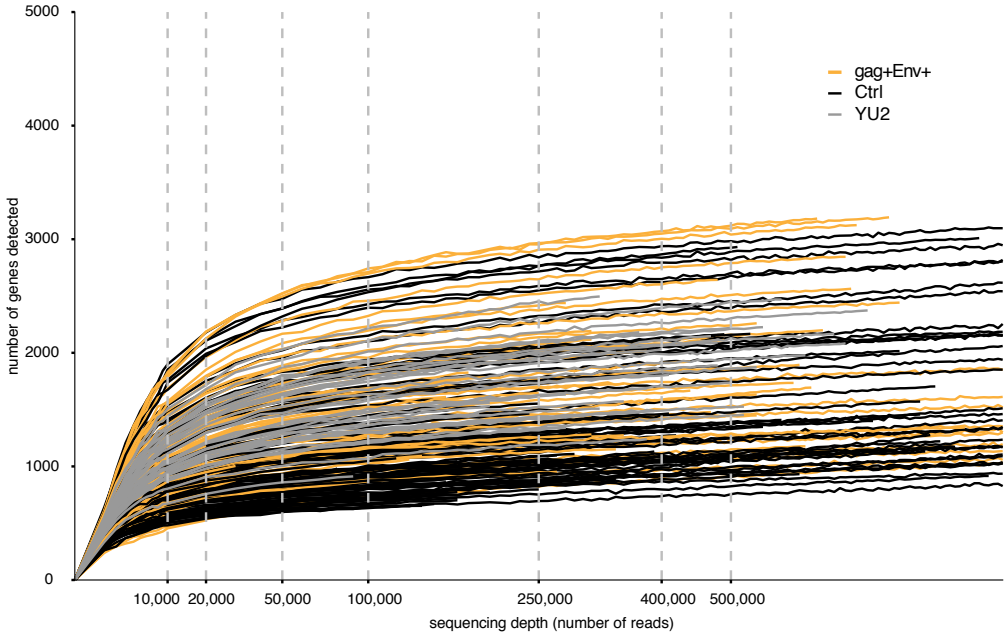


Figure 4.5 Number of genes detected per cell. Results of single cell RNASeq showing saturation of genes detected.

As expected based on the rarity of infected cells in the unfractionated, activated control cells ($\leq 1.6X10^{-5}$), HIV reads were not detectable in these samples (Figure 4.6). In contrast, cells captured by LURE and YU2 infected cells showed similar percentages of total mRNA reads mapping to the HIV-1 genome (3.8 and 4.5% respectively, as expected (Sherrill-Mix et al., 2015)) (Figure 4.6). We conclude that scRNASeq performed on reactivated latent cells captured by LURE contains RNA sequences mapping to the human genome and HIV-1.

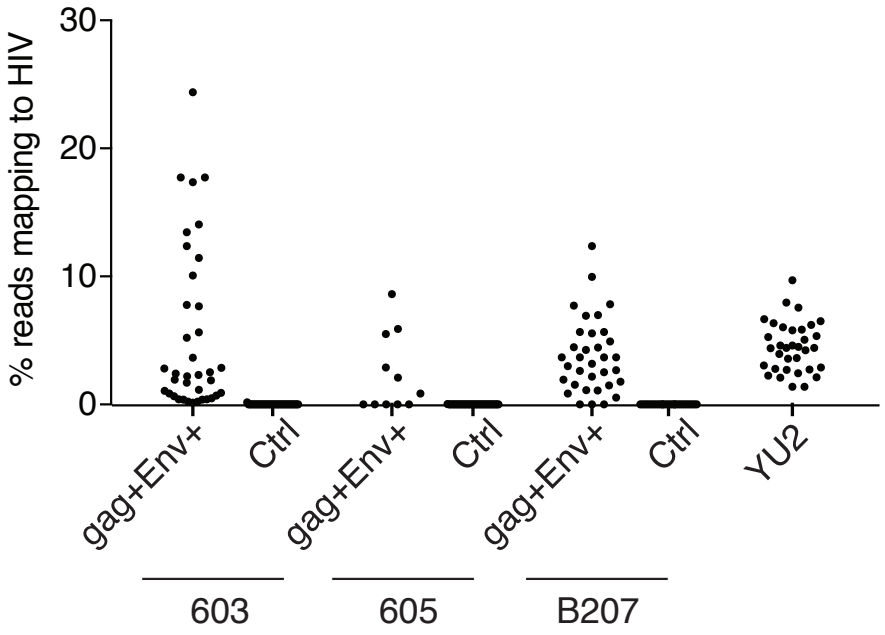


Figure 4.6 Frequency of HIV-1 reads detected in single cell RNA Seq libraries. Fraction of reads mapping to HIV-1 in unfractionated control, LURE purified gag+Env+, and YU2 infected scRNASeq libraries.

To determine whether the reactivated cells captured by LURE express intact viruses, we used Iterative Virus Assembler software to reconstruct the virus from scRNASeq reads in each individual CD4+ T cell (Hunt et al., 2015). HIV RNA recovered by scRNASeq was dependent on proviral transcription as indicated by analysis of HIV-1 splice variants (Figure 4.7).

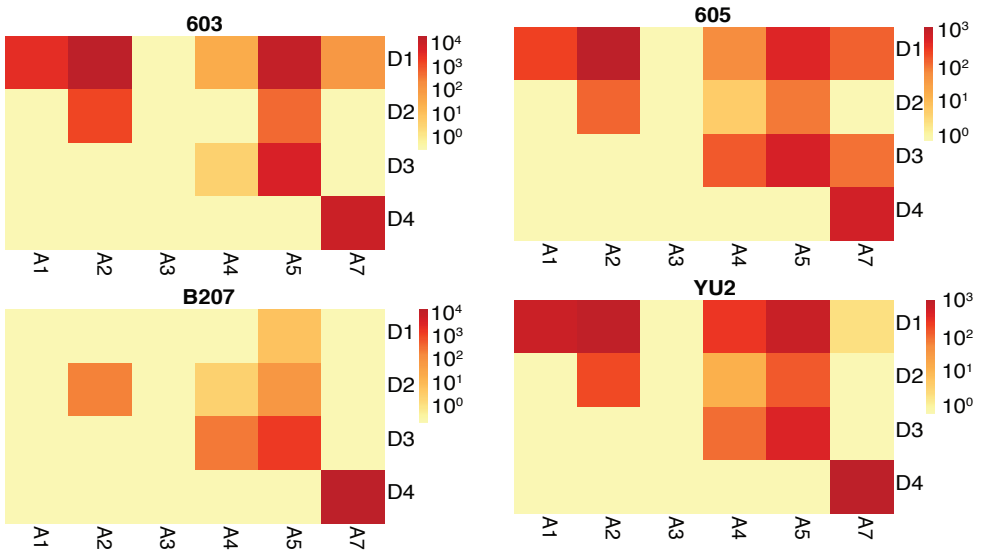


Figure 4.7 HIV-1 splice sites identified in single cell RNASeq libraries. Junctions between HIV splice donors and acceptors observed in RNASeq data. Acceptors are shown as the columns and donors as the rows with the coloring indicating the frequency of reads identified containing indicated splice junction.

We were able to fully reconstruct virus from 26 of 32 individual cells infected with YU2 (Figure 4.8a).

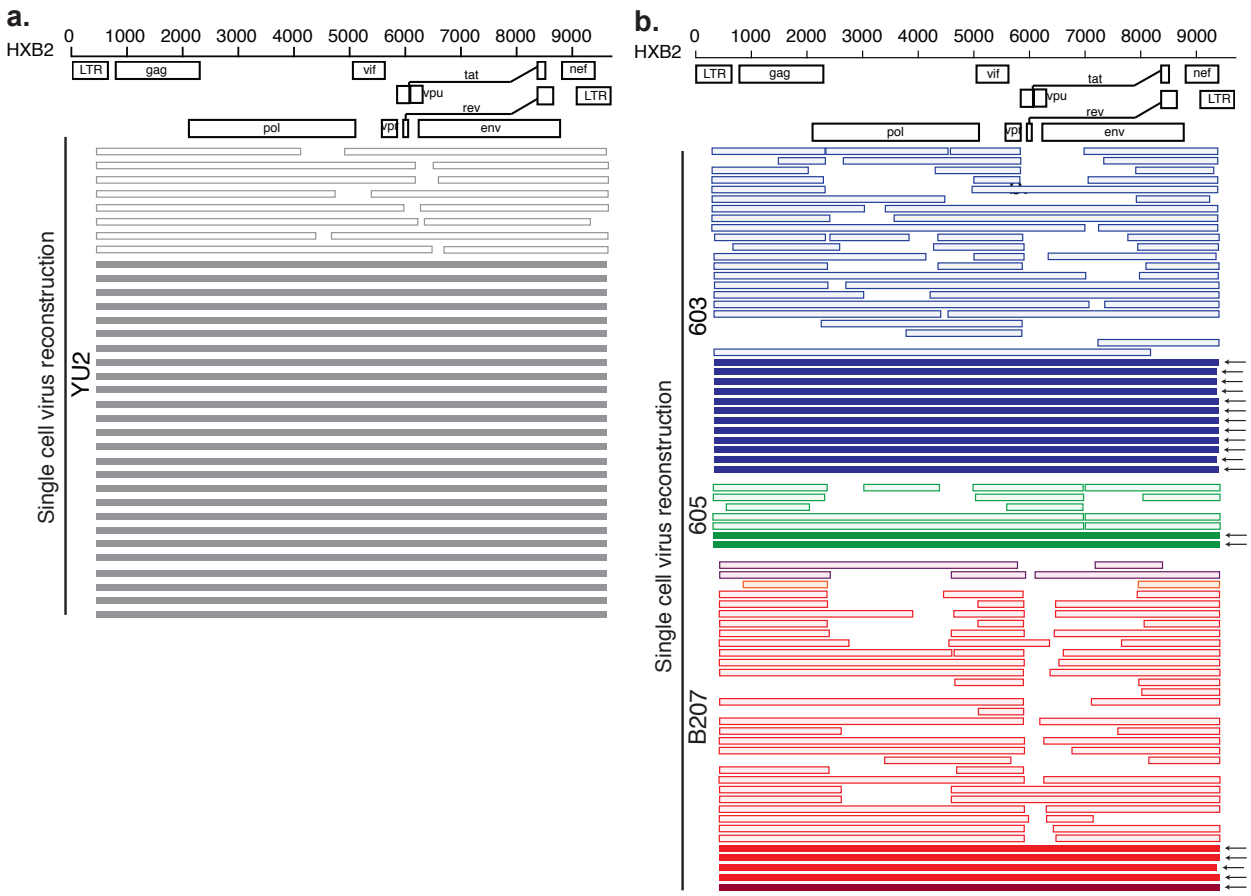


Figure 4.8 Full length virus sequences recovered by scRNASeq. Map of individual viruses reconstructed from scRNASeq. Each horizontal bar represents a single virus from an individual cell. Solid bars and arrows indicate that the entire virus was reconstructed from scRNASeq reads. Outlined, lighter colored bars indicate incomplete genome reconstruction. Different colors indicate different sequences. For participants 603 and 605, every virus identified was identical. For B207, we identified 4 unique viruses, with one clone (in red) predominating.

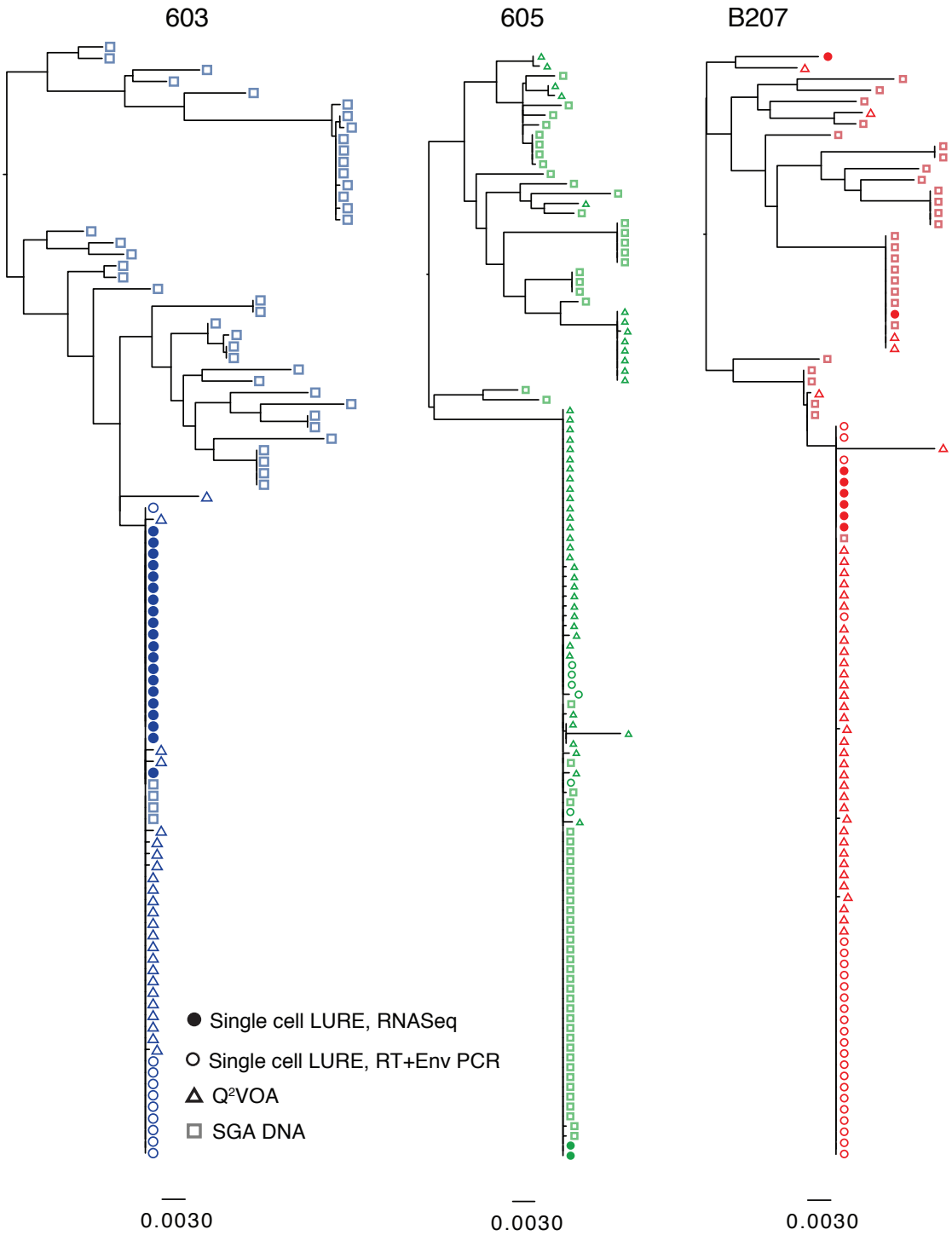
Among the reactivated cells captured by LURE, we were able to fully reconstruct viruses from 12 cells from 603, 2 from 605 and 5 from B207, while the viruses in the remaining LURE cells were partially reconstructed (Figure 4.8b). Every virus sequence obtained by scRNASeq in 603 and 605 belonged to a single expanded viral clone (Figure 4.8b). In

B207, we identified 4 different viruses: the first virus was found in 30 cells. From these 30 cells expressing identical virus, we were able to reconstruct the full virus from 4 cells while in the remaining 26 cells the virus was partially reconstructed; the second virus was found in a single cell and was fully reconstructed; the third virus was found in a single cell and was only partially reconstructed; the fourth virus was found in 2 cells and was also only partially reconstructed (Figure 4.8b). Finally, all of the fully reconstructed viruses obtained from scRNASeq libraries were completely intact when analyzed by Gene Cutter software. Thus, the combination of LURE and scRNASeq can be used to recover full length, intact HIV-1 from single reactivated latent cells.

Captured cells express functional virus

Replication competent latent viruses obtained in viral outgrowth cultures show *Env* sequences that segregate from the majority of defective proviruses found by single genome analysis (SGA) (Bui et al., 2017; Lorenzi et al., 2016). To determine whether the full-length viruses expressed in the purified single cells obtained by LURE correspond to the intact latent viruses that emerge in viral outgrowth assays, we compared their *Env* sequences (Figure 4.9). To do so, we performed quantitative and qualitative viral outgrowth assays (Q^2VOA) (Lorenzi et al., 2016), *Env* SGA on DNA isolated from CD4+ T cells, and compared these sequences to those found in LURE cells.

Figure 4.9 Captured cells express Env that is identical to latent virus emerging in Q2VOA. Maximum likelihood phylogenetic trees compare full length Env sequences derived from single cells capture by LURE (solid and open circles), DNA proviruses (open squares) and replication-competent single cell viral outgrowth cultures (Q2VOA) (open triangles) from participants 603, 605, and B207. Sequences from LURE cells were obtained either by recovery and assembly from RNASeq reads (closed circles) or from reverse transcription of RNA in single cells followed by specific Env PCR from single gag+Env+ LURE cells (open circles). Arrows indicate confirmed full-length sequences.


603
605
B207
Single cell LURE, RNASeq
Single cell LURE, RT+Env PCR
Q²VOA
SGA DNA
0.0030
0.0030
0.0030

Phylogenetic analysis of *Env* sequences revealed that in donors 603 and B207 the Env sequences obtained by LURE and Q^{2}VOA generally clustered together, were part of an expanded clone, and did not overlap significantly with sequences obtained by proviral DNA SGA (Figure 4.9). Participant 605 has an unusual distribution of DNA SGA proviral sequences in that there is a significant overlap with the Env sequences found in viral outgrowth cultures. Nevertheless, the majority of LURE derived Env sequences belong to the major viral outgrowth clone found in Q^{2}VOA (Figure 4.9) in all three individuals. We conclude that the Env sequences expressed by cells purified by LURE are typically identical to those found in viruses that emerge from latent cells in viral outgrowth cultures and therefore are replication competent.

Clones of infected cells in replication competent reservoir

The idea that latent cells harboring identical replication competent viruses arise by clonal expansion is supported by observations that latent cells can divide *in vitro*, identical replication competent viruses can arise from multiple cells, and by proviral integration site mapping (Bui et al., 2017; Cohn et al., 2015; Hosmane et al., 2017; Lee et al., 2017; Lorenzi et al., 2016; Maldarelli et al., 2014; Mullins and Frenkel, 2017; Simonetti et al., 2016; Wagner et al., 2014). However, a less likely alternative interpretation is that latency is established during a viral replicative burst and that identical viruses are integrated into the genome of a diverse group of T cells. To distinguish between clonal expansion and a possible replicative burst, we analyzed the T cell receptor (TCR) sequences obtained from single latent cells captured by LURE. CD4+ T cells express unique antigen receptors

produced by random TCR variable, diversity and joining gene segment (VDJ) recombination. T cells with identical TCRs are only produced by clonal expansion.

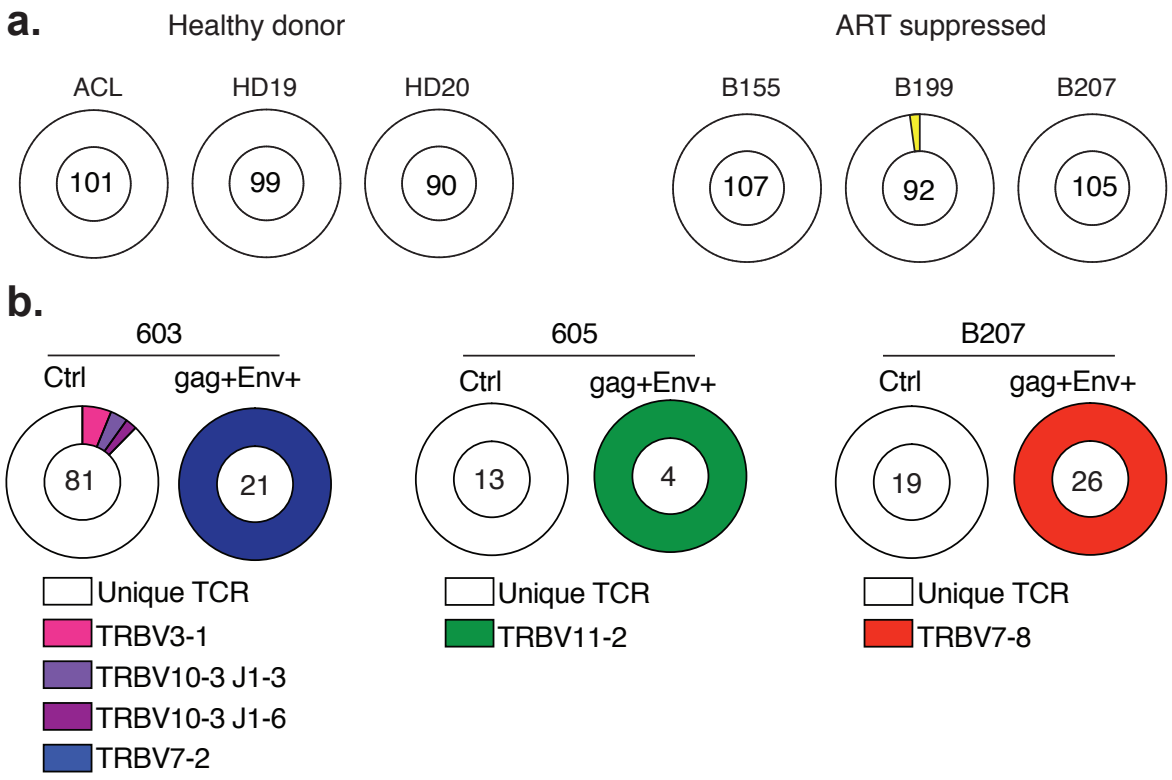


Figure 4.10 Captured cells represents clones of expanded CD4+ T cells. a) TCR sequences amplified by PCR in single sorted CD4+ T cells. The number in the center of the pie denotes the number of cells sequenced; yellow slice is a unique clone consisting of two members. The single clone in B199 was identified by shared TCR alpha and beta sequence. b) TCR sequences recovered from scRNASeq or amplified by PCR, for control (unfractionated pre-enrichment) and gag+Env+ LURE purified cells. The number in the center of the pie denotes the number of cells sequenced; slices are proportional to clone size showing unique TCRs (white slices) and clonal TCRs (colored slices). Clones were identified by their shared TCR alpha and beta sequences.

As a control, we obtained TCR sequences from nearly 600 single CD4+ T cells from 3 healthy and 3 ART treated HIV-1 infected donors. We found that 99.9% of all control TCR

sequences were unique, with only a single 2-member clone identified in 1 of the 6 individuals (Figure 4.10a). In contrast, the TCR sequences derived from the latent cells with identical proviruses captured by LURE (Figure 4.9) were entirely clonal in all 3 donors (Figure 4.10b). To rule out the possibility that the observed clonality was due in part to T cell division *in vitro,* we labeled cells with CFSE and found that there was no measurable T cell division in 36h under our culture conditions (data not shown). We conclude that groups of latent cells containing identical replication competent viruses are products of CD4+ T cell clonal expansion *in vivo*.

Distinct gene signature identified in reactivated latent cells

To further characterize the reactivated latent cells captured by LURE, we performed single-cell transcriptome analysis, and compared the results to unfractionated, PHA stimulated control cells from the same cultures, and to activated CD4+ T cells productively infected with YU2. We included these controls to ensure that any differentially expressed genes or gene signatures would be specific to reactivated latent cells, and not to differences between infected individuals, PHA activation, or active infection. We performed hierarchical clustering using a principal-component analysis (PCA) called Seurat (Satija et al., 2015) using gene expression data from the 227 cells. Seurat identifies significant cell clusters by performing density-based clustering on a two-dimensional t-distributed Stochastic Neighbor Embedding (t-SNE) map of the total gene expression data. This unbiased analysis identified three unique groups of genes that segregated the cells into three separate clusters. Each of these clusters was found to correspond to one of the three input groups: control, LURE, and YU2 infected cells (Figure

4.11). Additional analysis which employs unsupervised clustering using all gene expression data (Single-cell Consensus Clustering, or SC3), confirmed these results comparing control to LURE cells (Figure 4.12). Thus, reactivated latent cells captured by LURE cluster separately from uninfected (control) and actively infected CD4+ T cells by PCA and unsupervised clustering.

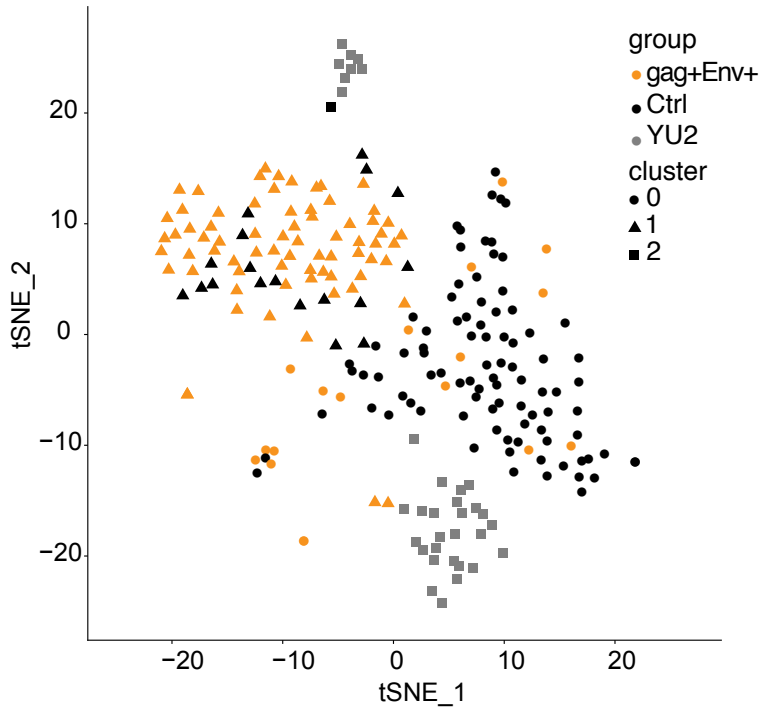


Figure 4.11 Principal components analysis (PCA) clusters cells by group. Shown is the Seurat t-SNE displayed output for the three groups. Plot shows single cells (Control (black), Env+ LURE (orange) and YU2 (gray)). Seurat analysis identified 3 distinct clusters of genes which define three groups of cells (circles (gene cluster 0), triangles (gene cluster 1) and squares (gene cluster 2)) by performing graph-based clustering over 6 principal components.

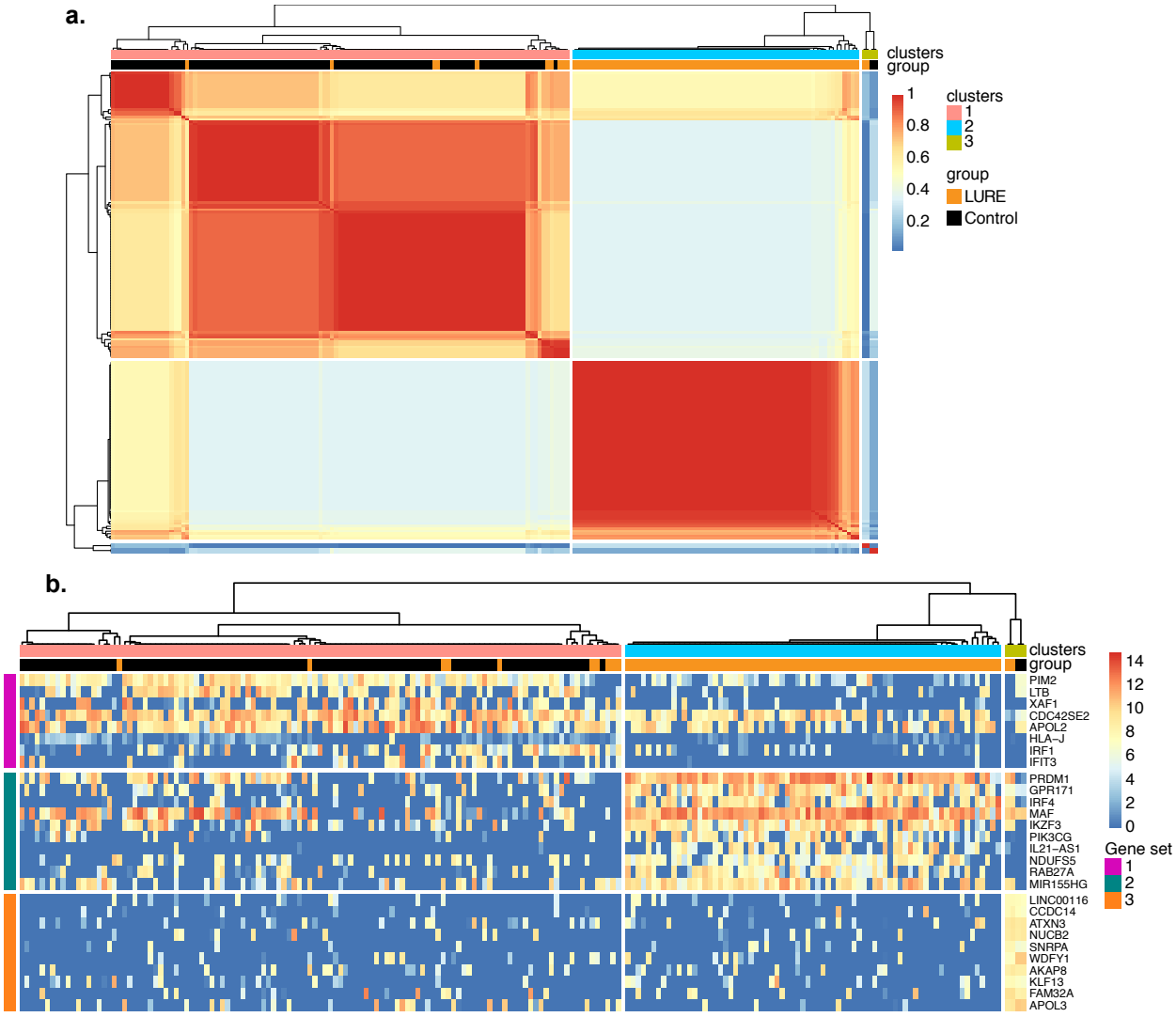


Figure 4.12 Single-cell clustering segregates control from LURE Env+gag+ cells. **a)** Single-cell consensus clustering (SC3) was used to cluster cells in an unsupervised manor. Color spectrum assigns cells to different clusters, with blue indicating assignment to a different cluster and red indicating cells in the same cluster. **b)** SC3-identified marker genes which are highlyexpressed in only one of the clusters and are able to distinguish it from all the remaining ones (blue: low expression, red: high expression of marker gene).

To further understand the transcriptional differences between the three groups of cells, we identified differentially expressed genes (DEG) ($p < 0.01$) between reactivated latent cells and PHA activated control cells. Using unsupervised clustering, we grouped the cells based on the expression of all significantly differentially expressed genes between LURE and control cell groups ($p<0.01$, 778 genes). We find that irrespective of donor, reactivated cells purified by LURE generally segregate from unfractionated, activated control cells in 2 of 3 individuals (Figure 4.13a), with cells from the third individual split between the LURE group and control group. Similar results were also obtained by comparison with YU2 infected cells (Figure 4.13b). We conclude that cells captured by LURE segregate from activated control cells and productively infected cells by three different methods of analysis.

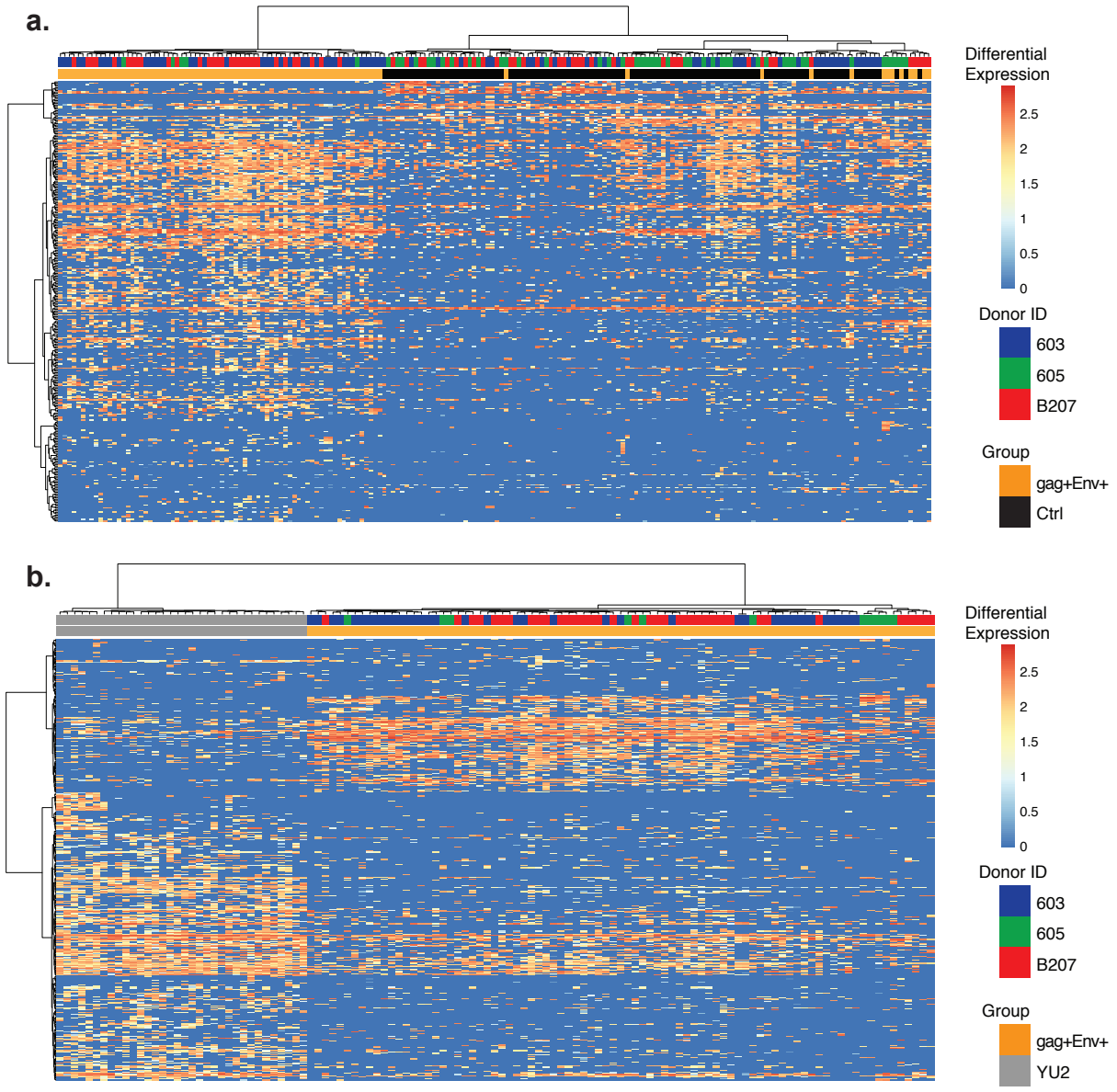


Figure 4.13 Differential gene expression clusters LURE cells from controls. Heat-maps show unsupervised clustering of differentially expressed genes between the gag+Env+ LURE purified group (orange bars) and control unfractionated group (black bars) and YU2 infected cells (gray bars). Cells from donor 603 are indicated in blue, 605 in green, and B207 in red. Color indicates the normalized level of expression.

Among the 240 genes which overlapped between the PCA and DEG ($p < 0.01$) (gene list in the Appendix), we find a number of genes highly expressed in the LURE cells which

have been shown by independent analyses to be associated with HIV-1 latency including microRNAs, chemokines, long non-coding RNAs and transcription factors (Figure 4.14a). For example, cell surface markers associated with latency Tigit (Baxter et al., 2016; Fromentin et al., 2016) and HLA-DR (Cockerham et al., 2014) were 140 and 76-fold up-regulated in cells purified by LURE (Figure 4.14a). CD32a (Descours et al., 2017) was not found by RNASeq or FACS analysis (Figure 4.14b). MiR-155, which inhibits TRIM32, prevents its interaction with HIV tat and reinforces viral latency (Ruelas et al., 2015), was 368 times more highly expressed in Env+Gag+ cells purified by LURE compared to controls. Chemokine CCL3, which is reported to have HIV-1 suppressive effects (Abdelwahab et al., 2003; Hudspeth et al., 2012), is expressed 795 times higher in Env+Gag+ cells purified by LURE compared to controls. Finally, a number of transcription factors were among the top 15 differentially expressed genes, including the top differentially expressed gene, PRDM1 (1365x). PRDM1 represses HIV-1 proviral transcription in memory CD4+ T cells by inhibition of HIV *tat* (Kaczmarek Michaels et al., 2015), and its overexpression is associated with lower levels of HIV-1 transcription in elite controllers (de Masson et al., 2014).

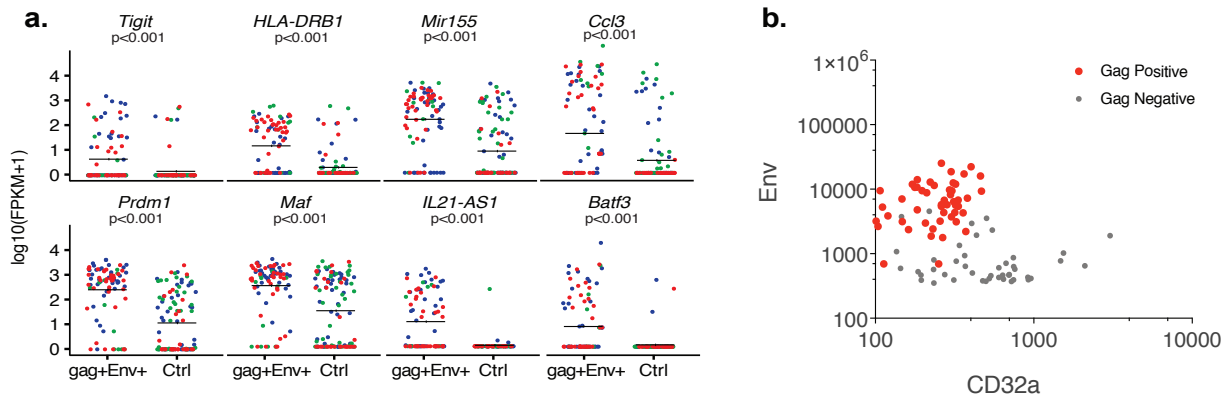


Figure 4.14 Selected gene expression in LURE cells compared to controls. a) Graphs show expression of selected genes in individual gag+Env+ LURE purified and control unfractionated cells as determined by MAST software in participants 603 (blue), 605 (green), B207 (red). **b)** Single cell index sorting was performed on Env+ enriched cells from participant B207 followed by RNA isolation, cDNA synthesis and gag qPCR. Gag+ cells were then examined for CD32a, Env and CD4 expression by index sorting. Shown is the mean fluorescence in gag+ and gag- cells from the same experiment.

To further examine the differences between LURE and control cells, we performed enrichment analysis using the Gene Ontology database with the 240 genes that overlapped between the DEG and PCA analyses. Among the top ten most significantly enriched biological processes, eight are related to immune system function, suggesting that PHA stimulated LURE and control cells differ in their expression of genes related to responses to pathogens. For example, LURE and control cells differ markedly in response to type I interferon and regulation of type I interferon production with control cells having higher expression of type I interferon responsive genes such as IFIT3, ISG20, IRF1, IFI6, RSAD2, STAT1, XAF1, CTNNB1 and UBE2L6. Consequently, the control cells also show a higher overall expression of genes involved in response to viruses such as CCL5, IFIT3,

ISG20, IRF1, SERINC5, IL2RA, RSAD2, DDIT4, STAT1, and PIM2. Thus, reactivated latent cells display a gene expression program that is consistent with their being less responsive to stimuli associated with viral infection than control cells obtained from the same cultures. Consistent with the altered gene expression program in reactivated latent cells, LURE and control cells show significant differences in the expression of genes that regulate transcription. For example, reactivated latent cells have higher levels of expression of transcriptional regulators PRDM1, MAF, IRF4, MTDH, IKZF3, and BATF3, whereas control cells have higher expression of PIM2, STAT1, HNRNPA2B, EZR, IRF1, CTNNB1 and NFKBIZ. We conclude that reactivated latent cells differ from control cells in a number of important ways many of which are related to the suppression of cellular anti-viral immunity.

CHAPTER 5:

DISCUSSION

The search for an HIV-1 cure has been ongoing for nearly 4 decades. Discovery and widespread use of antiretroviral therapy (ART) has dramatically slowed progression of AIDS in infected individuals (Arts and Hazuda, 2012), and is efficacious in preventing HIV transmission (Grant et al., 2010). However, existing therapies are not curative, due to the presence of latently infected cells which persist for the lifetime of infected individuals and are responsible for viral rebound if therapy is interrupted – requiring lifelong adherence to ART. Further consequences of nonadherence to antiretroviral drugs are the development of drug resistance mutations (Nachega et al., 2011) and possibly an increased latent reservoir size (Jain et al., 2013; Sarmati et al., 2015). Since the first description of the latent reservoir in 1997 (Chun et al., 1997a; Chun et al., 1997b), thousands of papers have been published attempting to describe, characterize, quantify and understand latently infected cells.

Latently infected cells are primarily CD4+ T cells, though other immune cell subsets and even non-immune cells in the brain have been described harboring HIV-1 DNA (Barton et al., 2016). In persons on suppressive ART, inducible latent cells in the blood number one per million CD4+ T cells. These cells are believed to harbor a quiescent provirus such that no viral RNAs, and therefore no viral proteins, are produced. Thus, the absence of any identifying surface marker, combined with their dramatic rarity has rendered isolation of these cells for study extremely difficult. Additionally, the

overabundance of proviruses with internal defects has confounded quantification of the replication competent reservoir.

In this thesis, I describe our efforts to characterize the latent reservoir in HIV-1 infected ART suppressed adults. We took three single cell approaches to describe latently infected cells. First, we used integration site sequencing to map large numbers of HIV-1 integration sites from single cells (Cohn et al., 2015). Upon the discovery that the majority of large expanded infected cell clones harbor defective virus, we performed a single-cell assay to characterize only replication competent viruses using the Quantitative and Qualitative Viral Outgrowth Assay (Q^2VOA) (Lorenzi et al., 2016). This method allowed the sequencing and phenotypic characterization of replication competent viruses arising from single cells. The discovery of identical replication competent sequences in multiple cells suggested that cells harboring replication competent virus may infrequently divide *in vivo*. The Q^2VOA characterizes of reservoir viruses, not the cells from which the virus originates as they are dead by the time virus is detectable by current methods. To characterize the latent cells themselves, we developed and isolation strategy which we call Latency Capture (LURE) (Cohn et al., In Press at the time of thesis printing). This strategy allowed us characterize the latent reservoir by simultaneously performing gene expression studies, TCR sequencing, and whole virus reconstruction from single recently reactivated latent cells.

Clonal expansion of infected cells *in vivo*

CD4+ T cells that are actively infected with HIV-1 are rapidly eliminated during anti-retroviral therapy, but this form of treatment is relatively ineffective in selecting against

latently infected CD4+ T cells, which have an estimated half-life of 44 months (Crooks et al., 2015; Finzi et al., 1999). Abolishing the latent reservoir is the current hurdle to finding a cure for HIV-1 infection. Although we have learned a great deal about the location of the latent compartment and its persistence during therapy, it has been difficult to uncover whether there are specific genomic features associated with latency (Siliciano and Greene, 2011). To further investigate the latent compartment, we used a high throughput method that uncovers sites of HIV-1 integration while enumerating clones of expanded T cells that bear identical integrations.

By comparing HIV-1 integration in controllers, untreated and treated progressors, including longitudinal samples obtained before and after therapy, we found that proliferating clones of infected cells accumulate over time. However, we were unable to detect intact, full-length viral sequences in these large, expanded clones. Instead, evidence from this study suggests that the reservoir resides primarily in cells that are selected against by ART in an integration specific manner, favoring the persistence of integrations in intergenic regions and silent genes, and with decay kinetics that argue against widespread cell division.

A number of different investigators have shown that HIV-1 prefers to integrate into the introns of highly expressed genes (Craigie and Bushman, 2012). This is true for all of the individuals in our study irrespective of their status as controllers or treatment with ART. Although the level of intrinsic viremic control has no detectable effect on integration site selection, therapy skews the integrated proviral population and selects against genic integrations. More specifically, therapy selects against integrations in highly expressed genes, when compared to untreated progressors or viremic controllers. Given that ART

seems to select for cells that bear silent proviruses, the results suggest that viruses integrated into genes are less likely to become latent than those found in intergenic regions. Moreover, the data indicate that among the proviruses integrated into genes, those that are found in genes expressed at low levels are also more likely to become latent. These findings are entirely consistent with *in vitro* experiments in cell lines showing that level of HIV-1 transcription is dependent in part on the status of surrounding chromatin (Jordan et al., 2003; Jordan et al., 2001; Sherrill-Mix et al., 2013).

HIV-1 integration has been studied in multiple cell types, but large libraries of integrations sites in primary infected T cells have only recently become available (Maldarelli et al., 2014; Wagner et al., 2014). Integration sites obtained from in vitro infected cell lines and primary T cells are distinct (Brady et al., 2009; Sherrill-Mix et al., 2013). Nevertheless, common features of HIV-1 integration have been defined including the observation that integration favors *Alu* repeats (Schroder et al., 2002). This association was thought to be dependent on the presence of these repeats in the introns of highly expressed genes (Schroder et al., 2002). However, we observed that integration preference into highly transcribed genes and into *Alu* repeats seem to be independently important.

The observation that HIV-1 prefers to integrate in the neighborhood of *Alu* repeats is consistent with the finding that different individuals have been reported to have multiple integrations in selected genes (Ikeda et al., 2007; Maldarelli et al., 2014; Schroder et al., 2002; Wagner et al., 2014). Our experiments define a group of overlapping hotspots for integration that share many of the features of all HIV-1 integrations including preference for introns of highly expressed genes and high density of *Alu* repeats. Viremic progressors

showed the highest levels of hotspot integration, possibly because persistent integration leads to over-representation of these favored sites, or because persistence of viruses integrated in non-hotspot sites are favored and become enriched over time on treatment. Alternatively, integration into hotspots might be positively selected by mechanisms that remain to be determined.

Individuals receiving ART show increasing numbers of cells with identical viral genomes by SGA analysis suggesting clonal expansion of a subset of cells bearing integrated proviruses (Buzon et al., 2014; Chomont et al., 2009; Wagner et al., 2013). Two independent groups have recently documented the long-term persistence of expanded clones of cells during therapy with ART (Maldarelli et al., 2014; Wagner et al., 2014). Our analysis confirms and extends these observations by showing that when considered as a group, expanded clones are less likely to occur when the provirus is in a genic region, and clones that are associated with genes tend to be in genes that are expressed at lower levels than single integrations. Thus, proviruses inserted into active regions of the genome, which would be more likely to support viral reactivation during T cell proliferation, are generally selected against during clonal expansion.

Why certain integration sites are permissive for clonal expansion is not known but finding that expanded clones with integrations occur in cancer related genes led to the suggestion that integration into genes that regulate cell division promotes proliferation (Wagner et al., 2014). While we also found a higher proportion of integrations in cancer-related genes as compared to random, this bias was not different from that observed for other highly expressed genes favored by HIV-1. Further, we do not see any differential bias for integration in cancer related genes in clonally expanded cells compared to single

integrations and an overall decrease in the number of integrations in cancer related genes during the course of therapy. Since the number and size of clones increases with time on therapy, the data indicate that integration into cancer genes is unlikely to be a general contributor to the proliferation of infected cells.

Our data show that ART positively selects for expanded clones and that viremic controllers resemble treated progressors in showing a higher proportion of expanded clones than untreated viremics. ART selects for clonal integrations irrespective of the location in the genome. This is in stark contrast to single, unique integrations, which are selected against by therapy. ART specifically favors the survival of single integrations in intergenic regions and is biased against genic regions with an overall half-life for single integrations of 127 months. The half-life of single integrations is not too dissimilar from the current estimate for the latent reservoir, which is believed to decay with a half-life of 44 months on ART (Finzi et al., 1999).

An important outstanding question after the discovery of clonally expanded cells with integrated HIV-1 is whether the virus from these large expanded clones of infected cells are the major contributor to the latent reservoir (Maldarelli et al., 2014; Wagner et al., 2014). Two different independent lines of evidence argue against this idea. First, whereas the reservoir appears to decay with time on ART, we find that clonally expanded integrations increase with time and do so irrespective of whether they are found in genes or intergenic regions. In contrast, single integrations in more active parts of the genome, which are more likely to support HIV-1 reactivation, are selected against with time on ART. Second, all 75 of the clonally expanded proviruses tested were defective, which is in agreement with the 2 examples in the literature (Imamichi et al., 2014; Josefsson et al.,

2013). Thus, we conclude intact virus is not enriched in clonally expanded infected cells. However, we cannot and do not rule out the possibility that a rare clone of cells contains an active virus. Nevertheless, the largest expanded clones in ART treated individuals are unlikely to be the major source of the rebounding latent reservoir. The data indicate that the majority of HIV-1 infected T cells that undergo clonal expansion are able to do so because their proviruses are defective and that the replication competent reservoir is more likely found in the subset of CD4+ T cells that remain relatively quiescent.

Identical replication competent viruses found in multiple cells

The key question arising from the above studies which I have addressed in this thesis is how clonal expansion of infected cells contributes to long-term persistence of the HIV-1 latent reservoir. Several additional mechanisms have been proposed to account for the persistence of this reservoir including low-level viral replication and the long half-life of latently infected cells.

Arguments against persistent low-level viral replication include the observation that further intensification of anti-retroviral therapy has no measurable effect (Dinoso et al., 2009; Vallejo et al., 2012). In addition, there is little evidence for viral evolution even after prolonged periods of antiretroviral therapy (Joos et al., 2008). In contrast, the idea that the reservoir is maintained, at least in part, by the proliferation of latently infected cells is supported by the observation of increasing proportions of identical proviral sequences in circulating CD4+ T cells (Wagner et al., 2013).

Archived, proviral integration site analysis, including the study above, provided further support for clonal expansion of infected T cells (Cohn et al., 2015; Maldarelli et al.,

2014; Wagner et al., 2014). In all cases examined, a large proportion of the archived integrated proviruses were found to belong to expanded clones of CD4+ T cells that share a unique integration site. However, as expected from the observation that the vast majority of integrated proviruses are defective (Bruner et al., 2016; Ho et al., 2013), many of the proviruses found in clonally expanded T cells are also defective (Cohn et al., 2015). The only exception was found in a singular individual suffering from metastatic squamous cell carcinoma that showed a large expanded clone of replication competent virus (Simonetti et al., 2016). However, the integration site of the provirus in this individual was ambiguous and could not be mapped with certainty. One clear caveat of the above experiments (Chapter 2) is our inability to purify cells harboring intact latent HIV-1 as opposed to cells containing defective viruses. Since the majority of integrated proviruses are defective (Ho et al., 2013), it is unsurprising that our experiments were unable to capture intact clones. Thus, the precise contribution of expanded cell clones bearing replication competent proviruses to maintaining the latent reservoir is not well defined.

To examine the diversity of the replication-competent latent reservoir we developed a modification of current quantitative viral outgrowth protocols to include a qualitative measure of the reservoir of replication-competent proviruses. Q^2VOA maximizes the yield of limiting dilution cultures to obtain viruses originating from a single cell. Consistent with the absence of viral evolution in individuals on suppressive ART, Q^2VOA revealed that the latent replication-competent viral population is similar between time points 4-6 months apart (Evering et al., 2012; Josefsson et al., 2013; Persaud et al., 2007). This is a relatively short interval in the life of an infected individual, and it will be interesting to evaluate the reservoir by Q^2VOA over longer time intervals and with various

treatment interventions such as antibody therapy or administration of latency reversal agents *in vivo*.

54% of the viruses emerging in Q^2VOA belonged to expanded cell clones bearing the same *env* sequences, and these viruses have a high probability of also being identical in the rest of their genomes (Laskey et al., 2016)). The isolation of identical sequences from multiple wells on different plates across 2 time-points suggests that these replicating viruses do not accumulate significant mutations during 14 days in culture. The lack of immunologic pressure under culture conditions may account for this finding.

While definitive proof that these identical viruses arise by cell division is lacking, the identical sequences found in different cells makes it a likely possibility. There are several non-mutually exclusive explanations for the prevalence of expanded CD4+ T cell clones carrying replication-competent viruses in the latent reservoir. Such cells could be clonally expanded as part of normal homeostatic processes, or in response to antigen, or as a result of HIV-1 integration into and disruption or activation of genes that regulate cell division (Maldarelli et al., 2014; Wagner et al., 2014). Alternatively, identical viral sequences could represent a burst of independent viral integration events into unrelated CD4+ T cells. Mapping the precise integration sites of the replication-competent proviruses or identifying the T cell receptor in cells harboring replication-competent proviruses is required to definitively distinguish between viral clones produced by CD4+ T cell clonal expansion or bursts of proviral integration.

Clonally expanded CD4+ T cells with shared proviral integration sites were detected in samples of 1-2×10^6 cells (Cohn et al., 2015; Maldarelli et al., 2014; Wagner et al., 2014), which is far fewer cells than the samples assayed in Q^2VOA. Only the largest

of these previously described expanded cell clones were amenable to further characterization, and all 75 proviruses that were examined molecularly showed gross defects that rendered them replication incompetent (Cohn et al., 2015). The finding that large proviral clones are typically defective is entirely consistent with the observation that proviral clone size is inversely related to the probability of reactivation. The larger the archival proviral clone, the less likely it was to be represented in the replication-competent latent reservoir. This observation is also in agreement with the finding that most clonally expanded proviruses integrated into CD4+ T cells are defective (Bruner et al., 2016; Cohn et al., 2015; Eriksson et al., 2013; Ho et al., 2013).

In conclusion, we find clones of replication-competent viruses in the latent reservoir. The viral sequences of these clones show limited overlap with archived proviral sequences in primary CD4+ T cells, suggesting that the two compartments may be under different types of selective pressure *in vivo*. Finally, if the reservoir is maintained by CD4+ T cell division, it may facilitate immunotherapy directed efforts for HIV-1 cure because T cell activation is thought to be associated with expression of HIV-1 antigens (Halper-Stromberg et al., 2014).

Single cell characterization of recently reactivated latent cells

Identification of clones of replication competent viruses in the latent reservoir raised the question whether these clones arise from division of infected cells or a homogenous viral burst with independent integration events. Additionally, though we and others made significant progress characterizing replication competent viruses, the field had yet to achieve latent cell isolation for phenotypic and functional characterization. We

set about to devise a strategy that would allow for the isolation and characterization of these rare CD4+ T cells from stably treated HIV-1 infected donors. We were guided by a method developed for the mouse to identify rare antigen binding cells (Pape et al., 2011), which allowed investigators to enrich low-frequency phycoerithrin (PE)-specific B cells from naïve mice. We adapted this strategy to capture rare reactivated *ex vivo* latent CD4+ T cells by combining *in vitro* latency reactivation with antibody-mediated enrichment of cells expressing surface viral Envelope in an assay we call Latency Capture (LURE). These cells are present in the average individual at low frequency, ranging from 0.1-10 cells per million peripheral CD4+ T cells. Whether this frequency is higher in tissues such as lymph nodes, is a current and ongoing focus of investigation by our group and others.

LURE allowed us to profile individual reactivated latent cells by single cell RNA sequencing. We compared the transcriptome of reactivated LURE cells to uninfected, activated cells from the same culture. Although the cells came from the same donors, were treated *in vitro* with the same activation stimulus and processed identically, unbiased clustering distinguished the majority of LURE cells from uninfected cells. Some of the differences in gene expression between the LURE group and control were large in magnitude – 28 genes had greater than 500-fold differential expression in LURE cells compared to the controls. We examined our gene list for enrichment using the Gene Ontology database and identified a number of enriched pathways of interest. LURE and control cells show significant differences in the expression of genes that regulate transcription. In the list containing the most significantly differently expressed genes, we found that expression of 13 transcription factors greatly differed between LURE cells and control cells. Unsurprisingly, differential expression of transcription factors resulted in

differential expression of many downstream gene pathways. Notably, of the top 10 significantly enriched pathways, eight were related to immune system function, suggesting that response to immune threats are differently regulated in LURE and control cells. Control cells express higher levels of interferon responsive genes and genes responsible for viral control, implying that they have a heightened response to viral infection compared to LURE cells. This difference could be due to either 1) inherent differences in LURE and control cells, such as the differential ability of specific T cell subsets to respond to pathogens, or 2) since the major certain difference between LURE and control cells is the presence of an actively transcribing provirus and consequently cytosolic HIV-1 RNA, perhaps this difference in gene expression is simply due to the cells' ability to sense proviral transcription. Additional experiments are ongoing to test the response of a latent cell to initial proviral transcription events. However, evidence of transcriptional changes upon reactivation from latency suggest that genes encoding proteasomes, histone deacetylases, and many transcription factors change in a time-dependent manner upon entry to the lytic viral lifecycle (Krishnan and Zeichner, 2004), suggesting that the cell likely senses early viral transcription after prolonged quiescence. Regardless of the mechanism, reactivated latent cells seem to harbor a gene expression program that is less responsive to stimuli associated with viral infection than control cells obtained from the same cultures. We conclude that the transcriptional profile in reactivated latent cells differs from control cells in a number of significant ways, many of which are implicated in suppression of immune function.

We characterized the transcriptome of 85 single reactivated latent cells from the blood of three ART suppressed donors with different sized reservoirs. As expected, our

ability to isolate latent cells depends, in part, on the number of cells sensitive to reactivation *in vitro*. Due to the relative resistance of some latent cells to reactivation (Bruner et al., 2016; Ho et al., 2013) LURE mirrors the viral outgrowth assay and is unable to capture the entirety of the latent reservoir. LURE purification of reactivated latent cells requires proviral activation to induce *Env* protein expression on the cell surface. Therefore, LURE captures a subset of latent cells with proviruses that can be reactivated in a single round of potent T cell stimulation (Bruner et al., 2016; Ho et al., 2013). Some reactivated latent cells are certainly lost during the multiple processing stages involved in the LURE protocol. Thus, the cells captured by LURE represent a fraction of the circulating latent reservoir that is closely related to and overlapping with the latent cells that emerge in traditional viral outgrowth assays. Our analysis is also limited to a single reactivation agent, PHA. PHA is a lectin that signals through the TCR to induce global T cell activation and simultaneously induces viruses out of latency (Laird et al., 2013). Many other reagents, used alone or in combination, can also be used to reactivate latent viruses (Laird et al., 2015), and will be important to test with LURE in future studies. Our analysis is limited to circulating CD4+ T cells that express Env proteins on the cell surface that are recognized by our antibody cocktail. We chose a combination of 3 broadly neutralizing antibodies, which together cover almost 90% of known circulating strains of HIV-1 envelope. However, for individuals whose virus is resistant to these antibodies, alternative cocktails of antibodies can be used to isolate cells harboring escaped viruses. Our analysis is limited to 3 individuals who were chosen based on sample availability and a range of reservoir sizes. Examination of additional individuals and methods of latent cell reactivation may reveal additional and or different genes and pathways involved in

maintaining latency. Finally, further experiments will be required to determine whether tissue resident latent cells have a similar gene program upon reactivation.

T cell division in response to antigen or mitogens like PHA and HIV-1 reactivation from latency are stimulated by shared metabolic and transcriptional pathways including NFκB (Siliciano and Greene, 2011). Once activated, productive HIV-1 infection typically leads to CD4+ T cell death by apoptosis or pyroptosis (Doitsh and Greene, 2016). However, cell death after latency reactivation *in vitro* appears to be stochastic with some cells being able to divide and survive after strong stimulation (Hosmane et al., 2017). Our finding that latent cells can survive upon cell division *in vivo* confirms *in vitro* experiments (Hosmane et al., 2017) and is also consistent with the observation that the latent compartment contains groups of CD4+ T cells that harbor proviruses with identical *Env* sequences (Hosmane et al., 2017; Lorenzi et al., 2016). Purification of reactivated latent cells by LURE and subsequent TCR sequencing provides definitive evidence that these cells arise by clonal expansion *in vivo*. The data is consistent with the idea that the protracted longevity of the latent compartment is due at least in part to cell division (Bui et al., 2017; Cohn et al., 2015; Hosmane et al., 2017; Lee et al., 2017; Lorenzi et al., 2016; Maldarelli et al., 2014; Mullins and Frenkel, 2017; Simonetti et al., 2016; Wagner et al., 2014). Finally, because the reservoir is stable over time (Crooks et al., 2015; Siliciano et al., 2003), the finding that latent cells divide implies that they are also dying at similar rate, and that the reservoir is a dynamic compartment.

Antibody binding to Env expressing cells *in vivo* leads to their accelerated clearance (Horwitz et al., 2017; Lu et al., 2016). Should latent cells undergoing clonal

expansion *in vivo* also express viral proteins, they too could be targeted for clearance by HIV-1 specific cytotoxic T cells, NK cells or by antibody dependent cellular cytotoxicity.

How does a subset of latent cells divide and still survive despite expression of HIV-1? Our single cell transcriptomic analysis of purified primary CD4+ T cells demonstrates that reactivated latent cells can express a distinct transcriptional program that includes muted responses to type I interferon and factors such as MiR-155 and PRDM1 that can suppress HIV-1 transcription (de Masson et al., 2014; Kaczmarek Michaels et al., 2015; Ruelas et al., 2015). We speculate that active HIV-1 suppression during CD4+ T cell division could be one of the mechanisms that maintains the latent reservoir. Further studies will be required to determine whether interference with these cellular safeguards could contribute to accelerating latent HIV-1 clearance.

Looking forward

Though we have made significant progress understanding the nature of the HIV-1 latent reservoir, there remain many fundamental unanswered questions relating to the persistence of HIV-1 *in vivo.* I am hopeful that many of these will be answered in the future and will have important impacts on therapeutic modalities. Looking forward, I want to address five major questions, outlined briefly below:

1. **Reactivation** – Advances in reservoir assays like Q^2VOA, TILDA, and LURE have increased our understanding of the reservoir, yet we rely on the use of known latency reversing agents to reactivate latent viruses hidden in CD4+ cells. Despite full T cell activation, some viruses do not reactivate (Hosmane et al., 2017). This is thought to be due to the stochastic nature of latency reversal, but that hypothesis

has yet to be proven. Indeed, we know very little about the way cells reactivate *in vitro* and even less about how they reactivate *in vivo.* To study reactivation, we need more reliable latency models and better assays to measure replication competent virus. Large, often unattainable, numbers of patient derived cells are required to study *ex vivo* latency and changes in the reservoir size are hard to measure with current methods. Since a large fraction of the latent reservoir is clonal in nature (Chapter 4), one insight into reactivation *in vivo* may lie in our ability to determine TCR specificity of the T cell clones harboring replication competent latent virus.

2. **TCR specificity** – naïve CD4+ T cells each have a unique receptor expressed on their surface which is responsible for interacting with specific peptide loaded MHC-class II. Upon recognition of cognate antigen presented by antigen presenting cells, naïve CD4+ T cells undergo robust clonal expansion to aid in the resolution of the immune response. We observe clones of CD4+ T cells harboring identical intact virus, indicating that upon infection, the latent cells have since divided. Since the latent reservoir exists in all individuals, has a long half-life, and is seemingly a dynamic compartment, uncovering the specificity of the T cell receptor in latent cells may help in our understanding of the maintenance of this reservoir of cells.
3. **Genes controlling latency** – While the latent compartment seems to be dynamic, it defies dogma because cell activation, division, and latency reactivation are controlled, in part, by the same transcription factors. Furthermore, it is thought that latency reactivation leads immediately to cell death due to the cytopathic effects of viral infection. Thus, investigation into genes responsible for maintenance of

latency during division will be fundamental to our understanding of persistence *in vivo.* Experiments performed on LURE cells have allowed the curation of a preliminary list of genes which could play a role in persistence, but further experiments to elucidate their function are required. Additionally, because LURE requires cell activation for purification, developing strategies to isolate and characterize gene expression in unstimulated latent cells directly *ex vivo* remains a top priority.

4. **Integration sites** – To date, it has been impossible to link integration site with replication-competent intact virus. Using LURE, there is great potential to discover integration sites that correspond to intact virus. Our data suggests that integrations in highly expressed genes will be unlikely candidates for long term persistence, but this important hypothesis remains to be definitely demonstrated.
5. **Contribution of tissue resident cells** – Finally, almost all experiments characterizing the latent reservoir in patients are performed on PBMCs from blood. However, T cells in the blood account for only a small fraction of the total CD4+ T cells in the body (Farber et al., 2014). Due to their non-invasive harvest compared with other tissues, they have been the primary focus of HIV-1 research. Tissue resident CD4+ T cells have become increasingly a focus for HIV-1 cure research due to their proximity to the site of infection and the presence of an inducible reservoir found in T follicular helper cells isolated from lymph nodes (Perreau et al., 2013). Most studies performed on tissue have been using the non-human primate model, but recently protocols to harvest human lymph tissue, gut biopsies and programs to harvest tissue *post mortem* are allowing for more robust

understanding of the contribution to persistence by tissue. I believe these studies are going to dramatically expand the way we think about HIV-1 persistence.

Collectively, the results described in this thesis provide evidence for the contribution of proliferating CD4+ T cells to the HIV-1 latent reservoir in ART suppressed patients. The data also suggests a mechanism for persistence through division whereby cells harboring intact latent virus express genes which may reduce HIV-1 transcription. These data illustrate the latent reservoir as dynamic target and therapeutic strategies addressing its dynamicity could have significant benefits in the future.

CHAPTER 6:

MATERIALS AND METHODS

Human sample collection

Human samples were collected after signed informed consent in accordance with Institutional Review Board (IRB)-reviewed protocols by all participating institutions. For the integration study, participants 1, 2, and 3 were selected from the Seattle HIV longitudinal cohort studies at Fred Hutchinson Cancer Research Center. Participants 4, 8 and 9 were recruited from the University of Cologne and samples were obtained at Rockefeller University. Participants 5, 6 and 7 were selected from the Rockefeller University HIV-1 antibody therapy clinical trial. Participants 10, 11, 12, and 13 were selected from a group of elite controllers that were followed at the Ragon Institute in Boston. For the Q^2VOA and LURE studies, all individuals were recruited by the Rockefeller Hospital and leukapheresis obtained and processed at Rockefeller.

Eligible participants were adults aged 18–65 years with HIV-1 infection and undetectable plasma HIV-1 RNA levels (<20 copies/ml) while on ART. PBMCs were isolated by Ficoll separation and frozen in aliquots. In all cases, HIV-1 infected participants on therapy were confirmed to be aviremic at the time of sample collection.

Integration Library

The method for integration library construction was adapted from TC-Seq (Klein et al., 2011).

CD4+ T cell isolation

CD4+ T cells were isolated from whole PBMC using anti-CD4 microbeads (Miltenyi Biotec). The percentage of live cells was determined by flow cytometry based on forward and side scatter. Purity of CD4+ T cells was determined by labeling isolated cells with anti-human CD3, CD4, CD8, CD19 and HLA-DR and gating on CD3, CD4 double positive cells. Isolated cells were used for library construction only if purity was >75%.

DNA preparation

0.2-2 million CD4+ T cells from HIV-1 infected individuals were lysed in Proteinase K buffer (100mM Tris [pH8], 0.2% SDS, 200mM NaCl, 5mM EDTA) and 5-10uL of 20mg/mL Proteinase K. Genomic DNA was extracted by phenol chloroform and fragmented by sonication (Covartis) to yield a 300-1000bp distribution of DNA fragments. DNA was blunted by End-It DNA Repair Kit (Epicenter), purified, then adenosine-tailed by Klenow fragment 3 → 5' exo- (NEB) and purified. Fragments were ligated to 200pmol of annealed linkers (Table S2). Virus sequences were eliminated by digestion with BglII (NEB) and fragments were purified.

Integration site amplification

Semi-nested ligation-mediated PCR was performed on linker-ligated DNA. Linkers made by annealing two oligos: GCAGCGGATAACAATTTCACACAGGACGTACTGTGG CGCGCCT and 5' phosphorylated, 3' dideoxycytosine GGCGCGCCACAGTACTTGAC

TGAGCTTTA. All PCRs were performed using the Phusion Polymerase system (Thermo). DNA was divided into 700ng aliquots and subjected to single-primer PCR with biotinylated LTR1 (5' bio: CTTAAGCCTCAATAAAGCTTGCCTTGAG) [1x(98C-1min) 12x (98C-15s, 62C-30s, 72C-30s) 1x(72C-5min)]. Each reaction was spiked with pLinker (GCAGCGGATAACAATTTCACACAGGAC) and subjected to additional cycles of PCR [1x(98C-1min) 25x(98C-15s, 62C-30s, 72C-30s) 1x(72C-5min)]. Products of 300-1000bp were isolated by agarose gel electrophoresis and magnetic streptavidin bead purification. Semi-nested PCR was performed on the magnetic beads first with a single primer LTR2 (AGACCCTTTTAGTCAGTGTGGAAAATC) [1x(98C-1min) 12x(98C-15s, 62C-30s, 72C-30s) 1x(72C-5min)] followed by spiking in pLinker and additional cycles [1x(98C-1min) 25x(98C-15s, 62C-30s, 72C-30s) 1x(72C-5min)] (Table S2). Products of 300-1000bp were isolated by agarose gel electrophoresis.

Paired-end library preparation

Linkers were digested by AscI such that a 6-nucleotide barcode (CGCGCC) was left on the DNA fragments, indicating linker-dependent amplification. Fragments were blunted by End-It DNA Repair Kit (Epicenter), purified with AmPure beads (Agencourt) and ligated to NextFlex paired-end adapters. Adaptor-ligated fragments were enriched by 35 cycles of PCR with NextFlex primers [1x(98C-1min) 35x(98C-15s, 66C-30s, 72C-30s) 1x(72C-5min)] and fragments between 300-1000bp were isolated by gel electrophoresis. Three libraries were mixed in equimolar ratios and sequenced by either 150bp paired-

end sequencing on Illumina MiSeq or 150bp or 100bp paired-end sequencing on an Illumina 2500.

Computational Analysis for Integration Site Identification

Read Alignment

Paired-end reads were mapped to the HIV-1 sequence (designated as a bait) using BLAT (Kent, 2002) with default settings. Reads that were mapped to the virus bait without mismatches, were checked for the linker barcode in the paired-end read, then barcode sequence was trimmed and the remainder was mapped to the human genome reference GRCh37/hg19) with *Bowtie* (Langmead et al., 2009). Only uniquely mapped reads (allowing for to 2 mismatches) were used as defined in the best alignment stratum (command line options: **-v2 -all -best -strata -m1**). Identical reads generated by PCR amplification were merged.

Integration determination

Once the paired-end reads (pair 1 and pair 2) were properly mapped in the bait and human genome reference respectively (see above), we determined the integration breakpoint by aligning the remaining nucleotide sequence that contained the 3' terminus of the HIV-1 LTR to the human genome using BLAT (default settings). Only uniquely mapped reads up to 1Kb away from its partner (pair 2) were kept. Adjacent (within 50 nucleotides) putative integrations sites were merged. Finally, the 5' end of pair 1 and pair 2 were used to deduce the integration and shear position sites in the human genome.

Hotspot detection

To detect preferred sites of HIV-1 integration genome-wide, we subjected our data set to *hot_scan* software analysis (Silva et al., 2014), which defines hotspots by scan statistics. Hotspots obtained by *hot_scan* were defined using different window widths (100, 200, 500, 1000, 2000, 5000, 10000, 20000, 50000 and 100000 bp).

Monte Carlo Simulation for virus integration and hotspots

To assess whether viral integration sites and hotspots are enriched around the INT-motif, we conducted a Monte Carlo simulation by shuffling the genomic locations of all virus integration sites or hotspot regions 10000 times using *bedtools shuffle* utility (Quinlan and Hall, 2010), and we counted the number of the randomized integrations which were 50bp or 1kb from a motif site. Then, we compared the observed number of motifs in hotspots with the median number of motifs in the randomized hotspots list. In both analysis, we assessed potential enrichments by *P-value* derived by counting the number of times that the number of observed events was equal or higher than the number of randomized events divided by N=10000.

Statistical analysis

All statistical analyses were done using R language.

Proportion test is the standard test for the difference between proportions, also known as a two-proportion z-test. We used R's implementation of this via the prop.test() function.

Integration library verification

To verify our integration sequencing strategy, we also constructed two libraries from DNA isolated from uninfected individuals. We recovered 13 sequences that mapped to integration sites. We subtracted these "integration sites" from all libraries before further analysis.

To test for the saturation of our method, two separate integration libraries were constructed from identical samples for three participants. We found that both libraries contained the same expanded clonal families, but the majority of single virus integrations were unique to each sample of cells used for library construction. Single viral integrations found in both libraries were less than 1% of observed viral integrations.

PCR verification

Genomic DNA isolated as described above was serially diluted and subjected to nested-PCR using genomic specific primers and primers LTR1 and LTR2 using HotStart Taq Polymerase (Qiagen) [1x(98C-14min) 40x(98C-30s, 55C-30s, 72C-30s) 1x(72C-5min)]. Products were isolated by gel electrophoresis and sequenced directly. Analysis of clones in this manner identified that integration sequencing underestimates the size of clones by 4-5 times (data not shown).

Virus sequencing

5'LTR

5' LTRs from large clones were amplified with nested genomic primers and LTR2Rev (Table S2) using Platinum High Fidelity Taq (Invitrogen) or HotStart Taq Polymerase (Qiagen) [1x(98C-14min) 40x(98C-30s, 55C-30s, 72C-1min) 1x(72C-5min)]. Products were isolated by gel electrophoresis and sequenced directly.

Full Length Virus

The virus sequencing method was adapted from Ho et al (Ho et al., 2013). Full length genomic DNA from infected individuals was isolated as described above and serially diluted into PCR tubes. Each well was filled to a final volume of 50μL with PCR reaction mixture (Platinum Taq MasterMix, Invitrogen) and primers to amplify virus from a specific integration site in the genome using touchdown cycling to increase specificity. Then, 2μL aliquots from the first PCR were subjected to nested genomic PCR and 1% gel electrophoresis. The positive wells were gel-purified and fragments were sequenced directly.

Q^2VOA

Viral outgrowth cultures were performed as previously described (Laird et al., 2016; van 't Wout et al., 2008). PBMCs from virologically suppressed chronically HIV-infected individuals on ART were obtained by leukapheresis. Briefly, leukapheresis products from each study participant at each time point were processed and PBMCs were isolated by density centrifugation on Ficoll (Thermo Scientific). Cryopreserved PBMCs

were then used to isolate total CD4+ T lymphocytes using negative selection by magnetic beads (Miltenyi). Purified CD4+ T lymphocytes were cultured at $0.3x10^6$ cells per well in 24-well plates in 1 ml of RPMI 1640 (RPMI; Gibco) supplemented with 10% fetal bovine serum (FBS; HyClone, Thermo Scientific), 1% Penicillin/Streptomycin (Gibco), 1 μg/ml Phytohemagglutinin (PHA; Life Technologies), and 100 U/ml of IL-2 (Peprotech) at 37°C and 5% CO2. Each well also received 1ml of medium containing $2.5x10^6$ irradiated heterologous PBMCs. After 24 hours 1.5ml of medium was discarded and 10^6 $CD8^+$-depleted lymphoblasts from an HIV-negative donor were added to each well as target cells. At day 9, 1 ml of medium was removed and another 10^6 $CD8^+$ depleted CD4+ T lymphoblasts were added to each well. At day 14 the supernatant of each well was tested for HIV-1 production using the Lenti-X p24 Rapid Titer Kit (Clontech) according to the manufacturer's instructions. Under these conditions, using $0.3x10^6$ donor cells, less than 30% of all cultures were positive for all individuals.

Bulk Cultures

Bulk cultures were performed as previously described (Caskey et al., 2015). Sequence analysis on bulk culture was also performed as previously described (Scheid et al., 2016).

Sequence analysis for Q^{2}VOA

The supernatant from p24-positive cultures was extracted using Qiagen MinElute Virus Spin kit QIAcube. cDNA was produced using 1:10 diluted RNA with the SuperScript III reverse transcriptase (Invitrogen Life Technologies) and the antisense primer env3out 5'–TTGCTACTTGTGATTGCTCCATGT–3' followed by RNase H digestion (Invitrogen Life Technologies) for 20 minutes at 37 °C. The full gp160 *env* was amplified from 1:40 diluted cDNA using envB5out 5'–TAGAGCCCTGGAAGCATCCAGGAAG–3' and envB3out 5'–TTGCTACTTGTGATTGCTCCATGT–3' in the first round and second round nested primers envB5in 5'–TTAGGCATCTCCTATGGCAGGAAGAAG–3' and envB3in 5'– GTCTCGAGATACTGCTCCCACCC-3'.

PCRs were performed using a High Fidelity Platinum Taq (Invitrogen) at 94°C, 2 min; (94°C, 15 sec; 55°C 30 sec; 68°C, 4 min) × 35; 68°C, 15 min. Second round nested PCR was performed using 1 μl of PCR1 product as template and High Fidelity Platinum Taq at 94°C, 2 min; (94°C, 15 sec; 57°C 30 sec; 68°C, 4 min) × 45; 68°C, 15 min. PCR2 products were checked using 1% 96-well E-Gels (Invitrogen). PCR bands with expected HIV envelope size were quantified and subjected to library preparation using the Illumina Nextera DNA Sample Preparation Kit (Illumina) as described (Kryazhimskiy et al., 2014). Briefly, 10 ng of DNA per band were subjected to tagmentation, ligated to barcoded sequencing adapters using the Illumina Nextera Index Kit and then purified using AmPure Beads XP (Agencourt). 96 different purified samples were pooled into one library and then subjected to paired-end sequencing using Illumina MiSeq Nano 300 (Illumina) cycle kits at a concentration of 15 pM.

Sequence adapters were removed using Cutadapt v1.8.3. Read assembly for each

virus was performed in three steps. First, de novo assembly was performed using Spades v3.6.1 to yield long contig files. Contigs longer than 255bp were subsequently aligned to an HIV envelope reference sequence and a consensus sequence was generated using Geneious 8. Finally, reads were re-aligned to the consensus sequence to close gaps and a final read consensus was generated for each sequence. Sequences with double peaks (cutoff consensus identity for any residue <75%), stop codons, or shorter than the expected envelope size were omitted from downstream analyses.

Proviral Single Genome Amplification

DNA from 1-10×10^6 CD4+ cells from HIV-1-infected individuals was prepared as previously described (Klein et al., 2011). Briefly, cells were collected after magnetic isolation and lysed in Proteinase K buffer (100 mM Tris [pH 8], 0.2% SDS, 200 mM NaCl, 5 mM EDTA) and 20mg/ml of Proteinase K at 56°C for 12 hours. Genomic DNA was extracted by phenol chloroform precipitation. Aliquots of the resulting DNA were diluted and used as template for full-length gp160 PCRs. All PCRs were performed using a High Fidelity Platinum Taq (Invitrogen). The first PCR at 94°C, 2 min; (94°C, 15 sec; 58.5°C 30 sec; 68°C, 4 min) × 35; 68°C, 15 min. Second round nested PCR was then performed using 1 μl of PCR1 product as template and High Fidelity Platinum Taq at 94°C, 2 min; (94°C, 15 sec; 61°C 30 sec; 68°C, 4 min) × 45; 68°C, 15 min. PCR2 products were checked using 1% 96-well E-Gels (Invitrogen). Bands with expected size of the HIV-1 envelope obtained from diluted DNA samples that showed an amplification efficiency of less than 30% were subjected to library preparation and sequencing using the Illumina

platform as described above. Hypermutants, sequences with double peaks (cutoff consensus identity for any residue <75%), stop codons, or shorter than the expected envelope size were omitted from downstream analyses.

Genealogical Sorting Index

The Genealogical Sorting Index (GSI) (Cummings et al., 2008) was used to calculate the degree of phylogenetic association of the *env* gene sequences. Phylogenies were inferred using PhyML version 3.0 (Guindon et al., 2010). Multiple bifurcations with intervening zero-length branches were collapsed to polytomies using the di2multi method implemented in the ape package for R (Paradis et al., 2004). These phylogram topologies were used to calculate the gsi values and p-values were determined using 10,000 replicate permutations (Cummings et al., 2008).

Computational Analyses

Bayesian inference was implemented in Stan (Carpenter et al., 2016) to jointly estimate the frequency of integrated proviruses and their reactivation probability. Data from proviral DNA sequencing and viral outgrowth assays were combined across multiple visits, incorporating a prior probability for the gradual decay of the reservoir (Crooks et al., 2015; Siliciano et al., 2003). We assumed a weakly informative prior for the frequency of integrated provirus and a uniform prior for the probability of reactivation. Parameters were estimated individually for viruses with identical sequences observed at least three times across both visits (large clones), while sequences observed less frequently were

grouped together (small clones). The observed relationship between clone frequency and reactivation probability also holds in simpler models incorporating data from single visits only, and when clones that were not observed in both assays during the same visit are excluded.

Latency Capture

Cells were cultured at $2x10^6$/mL in R10 (RPMI supplemented with 10% heat inactivated FCS, 10mM HEPES, 100U/mL PenStrep), and 25% volume conditioned media. Conditioned media was made by culturing healthy PBMCs in R10 with PHA and IL-2 for 2 days, followed by a wash and 5 days in culture with IL-2 alone. The conditioned media was then collected and frozen at -80C until use. 100U/mL IL-2 (Peprotech), 1ug/mL PHA (Sigma), 10uM Z-VAD-FKM (R&D), 10uM Ritonavir, 10uM Dolutegravir, 10uM Emtricitabine, 5uM Tenofovir, and 10uM Maraviroc (all Selleckchem) were added to the media. 36h later, cells were labeled with 5ug/mL each of biotinylated 3BNC117, 10-1074, PG16, followed by Streptavidin PE (1:500, BD) and anti-PE magnetic beads (Miltenyi Biotech). Cells were then passed over a magnetic column and bound cells were eluted for downstream analysis. For FACS sorting, cells were labeled with the following antibodies, all Biolegend: CD1c (cat. no. 331510), CD3 (cat. no. 300430), CD4 (cat. no. 317444), CD8 (cat. no. 344726), CD14 (cat. no. 301812), CD20 (cat. no. 302318), CD32a (cat. no. 303204), and CD56 (cat. no. 318314).

Gag bulk qPCR

RNA was extracted from equivalent numbers of cells irrespective of enrichment. *Gag* qPCR was performed using RNA-to-CT one-step RT-PCR mix (ThermoFisher) and previously described primers (Palmer et al., 2003).

Single Cell sorting

All sorts were performed on BD FACS Aria into 96-well plates containing guanidine thiocyanate buffer (Qiagen) supplemented with 1% β-mercaptoethanol. Plates were immediately frozen on dry ice and transferred to long-term storage at -80C. LURE cells were gated on live, CD1c, CD8, CD14, CD20, and CD56 negative, CD3 positive and sorted based on Env staining. Control cells were gated on live, CD1c, CD8, CD14, CD20, and CD56 negative and sorted CD3 positive cells.

Single Cell gag qPCR and ENV PCR

Nucleic acids were isolated by SPRI bead cleanup as described (Tas et al., 2016). RNA was reverse-transcribed into cDNA using an oligo(dT) primer. *Gag* qPCR was performed on one-fifth of the cDNA (Palmer et al., 2003). Gag+Env+ cells were selected based on the presence of cell-associated gag RNA measured by qPCR. Control cells were assayed for gag RNA and none was detected. Nested Env PCR was performed on one-fifth of the cDNA as described above.

YU2 infection and sorting

CD4+ T cells were activated and infected with YU2 and labeled as previously described (Lu et al., 2016). CD4lo, Envelope positive cells were sorted.

Single Cell RNASeq

RNASeq libraries were constructed based on Trombetta et al. (Trombetta et al., 2014) using primers from Islam et al. (Islam et al., 2014) Briefly, RNA was converted to full-length cDNA using oligo(dT) priming (Bio-AATGATACGGCGACC ACCGATCGTTTTTTTTTTTTTTTTTTTTTTTTTTTTTT) and SMART template switching technology (all RNA oligo: Bio-AAUGAUACGGCGACCACCGAUNNNNNGGG) followed by 24 cycles of PCR preamplification of cDNA (primer Bio-GAATGATACGGCGACCACCGAT). We used the amplified cDNA to construct standard Illumina sequencing libraries Nextera XT library preparation kit. Samples were sequenced by Illumina NextSeq.

RNASeq Analysis

The quality of the RNASeq libraries was evaluated using the *f*astQC1 (Anders et al., 2015). We used STAR (2.4.1d) (Dobin et al., 2013) aligner to map the raw paired-end reads to the reference genome GRCh37/hg19. The gene-level counts were obtained using *H*TSEQ (Anders et al., 2015). We performed a saturation analysis to detect the number of detected genes and filtered out the outlier cells as in Gaublomme et al (Gaublomme et al., 2015). Briefly, we excluded cells with number of aligned reads

<25,000 and percentage of identified genes <20% of the group maximum. Normalized expression values were calculated using the *s*cran package (Lun et al., 2016). Heatmaps and dotplots were generated in R. The gene counts were used to infer the differentially expressed genes (DEG) in the data by MAST (v1.2.1) (Finak et al., 2015).

HIV Splice variant analysis

We recovered the reads which failed to map to the human genome and mapped these reads to annotated junctions between HIV splice donors and acceptors to reconstruct the splice variants present in the scRNASeq data.

HIV reads alignment and reconstruction

We carried out HIV assembly analysis on the all reads which failed to map to the human genome by the IVA de novo assembler (v1.0.7) (Hunt et al., 2015).

TCR identification

TraceR (Stubbington et al., 2016) was used to reconstruct full-length, paired T cell receptor (TCR) sequences.

TCR sequences unable to be recovered from RNASeq reads were amplified as previously described (Han et al., 2014).

PCA Seurat

We used the Seurat package (v1.4.0.16) to identify variable genes, principal components (PCs), clusters and gene markers as described (Satija et al., 2015). Briefly, the software identifies highly variably expressed genes using a normalized z-score, performs linear dimensional reduction (PCA) on the filtered genes, obtains additional transcriptome PCA loading genes using projection of these principal components to the entire dataset, determines groups by density clustering of the t-SNE significant principal component scores and performs gene marker discovery. We also used the Improved Stochastic Ranking Evolution Strategy algorithm (Runarsson and Yao, 2005) implemented by *N*Lopt, to find the optimal set of PCs and parameters, and to find the optimal set of clusters that best correlate with each group of cells.

Single Cell Consensus Clustering

Single-Cell Consensus Clustering (SC3) tool (Kiselev et al., 2017) (default settings) was used for unsupervised clustering of single cells in this study. SC3 consistently integrates different clustering solutions through a consensus approach and identifies marker genes which are highly expressed in only one of the clusters and are able to distinguish it from all the remaining ones.

We have tested combinations of clustering settings (**k**=2, 3 and 4) and used a quantitative measure of the diagonality of the consensus matrix to select the **k** in which the measure is closest to 1 (k=3). We then used SC3 (AUROC>0.6 and FDR < 0.1) to identify marker genes which are highly expressed in only one of the clusters and are able to distinguish it from all the remaining clusters.

Data availability

The data reported in this thesis is archived at the following databases:

Data from Chapter 2 are accessible via NCBI SRA using the accession number SRP045822.

Data from Chapter 3 have been deposited in the GenBank database using accession numbers KY113379–KY114054.

Data from Chapter 4 are available via multiple databases: Single cell RNASeq data is available at NCBI GEO (GSM2801437); Envelope sequences are available in the Genebank database (MG196359 - MG196639); TCR sequences are available in the Genebank database (MG192535-MG193127).

APPENDIX

240 genes which overlap between principal components analysis and differential expression analysis (LURE vs control). Positive fold change indicates genes highly expressed in LURE cells. Negative fold change indicates genes highly expressed in control cells.

Gene Name	p-value	Fold Change
PRDM1	1.28E-06	1365.04
MAF	0.000931106	1222.00
CCL3	3.31E-05	795.63
GPR171	3.09E-10	741.65
ZBED2	0.000178928	729.34
SRGN	8.80E-07	660.79
IRF4	1.35E-10	626.10
RGS1	0.001195892	614.30
LINC-PINT_dup2	7.66E-05	485.05
MIR155HG	1.32E-06	368.48
MTDH	1.45E-05	350.12
IL21-AS1	4.19E-14	337.92
ATP1B3	0.000131055	298.96
PSMA7	0.006644358	290.93
RDH10	0.000686327	282.16
CD96	0.001107027	279.72
TNFAIP8	8.53E-07	270.69
STX11	0.000925105	267.24
RAB27A	1.04E-05	266.35
RNR1	0.001522231	258.26
IQCG	5.85E-05	257.99
STOM	2.16E-06	256.51
MT2A	0.000596151	254.52
EXOC2	0.000199389	254.23
VMP1	1.34E-05	249.53
AKAP13	1.27E-08	237.65
PIK3CG	1.59E-09	190.23
SAE1	0.001266734	189.14
GSTP1	4.41E-06	187.08
MCOLN2	5.23E-08	185.02
OAZ1	0.004028392	183.23
IKZF3	1.09E-06	179.89
TMEM173	0.001445032	170.19

PSMB3	6.11E-05	166.34
PSMD12	0.000531619	162.36
BATF3	3.28E-09	158.94
NDUFAB1	0.000522233	156.25
CCT3	2.98E-05	149.31
CTSB	6.46E-05	148.74
GPR183	0.000444313	145.95
NDUFS5	1.65E-06	142.89
PEBP1	0.009299749	142.84
HUWE1	0.000364271	139.87
TIGIT	1.44E-06	137.58
BST2	0.0070902	134.12
INPP4B	3.81E-10	129.73
ST8SIA4	0.000119258	129.15
CD53	0.001193214	126.33
SMARCA2	9.33E-07	125.78
ARAP2	0.008208236	123.72
IL12RB2	1.40E-06	108.10
NDUFA8	0.000864071	104.91
GTF2A2	0.001514079	103.66
CRIM1	1.45E-07	103.03
TMBIM6	0.000683334	101.61
C4orf3	0.000352459	99.83
PPP3CA	0.001166999	98.07
RNASEH2C	0.004424494	97.84
HLA-DQB1	7.66E-09	96.91
SUSD6	0.009646603	96.59
TRPS1	7.90E-08	96.56
NIT2	0.000808645	93.11
CREM	6.46E-05	93.04
COPS5	0.002519311	92.93
SDF4	0.008562895	92.31
ACTA2	1.56E-09	91.12
RAPGEF2	1.17E-05	90.23
ARID5B	4.50E-05	90.16
TPM4	0.003295339	88.84
CYTOR	0.006228062	87.10
ZEB2	9.95E-05	86.50
MT1X	0.001493762	80.13
NDUFA11	0.00711934	78.81
RANBP2	0.002914352	78.40
DNTTIP2	0.002506043	78.13
FABP5	2.40E-06	77.89
FBXW5	0.002510012	77.80

COX5B	1.16E-05	76.66
HLA-DRB1	6.53E-09	74.58
RORA	3.69E-05	74.14
ZBTB38	2.53E-05	73.86
RNF19A	0.001374051	70.37
PRDX6	0.000426547	64.34
C17orf62	0.00039074	60.27
NME1	0.00251826	60.08
SEC63	0.009747048	57.31
TNFRSF9	0.000166316	56.89
MSC	5.66E-07	55.70
ADRM1	0.00104375	54.83
HLA-DPA1	1.09E-06	53.40
KRT10	0.001280249	50.73
NAA38	0.004901974	50.16
CD226	8.53E-07	49.78
S100A10	5.81E-09	48.84
SRI	0.000196042	47.48
DGKH	7.22E-05	41.28
IPCEF1	0.000188963	40.85
NDUFB6	0.002733726	40.01
TOX	7.04E-11	39.94
C12orf57	0.00440838	39.51
MIR4435-2HG	0.000257797	38.76
SFT2D1	0.000142357	33.76
NAMPT	6.13E-08	32.98
PSMD4	0.002610897	32.64
PSMD11	4.25E-06	32.52
UBE2D1	0.00481846	32.44
COA6	9.25E-05	30.95
APOL1	0.001017343	30.69
RBPJ	5.18E-05	27.76
UBAC2	0.002421166	26.68
STRIP2	2.87E-07	25.48
BTF3	0.000129221	22.04
SMIM15	0.00130664	18.33
AHR	2.76E-07	17.54
SEC61G	0.008242902	16.00
MLEC	0.002695913	15.38
CASK	0.000342703	12.29
EVI5	0.00279202	11.74
ELL2	0.000343087	10.73
TMEM70	7.40E-07	10.36
GTF3C6	0.002807797	9.43

PSMD7	0.007641893	9.10
OSTF1	0.000316277	6.58
OIP5	0.001004668	4.76
PCMT1	0.000785253	1.78
MIR6723	0.001548566	1.71
F2R	7.07E-06	1.04
YWHAQ	0.008488231	-0.68
CHMP1B	0.000680176	-0.71
HIPK2	0.004757124	-1.63
YWHAB	0.004221649	-2.05
TBK1	0.000146177	-7.71
CD2BP2	0.004625936	-7.80
PRKCQ	0.0022731	-7.91
TIMM17A	0.007815617	-11.81
SH2D1A	0.000329061	-13.34
CASP4	0.001320572	-13.40
EED	0.000371361	-15.30
TNFRSF18	0.001768085	-16.19
TRIB2	0.00462457	-16.20
DCXR	0.004391609	-22.11
NDRG3	0.008790155	-24.33
GALM	5.04E-06	-25.53
LIMS1	1.88E-06	-26.21
NDUFA6	0.008180147	-26.40
CTLA4	0.000619876	-26.71
RDX	9.90E-05	-29.57
PLAC8	1.65E-07	-30.64
FAM162A	0.00079396	-30.89
ETV6	0.00658202	-36.00
CST7	0.008462729	-36.16
RASA2	0.003317739	-36.22
NDUFA4	0.000102627	-36.91
CDK2AP2	0.002775209	-37.34
MMD	0.00375033	-37.66
PRKCA	0.008697013	-41.96
MAGOH	0.005135903	-42.15
PDIA6	0.000218341	-42.94
LPAR6	0.009600017	-43.81
SOCS3	0.001998066	-50.06
SLA	0.006174637	-50.88
LAG3	0.009802407	-53.23
POMP	0.008302933	-54.65
MAL	0.000442452	-56.70
CCL5	5.19E-07	-56.82

BBX	0.008655036	-59.35
COX5A	0.002052359	-61.34
DNAJB11	0.002703867	-61.95
ARPP19	0.000129775	-62.31
STYX	0.002902084	-63.07
ITGB1BP1	0.009295115	-65.51
AKR1B1	0.000496086	-66.57
GHITM	0.007645077	-78.61
PSMD1	0.004990284	-78.67
HERPUD2	0.00202364	-81.67
S100A4	0.00093742	-83.69
PSMC1	0.003123379	-87.12
MAP3K8	0.002806797	-90.45
CNN2	0.000413875	-90.68
CASP10	0.000124689	-94.16
GBP4	0.003172026	-96.15
UQCRC2	0.009705553	-97.62
RABAC1	0.003594757	-99.15
ARL6IP4	0.00327542	-112.93
G3BP2	0.00345263	-117.79
EMP3	0.001016469	-119.19
RAP1A	0.005492263	-122.49
GIMAP7	0.000167305	-124.57
IFIT3	0.000847454	-127.10
EIF3M	0.002866585	-134.66
NFKBIZ	0.008591639	-135.56
CTNNB1	0.002914301	-139.35
SEMA4D	0.003775067	-141.50
BCCIP	0.003608674	-142.45
TMED10	0.000716056	-144.23
HSPA9	0.006574652	-148.39
VCP	5.15E-05	-155.23
TALDO1	0.001231597	-157.09
UBE2L6	0.00729835	-159.37
ARGLU1	0.001840139	-161.74
SCD	0.004039508	-163.27
SYNE2	0.001555763	-166.54
ISG20	0.000162219	-176.05
SKIL	3.61E-06	-184.26
PSMB6	0.004321897	-184.26
PIK3CD	0.003799058	-191.10
IRF1	0.001345965	-208.58
SH3KBP1	0.007169478	-211.86
S100A11	0.003714074	-219.94

NOP58	0.000654669	-226.08
IFI6	4.76E-06	-263.15
TPI1	0.003160064	-263.92
PSMA6	0.00232043	-269.09
SERINC5	0.000246139	-278.29
LGALS1	0.000160923	-280.41
IL2RA	0.003065813	-295.42
EZR	0.00394443	-330.43
RSAD2	0.000159317	-338.85
DDIT4	0.001158105	-339.04
RPS9	0.00073819	-344.22
IFI44L	0.000330539	-347.29
HNRNPA2B1	0.009409872	-373.48
STAT1	0.006105408	-413.06
CCND2	0.002017967	-413.66
XAF1	5.09E-06	-415.07
UBE2B	0.000596287	-456.00
ARPC1B	3.22E-05	-485.98
GBP5	1.08E-06	-488.05
CORO1A	0.000552011	-490.77
PGK1	0.005921922	-564.96
PIM2	5.88E-14	-580.10
PARP14	0.006204162	-671.70
IL32	0.000683834	-709.68
LTB	1.26E-11	-732.60
PARP9	4.37E-07	-794.81
YWHAZ	0.008959673	-885.59
DTX3L	0.0053721	-891.16
APOL2	1.07E-06	-1104.25
PPP1CB	0.004961738	-1154.06
RPL3	0.000396375	-2101.35

REFERENCES

Abdelwahab, S.F., F. Cocchi, K.C. Bagley, R. Kamin-Lewis, R.C. Gallo, A. DeVico, and G.K. Lewis. 2003. HIV-1-suppressive factors are secreted by CD4+ T cells during primary immune responses. *Proceedings of the National Academy of Sciences of the United States of America* 100:15006-15010.

Abram, M.E., A.L. Ferris, W. Shao, W.G. Alvord, and S.H. Hughes. 2010. Nature, position, and frequency of mutations made in a single cycle of HIV-1 replication. *Journal of virology* 84:9864-9878.

Anders, S., P.T. Pyl, and W. Huber. 2015. HTSeq--a Python framework to work with high-throughput sequencing data. *Bioinformatics* 31:166-169.

Anderson, J.A., N.M. Archin, W. Ince, D. Parker, A. Wiegand, J.M. Coffin, J. Kuruc, J. Eron, R. Swanstrom, and D.M. Margolis. 2011. Clonal sequences recovered from plasma from patients with residual HIV-1 viremia and on intensified antiretroviral therapy are identical to replicating viral RNAs recovered from circulating resting CD4+ T cells. *Journal of virology* 85:5220-5223.

Archin, N.M., N.K. Vaidya, J.D. Kuruc, A.L. Liberty, A. Wiegand, M.F. Kearney, M.S. Cohen, J.M. Coffin, R.J. Bosch, C.L. Gay, J.J. Eron, D.M. Margolis, and A.S. Perelson. 2012. Immediate antiviral therapy appears to restrict resting CD4+ cell HIV-1 infection without accelerating the decay of latent infection. *Proceedings of the National Academy of Sciences of the United States of America* 109:9523-9528.

Arts, E.J., and D.J. Hazuda. 2012. HIV-1 antiretroviral drug therapy. *Cold Spring Harbor perspectives in medicine* 2:a007161.

Bailey, J.R., A.R. Sedaghat, T. Kieffer, T. Brennan, P.K. Lee, M. Wind-Rotolo, C.M. Haggerty, A.R. Kamireddi, Y. Liu, J. Lee, D. Persaud, J.E. Gallant, J. Cofrancesco, Jr., T.C. Quinn, C.O. Wilke, S.C. Ray, J.D. Siliciano, R.E. Nettles, and R.F. Siliciano. 2006. Residual human immunodeficiency virus type 1 viremia in some patients on antiretroviral therapy is dominated by a small number of invariant clones rarely found in circulating CD4+ T cells. *Journal of virology* 80:6441-6457.

Barton, K., A. Winckelmann, and S. Palmer. 2016. HIV-1 Reservoirs During Suppressive Therapy. *Trends in microbiology* 24:345-355.

Baxter, A.E., J. Niessl, R. Fromentin, J. Richard, F. Porichis, R. Charlebois, M. Massanella, N. Brassard, N. Alsahafi, G.G. Delgado, J.P. Routy, B.D. Walker, A. Finzi, N. Chomont, and D.E. Kaufmann. 2016. Single-Cell Characterization of Viral Translation-Competent Reservoirs in HIV-Infected Individuals. *Cell Host Microbe* 20:368-380.

Berger, E.A., R.W. Doms, E.M. Fenyo, B.T. Korber, D.R. Littman, J.P. Moore, Q.J. Sattentau, H. Schuitemaker, J. Sodroski, and R.A. Weiss. 1998. A new classification for HIV-1. *Nature* 391:240.

Berry, C.C., N.A. Gillet, A. Melamed, N. Gormley, C.R. Bangham, and F.D. Bushman. 2012. Estimating abundances of retroviral insertion sites from DNA fragment length data. *Bioinformatics* 28:755-762.

Binley, J.M., H.J. Ditzel, C.F. Barbas, 3rd, N. Sullivan, J. Sodroski, P.W. Parren, and D.R. Burton. 1996. Human antibody responses to HIV type 1 glycoprotein 41 cloned in phage display libraries suggest three major epitopes are recognized and give evidence for conserved antibody motifs in antigen binding. *AIDS Res Hum Retroviruses* 12:911-924.

Brady, T., L.M. Agosto, N. Malani, C.C. Berry, U. O'Doherty, and F. Bushman. 2009. HIV integration site distributions in resting and activated CD4+ T cells infected in culture. *Aids* 23:1461-1471.

Brenchley, J.M., T.W. Schacker, L.E. Ruff, D.A. Price, J.H. Taylor, G.J. Beilman, P.L. Nguyen, A. Khoruts, M. Larson, A.T. Haase, and D.C. Douek. 2004. CD4+ T cell depletion during all stages of HIV disease occurs predominantly in the gastrointestinal tract. *J Exp Med* 200:749-759.

Bruner, K.M., A.J. Murray, R.A. Pollack, M.G. Soliman, S.B. Laskey, A.A. Capoferri, J. Lai, M.C. Strain, S.M. Lada, R. Hoh, Y.C. Ho, D.D. Richman, S.G. Deeks, J.D. Siliciano, and R.F. Siliciano. 2016. Defective proviruses rapidly accumulate during acute HIV-1 infection. *Nature medicine* 22:1043-1049.

Bui, J.K., M.D. Sobolewski, B.F. Keele, J. Spindler, A. Musick, A. Wiegand, B.T. Luke, W. Shao, S.H. Hughes, J.M. Coffin, M.F. Kearney, and J.W. Mellors. 2017. Proviruses with identical sequences comprise a large fraction of the replication-competent HIV reservoir. *PLoS pathogens* 13:e1006283.

Buzon, M.J., H. Sun, C. Li, A. Shaw, K. Seiss, Z. Ouyang, E. Martin-Gayo, J. Leng, T.J. Henrich, J.Z. Li, F. Pereyra, R. Zurakowski, B.D. Walker, E.S. Rosenberg, X.G. Yu, and M. Lichterfeld. 2014. HIV-1 persistence in CD4+ T cells with stem cell-like properties. *Nature medicine* 20:139-142.

Carpenter, B., A. Gelman, M. Hoffman, D. Lee, B. Goodrich, M. Betancourt, M.A. Brubaker, J. Guo, P. Li, and A. Riddell. 2016. Stan: A probabilistic programming language. *J Stat Softw* In press:

Caskey, M., F. Klein, J.C. Lorenzi, M.S. Seaman, A.P. West, Jr., N. Buckley, G. Kremer, L. Nogueira, M. Braunschweig, J.F. Scheid, J.A. Horwitz, I. Shimeliovich, S. Ben-Avraham, M. Witmer-Pack, M. Platten, C. Lehmann, L.A. Burke, T. Hawthorne, R.J. Gorelick, B.D. Walker, T. Keler, R.M. Gulick, G. Fatkenheuer, S.J.

Schlesinger, and M.C. Nussenzweig. 2015. Viraemia suppressed in HIV-1-infected humans by broadly neutralizing antibody 3BNC117. *Nature* 522:487-491.

Cellerai, C., A. Harari, H. Stauss, S. Yerly, A.M. Geretti, A. Carroll, T. Yee, J. Ainsworth, I. Williams, J. Sweeney, A. Freedman, M. Johnson, G. Pantaleo, and S. Kinloch-de Loes. 2011. Early and prolonged antiretroviral therapy is associated with an HIV-1-specific T-cell profile comparable to that of long-term non-progressors. *PloS one* 6:e18164.

Chomont, N., M. El-Far, P. Ancuta, L. Trautmann, F.A. Procopio, B. Yassine-Diab, G. Boucher, M.R. Boulassel, G. Ghattas, J.M. Brenchley, T.W. Schacker, B.J. Hill, D.C. Douek, J.P. Routy, E.K. Haddad, and R.P. Sekaly. 2009. HIV reservoir size and persistence are driven by T cell survival and homeostatic proliferation. *Nature medicine* 15:893-900.

Chun, T.W., L. Carruth, D. Finzi, X. Shen, J.A. DiGiuseppe, H. Taylor, M. Hermankova, K. Chadwick, J. Margolick, T.C. Quinn, Y.H. Kuo, R. Brookmeyer, M.A. Zeiger, P. Barditch-Crovo, and R.F. Siliciano. 1997a. Quantification of latent tissue reservoirs and total body viral load in HIV-1 infection. *Nature* 387:183-188.

Chun, T.W., D. Engel, M.M. Berrey, T. Shea, L. Corey, and A.S. Fauci. 1998. Early establishment of a pool of latently infected, resting CD4(+) T cells during primary HIV-1 infection. *Proceedings of the National Academy of Sciences of the United States of America* 95:8869-8873.

Chun, T.W., L. Stuyver, S.B. Mizell, L.A. Ehler, J.A. Mican, M. Baseler, A.L. Lloyd, M.A. Nowak, and A.S. Fauci. 1997b. Presence of an inducible HIV-1 latent reservoir during highly active antiretroviral therapy. *Proceedings of the National Academy of Sciences of the United States of America* 94:13193-13197.

Cockerham, L.R., J.D. Siliciano, E. Sinclair, U. O'Doherty, S. Palmer, S.A. Yukl, M.C. Strain, N. Chomont, F.M. Hecht, R.F. Siliciano, D.D. Richman, and S.G. Deeks. 2014. CD4+ and CD8+ T cell activation are associated with HIV DNA in resting CD4+ T cells. *PloS one* 9:e110731.

Cohn, L.B., I.T. Silva, T.Y. Oliveira, R.A. Rosales, E.H. Parrish, G.H. Learn, B.H. Hahn, J.L. Czartoski, M.J. McElrath, C. Lehmann, F. Klein, M. Caskey, B.D. Walker, J.D. Siliciano, R.F. Siliciano, M. Jankovic, and M.C. Nussenzweig. 2015. HIV-1 integration landscape during latent and active infection. *Cell* 160:420-432.

Collaborators, G.H. 2016. Estimates of global, regional, and national incidence, prevalence, and mortality of HIV, 1980-2015: the Global Burden of Disease Study 2015. *Lancet HIV* 3:e361-e387.

Craigie, R., and F.D. Bushman. 2012. HIV DNA integration. *Cold Spring Harbor perspectives in medicine* 2:a006890.

Crooks, A.M., R. Bateson, A.B. Cope, N.P. Dahl, M.K. Griggs, J.D. Kuruc, C.L. Gay, J.J. Eron, D.M. Margolis, R.J. Bosch, and N.M. Archin. 2015. Precise Quantitation of the Latent HIV-1 Reservoir: Implications for Eradication Strategies. *The Journal of infectious diseases* 212:1361-1365.

Cummings, M.P., M.C. Neel, and K.L. Shaw. 2008. A genealogical approach to quantifying lineage divergence. *Evolution* 62:2411-2422.

de Masson, A., A. Kirilovsky, R. Zoorob, V. Avettand-Fenoel, V. Morin, A. Oudin, B. Descours, C. Rouzioux, and B. Autran. 2014. Blimp-1 overexpression is associated with low HIV-1 reservoir and transcription levels in central memory CD4+ T cells from elite controllers. *Aids* 28:1567-1577.

Descours, B., G. Petitjean, J.L. Lopez-Zaragoza, T. Bruel, R. Raffel, C. Psomas, J. Reynes, C. Lacabaratz, Y. Levy, O. Schwartz, J.D. Lelievre, and M. Benkirane. 2017. CD32a is a marker of a CD4 T-cell HIV reservoir harbouring replication-competent proviruses. *Nature* 543:564-567.

Dinoso, J.B., S.Y. Kim, A.M. Wiegand, S.E. Palmer, S.J. Gange, L. Cranmer, A. O'Shea, M. Callender, A. Spivak, T. Brennan, M.F. Kearney, M.A. Proschan, J.M. Mican, C.A. Rehm, J.M. Coffin, J.W. Mellors, R.F. Siliciano, and F. Maldarelli. 2009. Treatment intensification does not reduce residual HIV-1 viremia in patients on highly active antiretroviral therapy. *Proceedings of the National Academy of Sciences of the United States of America* 106:9403-9408.

Dobin, A., C.A. Davis, F. Schlesinger, J. Drenkow, C. Zaleski, S. Jha, P. Batut, M. Chaisson, and T.R. Gingeras. 2013. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* 29:15-21.

Doitsh, G., and W.C. Greene. 2016. Dissecting How CD4 T Cells Are Lost During HIV Infection. *Cell Host Microbe* 19:280-291.

Doores, K.J. 2015. The HIV glycan shield as a target for broadly neutralizing antibodies. *FEBS J* 282:4679-4691.

Doria-Rose, N.A., C.A. Schramm, J. Gorman, P.L. Moore, J.N. Bhiman, B.J. DeKosky, M.J. Ernandes, I.S. Georgiev, H.J. Kim, M. Pancera, R.P. Staupe, H.R. Altae-Tran, R.T. Bailer, E.T. Crooks, A. Cupo, A. Druz, N.J. Garrett, K.H. Hoi, R. Kong, M.K. Louder, N.S. Longo, K. McKee, M. Nonyane, S. O'Dell, R.S. Roark, R.S. Rudicell, S.D. Schmidt, D.J. Sheward, C. Soto, C.K. Wibmer, Y. Yang, Z. Zhang, N.C.S. Program, J.C. Mullikin, J.M. Binley, R.W. Sanders, I.A. Wilson, J.P. Moore, A.B. Ward, G. Georgiou, C. Williamson, S.S. Abdool Karim, L. Morris, P.D. Kwong, L. Shapiro, and J.R. Mascola. 2014. Developmental pathway for potent V1V2-directed HIV-neutralizing antibodies. *Nature* 509:55-62.

Douek, D.C., J.M. Brenchley, M.R. Betts, D.R. Ambrozak, B.J. Hill, Y. Okamoto, J.P. Casazza, J. Kuruppu, K. Kunstman, S. Wolinsky, Z. Grossman, M. Dybul, A. Oxenius, D.A. Price, M. Connors, and R.A. Koup. 2002. HIV preferentially infects HIV-specific CD4+ T cells. *Nature* 417:95-98.

Eriksson, S., E.H. Graf, V. Dahl, M.C. Strain, S.A. Yukl, E.S. Lysenko, R.J. Bosch, J. Lai, S. Chioma, F. Emad, M. Abdel-Mohsen, R. Hoh, F. Hecht, P. Hunt, M. Somsouk, J. Wong, R. Johnston, R.F. Siliciano, D.D. Richman, U. O'Doherty, S. Palmer, S.G. Deeks, and J.D. Siliciano. 2013. Comparative analysis of measures of viral reservoirs in HIV-1 eradication studies. *PLoS pathogens* 9:e1003174.

Escolano, A., P. Dosenovic, and M.C. Nussenzweig. 2017. Progress toward active or passive HIV-1 vaccination. *J Exp Med* 214:3-16.

Evering, T.H., S. Mehandru, P. Racz, K. Tenner-Racz, M.A. Poles, A. Figueroa, H. Mohri, and M. Markowitz. 2012. Absence of HIV-1 evolution in the gut-associated lymphoid tissue from patients on combination antiviral therapy initiated during primary infection. *PLoS Pathog* 8:e1002506.

Farber, D.L., N.A. Yudanin, and N.P. Restifo. 2014. Human memory T cells: generation, compartmentalization and homeostasis. *Nat Rev Immunol* 14:24-35.

Finak, G., A. McDavid, M. Yajima, J. Deng, V. Gersuk, A.K. Shalek, C.K. Slichter, H.W. Miller, M.J. McElrath, M. Prlic, P.S. Linsley, and R. Gottardo. 2015. MAST: a flexible statistical framework for assessing transcriptional changes and characterizing heterogeneity in single-cell RNA sequencing data. *Genome biology* 16:278.

Finzi, D., J. Blankson, J.D. Siliciano, J.B. Margolick, K. Chadwick, T. Pierson, K. Smith, J. Lisziewicz, F. Lori, C. Flexner, T.C. Quinn, R.E. Chaisson, E. Rosenberg, B. Walker, S. Gange, J. Gallant, and R.F. Siliciano. 1999. Latent infection of CD4+ T cells provides a mechanism for lifelong persistence of HIV-1, even in patients on effective combination therapy. *Nature medicine* 5:512-517.

Finzi, D., M. Hermankova, T. Pierson, L.M. Carruth, C. Buck, R.E. Chaisson, T.C. Quinn, K. Chadwick, J. Margolick, R. Brookmeyer, J. Gallant, M. Markowitz, D.D. Ho, D.D. Richman, and R.F. Siliciano. 1997. Identification of a reservoir for HIV-1 in patients on highly active antiretroviral therapy. *Science* 278:1295-1300.

Fromentin, R., W. Bakeman, M.B. Lawani, G. Khoury, W. Hartogensis, S. DaFonseca, M. Killian, L. Epling, R. Hoh, E. Sinclair, F.M. Hecht, P. Bacchetti, S.G. Deeks, S.R. Lewin, R.P. Sekaly, and N. Chomont. 2016. CD4+ T Cells Expressing PD-1, TIGIT and LAG-3 Contribute to HIV Persistence during ART. *PLoS pathogens* 12:e1005761.

Galvin, S.R., and M.S. Cohen. 2004. The role of sexually transmitted diseases in HIV transmission. *Nature reviews. Microbiology* 2:33-42.

Gaublomme, J.T., N. Yosef, Y. Lee, R.S. Gertner, L.V. Yang, C. Wu, P.P. Pandolfi, T. Mak, R. Satija, A.K. Shalek, V.K. Kuchroo, H. Park, and A. Regev. 2015. Single-Cell Genomics Unveils Critical Regulators of Th17 Cell Pathogenicity. *Cell* 163:1400-1412.

Grant, R.M., J.R. Lama, P.L. Anderson, V. McMahan, A.Y. Liu, L. Vargas, P. Goicochea, M. Casapia, J.V. Guanira-Carranza, M.E. Ramirez-Cardich, O. Montoya-Herrera, T. Fernandez, V.G. Veloso, S.P. Buchbinder, S. Chariyalertsak, M. Schechter, L.G. Bekker, K.H. Mayer, E.G. Kallas, K.R. Amico, K. Mulligan, L.R. Bushman, R.J. Hance, C. Ganoza, P. Defechereux, B. Postle, F. Wang, J.J. McConnell, J.H. Zheng, J. Lee, J.F. Rooney, H.S. Jaffe, A.I. Martinez, D.N. Burns, D.V. Glidden, and T. iPrEx Study. 2010. Preexposure chemoprophylaxis for HIV prevention in men who have sex with men. *The New England journal of medicine* 363:2587-2599.

Gray, E.S., M.C. Madiga, T. Hermanus, P.L. Moore, C.K. Wibmer, N.L. Tumba, L. Werner, K. Mlisana, S. Sibeko, C. Williamson, S.S. Abdool Karim, L. Morris, and C.S. Team. 2011. The neutralization breadth of HIV-1 develops incrementally over four years and is associated with CD4+ T cell decline and high viral load during acute infection. *Journal of virology* 85:4828-4840.

Guindon, S., J.F. Dufayard, V. Lefort, M. Anisimova, W. Hordijk, and O. Gascuel. 2010. New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. *Syst Biol* 59:307-321.

Halper-Stromberg, A., C.L. Lu, F. Klein, J.A. Horwitz, S. Bournazos, L. Nogueira, T.R. Eisenreich, C. Liu, A. Gazumyan, U. Schaefer, R.C. Furze, M.S. Seaman, R. Prinjha, A. Tarakhovsky, J.V. Ravetch, and M.C. Nussenzweig. 2014. Broadly neutralizing antibodies and viral inducers decrease rebound from HIV-1 latent reservoirs in humanized mice. *Cell* 158:989-999.

Han, A., J. Glanville, L. Hansmann, and M.M. Davis. 2014. Linking T-cell receptor sequence to functional phenotype at the single-cell level. *Nat Biotechnol* 32:684-692.

Han, Y., K. Lassen, D. Monie, A.R. Sedaghat, S. Shimoji, X. Liu, T.C. Pierson, J.B. Margolick, R.F. Siliciano, and J.D. Siliciano. 2004. Resting CD4+ T cells from human immunodeficiency virus type 1 (HIV-1)-infected individuals carry integrated HIV-1 genomes within actively transcribed host genes. *Journal of virology* 78:6122-6133.

Hiener, B., B.A. Horsburgh, J.S. Eden, K. Barton, T.E. Schlub, E. Lee, S. von Stockenstrom, L. Odevall, J.M. Milush, T. Liegler, E. Sinclair, R. Hoh, E.A. Boritz,

D. Douek, R. Fromentin, N. Chomont, S.G. Deeks, F.M. Hecht, and S. Palmer. 2017. Identification of Genetically Intact HIV-1 Proviruses in Specific CD4(+) T Cells from Effectively Treated Participants. *Cell Rep* 21:813-822.

Ho, Y.C., L. Shan, N.N. Hosmane, J. Wang, S.B. Laskey, D.I. Rosenbloom, J. Lai, J.N. Blankson, J.D. Siliciano, and R.F. Siliciano. 2013. Replication-competent noninduced proviruses in the latent reservoir increase barrier to HIV-1 cure. *Cell* 155:540-551.

Horwitz, J.A., Y. Bar-On, C.L. Lu, D. Fera, A.A.K. Lockhart, J.C.C. Lorenzi, L. Nogueira, J. Golijanin, J.F. Scheid, M.S. Seaman, A. Gazumyan, S. Zolla-Pazner, and M.C. Nussenzweig. 2017. Non-neutralizing Antibodies Alter the Course of HIV-1 Infection In Vivo. *Cell* 170:637-648 e610.

Hosmane, N.N., K.J. Kwon, K.M. Bruner, A.A. Capoferri, S. Beg, D.I. Rosenbloom, B.F. Keele, Y.C. Ho, J.D. Siliciano, and R.F. Siliciano. 2017. Proliferation of latently infected CD4+ T cells carrying replication-competent HIV-1: Potential role in latent reservoir dynamics. *J Exp Med* 214:959-972.

Hu, W.S., and S.H. Hughes. 2012. HIV-1 reverse transcription. *Cold Spring Harbor perspectives in medicine* 2:

Hudspeth, K., M. Fogli, D.V. Correia, J. Mikulak, A. Roberto, S. Della Bella, B. Silva-Santos, and D. Mavilio. 2012. Engagement of NKp30 on Vdelta1 T cells induces the production of CCL3, CCL4, and CCL5 and suppresses HIV-1 replication. *Blood* 119:4013-4016.

Hunt, M., A. Gall, S.H. Ong, J. Brener, B. Ferns, P. Goulder, E. Nastouli, J.A. Keane, P. Kellam, and T.D. Otto. 2015. IVA: accurate de novo assembly of RNA virus genomes. *Bioinformatics* 31:2374-2376.

Hunt, P.W. 2014. HIV and aging: emerging research issues. *Current opinion in HIV and AIDS* 9:302-308.

Ikeda, T., J. Shibata, K. Yoshimura, A. Koito, and S. Matsushita. 2007. Recurrent HIV-1 integration at the BACH2 locus in resting CD4+ T cell populations during effective highly active antiretroviral therapy. *The Journal of infectious diseases* 195:716-725.

Imamichi, H., V. Natarajan, J.W. Adelsberger, C.A. Rehm, R.A. Lempicki, B. Das, A. Hazen, T. Imamichi, and H.C. Lane. 2014. Lifespan of effector memory CD4+ T cells determined by replication-incompetent integrated HIV-1 provirus. *Aids*

Islam, S., A. Zeisel, S. Joost, G. La Manno, P. Zajac, M. Kasper, P. Lonnerberg, and S. Linnarsson. 2014. Quantitative single-cell RNA-seq with unique molecular identifiers. *Nat Methods* 11:163-166.

Jain, V., W. Hartogensis, P. Bacchetti, P.W. Hunt, H. Hatano, E. Sinclair, L. Epling, T.H. Lee, M.P. Busch, J.M. McCune, C.D. Pilcher, F.M. Hecht, and S.G. Deeks. 2013. Antiretroviral therapy initiated within 6 months of HIV infection is associated with lower T-cell activation and smaller HIV reservoir size. *The Journal of infectious diseases* 208:1202-1211.

Janovitz, T., I.A. Klein, T. Oliveira, P. Mukherjee, M.C. Nussenzweig, M. Sadelain, and E. Falck-Pedersen. 2013. High-throughput sequencing reveals principles of adeno-associated virus serotype 2 integration. *Journal of virology* 87:8559-8568.

Joos, B., M. Fischer, H. Kuster, S.K. Pillai, J.K. Wong, J. Boni, B. Hirschel, R. Weber, A. Trkola, H.F. Gunthard, and H.I.V.C.S. Swiss. 2008. HIV rebounds from latently infected cells, rather than from continuing low-level replication. *Proceedings of the National Academy of Sciences of the United States of America* 105:16725-16730.

Jordan, A., D. Bisgrove, and E. Verdin. 2003. HIV reproducibly establishes a latent infection after acute infection of T cells in vitro. *The EMBO journal* 22:1868-1877.

Jordan, A., P. Defechereux, and E. Verdin. 2001. The site of HIV-1 integration in the human genome determines basal transcriptional activity and response to Tat transactivation. *The EMBO journal* 20:1726-1738.

Josefsson, L., S. von Stockenstrom, N.R. Faria, E. Sinclair, P. Bacchetti, M. Killian, L. Epling, A. Tan, T. Ho, P. Lemey, W. Shao, P.W. Hunt, M. Somsouk, W. Wylie, D.C. Douek, L. Loeb, J. Custer, R. Hoh, L. Poole, S.G. Deeks, F. Hecht, and S. Palmer. 2013. The HIV-1 reservoir in eight patients on long-term suppressive antiretroviral therapy is stable with few genetic changes over time. *Proceedings of the National Academy of Sciences of the United States of America* 110:E4987-4996.

Kaczmarek Michaels, K., M. Natarajan, Z. Euler, G. Alter, G. Viglianti, and A.J. Henderson. 2015. Blimp-1, an intrinsic factor that represses HIV-1 proviral transcription in memory CD4+ T cells. *Journal of immunology* 194:3267-3274.

Kauder, S.E., A. Bosque, A. Lindqvist, V. Planelles, and E. Verdin. 2009. Epigenetic regulation of HIV-1 latency by cytosine methylation. *PLoS pathogens* 5:e1000495.

Keele, B.F., E.E. Giorgi, J.F. Salazar-Gonzalez, J.M. Decker, K.T. Pham, M.G. Salazar, C. Sun, T. Grayson, S. Wang, H. Li, X. Wei, C. Jiang, J.L. Kirchherr, F. Gao, J.A. Anderson, L.H. Ping, R. Swanstrom, G.D. Tomaras, W.A. Blattner, P.A. Goepfert, J.M. Kilby, M.S. Saag, E.L. Delwart, M.P. Busch, M.S. Cohen, D.C. Montefiori, B.F. Haynes, B. Gaschen, G.S. Athreya, H.Y. Lee, N. Wood, C. Seoighe, A.S. Perelson, T. Bhattacharya, B.T. Korber, B.H. Hahn, and G.M. Shaw. 2008. Identification and characterization of transmitted and early founder virus envelopes in primary HIV-1 infection. *Proceedings of the National Academy of Sciences of the United States of America* 105:7552-7557.

Kent, W.J. 2002. BLAT--the BLAST-like alignment tool. *Genome research* 12:656-664.

Kiselev, V.Y., K. Kirschner, M.T. Schaub, T. Andrews, A. Yiu, T. Chandra, K.N. Natarajan, W. Reik, M. Barahona, A.R. Green, and M. Hemberg. 2017. SC3: consensus clustering of single-cell RNA-seq data. *Nat Methods* 14:483-486.

Klein, F., H. Mouquet, P. Dosenovic, J.F. Scheid, L. Scharf, and M.C. Nussenzweig. 2013. Antibodies in HIV-1 vaccine development and therapy. *Science* 341:1199-1204.

Klein, I.A., W. Resch, M. Jankovic, T. Oliveira, A. Yamane, H. Nakahashi, M. Di Virgilio, A. Bothmer, A. Nussenzweig, D.F. Robbiani, R. Casellas, and M.C. Nussenzweig. 2011. Translocation-capture sequencing reveals the extent and nature of chromosomal rearrangements in B lymphocytes. *Cell* 147:95-106.

Krishnan, V., and S.L. Zeichner. 2004. Host cell gene expression during human immunodeficiency virus type 1 latency and reactivation and effects of targeting genes that are differentially expressed in viral latency. *Journal of virology* 78:9458-9473.

Kryazhimskiy, S., D.P. Rice, E.R. Jerison, and M.M. Desai. 2014. Microbial evolution. Global epistasis makes adaptation predictable despite sequence-level stochasticity. *Science* 344:1519-1522.

Laird, G.M., C.K. Bullen, D.I. Rosenbloom, A.R. Martin, A.L. Hill, C.M. Durand, J.D. Siliciano, and R.F. Siliciano. 2015. Ex vivo analysis identifies effective HIV-1 latency-reversing drug combinations. *The Journal of clinical investigation* 125:1901-1912.

Laird, G.M., E.E. Eisele, S.A. Rabi, J. Lai, S. Chioma, J.N. Blankson, J.D. Siliciano, and R.F. Siliciano. 2013. Rapid quantification of the latent reservoir for HIV-1 using a viral outgrowth assay. *PLoS pathogens* 9:e1003398.

Laird, G.M., D.I. Rosenbloom, J. Lai, R.F. Siliciano, and J.D. Siliciano. 2016. Measuring the Frequency of Latent HIV-1 in Resting CD4(+) T Cells Using a Limiting Dilution Coculture Assay. *Methods Mol Biol* 1354:239-253.

Langmead, B., C. Trapnell, M. Pop, and S.L. Salzberg. 2009. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome biology* 10:R25.

Laskey, S.B., C.W. Pohlmeyer, K.M. Bruner, and R.F. Siliciano. 2016. Evaluating Clonal Expansion of HIV-Infected Cells: Optimization of PCR Strategies to Predict Clonality. *PLoS pathogens* 12:e1005689.

Lassen, K.G., K.X. Ramyar, J.R. Bailey, Y. Zhou, and R.F. Siliciano. 2006. Nuclear retention of multiply spliced HIV-1 RNA in resting CD4+ T cells. *PLoS pathogens* 2:e68.

Lee, G.Q., N. Orlova-Fink, K. Einkauf, F.Z. Chowdhury, X. Sun, S. Harrington, H.H. Kuo, S. Hua, H.R. Chen, Z. Ouyang, K. Reddy, K. Dong, T. Ndung'u, B.D. Walker, E.S. Rosenberg, X.G. Yu, and M. Lichterfeld. 2017. Clonal expansion of genome-intact HIV-1 in functionally polarized Th1 CD4+ T cells. *The Journal of clinical investigation* 127:2689-2696.

Lenasi, T., X. Contreras, and B.M. Peterlin. 2008. Transcriptional interference antagonizes proviral gene expression to promote HIV latency. *Cell Host Microbe* 4:123-133.

Liao, H.X., R. Lynch, T. Zhou, F. Gao, S.M. Alam, S.D. Boyd, A.Z. Fire, K.M. Roskin, C.A. Schramm, Z. Zhang, J. Zhu, L. Shapiro, N.C.S. Program, J.C. Mullikin, S. Gnanakaran, P. Hraber, K. Wiehe, G. Kelsoe, G. Yang, S.M. Xia, D.C. Montefiori, R. Parks, K.E. Lloyd, R.M. Scearce, K.A. Soderberg, M. Cohen, G. Kamanga, M.K. Louder, L.M. Tran, Y. Chen, F. Cai, S. Chen, S. Moquin, X. Du, M.G. Joyce, S. Srivatsan, B. Zhang, A. Zheng, G.M. Shaw, B.H. Hahn, T.B. Kepler, B.T. Korber, P.D. Kwong, J.R. Mascola, and B.F. Haynes. 2013. Co-evolution of a broadly neutralizing HIV-1 antibody and founder virus. *Nature* 496:469-476.

Lorenzi, J.C., Y.Z. Cohen, L.B. Cohn, E.F. Kreider, J.P. Barton, G.H. Learn, T. Oliveira, C.L. Lavine, J.A. Horwitz, A. Settler, M. Jankovic, M.S. Seaman, A.K. Chakraborty, B.H. Hahn, M. Caskey, and M.C. Nussenzweig. 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 of the United States of America* 113:E7908-E7916.

Lu, C.L., D.K. Murakowski, S. Bournazos, T. Schoofs, D. Sarkar, A. Halper-Stromberg, J.A. Horwitz, L. Nogueira, J. Golijanin, A. Gazumyan, J.V. Ravetch, M. Caskey, A.K. Chakraborty, and M.C. Nussenzweig. 2016. Enhanced clearance of HIV-1-infected cells by broadly neutralizing antibodies against HIV-1 in vivo. *Science* 352:1001-1004.

Lun, A.T., D.J. McCarthy, and J.C. Marioni. 2016. A step-by-step workflow for low-level analysis of single-cell RNA-seq data with Bioconductor. *F1000Res* 5:2122.

Maartens, G., C. Celum, and S.R. Lewin. 2014. HIV infection: epidemiology, pathogenesis, treatment, and prevention. *Lancet* 384:258-271.

Maddon, P.J., A.G. Dalgleish, J.S. McDougal, P.R. Clapham, R.A. Weiss, and R. Axel. 1986. The T4 gene encodes the AIDS virus receptor and is expressed in the immune system and the brain. *Cell* 47:333-348.

Mahnke, Y.D., T.M. Brodie, F. Sallusto, M. Roederer, and E. Lugli. 2013. The who's who of T-cell differentiation: human memory T-cell subsets. *Eur J Immunol* 43:2797-2809.

Maldarelli, F., X. Wu, L. Su, F.R. Simonetti, W. Shao, S. Hill, J. Spindler, A.L. Ferris, J.W. Mellors, M.F. Kearney, J.M. Coffin, and S.H. Hughes. 2014. HIV latency. Specific HIV integration sites are linked to clonal expansion and persistence of infected cells. *Science* 345:179-183.

Marini, B., A. Kertesz-Farkas, H. Ali, B. Lucic, K. Lisek, L. Manganaro, S. Pongor, R. Luzzati, A. Recchia, F. Mavilio, M. Giacca, and M. Lusic. 2015. Nuclear architecture dictates HIV-1 integration site selection. *Nature* 521:227-231.

Markowitz, M., M. Louie, A. Hurley, E. Sun, M. Di Mascio, A.S. Perelson, and D.D. Ho. 2003. A novel antiviral intervention results in more accurate assessment of human immunodeficiency virus type 1 replication dynamics and T-cell decay in vivo. *Journal of virology* 77:5037-5038.

McDougal, J.S., M.S. Kennedy, J.M. Sligh, S.P. Cort, A. Mawle, and J.K. Nicholson. 1986. Binding of HTLV-III/LAV to T4+ T cells by a complex of the 110K viral protein and the T4 molecule. *Science* 231:382-385.

Mitchell, R.S., B.F. Beitzel, A.R. Schroder, P. Shinn, H. Chen, C.C. Berry, J.R. Ecker, and F.D. Bushman. 2004. Retroviral DNA integration: ASLV, HIV, and MLV show distinct target site preferences. *PLoS biology* 2:E234.

Mouquet, H., L. Scharf, Z. Euler, Y. Liu, C. Eden, J.F. Scheid, A. Halper-Stromberg, P.N.P. Gnanapragasam, D.I.R. Spencer, M.S. Seaman, H. Schuitemaker, T. Feizi, M.C. Nussenzweig, and P.J. Bjorkman. 2012. Complex-type N-glycan recognition by potent broadly neutralizing HIV antibodies. *Proceedings of the National Academy of Sciences of the United States of America* 109:E3268-E3277.

Mullins, J.I., and L.M. Frenkel. 2017. Clonal Expansion of Human Immunodeficiency Virus-Infected Cells and Human Immunodeficiency Virus Persistence During Antiretroviral Therapy. *The Journal of infectious diseases* 215:S119-S127.

Murray, A.J., K.J. Kwon, D.L. Farber, and R.F. Siliciano. 2016. The Latent Reservoir for HIV-1: How Immunologic Memory and Clonal Expansion Contribute to HIV-1 Persistence. *Journal of immunology* 197:407-417.

Nabel, G., and D. Baltimore. 1987. An inducible transcription factor activates expression of human immunodeficiency virus in T cells. *Nature* 326:711-713.

Nachega, J.B., V.C. Marconi, G.U. van Zyl, E.M. Gardner, W. Preiser, S.Y. Hong, E.J. Mills, and R. Gross. 2011. HIV treatment adherence, drug resistance, virologic failure: evolving concepts. *Infect Disord Drug Targets* 11:167-174.

Oxenius, A., S. Fidler, M. Brady, S.J. Dawson, K. Ruth, P.J. Easterbrook, J.N. Weber, R.E. Phillips, and D.A. Price. 2001. Variable fate of virus-specific CD4(+) T cells during primary HIV-1 infection. *Eur J Immunol* 31:3782-3788.

Palmer, S., A.P. Wiegand, F. Maldarelli, H. Bazmi, J.M. Mican, M. Polis, R.L. Dewar, A. Planta, S. Liu, J.A. Metcalf, J.W. Mellors, and J.M. Coffin. 2003. New real-time reverse transcriptase-initiated PCR assay with single-copy sensitivity for human immunodeficiency virus type 1 RNA in plasma. *J Clin Microbiol* 41:4531-4536.

Pape, K.A., J.J. Taylor, R.W. Maul, P.J. Gearhart, and M.K. Jenkins. 2011. Different B cell populations mediate early and late memory during an endogenous immune response. *Science* 331:1203-1207.

Paradis, E., J. Claude, and K. Strimmer. 2004. APE: Analyses of Phylogenetics and Evolution in R language. *Bioinformatics* 20:289-290.

Parrish, N.F., F. Gao, H. Li, E.E. Giorgi, H.J. Barbian, E.H. Parrish, L. Zajic, S.S. Iyer, J.M. Decker, A. Kumar, B. Hora, A. Berg, F. Cai, J. Hopper, T.N. Denny, H. Ding, C. Ochsenbauer, J.C. Kappes, R.P. Galimidi, A.P. West, Jr., P.J. Bjorkman, C.B. Wilen, R.W. Doms, M. O'Brien, N. Bhardwaj, P. Borrow, B.F. Haynes, M. Muldoon, J.P. Theiler, B. Korber, G.M. Shaw, and B.H. Hahn. 2013. Phenotypic properties of transmitted founder HIV-1. *Proceedings of the National Academy of Sciences of the United States of America* 110:6626-6633.

Perreau, M., A.L. Savoye, E. De Crignis, J.M. Corpataux, R. Cubas, E.K. Haddad, L. De Leval, C. Graziosi, and G. Pantaleo. 2013. Follicular helper T cells serve as the major CD4 T cell compartment for HIV-1 infection, replication, and production. *J Exp Med* 210:143-156.

Persaud, D., S.C. Ray, J. Kajdas, A. Ahonkhai, G.K. Siberry, K. Ferguson, C. Ziemniak, T.C. Quinn, J.P. Casazza, S. Zeichner, S.J. Gange, and D.C. Watson. 2007. Slow human immunodeficiency virus type 1 evolution in viral reservoirs in infants treated with effective antiretroviral therapy. *AIDS Res Hum Retroviruses* 23:381-390.

Pope, M., and A.T. Haase. 2003. Transmission, acute HIV-1 infection and the quest for strategies to prevent infection. *Nature medicine* 9:847-852.

Procopio, F.A., R. Fromentin, D.A. Kulpa, J.H. Brehm, A.G. Bebin, M.C. Strain, D.D. Richman, U. O'Doherty, S. Palmer, F.M. Hecht, R. Hoh, R.J. Barnard, M.D. Miller, D.J. Hazuda, S.G. Deeks, R.P. Sekaly, and N. Chomont. 2015. A Novel Assay to Measure the Magnitude of the Inducible Viral Reservoir in HIV-infected Individuals. *EBioMedicine* 2:874-883.

Quinlan, A.R., and I.M. Hall. 2010. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* 26:841-842.

Ruelas, D.S., J.K. Chan, E. Oh, A.J. Heidersbach, A.M. Hebbeler, L. Chavez, E. Verdin, M. Rape, and W.C. Greene. 2015. MicroRNA-155 Reinforces HIV Latency. *J Biol Chem* 290:13736-13748.

Runarsson, T. and Yao, X. 2005. Search biases in constrained evolutionary optimization. *IEEE Trans. on Systems, Man, and Cybernetics Part C: Applications and Reviews,* 35(3):233– 243.

Rusert, P., R.D. Kouyos, C. Kadelka, H. Ebner, M. Schanz, M. Huber, D.L. Braun, N. Hoze, A. Scherrer, C. Magnus, J. Weber, T. Uhr, V. Cippa, C.W. Thorball, H. Kuster, M. Cavassini, E. Bernasconi, M. Hoffmann, A. Calmy, M. Battegay, A. Rauch, S. Yerly, V. Aubert, T. Klimkait, J. Boni, J. Fellay, R.R. Regoes, H.F. Gunthard, A. Trkola, and H.I.V.C.S. Swiss. 2016. Determinants of HIV-1 broadly neutralizing antibody induction. *Nature medicine* 22:1260-1267.

Sallusto, F. 2016. Heterogeneity of Human CD4(+) T Cells Against Microbes. *Annual review of immunology* 34:317-334.

Sarmati, L., G. D'Ettorre, S.G. Parisi, and M. Andreoni. 2015. HIV Replication at Low Copy Number and its Correlation with the HIV Reservoir: A Clinical Perspective. *Curr HIV Res* 13:250-257.

Satija, R., J.A. Farrell, D. Gennert, A.F. Schier, and A. Regev. 2015. Spatial reconstruction of single-cell gene expression data. *Nat Biotechnol* 33:495-502.

Scheid, J.F., J.A. Horwitz, Y. Bar-On, E.F. Kreider, C.L. Lu, J.C. Lorenzi, A. Feldmann, M. Braunschweig, L. Nogueira, T. Oliveira, I. Shimeliovich, R. Patel, L. Burke, Y.Z. Cohen, S. Hadrigan, A. Settler, M. Witmer-Pack, A.P. West, Jr., B. Juelg, T. Keler, T. Hawthorne, B. Zingman, R.M. Gulick, N. Pfeifer, G.H. Learn, M.S. Seaman, P.J. Bjorkman, F. Klein, S.J. Schlesinger, B.D. Walker, B.H. Hahn, M.C. Nussenzweig, and M. Caskey. 2016. HIV-1 antibody 3BNC117 suppresses viral rebound in humans during treatment interruption. *Nature* 535:556-560.

Scheid, J.F., H. Mouquet, B. Ueberheide, R. Diskin, F. Klein, T.Y.K. Oliveira, J. Pietzsch, D. Fenyo, A. Abadir, K. Velinzon, A. Hurley, S. Myung, F. Boulad, P. Poignard, D.R. Burton, F. Pereyra, D.D. Ho, B.D. Walker, M.S. Seaman, P.J. Bjorkman, B.T. Chait, and M.C. Nussenzweig. 2011. Sequence and Structural Convergence of Broad and Potent HIV Antibodies That Mimic CD4 Binding. *Science* 333:1633-1637.

Schroder, A.R., P. Shinn, H. Chen, C. Berry, J.R. Ecker, and F. Bushman. 2002. HIV-1 integration in the human genome favors active genes and local hotspots. *Cell* 110:521-529.

Sherrill-Mix, S., M.K. Lewinski, M. Famiglietti, A. Bosque, N. Malani, K.E. Ocwieja, C.C. Berry, D. Looney, L. Shan, L.M. Agosto, M.J. Pace, R.F. Siliciano, U. O'Doherty,

J. Guatelli, V. Planelles, and F.D. Bushman. 2013. HIV latency and integration site placement in five cell-based models. *Retrovirology* 10:90.

Sherrill-Mix, S., K.E. Ocwieja, and F.D. Bushman. 2015. Gene activity in primary T cells infected with HIV89.6: intron retention and induction of genomic repeats. *Retrovirology* 12:79.

Siliciano, J.D., J. Kajdas, D. Finzi, T.C. Quinn, K. Chadwick, J.B. Margolick, C. Kovacs, S.J. Gange, and R.F. Siliciano. 2003. Long-term follow-up studies confirm the stability of the latent reservoir for HIV-1 in resting CD4+ T cells. *Nature medicine* 9:727-728.

Siliciano, R.F., and W.C. Greene. 2011. HIV latency. *Cold Spring Harbor perspectives in medicine* 1:a007096.

Silva, I.T., R.A. Rosales, A.J. Holanda, M.C. Nussenzweig, and M. Jankovic. 2014. Identification of cromosomal translocation hotspots via scan statistics. *Bioinformatics*

Simonetti, F.R., M.D. Sobolewski, E. Fyne, W. Shao, J. Spindler, J. Hattori, E.M. Anderson, S.A. Watters, S. Hill, X. Wu, D. Wells, L. Su, B.T. Luke, E.K. Halvas, G. Besson, K.J. Penrose, Z. Yang, R.W. Kwan, C. Van Waes, T. Uldrick, D.E. Citrin, J. Kovacs, M.A. Polis, C.A. Rehm, R. Gorelick, M. Piatak, B.F. Keele, M.F. Kearney, J.M. Coffin, S.H. Hughes, J.W. Mellors, and F. Maldarelli. 2016. Clonally expanded CD4+ T cells can produce infectious HIV-1 in vivo. *Proceedings of the National Academy of Sciences of the United States of America* 113:1883-1888.

Stubbington, M.J.T., T. Lonnberg, V. Proserpio, S. Clare, A.O. Speak, G. Dougan, and S.A. Teichmann. 2016. T cell fate and clonality inference from single-cell transcriptomes. *Nat Methods* 13:329-332.

Swain, S.L., K.K. McKinstry, and T.M. Strutt. 2012. Expanding roles for CD4(+) T cells in immunity to viruses. *Nat Rev Immunol* 12:136-148.

Tas, J.M., L. Mesin, G. Pasqual, S. Targ, J.T. Jacobsen, Y.M. Mano, C.S. Chen, J.C. Weill, C.A. Reynaud, E.P. Browne, M. Meyer-Hermann, and G.D. Victora. 2016. Visualizing antibody affinity maturation in germinal centers. *Science* 351:1048-1054.

Tomaras, G.D., N.L. Yates, P. Liu, L. Qin, G.G. Fouda, L.L. Chavez, A.C. Decamp, R.J. Parks, V.C. Ashley, J.T. Lucas, M. Cohen, J. Eron, C.B. Hicks, H.X. Liao, S.G. Self, G. Landucci, D.N. Forthal, K.J. Weinhold, B.F. Keele, B.H. Hahn, M.L. Greenberg, L. Morris, S.S. Karim, W.A. Blattner, D.C. Montefiori, G.M. Shaw, A.S. Perelson, and B.F. Haynes. 2008. Initial B-cell responses to transmitted human immunodeficiency virus type 1: virion-binding immunoglobulin M (IgM) and IgG antibodies followed by plasma anti-gp41 antibodies with ineffective control of initial viremia. *Journal of virology* 82:12449-12463.

Trombetta, J.J., D. Gennert, D. Lu, R. Satija, A.K. Shalek, and A. Regev. 2014. Preparation of Single-Cell RNA-Seq Libraries for Next Generation Sequencing. *Curr Protoc Mol Biol* 107:4 22 21-17.

Vallejo, A., C. Gutierrez, B. Hernandez-Novoa, L. Diaz, N. Madrid, M. Abad-Fernandez, F. Dronda, M.J. Perez-Elias, J. Zamora, E. Munoz, M.A. Munoz-Fernandez, and S. Moreno. 2012. The effect of intensification with raltegravir on the HIV-1 reservoir of latently infected memory CD4 T cells in suppressed patients. *AIDS* 26:1885-1894.

van 't Wout, A.B., H. Schuitemaker, and N.A. Kootstra. 2008. Isolation and propagation of HIV-1 on peripheral blood mononuclear cells. *Nat Protoc* 3:363-370.

Van Lint, C., S. Bouchat, and A. Marcello. 2013. HIV-1 transcription and latency: an update. *Retrovirology* 10:67.

Vogelstein, B., N. Papadopoulos, V.E. Velculescu, S. Zhou, L.A. Diaz, Jr., and K.W. Kinzler. 2013. Cancer genome landscapes. *Science* 339:1546-1558.

von Stockenstrom, S., L. Odevall, E. Lee, E. Sinclair, P. Bacchetti, M. Killian, L. Epling, W. Shao, R. Hoh, T. Ho, N.R. Faria, P. Lemey, J. Albert, P. Hunt, L. Loeb, C. Pilcher, L. Poole, H. Hatano, M. Somsouk, D. Douek, E. Boritz, S.G. Deeks, F.M. Hecht, and S. Palmer. 2015. Longitudinal Genetic Characterization Reveals That Cell Proliferation Maintains a Persistent HIV Type 1 DNA Pool During Effective HIV Therapy. *The Journal of infectious diseases* 212:596-607.

Wagner, T.A., J.L. McKernan, N.H. Tobin, K.A. Tapia, J.I. Mullins, and L.M. Frenkel. 2013. An increasing proportion of monotypic HIV-1 DNA sequences during antiretroviral treatment suggests proliferation of HIV-infected cells. *Journal of virology* 87:1770-1778.

Wagner, T.A., S. McLaughlin, K. Garg, C.Y. Cheung, B.B. Larsen, S. Styrchak, H.C. Huang, P.T. Edlefsen, J.I. Mullins, and L.M. Frenkel. 2014. Proliferation of cells with HIV integrated into cancer genes contributes to persistent infection. *Science*

Walker, L.M., S.K. Phogat, P.Y. Chan-Hui, D. Wagner, P. Phung, J.L. Goss, T. Wrin, M.D. Simek, S. Fling, J.L. Mitcham, J.K. Lehrman, F.H. Priddy, O.A. Olsen, S.M. Frey, P.W. Hammond, G.P.I. Protocol, S. Kaminsky, T. Zamb, M. Moyle, W.C. Koff, P. Poignard, and D.R. Burton. 2009. Broad and potent neutralizing antibodies from an African donor reveal a new HIV-1 vaccine target. *Science* 326:285-289.

Wang, G.P., A. Ciuffi, J. Leipzig, C.C. Berry, and F.D. Bushman. 2007. HIV integration site selection: analysis by massively parallel pyrosequencing reveals association with epigenetic modifications. *Genome research* 17:1186-1194.

Ward, A.B., and I.A. Wilson. 2015. Insights into the trimeric HIV-1 envelope glycoprotein structure. *Trends Biochem Sci* 40:101-107.

Wei, X., J.M. Decker, S. Wang, H. Hui, J.C. Kappes, X. Wu, J.F. Salazar-Gonzalez, M.G. Salazar, J.M. Kilby, M.S. Saag, N.L. Komarova, M.A. Nowak, B.H. Hahn, P.D. Kwong, and G.M. Shaw. 2003. Antibody neutralization and escape by HIV-1. *Nature* 422:307-312.

West, M.J., A.D. Lowe, and J. Karn. 2001. Activation of human immunodeficiency virus transcription in T cells revisited: NF-kappaB p65 stimulates transcriptional elongation. *Journal of virology* 75:8524-8537.

Wilen, C.B., J.C. Tilton, and R.W. Doms. 2012. HIV: cell binding and entry. *Cold Spring Harbor perspectives in medicine* 2:

Williams, S.A., and W.C. Greene. 2007. Regulation of HIV-1 latency by T-cell activation. *Cytokine* 39:63-74.

Yoon, H., J. Macke, A.P. West, Jr., B. Foley, P.J. Bjorkman, B. Korber, and K. Yusim. 2015. CATNAP: a tool to compile, analyze and tally neutralizing antibody panels. *Nucleic acids research* 43:W213-219.

Zhao, M., J. Sun, and Z. Zhao. 2013. TSGene: a web resource for tumor suppressor genes. *Nucleic acids research* 41:D970-976.

Zhu, J., H. Yamane, and W.E. Paul. 2010. Differentiation of effector CD4 T cell populations (*). *Annual review of immunology* 28:445-489.