1-31-2019 2:00 PM

Novel insights into the genomic integration site landscape of HIV-1 and other retrovirus genera

Hinissan P. Kohio, *The University of Western Ontario*

Supervisor: Barr, Stephen D., *The University of Western Ontario*

A thesis submitted in partial fulfillment of the requirements for the Doctor of Philosophy degree in Microbiology and Immunology

Abstract

An important event during infection by retroviruses such as human immunodeficiency virus type 1 (HIV-1) is the permanent integration of the viral genome into the host genome. This event leads to life-long infection and is accompanied by a period of quiescence/latency ranging from a few years to >10 years where HIV-1 expression is barely detectable or undetectable. Despite the use of combination antiretroviral therapy (cART) which controls HIV-1 infection, quiescent/latent virus presents a major obstacle towards a functional cure. Integration site location in the genome is thought to contribute to latent infections and has the potential to confound anti-latency treatments, necessitating a greater understanding of the effects of integration site location on latency.

To examine the global preference for integration location, we performed an extensive bioinformatics analysis on the integration site profile of HIV-1 and other retroviruses. We found that HIV-1 integration sites and that of other retroviruses are enriched in and/or near non-B DNA motifs. Non-B DNA are secondary structures in our genome formed by specific nucleotide sequences that exhibit non-canonical DNA base pairing. We demonstrated a strong correlation between integration sites in and near guanine-quadruplex (G4) motifs, a type of non-B DNA associated with transcriptional silencing, and reactivation of latent proviruses with latency reversal agents. Additionally, integration site studies have focused on HIV-1 subtype B infections; however, infections with other subtypes exist worldwide. A comparative analysis of 62 infected individuals with different HIV-1 subtypes showed significant differences in the integration site profiles between different subtypes, which was further altered by cART. Finally, we examined HIV-1 integration site profiles in anatomical sites and showed distinct integration profiles from peripheral blood, brain, and the gastrointestinal tract.

Overall, our findings identified similarities and differences in the integration site profiles among evolutionarily diverse retroviruses. Notably, we have implicated non-B DNA as a new factor that influences integration site targeting and may play an important role in the establishment of HIV-1 latency and/or disease progression.

Keywords

HIV-1, HIV-1 subtype A, B, C and D, sanctuary sites, integration sites, latency, non-B DNA motifs, guanine-quadruplex/G4, retroviruses.

Co-Authorship Statement

Chapter 2: All experiments related to guanine-quadruplex (G4) compounds treatments including MTT assays were performed by Hinissan P. Kohio. Preparation of samples for sequencing was performed by Hinissan P. Kohio with the assistance of Macon Coleman. Sequencing analyses for the G4 compounds treatments were conducted by Hinissan P. Kohio. All other analyses were performed by Dr. Stephen Barr and Dr. Hannah O. Ajoge. Dr. Hannah O. Ajoge was responsible for developing the bioinformatics pipeline leading to data generation for all figures in chapter 2.

Chapter 3: All Uganda and Zimbabwe patients' samples were prepared for sequencing by Hinissan P. Kohio with the assistance of Macon Coleman. All experimental analyses were performed by Hinissan P. Kohio. Dr. Hannah O. Ajoge was responsible for developing the bioinformatics pipeline leading to data generation for all figures in chapter 3.

Chapter 4: All tissue samples were prepared for sequencing by Hinissan P. Kohio with the assistance of Macon Coleman. All experimental analyses were performed by Hinissan P. Kohio and Dr. Hannah O. Ajoge. Dr. Hannah O. Ajoge was responsible for developing the bioinformatics pipeline leading to data generation for all figures in chapter 4.

Acknowledgments

First and foremost, I would like to express my deepest gratitude to my supervisor Dr. Stephen Barr for giving me this amazing opportunity to be part of his laboratory and to conduct my thesis research in his laboratory. I thank you for your continuous guidance, scientific input, motivation and patience. Your positive attitude towards science and research have greatly helped me to overcome the different challenges I encountered throughout my research as a Ph.D. student. You have taught me many skills for being an effective scientist. Your mentorship has allowed me to succeed throughout my graduate studies and has further prepare me for a successful career in the scientific field. It has been an extreme pleasure, and privilege to work with you.

My gratitude also goes towards my advisory committee members, Dr. Joe Mymryk, Dr. Eric Arts and Dr. Jimmy Dikeakos. Thank you for the time and effort you have invested in me and my project during the past 4.5 years. You all provided me with insightful comments and constructive feedback which were instrumental in the completion of my project. I would also like to thank the past and present members of the Barr laboratory for their assistance with my project and for helping make my laboratory experience a meaningful one.

Finally, I am deeply indebted to my amazing family who strongly supported me throughout this journey. I am forever grateful to my mother and my father for their endless prayers and encouragement and for all the sacrifices they made. A special thank you to my uncle Dr. Yazoume Ye who has always imparted in me his wisdom and for keeping me motivated throughout my studies. Thank you to my siblings Flore, Thierry and Harold for your love and for always believing in me. Thank you to my friends for all your support.

Table of Contents

List of Tables

List of Figures

List of Abbreviations

AIDS	Acquired immunodeficiency syndrome
ASLV	Avian sarcoma leukosis virus
AZT	Azidothymidine
BAF	Barrier-to-autointegration factor
bp	Base pair
BRACO19	N, N′-(9-(4-(Dimethylamino) phenylamino) acridine-3, 6-diyl) bis (3-(pyrrolidin-1-yl) propanamide)
CA	Capsid
cART	Combination antiretroviral therapy
CCD	Catalytic core domain
CCR5	Chemokine receptor type 5
CD3	Cluster of differentiation 3
CD4	Cluster of differentiation 4
CNS	Central nervous system
CPSF6	Cleavage and polyadenylation specificity factor 6
CRFs	Circulating recombinant forms
CSF	Cerebrospinal fluid
CTD	C-terminal domain
CXCR4	Chemokine receptor type 4
ddI	Didanosine

DHS	DNaseI hypersensitivity sites
DMEM	Dulbecco’s modified eagle medium
DMSO	Dimethyl sulfoxide
DNA	Deoxyribonucleic acid
dNTP	Deoxynucleotide triphosphate
ER	Endoplasmic reticulum
ERVs	Endogenous retroviruses
FBS	Fetal bovine serum
FIV	Feline immunodeficiency virus
FV	Foamy virus
G4	Guanine-quadruplex
GALT	Gut associated lymphoid tissues
GIT	Gastrointestinal tract
GU	Genitourinary
HDACi	Histone deacetylase inhibitor
HEK 293T	Human embryonic kidney 293T
HGMGA1	High mobility group chromosomal protein A1
HIV-1	Human immunodeficiency virus type 1
HIV-2	Human immunodeficiency virus type 2
HTLV-1	Human T-lymphotropic virus 1

IBD	Integrase binding domain
IN	Integrase
INI1	Integrase interactor 1
LADs	Lamina associated domains
LEDGF/p75	Lens epithelium-derived growth factor and co-factor p75
LINE	Long interspersed nuclear element
LRAs	Latency reversal agents
LTR	Long terminal repeat
MA	Matrix
MHC I	Major histocompatibility complex class I
MLV	Murine leukemia virus
MMTV	Mouse mammary tumor virus
MoMLV	Moloney murine leukemia virus
MRC	Matched random control
NC	Nucleocapsid
NFAT	Nuclear factor of activated T-cells
NF-κB	Nuclear factor-kappa B
NMR	Nuclear magnetic resonance
NRLIC	Non-reactivated latently-infected cells
NRTIs	Nucleoside reverse transcriptase inhibitors

NTD	N-terminal domain
PBLs	Peripheral blood lymphocytes
PBMCs	Peripheral blood mononuclear cells
PCR	Polymerase chain reaction
PIC	Productively-infected cells
PR	Protease
pTEFb	Positive transcription elongation factor b
PWWP	Proline-Tryptophan-Tryptophan-Proline
RLIC	Reactivated latently-infected cells
RNA	Ribonucleic acid
RT	Reverse transcriptase
SAHA	Suberoylamide hydroxamic acid
SINE	Short interspersed nuclear element
SIV	Simian immunodeficiency virus
STR	Short tandem repeat
TFO	Triplex forming oligonucleotide
TMPyP4	5, 10, 15, 20-tetra (N-methyl-4-pyridyl) porphin
TNPO3	Transportin3
TSS	Transcription start site
UNAIDS	The Joint United Nations Programme on HIV and AIDS

URF	Unique recombinant form

Chapter 1

1 Introduction

Viruses are obligate parasitic microganisms that can hijack host cellular pathways and machineries for their replication and persistence; retroviruses are no exception. The *Retroviridae* or retrovirus family encompass a diverse group of small enveloped viruses capable of spreading and causing severe diseases. All retroviruses have a positive sense single-stranded ribonucleic acid (RNA) genome ranging from 7 to 12 kilobases (kb) in size [1]. This family of viruses is divided into 7 genera that include: the *alpha-*, *beta-*, *gamma-*, *delta-*, *epsilon-* retroviruses, the *spumavirus* and the *lentivirus* [1,2]. Retroviruses are further classified into 2 categories comprising the simple and complex retroviruses. The main difference between the simple and complex retroviruses lies in their genomic organization [1]. More specifically, simple and complex retroviruses encode for three major polyprotein genes: the group specific antigen *(gag),* the polymerase *(pol),* and the envelope *(env)* gene [1,2]. However, contrary to the simple retroviruses, the complex retroviruses code for other regulatory and accessory genes in addition to the three major genes [1,2]. **Table 1. 1** gives a list of identified retrovirus genera with examples of species for each. Retroviruses have a unique life cycle that involves conversion of their genomic RNA into linear double-stranded deoxyribonucleic acid (DNA) and integration of the double-stranded DNA into the chromosomal host DNA [1,2]. These steps of their life cycle are hallmarks of the *Retroviridae* family. Additionally, the ability of retroviruses to permanently integrate their viral DNA into the chromosomal host DNA allows these viruses to maintain a persistent life-long infection within diverse vertebrate organisms [1]. One of the most studied and clinically prevalent retroviruses is the human immunodeficiency virus (HIV). HIV is a complex retrovirus belonging to the lentivirus genus. Lentiviruses represent a genus of viruses that cause slow and chronic disease. HIV is the causative agent of Acquired Immunodeficiency Syndrome (AIDS) a chronic disease characterized by the depletion of $CD4^+$ T-lymphocytes ($CD4^+$ T cells) [3].

Table 1.1: List of retrovirus genera (adapted from reference [2]).

Genus Name	Species Examples	Genome Characteristic
Alpharetrovirus	Avian sarcoma leukosis virus Avian myeloblastosis virus Rous Sarcoma virus	Simple
Betaretrovirus	Mason-Pfizer monkey virus Mouse mammary tumor virus Langur virus	Simple
Gammaretrovirus	Murine leukemia virus Moloney murine sarcoma virus Feline leukemia virus	Simple
Deltaretrovirus	Human T-lymphotropic virus 1 Human T-lymphotropic virus 2 Bovine leukemia virus	Complex
Epsilonretrovirus	Walleye epidermal hyperplasia virus 1 Walleye epidermal hyperplasia virus 2 Walleye dermal sarcoma virus	Complex
Spumavirus	Feline foamy virus Equine foamy virus Bovine foamy virus	Complex
Lentivirus	Human immunodeficiency virus type 1 Human immunodeficiency virus type 2 Simian immunodeficiency virus	Complex

Two types of HIV have been identified and are classified as HIV type 1 (HIV-1) and HIV type 2 (HIV-2) [3,4]. Both HIV-1 and HIV-2 share a similar genomic organization but differ in their pathogenicity. In fact, HIV-1 is the main agent of the HIV/AIDS pandemic while HIV-2 infection is confined to regions in Western and Central Africa [3]. Currently, more than 36 million individuals are infected with HIV-1 worldwide with approximately 2 million new infections occurring annually [5]. In this thesis, the focus will be on HIV-1 infection.

Since the discovery of HIV-1 in the early 1980's [3,6] the scientific community has made great efforts towards developing effective therapeutic drugs that control HIV-1 infections. However, advances in the development of combination antiretroviral therapy (cART) can only help control HIV-1 replication in infected individuals and fail to eradicate the virus [7].

Early during infection (within hours to days), HIV-1 may actively replicate leading to productive infection while in some cases, HIV-1 can become quiescent/latent [8,9,10,11]. HIV-1 viral latency is characterized by the low expression levels of viral transcripts (which is undetectable by most sensitive assays) or no expression of viral transcripts [12]. Therefore, in this thesis, HIV-1 latency is defined as having undetectable and no expression of viral transcripts/proteins.

Latent viruses can remain inactive for years without producing viral proteins. This allows latently infected cells to become undetectable by the immune system and escape cytopathic effect [12]. Additionally, cART is only effective against replicating viruses and are ineffective against latent viruses [12]. However, latent viruses can replicate and produce infectious particles when cART treatment is discontinued [13,14]. Thus, a cure for HIV-1 infection requires the complete elimination of latently-infected cells. Latently infected cells present a challenge for HIV-1/AIDS eradication, which remains an incurable disease and a major public health concern worldwide. Previous studies reported an association between HIV-1 integration sites in the human genome and disease persistence/latency, but the mechanisms underlying this association are unclear [15]. Therefore, this thesis investigates the integration site selection profile primarily in the context of HIV-1 infection and how

integration site selection in the genome may contribute to a persistent live-long infection of the virus.

1.1 A brief history of HIV-1/AIDS

In the early 1980s, cases of a new human epidemic began to emerge. Infected individuals presented unusual symptoms of immune dysfunction [16]. In 1981, AIDS was recognized by the scientific community. AIDS manifested itself with a rapid decrease in the $CD4^+$ T cell count, usually below 200 cells/mm^3 [17]. During this stage, individuals succumbed to otherwise rare opportunistic infections and unusual cancers. Most notably, the same type of T cells are targeted by the human T-lymphotropic virus 1 (HTLV-1). HTLV-1 was isolated in 1980 by Dr. Robert Gallo and was reported as the first pathogenic human retrovirus [18]. However, HTLV-1 transforms $CD4^+$ T cells into T-cell leukemia and does not cause depletion of $CD4^+$ T cells. This suggested that a new, unknown retrovirus was responsible for the epidemic seen at the time. In 1983, Dr. Luc Montagnier and his colleagues at the Pasteur Institute isolated the virus from the lymph nodes of patients with acute lymphadenopathy [6] . The virus was first known at the time as the lymphoadenopathy-associated virus and was suspected to have been the cause of AIDS. One year later, Dr. Gallo and his collaborators at the National Institute of Health confirmed this new virus has been the causative agent of AIDS [19,20] . In 1986, the newly discover human retrovirus was officially termed HIV-1 [21].

1.2 The human immunodeficiency virus type 1 (HIV-1)

Following the discovery of HIV-1, great progress had been made in understanding more about HIV-1. Notably, these advances include a detailed understanding of the HIV-1 modes of transmission, pathogenesis, structure, complete sequencing of the HIV-1 genome and isolation of different HIV-1 subtypes.

1.2.1 Classification of HIV-1 groups and subtypes

With the rise of the polymerase chain reaction (PCR), amplification of viral genomes was made possible. This was followed by advances in genome sequencing that further helped establish the sequences of diverse HIV-1 isolates throughout the world [3]. The identified

HIV-1 isolates/strains are currently divided into three major groups. This include group M ("M" stands for main), group N ("N" stands for non M or O), group O ("O" stands for outlier) [3,22,23]. A new isolate that is divergent from the major groups has also been identified and is classified as group "P" [24] . The HIV-1 M group which constitutes 95% of all isolated HIV-1 strains is further subdivided into 9 distinct clades or subtypes [25]. The M group subtypes are designated A, B, C, D, F, G, H, J, and K. Viruses of the M group dominate most HIV-1 infections worldwide. Subtypes within the N group have not been fully determined. Nevertheless, only a few isolates of the N group have been sequenced [26]. On the other hand, no subtypes have been defined for group O and P. Additionally, recombinant forms of HIV-1 have also been isolated. Recombinant forms occur as a result of a recombination event between the genome of identical subtypes or different subtypes. These recombinant viruses are known as circulating recombinant forms (CRFs). As of 2018, more than 90 CRFs have been characterized [26].

1.2.2 HIV-1 virion structure and composition

The HIV-1 virion has an average diameter of 100 nm with a spherical to conical shape [4]. Each virion is surrounded by a host derived envelope membrane [23] **(Figure 1.1).** The envelope membrane anchors surface glycoproteins (gp120 and gp41) which aid viral entry [4,23]. The envelope membrane is further surrounded by an inner layer of the viral matrix (MA) proteins. Additionally, the envelope encases a cone-shaped core composed of the capsid (CA) proteins. The conical core capsid harbors the copies of viral RNA genomes and the nucleocapsid (NC) protein that form a complex with the viral RNA genomes [3,4,23,27]. The virion also encloses three essential viral enzymes: reverse transcriptase (RT), integrase (IN) and protease (PR)[23]. Accessory and regulatory proteins are also present within the virion. These include viral infectivity factor (Vif), virus protein R (Vpr), viral protein U (Vpu), negative regulator factor (Nef), and RNA splicing-regulator (Rev) proteins [3].

Figure 1.1

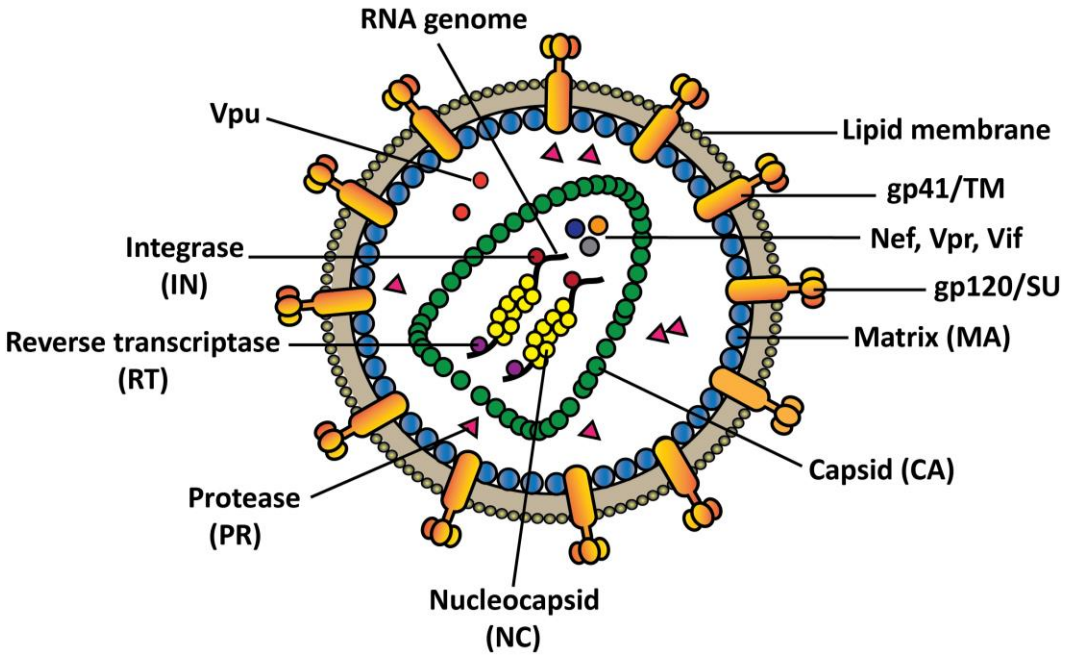


Figure 1.1: Illustration of the HIV-1 virion (adapted from reference [27]).

The size of the HIV-1 virion is ~100nm in diameter and is enveloped by a host-derived lipid membrane. The gp120-gp41 glycoprotein complexes are embedded in the lipid membrane. The matrix proteins line the inner membrane of the envelope. The capsids make up the conical core which contains the two single-stranded RNA genomes. The viral RNA genomes are surrounded by nucleocapsid proteins. Functional and accessory proteins such as integrase (IN), reverse transcriptase (RT), protease (PR), Vpr, Vif, Nef, and Vpu as well as host proteins are packaged into the virion.

1.2.3 HIV-1 genomic organization and gene functions

The HIV-1 genome consists of positive sense single-stranded RNA. The size of the HIV-1RNA molecules are about 9.2 kb. The HIV-1 genome contains 9 genes that encode 15 viral proteins. **Figure 1.2** shows a description of the HIV-1 genome.

The *gag* gene produces a polyprotein (Pr55Gag) that encodes all structural proteins and is proteolytically cleaved into the capsid (CA/p24), the matrix (MA/p17), the nucleocapsid (NC/p7) and the particle release protein (p6) [28]. The *pol* gene codes for 3 viral enzymes that are essential for viral replication. These enzymes include the RT, IN and PR. RT (p66/p51) is responsible for the reverse transcription of the viral RNA genome into DNA [28]. IN (p32) mediates the integration of the reverse transcribed viral DNA into the host DNA[29]. PR (p10) enzyme is essential for the cleavage of the polyprotein during maturation of the viral particle [28]. The *env* gene encodes the viral envelope glycoproteins 120 and 41 (gp120 and gp41). Both glycoproteins mediate viral entry into the host cell. Gp120 is the surface (SU) glycoprotein that mediates viral attachment to the target cell [28]. Gp41 is the transmembrane (TM) glycoprotein that anchors fusion of the viral and cell membrane [28]. These two proteins arise from the glycosylation and proteolytic cleavage of the full-length gp160.

As a complex retrovirus, HIV-1 encodes 6 other regulatory and accessory genes in addition to the three major genes. The regulatory genes include transactivation of transcription *(tat)* and RNA splicing-regulator *(rev)*. The Tat protein induces an increase in transcription and promotes full-length elongation of the viral transcripts [30,31]. The Rev protein helps facilitate the transport of unspliced and incompletely spliced messenger RNAs from the nucleus to the cytoplasm [32]. The accessory genes of HIV-1 are the viral infectivity factor *(vif)*, the virus protein R *(vpr)*, the virus protein U *(vpu)* and the negative regulator factor *(nef)*. Examples of some of the functions of each accessory protein are described as follows: Vif (p23) is the accessory protein that modulates and enhances HIV-1 infectivity in certain target cells such as lymphocytes and macrophages [33]. Vpr (p15) facilitates the transport of the viral DNA in the nucleus for integration [34–36]. Vpr also promotes cell cycle arrest at the G2 phase [37]. The Vpu (p16) protein enhances viral release during budding [38] and also mediates degradation of CD4 through ubiquitination [39].

Figure 1.2

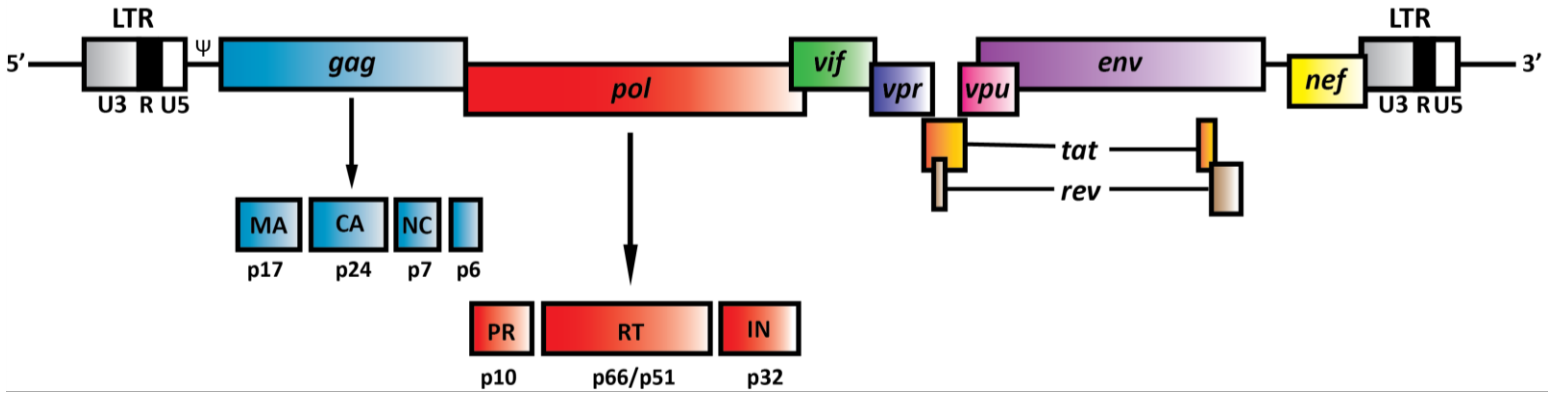


Figure 1.2: HIV-1 proviral DNA structure (adapted from reference [27]).

The HIV-1 genome has 9 open reading frames that code for 15 viral proteins and is flanked by the long terminal repeat (LTR) at both 5' and 3' ends.

Degradation of CD4 helps release the Env protein from CD4-env complex in the endoplasmic reticulum (ER). Nef (p27) mediates down regulation of cell surface expression of CD4 and the major histocompatibility complex class I (MHC I) [40,41]. Nef is also involved in modulating HIV-1 replication and enhances infectivity of the virion [42].

Furthermore, the integrated proviral DNA is flanked at both 5'and 3'ends of the viral genome by the long terminal repeat (LTR) region. The LTR regions contain promoter sites, enhancers sites, transcription termination sites/polyadenylation signal and other regulatory signals that interact with the host transcriptional machinery. Each of the LTR sequences is composed of 3 regions that include the U3 (3' unique), R (repeated sequence) and the U5 (5'unique) region [43]. More specifically, the 5'LTR contains binding sites for cellular transcription factors, enhancers and cellular RNA polymerase [28]. The 5' LTR region also contains the promoter binding site for the HIV-1 Tat protein [28] as well as the viral RNA packaging signal sequence denoted as the ψ (Psi) signal [44]. On the other end, the 3'LTR acts as a transcription termination and polyadenylation site [45,46].

1.2.4 HIV-1 replication and disease progression

HIV-1 infection starts with the attachment of the virion to its target cell and subsequent fusion of the viral and target cell membrane **(Figure 1.3)** [47]. The viral envelope, that harbors the gp120 and gp41 heterodimer, mediates viral entry into the host cell. Viral entry is initiated by the surface envelope glycoprotein gp120 through binding to its cellular receptor CD4 [48–50]. Specifically, gp120 binds the CD4 receptor via its C4 domain [51]. Host cells that are CD4 positive such as helper T cells, macrophages, dendritic cells, microglial cells and astrocytes are targets for HIV-1 infection. Following this initial attachment, the gp120 undergoes a conformational change exposing a conserved region within the third variable loop in the gp120 [52]. This allows binding of gp120 to its co-receptor. The co-receptors for HIV-1 infection are chemokine receptor type 5 (CCR5) or chemokine receptor type 4 (CXCR4) [53]. The binding of gp120 to CD4 and to the co-receptor induces an additional structural change in gp120 with a subsequent change in gp41 conformation. These changes in gp41 conformation lead to its insertion into the cell membrane via it's N-terminal fusion peptide, and fusion of the host cell membrane and viral envelope occurs [54].

Figure 1.3

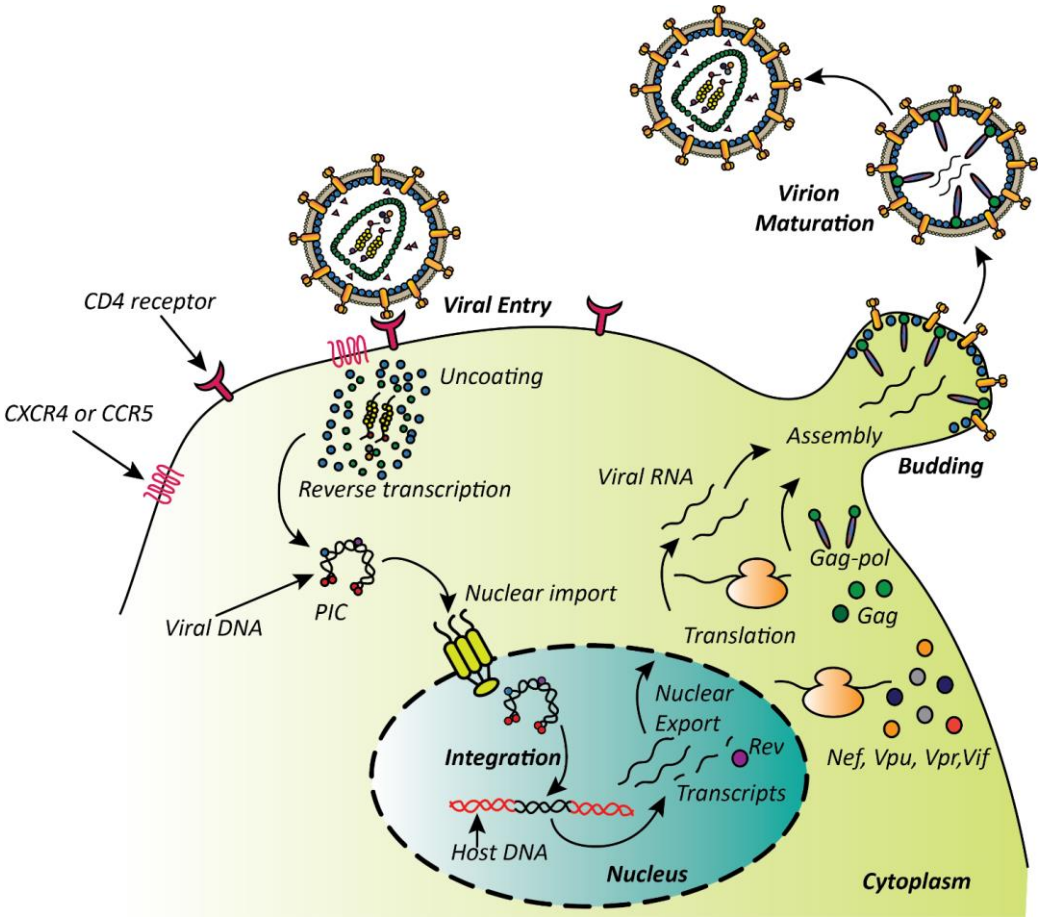


Figure 1.3: HIV-1 replication cycle (adapted from reference [47]).

HIV-1 infection begins with attachment of the gp120 and gp41 to the CD4 cell surface membrane receptor and the CXCR4 or CCR5 co-receptor, respectively. Following entry into cells and uncoating of the core shell, the viral RNA is reverse transcribed into the double-stranded viral cDNA. The viral cDNA interacts with other host and viral proteins forming the pre-integration complex. The pre-integration complex then gets imported into the nucleus. The viral integrase enzyme facilitates the integration of the viral DNA into the host DNA. Following integration, more viral RNAs are transcribed and are translated into viral proteins. Viral RNAs and viral proteins assemble at the plasma membrane. The immature virions bud from the cell. The viral polyproteins are proteolytically cleaved generating mature virions that are able to infect new cells.

Following this fusion event, the viral core is translocated into the cytoplasm with a subsequent uncoating of the capsid and release of the viral RNA genome and proteins in the cytoplasm. In the cytoplasm, the viral RNA is reverse transcribed into a double-stranded complimentary DNA (cDNA) via RT activity. The viral cDNA then associates with the IN and several viral and host proteins forming the pre-integration complex [55,56] .

After the formation of the pre-integration complex, it is actively transported into the nucleus through microtubules [57] and microfilaments [58]. In the nucleus, the viral DNA gets incorporated into the host cell's genome via the viral IN in a process called integration. The integration process will be further described in section 1.4.1. The permanently integrated viral DNA is referred to as a provirus. Once integrated, the proviral DNA can become actively expressed for the production of viral progeny or remain silent during latent infection. All viral genes are transcribed by the host RNA polymerase II and initiate in the LTR. Completely spliced messenger RNAs of the *tat* and *rev gene* are expressed during the first stage of infection. This event is followed by the expression of incompletely spliced messenger RNAs encoding the *env, vpr, vif* and *vpu* genes [59].

Later during infection, full-length unspliced messenger RNAs encoding the full-length viral RNA and the Gag-Pol polyprotein are transcribed. Transport of unspliced and incompletely spliced transcripts into the cytoplasm is mediated by the Rev protein [32,59]. Lastly, newly generated viral particles bud from the host cell membrane followed by maturation of the particles due to PR cleavage of the Gag and Gag-Pol polyproteins. Mature viral particles can now infect other cells.

1.2.5 The course of HIV-1 infection

The immunological and virological factors used to determine HIV-1 progression in infected individuals are the $CD4^+$ T-cell count and the RNA viral load in the plasma [60]. HIV-1 disease progression can be divided into 3 stages. These include: 1) acute infection, 2) chronic asymptomatic stage, and 3) AIDS [61,62]. These stages are usually seen in patients not receiving anti-retroviral treatment [17]. Acute infection is characterized by a drastic increase in the level of circulating virus and a decrease in $CD4^+$ T-cell count in the blood and peripheral lymphoid tissues. The acute stage usually occurs during the first 2-10 weeks

of infection [17]. Following this period of primary infection, there is a decline in viral load and an increase in $CD4^+$ T-cells [17]. The decline in viral load occurs due to self-limiting infection and an elevated immune response/rise in $CD8^+$ T-lymphocytes ($CD8^+$ T-cells) levels; however, the virus is not fully contained by the immune response [63].

This is then accompanied by a chronic or asymptomatic stage that can last 7-10 years without the patient exhibiting major symptoms of disease progression [17]. Although no apparent symptoms occur, the virus still replicates and infects new cells causing a progressive decline in $CD4^+$ T-cells. Decline in $CD4^+$ T-cells may result from cell death during productive infection *in vivo* and *in vitro* [64–66] . Additional, it has been reported that a decline in $CD4^+$ T-cells may be due to pyroptosis of non-productively infected cells as demonstrated *in vitro* [67,68] . After the chronic stage, an increase in the viral load occurs and the level of $CD4^+$ T-cells drops below 200 cells/mm^3 which can lead to the onset of opportunistic infections and is characteristic of AIDS progression [17].

1.3 Antiretroviral therapy and HIV-1 persistence

1.3.1 Antiretroviral therapy

Antiretroviral therapy was first introduced in the early 1990's [7]. Azidothymidine (AZT)/ Zidovudine mono-therapy, a nucleoside reverse transcriptase inhibitor, was the treatment of choice to prevent HIV-1 replication and slow disease progression. However, viral replication occurs rapidly with a high mutation rate of the virus leading to the occurrence of drug resistant viruses [69]. Consequently, mono-therapy became quickly ineffective as HIV-1 became resistant to treatment [70,71]. The current treatment option involves combinations of antiretroviral drugs commonly known as combination antiretroviral therapy (cART). These combinations of drugs simultaneously target different stages of the virus life cycle, thus optimizing their effectiveness. There are ~30 approved antiretroviral drugs categorized as: 1) reverse transcriptase inhibitors that consist of nucleoside or non-nucleoside inhibitors, 2) integrase inhibitors that interfere with the strand transfer activity of the viral integrase enzyme, 3) protease inhibitors and 4) viral entry inhibitors such as fusion inhibitors, CCR5 co-receptor antagonists and attachment inhibitors [7].

cART has been more effective in suppressing viral replication than AZT mono-therapy in general as the viral load drops below 50 RNA copies/ml while reducing the mortality rate related to HIV-1/AIDS [7]. This led to the hope that cART could potentially eradicate the virus. Nonetheless, once treatment is discontinued the virus can replicate and produce infectious particles leading to a rapid rebound in viremia [72]. Unfortunately, it is now evident that cART cannot completely clear the virus from infected individuals. In fact, cART is mostly effective against replicating virus and preventing new infection of cells [73]. This further confirms that a replication competent quiescent/latent reservoir of infected cells exist and can persist despite therapy.

A viral reservoir could be defined as any subset of cells or anatomical sites that harbor a replication competent form of the virus that persists for a very long time compared to the pool of actively replicating virus [10]. The main cellular reservoir of infected latent cells are $CD4^+$ memory T cells. This will be further discussed in section 1.3.2. Macrophages are also potential latent cellular reservoirs of HIV-1 [74]. More on anatomical reservoirs will be discussed in section 1.3.3.

1.3.2 HIV-1 viral latency

HIV-1 viral latency is characterized by the low expression levels of viral transcripts or no expression of viral transcripts where HIV-1 can remain in a long-lived quiescent /latent state within infected cells [10,11,12]. Latency mainly occurs as a result of a transcriptional block in HIV-1 expression and is characterized by little to no detectable expression of viral transcripts as previously described [12,75]. Chromatin modifications and epigenetic regulations can also lead to HIV-1 latency. These multiple restrictions on HIV-1 expression are further described below.

HIV-1 latency can occur in two distinct forms: pre-integration or post-integration latency [76]. It is unclear how early latency is established. However, it was shown that early administration of cART (within 10 days) following the occurrence of symptoms related to primary infection could not prevent the production of latently infected cells in infected

individuals [8]. Additionally, studies from nonhuman primate suggested that latency can occur as early as 3 days post infection despite early administration of cART [9].

During pre-integration latency, the virus enters non-dividing cells where reverse transcription of the viral RNA genome occurs. The viral DNA only gets integrated into the host cell genome when those non-dividing cells become activated [76]. However, since the pre-integrated complex has a very short half-life of ~1 day, pre-integration latency of the viral DNA is less likely to be the major contributing factor to the long-term persistence of HIV-1 infection [77,78,79,80].

Post-integration latency results from the viral DNA integrating into the host genome where viral gene expression is impeded. Contrary to pre-integration latency, post-integration latency is highly stable and can persist for a life-time. The best characterized reservoir for post-integration latency are the resting memory $CD4^+$ T cells. In their resting state, these cells have a very low metabolic rate and are transcriptionally inactive. Therefore, the integrated proviral DNA can remain transcriptionally silent and the infection is not targeted by the immune system or cART [81,82]. Upon activation of the infected resting memory cells, viral production can resume as latency is reversed [83].

HIV-1 latency is thought to be first established when activated $CD4^+$ cells get infected. Some of the infected and active $CD4^+$ cells that are not killed by the cytopathic effects of viral replication and the immune system revert back as resting memory cells [82]. The result is a stably integrated latent virus. Additionally, resting memory $CD4^+$ T cells is a stable latent reservoir for HIV-1 infection. The slow decay rate and long half-life of infected $CD4^+$ T cells contribute to the stability of the latent reservoir [84,85].

Another contributing factor to the stability of the latent reservoir involves the proliferation of infected cells [86,87,88]. T cells proliferation or expansion is usually driven by different stimuli such as antigen and cytokine driven homeostatic proliferation [89]. Antigen-induced proliferation leads to a rapid and transient cell division and amplification of T cell clones/clonal expansion in response to activation [90,91]. Once antigen exposure is cleared during antigen-induced proliferation, the majority of T cells die with a small subset that

revert into resting memory T cells. Homeostatic proliferation leads to clonal expansion which contributes to the persistence of the viral reservoir [90,91]. Homeostatic proliferation is a process driven by cytokines that is important for the normal maintenance of size and diversity in the total pool of T cells which enables T cell clones to either maintain their numbers or expand over time. Recent work further revealed the presence of identical HIV-1 integration site positions in the human genome within a large portion of infected $CD4^+$ T cells [86,87,92,88]. Interestingly, clonally expanded cells showed integration into genes associated with cell proliferation and growth [87,88] . Expanded cells were also shown to carry replication-competent latent virus *in vivo* [93].

Although, HIV-1 latency is first established through infection of activated $CD4^+$ cells before they revert into memory cells, the molecular mechanisms that maintain latency are not well understood. It is suggested that HIV-1 latency is a multifactorial process and is thought to be maintained by: 1) the site and orientation of integration within actively transcribed genes that can interfere with HIV-1 gene expression [94,95]; 2) epigenetic changes in chromatin structure that prevent the action of transcription factors on the HIV-1 promoter region [82,75]; 3) sequestration of cellular factors such as the nuclear factor-kappa B (NF-κB) and nuclear factor of activated T-cells (NFAT) which are essential for HIV-1 transcription and are sequestered in the cytoplasm due to the absence of signaling in resting $CD4^+$ T cells [75,96]. The positive transcription elongation factor b (pTEFb) which associates with HIV-1 Tat protein to promote elongation of the viral transcripts is also sequestered in resting $CD4^+$ T cells by cellular regulatory complexes [75,97] ; and 4) microRNA which may bind to the viral messenger RNA and prevent viral translation [98] .

1.3.3 Sanctuary reservoirs

HIV-1 may persist in anatomical sites or compartments where replication can still occur due to the limited penetration of cART in these sites. Throughout untreated infection, the majority of HIV-1 infection occurs in the lymphoid organs such as the gut associated lymphoid tissue (GALT), the lymph nodes and the spleen [99,100]. In the GALT, a high frequency of infected cells was observed compared to infection in the circulating blood despite long-term antiretroviral therapy [101]. This was further associated with cross-

infection between the GALT compartment and the blood [101]. This suggests that the GALT is a reservoir for HIV-1 infection. Other anatomical sites such as the central nervous system (CNS), and the genitourinary (GU) tract can also be sites of HIV-1 infection. In the CNS, HIV-1 primarily infects perivascular macrophages and microglial cells [102,103]. In untreated patients, there is clear evidence that infection in the CNS is compartmentalized. Specifically, virus isolated from the peripheral blood is distinct from those isolated in the CNS/ cerebrospinal spinal fluid [104]. In the GU, HIV-1 has been found in the seminal fluid either as unintegrated and integrated virus in latently infected cells [105–107].

1.3.4 Targeting the latent reservoir

HIV-1 latency is multifactorial. Multiple strategies have been proposed to reactivate the latent reservoir. The most explored strategies are the "Shock and kill" methods. These methods involve reactivation of the latent virus using latency reversal agents (LRAs) which allows depletion of the virus through the immune response against infection or therapeutic means. Some of the earlier LRAs that were investigated involved the use of anti-CD3 antibodies, interferon gamma and interleukin 2 to induce immune activation; however, these approaches were ineffective as patients experienced global T cell and cytokine activation [108]. Other methods of reactivation involved the use of histone deacetylase inhibitors (HDACi) such as suberoylamide hydroxamic acid (SAHA/vorinostat), romidepsin and panobinostat that counteract chromatin mediated repression [109–111, 112,113–115]. Phorbol esters such as bryostatin-1 and prostratin that induce HIV-1 transcription via activation of the host cellular protein kinase C pathway were also used [116–118]. Disulfiram is another LRA that depletes expression of the phosphatase tension homolog (PTEN) protein which leads to the activation or the phosphorylation of the protein kinase B (Akt) signaling pathway [119] . Activation of AkT signaling pathway results in HIV-1 reactivation in an NF-κB-dependent manner [119]. Additionally, Toll-like receptor (TLR) agonists have been used to activate HIV-1 expression in latently infected cells [120]. GS-9620 is a TLR7 agonist that was shown to reactivate HIV-1 expression from cells of infected individuals on suppressive cART [121] . Other TLR agonists that has been investigated *in vivo* are the CPG 7909 and MGN1703 which are both TLR9 agonists [122,123]. Combination of current LRAs have been administered simultaneously with the hope of enhancing reactivation of

latently infected cell [124,125] . However, both single use of LRA and combination of LRAs fail to reactivate the entire pool of latently-infected cells, thus failing to completely purge the virus. A major factor contributing to this failure is genomic location-driven differences in HIV-1 expression. These findings suggest that integration location has the potential to confound these anti-latency treatments and that integration site selection may be a major contributing factor to latency by influencing proviral gene expression.

1.4 HIV-1 integration process and the viral integrase structure

HIV-1 must integrate its newly synthesized DNA into the chromosomal DNA for successful production of new viral progeny. Integration is a permanent event. Thus, the viral genetic information can be transferred into daughter cells during cell division. Furthermore, the ability of HIV-1 to integrate into the chromosomal DNA presents a great challenge for eradication. Once integrated, HIV-1 persists and establishes a latent reservoir of infected cells that cannot be eliminated with cART. The viral IN enzyme is the key enzyme that catalyzes the integration reaction.

The integration process is common to other retroviruses and involves 3 major steps: 1) 3’ processing of the reverse transcribed viral DNA, 2) DNA strand transfer or joining of viral DNA to the target DNA, 3) end repair process [126,127]. It should be noted that the IN alone cannot execute the entire integration process. In fact, IN is part of the pre-integration complex where the viral DNA associates with other viral and host factors including the IN itself.

1.4.1 The integration reaction

Figure 1.4 illustrates the 3 main steps of the integration process. Integration into the host genome is initiated through recognition of the viral DNA by the IN in the cytoplasm. First, the IN binds to the viral DNA at the viral attachment (att) sites located on both 5’and 3’ ends of the LTR of the viral DNA[55,128,129]. The IN removes 2 to 3 nucleotides, usually pGT nucleotides from both 3’ ends of the LTR at the complimentary CA dinucleotide site [130,131,132]. This reaction is known as 3’-processing and occurs in the cytoplasm within the pre-integration complex. The 3’processing reaction exposes a 3’ hydroxyl (3’-OH) group

at both ends of the viral DNA[55] . As mentioned in section 1.2.4, the viral DNA associated with IN and other factors forms the pre-integration complex, which is transported into the nucleus for the strand transfer reaction to take place. The IN mediates a nucleophilic attack by the 3'-OH groups on phosphodiester bonds of the target DNA. This results in the cleavage of the target host DNA and the simultaneous 3'end joining of the viral DNA and the 5'end of the target DNA [55,133,132]. The integration sites on the two strands of the target DNA are separated by 5 nucleotides leaving single-stranded gaps.

Following the strand transfer reaction, two unpaired/overhanging nucleotides at the 5' ends of the viral DNA are removed. The single stranded gaps are filled and ligated, possibly by DNA damage repair enzymes [134,135].

1.4.2 HIV-1 integrase structure

The HIV-1 IN protein is the main player for catalyzing the integration process and is a 288 amino acid protein (32kDa) that is proteolytically cleaved from the Gag-Pol polyprotein precursor [136]. The IN is composed of three major structural and functional domains: the N-terminal domain (NTD), the C-terminal domain (CTD) and a central catalytic core domain (CCD) **(Figure 1.5)** [136,55]. The structure of the HIV-1 IN domains have been characterized by X-ray crystallography and nuclear magnetic resonance (NMR) spectroscopy [137,138].

The NTD structure (amino acid 1-49) consists of 4 α-helices coordinated by conserved histidine and cysteine amino acid residues. Specifically, His12, His16, Cys40 and Cys43 (HH CC) binds to a single zinc ion (Zn^{2+}) stabilizing the folded alpha helical structure of the NTD, which also promotes multimerization of the IN for its activity [139,140].

The CCD (amino acid 50-212) is composed of 6 α-helices and 5 β-sheets [141] and is conserved among the different retroviruses. Additionally, the CCD consists of a triad of highly conserved amino acid residues commonly referred to as the D, D, E motif (Asp-64, Asp-116 and Glu-152).

The DDE residue of the CCD is the catalytic site of the IN enzyme. The DDE residues coordinates two divalent metal ions (e.g.: Magnesium/Mg^{2+} or Manganese/Mn^{2+}) that

Figure 1.4

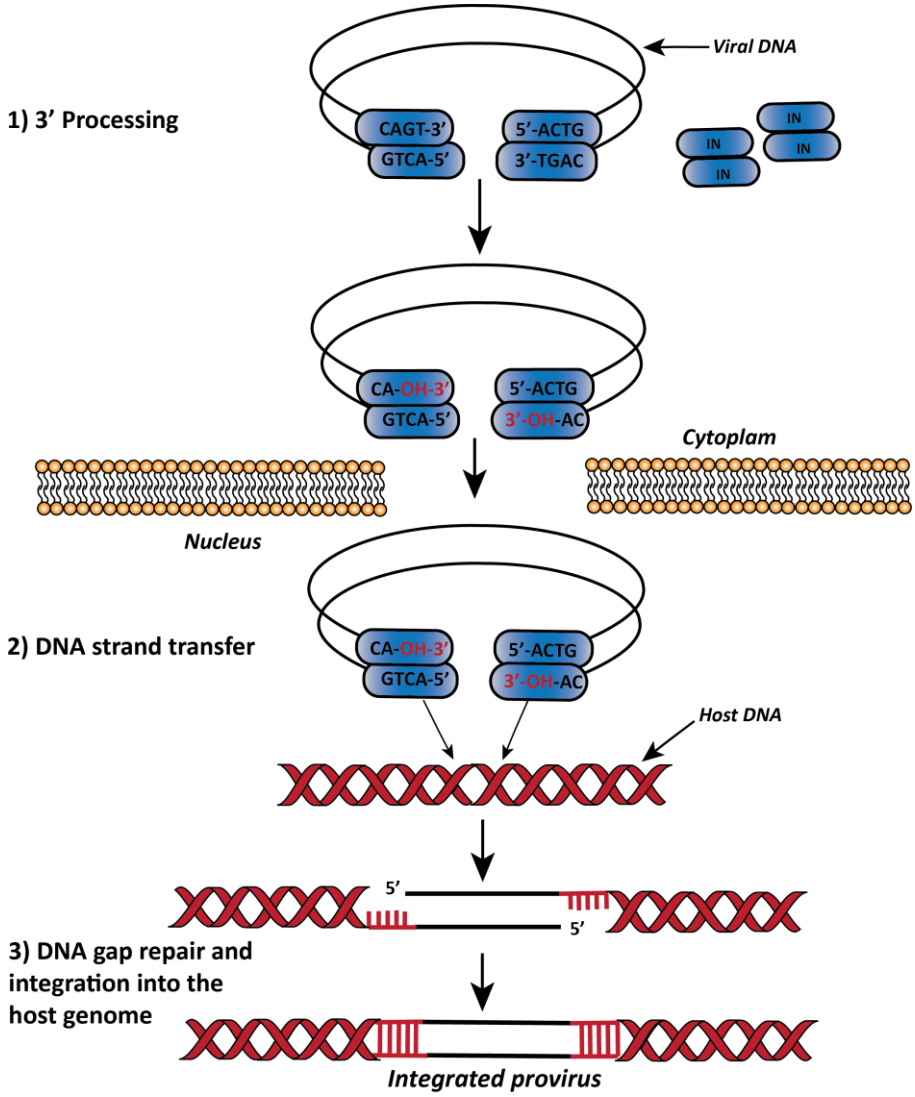


Figure 1.4: Steps of the integration reaction (adapted from reference [126]).

The reverse transcribed viral DNA associates with the integrase (IN) in the pre-integration complex. During the 3'processing reaction, the IN removes 2-3 nucleotides from both 3'ends of the viral DNA exposing hydroxyl groups. In the strand transfer reaction, the IN catalyzes 3' OH group nucleophilic attack on the host DNA. The 3' end of the viral DNA and the 5' end of the host DNA simultaneously link together. The unpaired gaps at the viral-host DNA junction are filled by host repair enzymes during the gap repair step. The fully integrated viral DNA is known as the provirus.

Figure 1.5

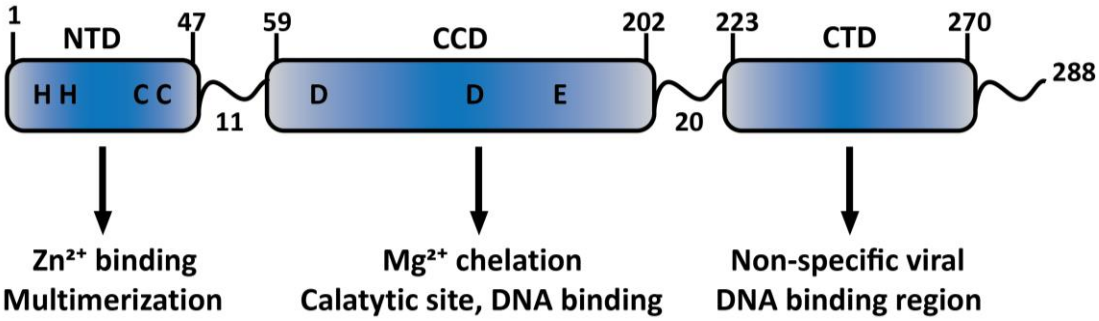


Figure 1.5: HIV-1 integrase structure (adapted from reference [127]).

HIV-1 integrase (IN) enzyme is a 288 amino acids protein. The IN is composed of three functional domains: the N-terminal domain (NTD), the catalytic core domain (CCD) and the C-terminal domain (CTD). The NTD contains a highly conserved HHCC motif (H for histidine and C for cysteine) that mediates Zn^{2+} binding. The CCD domain contains a conserved DDE motif that is part of the catalytic activity of the IN and binds to Mg^{2+}. The CTD is a less conserved domain and exerts non-specific DNA binding.

catalyze the 3'processing and DNA strand transfer step of the integration process [130,142,143]. Specifically, the metal ion binds the D64 and D116 residues of HIV-1 IN [143]. The CCD has also been shown to be involved in viral DNA recognition and target DNA binding [144,145]. On the other hand, the CTD (amino acids 213-288) is less conserved among the retrovirus family and is thought to be involved in DNA binding [131,146]. Overall, the HIV-1 IN can catalyze both the 3'processing and DNA transfer reaction alone and requires the use of metal ions for its activity. The role of the IN in the end repair reaction of the integration process has yet to be shown.

The retroviral integration process can be reproduced *in vitro* [147,148] . Integration studies were performed via isolation of the pre-integration complex from infected cells or purification of the IN [29,149]. The HIV-1 pre-integration complex, was shown to not only contain the IN, but also included the viral MA, RT, NC and Vpr proteins [56,150]. However, it was also found that a number of host cellular proteins may assist the virus during integration. In particular, host cellular proteins can have a profound role in integration site targeting. Currently identified host factors involved in integration are further discussed below in section 1.4.3.

1.4.3 Host proteins interacting with HIV-1 integrase

A number of approaches have been used to identify candidate host proteins that are binding partners of HIV-1 IN during integration. These approaches included the yeast-two hybrid assay for protein-protein interactions [151,152], co-immunoprecipitation[153,154], and the *in vitro* reconstitution of enzymatic activity of salt-stripped pre-integration complex [155,156].

1.4.3.1 Integrase interactor 1 (IN 1)

Through the application of the yeast-two hybrid assay, the first IN binding protein was identified [152]. This protein was named integrase interactor 1 (INI1) complex and is the human homolog of the yeast chromatin remodeling activator SNF5. SNF5 is a transcription activator and part of the chromatin remodeling SWI/SNF complex [157,158]. INI1 is comprised of three highly conserved regions. These regions comprise 2 direct imperfect

repeats, repeat 1 (Rpt1) and repeat 2 (Rpt2), a C-terminal coiled-coiled domain and a homology region [159]. The Rpt1 region was shown to be necessary and sufficient for the HIV-1 IN interaction [159]. *In vitro* studies also demonstrated that INI1 stimulates the strand-transfer activity of HIV-1 IN [152]. However, no strong evidence has been found to support INI1 function in HIV-1 integration *in vivo*. Other studies found that INI1 can also bind with different host and viral proteins such as the cMYC [160] and p53 [161] host proteins, the human papilloma virus E1 protein [162] and the Epstein-Barr virus nuclear antigen 2 protein (EBNA2)[163]. Other studies have demonstrated a potential role of INI1 during HIV-1 replication as well. Specifically, a fragment containing the minimal IN binding domain of INI1, located between residues 183-294, induced a substantial decrease in HIV-1 replication/production and release in a transdominant manner [164]. Moreover, INI1 was found to be incorporated into HIV-1 virions [164,165] and is necessary for the efficient production of infectious viral particles [164]. The Rpt2 region of INI1 was also shown to have a masked nuclear export domain [166]. Currently, it is still unclear whether INI1 is required for HIV-1 replication or if it is involved during integration *in vivo.*

1.4.3.2 The barrier-to-autointegration factor (BAF)

The barrier-to-autointegration factor (BAF) is another IN host binding protein that was first identified in Moloney murine leukemia virus (MoMLV) pre-integration complexes. BAF is a highly conserved 89 amino acid protein that can bind to double-stranded DNA [167,168]. BAF has also been known to condense DNA structure [169]. BAF was shown to prevent autointegration of the viral DNA thus averting suicidal integration [155]. It was proposed that BAF prevents autointegration by coating the viral DNA as well as inducing changes in the viral DNA structure through condensation of the viral DNA [155]. With the use of co-immunoprecipitation experiment and the use of anti-BAF antibodies, the presence of BAF was established in HIV-1 as BAF co-immunoprecipitated with the HIV-1 pre-integration complex [153]. BAF also restored HIV-1 integration activity in experiments where the pre-integration complexes lost their function following salt-stripped inactivation of the complex [170]. Additionally, it was shown that BAF associated with the lamina associated polypeptide LAP-2α, which is involved in chromatin and nuclear structure reorganization [171]. In addition, LAP-2α seems to assist BAF recruitment to the pre-integration complex

[172]. The fact that BAF also interacts with LAP-2α, suggests it potential role in chromatin reorganization. So far, BAF involvement in stimulating integration has only been observed in *in vitro* experiments.

1.4.3.3 High mobility group chromosomal protein A1 (HMGA1)

The high mobility group chromosomal protein A1 (HMGA1, formerly HMGI (Y)) is a non-histone DNA binding protein that can also interact with other proteins. HMGA1 is known to control transcription and modulate chromatin structure [156]. In the case of HIV-1, HMGA1 associated with HIV-1 pre-integration complexes following purification of the pre-integration complexes from infected cells [156]. Moreover, when HMGA1 was added to salt-stripped pre-integration complexes, recombinant HMGA1 restored the integration activity *in vitro* [156]. However, HMGA1 seems to show a lower stimulatory effect than BAF when added to salt-stripped pre-integration complexes to restore integration activity [156,170]. As a DNA binding protein, it was proposed that HMGA1 will interact with the viral DNA, thus bringing both LTR ends into close proximity and enabling IN binding by unwinding the ends of the viral DNA [173,174]. However, other studies suggested that HMGA1 is not required for retroviral integration [175]. Hence, a role for HMGA1 in HIV-1 integration is still a matter of debate.

1.4.3.4 The lens epithelium-derived growth factor and co-factor p75 (LEDGF/p75)

The lens epithelium-derived growth factor and co-factor p75 (LEDGF/p75) is a 76 kDa transcriptional regulatory protein and a member of the hepatoma-derived growth factor (HDGF) family. LEDGF/p75 is a ubiquitously expressed nuclear protein and mainly functions in cell growth and protecting cells from stress-induced cell death [176,177]. LEDGF/p75 accomplishes its protective function by transcriptionally activating anti-apoptotic genes and stress related proteins, such as the heat shock proteins [177]. LEDGF/p75 is widely accepted as a binding partner of HIV-1 IN. Through co-immunoprecipitation studies, LEDGF/p75 was found to interact with HIV-1 IN in cells overexpressing IN [154]. This interaction was further confirmed by another study through yeast two-hybrid experiments [151]. LEDGF/p75 also stimulated the IN strand-transfer activity by binding to

the IN [178]. Most importantly, LEDGF/p75 interaction with IN was mapped to a conserved ~80 amino acids residues at the C-terminus of LEDGF/p75, which has hence been named the integrase binding domain (IBD) [178–180]. The IBD of LEDGF/p75 interacts with the CCD and NTD of IN. More specifically, the CCD of IN is sufficient for this interaction. However, additional binding of IBD to the NTD of IN increased the affinity of the interaction. These interactions were confirmed via protein crystallography [181,182]. The N-terminus of LEDGF/p75 functions as a chromatin binding region and contains several domains: 1) the PWWP (Pro-Trp-Trp-Pro) domain that functions as a protein-protein and/or DNA-binding domain [183], 2) a nuclear localization signal [184] and 3) a AT-hook binding domain [185]. As such, the N-terminus of LEDGF/p75 binds to chromatin and its C-terminus interacts with IN. Thus, LEDGF/p75 has been shown to function as a tethering factor that may recruit the IN and other IN binding partners to the chromatin [154,179] . LEDGF/p75 will be further discussed in section 1.5.3.

1.4.3.5 Other HIV-1 integrase binding proteins

The transportin 3 (TNPO3) protein was identified to be a binding partner of HIV-1 IN via yeast-two hybrid experiments [186,187]. TNPO3 appeared to be essential in facilitating the transport of the pre-integration complex into the nucleus [186,187]. However, subsequent studies reported that the nuclear transport of the pre-integration complex likely functions through TNPO3 interaction with the viral CA and not with the IN [187,188]. Additionally, the host DNA repair protein Ku70, was shown to directly bind to HIV-1 IN [189,190]. This interaction was shown to protect the IN from proteosomal degradation by preventing the IN from ubiquitination [190]. A decrease in HIV-1 integration and replication was also reported following depletion of the Ku70 protein [190]. Ku70 protein was further detected in HIV-1 virions [190]. Additionally, cleavage and polyadenylation specificity factor 6 (CPSF6), a 68 kDa protein member of the pre-messenger RNA splicing factors, has been reported to be involved in the transport of the pre-integration complex from the cytoplasm to the nucleus through its interaction with the viral CA protein[191,192, 193] . Truncation of the C-terminal domain of CPSF6 impeded the nuclear transport of the pre-integration complex [193,194] .

1.5 Genomic profile of HIV-1 integration and factors affecting HIV-1 integration site selection

Understanding HIV-1 integration site selection is of paramount importance especially when integration site selection can influence proviral gene expression and latency. Previous studies proposed that specific host DNA sequences could act as a target for HIV-1 integration. Thus, the DNA sequence adjacent to the integrated virus was assessed. Using cell line models, it has been found that HIV-1 preferentially integrates within the transcription units of active genes [195,196]. These integrations sites are associated with regions of high G/C content, high gene density, high CpG island density, short introns, high frequencies of Alu repeats, low frequencies of long interspersed nuclear element (LINE) repeats, and characteristic epigenetic modifications [195,196,197,198,]. Integration in active transcription units has been shown during acute infection in different cells types [195,196]. It is important to note that integration site preference differs among the retroviral family. For example, the gammaretrovirus murine leukemia virus (MLV) is primarily found integrated at transcription start sites and CpG islands [199,200]. In contrast, alpharetroviruses, such as the avian sarcoma leukosis virus (ASLV), deltaretroviruses, such as the human T-lymphotropic virus 1 (HTLV-1), and betaretroviruses, such as the mouse mammary tumor virus (MMTV), showed no preference for integration within transcription units [201,202,203,204]. Multiple mechanisms/models have been proposed to address integration site selection. The following three mechanisms/models, none of which are mutually exclusive, have been proposed to address integration site selection: 1) chromatin remodeling/accessibility model 2) the cell cycle model and 3) the host factors/proteins tethering model.

1.5.1 Chromatin remodeling and accessibility model

In the nucleus, eukaryotic DNA is tightly wrapped around histones thus forming chromatin structures and complexes known as nucleosomes. DNA structure has a propensity to change during transcription and cell cycle phases, allowing host factors to interact with the DNA. Therefore, it seemed likely that the virus will integrate in regions that are more accessible, such as in euchromatin regions. It was then suggested that DNA wrapping into nucleosomes will alter its accessibility to the pre-integration complex, thus influencing

integration site selection. In fact, it was found that DNA compaction around nucleosomes creates hotspots for integration at sites of DNA distortion/bending [205–208]. Retroviral integration was also favored on the major grooves of the DNA [206]. More specifically, integration was predicted to occur on the major grooves of DNA facing outward from the nucleosome [197]. Although the role of DNA wrapping around the nucleosome has been demonstrated to influence and facilitate integration, chromatin accessibility cannot solely explain the difference in integration site selection among different retroviruses such as HIV-1 and MLV.

1.5.2 Cell cycle model

As a lentivirus, HIV-1 is capable of infecting non-dividing as well as dividing cells [209,210,211]. Infection of non-dividing cells can be accomplished through the active nuclear import of the HIV-1 pre-integration complex [209]. Contrary to HIV-1, a lentivirus, gamma-retroviruses such as MLV can only infect dividing cells [212]. Thus, MLV requires the disruption of the nuclear envelope to integrate it viral DNA into the host DNA. As such, it was proposed that cell division/mitosis could contribute to the differences in integration site selection seen between HIV-1 and MLV. Since remodeling of the chromatin occurs during DNA replication, it was further suggested that cell division could lead to an increase in integration into certain regions as opposed to other sites.

To investigate this hypothesis, studies were performed to assess the integration site distribution in non-dividing and dividing primary lung fibroblasts cells [195]. The integration profile into other non-dividing cells such as human macrophages was also assessed [213]. It was shown that cell cycle stage did not have a major effect on HIV-1 site distribution [195,213]. As an alternative, it has been proposed that cellular host proteins that bind to the pre-integration complexes and the chromosome act as tethering factors for the pre-integration complexes.

1.5.3 Host factors/proteins tethering model

In this model, it has been proposed that cellular host proteins would interact with the pre-integration complex thus targeting the pre-integration complex to specific regions of the host chromatin. As previously discussed, HIV-1 IN which is also part of the pre-integration

complex interacts with several host factors such as LEDGF/p75. LEDGF/p75 is the only known *bona fide* tethering factor of HIV-1 and other lentiviruses. As a chromatin-associated protein, LEDGF/p75 was shown to be involved in targeting HIV-1 integration within actively transcribed regions/transcription units[214–216].

The role of LEDGF/p75 as a determinant for HIV-1 integration site selection was further confirmed through knockdown studies. Knockdown of LEDGF/p75 in human cell lines led to a significant decrease in integration in transcription units and HIV-1 replication [215–217]. However, in the absence of LEDGF/p75, HIV-1 integration was redirected to CpG island and transcription start sites [217]. This new integration site selection is similar to the integration site targeting of gamma-retroviruses. This implies that yet other host factors are involved in integration at the alternative chromosomal locations.

1.6 Non-B DNA structures are new factors influencing HIV-1 integration site targeting

Primary sequences at around 5-10 bases immediately flanking HIV-1 integration were used to determine the sequence region surrounding integration site targeted by HIV-1. Through these *in vitro* integration site assays, it was found that these short primary sequences had only minor influences on HIV-1 integration site selection [218–220]. One major question that arose is whether analysis of a larger sequence window would provide more information on HIV-1 site selection.

We recently characterized the integration site of an HIV-1 based lentivector in the murine brain by analyzing a larger window surrounding integration sites, up to 40 bases downstream and upstream of the integration sites. We identified two strong consensus guanine-quadruplex forming motifs (G4 motifs; also known as tetraplex) flanking the integration sites [221]**.** These findings identified a new cis-acting factor affecting lentiviral/HIV-1 integration site selection. The G4 motif is a member of non-B DNA forming structures/ motifs. Non-B DNA motifs are DNA structures formed from non-canonical Watson-Crick base pairing with contorted bond angles or unpaired nucleotides compared with the orthodox B-DNA form [222].

Further analysis of data from our previous study showed that HIV-1 preferentially integrates in or near a variety of non-B DNA motifs in different cell lines including murine brain cells and human cells , such as Jurkat, SupT1, HEK 293, HeLa, and HOS cells as well as primary human cells such as macrophages, peripheral blood mononuclear cells (PBMCs) [221]. Taken together, this data demonstrates that pre-integration complexes are attracted to non-B DNA. Moreover, some of the non-B DNA motifs, such as G4 motifs, are known to promote recombination and influence genomic stability and cellular processes such as transcription [223,224]. These recent findings are of great interest to our laboratory as non-B DNA can not only influence HIV-1 integration site selection, but also potentially influence the establishment and/or maintenance of latency potentially by impeding the RNA polymerase processivity. These findings also set the foundation for this thesis. Thus, non-B DNA and HIV-1 integration site selection are the major focus for this dissertation. More discussion on the canonical B-DNA structure and non-B DNA motifs will be presented in section 1.6.1 and 1.6.2 respectively.

1.6.1 The canonical B-DNA structure

DNA was first isolated by Friedrich Miescher in 1869 and is the ideal molecule for the storage of genetic information [225]. More than 80 years following the discovery of DNA, Rosalind Franklin and Maurice Wilkins demonstrated that DNA forms a repeated helical structure in 1953 via X-ray analysis [226,227]. During that same period following Franklin and Wilkins' study, James Watson and Francis Crick elucidated the three dimensional molecular structure of DNA in 1953 [228]. These findings paved the way for a better understanding of the biological function of DNA.

DNA is a polymer of molecules called nucleotides and is commonly found as a double-stranded helix structure in the cell. Each nucleotide is composed of a nitrogen base linked to a 5 carbon sugar molecule and a phosphate group that is attached to the sugar molecule [229]. The sugar molecule in DNA is referred to as deoxyribose. There are four different bases derived from purine and pyrimidine that make up the nucleotides of DNA. The purine bases are Adenine (A) and Guanine (G). The pyrimidine bases are Thymine (T) and Cytosine (C). A complementary base interaction exists between the 2 strands of the double helix DNA where A always pair with T and G pairing with C [229] . These base pairs are further

stabilized by Watson and Crick hydrogen bonds. Overall, DNA is often found as a double-stranded structure where the nucleotides are linked together by phosphodiester bonds through the sugars and phosphates forming a chain of alternating sugar-phosphate backbone. The most commonly described and biological form of DNA is B-DNA.

B-DNA structure consists of two antiparallel polynucleotide chains. The two polynucleotide chains are held together in the center through hydrogen bonding between complementary bases. Therefore, the bases occupy the interior of the double helix and the sugar-phosphate backbones are found on the outside of the helix structure [226,227]. This molecular organization creates a wide major groove and narrow minor groove in the DNA. The complementary bases pairings of A-T and G-C as described by Watson and Crick is also a common feature of B-DNA. With this arrangement of the complementary base paring in the center, the two sugar-phosphate backbones wind around forming right-handed double helix structure **(Figure 1.6 A)**.

The structure of the double helix B-DNA can shift to adopt several distinct conformations based on many factors, including non-canonical base pairing. These non-canonical forms of DNA are known as non-B DNA structures or motifs.

1.6.2 Non-B DNA structures

In its inactive and non-transcribed state, the DNA structure primarily exists in the form of the right-handed B-DNA[228]. However, DNA is dynamic and can assume several alternative non-B DNA conformations under certain physiological conditions [230,231]. Non-B DNA structures occur within specific genomic sequences and are in higher energy states. Non-B DNA are believed to form at repetitive sequence motifs by the free energy generated from negative supercoiling of the DNA during replication or transcription, as the DNA partially unwinds, as well as during protein binding [232,233,234]. Overall, non-B DNA occurs when DNA encounters a high level of torsion or stress. In the human genome, the repeat DNA sequences that have the potential to fold and form non-B DNA comprise 50% of the genome. On the other hand, simple sequence repeats account for only about 3% of the total genomic DNA [235]. Currently more than 10 different types of non-B DNA structures have been identified. This include A-phased repeats, inverted repeats, direct repeats, cruciform

motifs, slipped motifs, mirror repeats, short tandem repeats, triplex repeats, G4 motifs, and Z-DNA motifs [222]. Examples of non-B DNA structures are presented in **Figure 1.6 B**.

Non-B DNA appears to play a significant role in several biologically important processes [236]. Specifically, non-B DNA are known to affect DNA replication, transcription, transcription factor recruitment, initiation repression, stalling of polymerase [223,224,237]. They can also induce genetic instability leading to certain human diseases [234,238,239]. Each non-B DNA motif is described below, followed by a brief description of their potential biological impact, if known.

- **A-phased motifs:** A-phased motifs are usually formed by 3 or more tracts of four to nine adenines or adenines succeeding by a thymidine [240]. They are separated with a central region containing 11-12 nucleotides.

- **Inverted repeats:** Inverted repeat sequences are single stranded nucleotide sequences that are arranged in opposing orientation. Inverted repeats are known to induce genome instability through excision of the repeat-associated region [241]. They have also been implicated in gene amplification [242].

- **Direct repeats:** Direct repeats are DNA sequences that are repeated two or more times downstream of each other. Direct repeats are shown to flank DNA deletion breakpoints [243,244].

- **Cruciform motifs:** Cruciform DNA forms at inverted repeat sequences and requires at least a 6 nucleotide inverted repeat sequence [232]. Cruciform DNA are similar in structure to the Holliday junction. Cruciform structures are typically located near break point junctions, replication origins and promoter regions [245,246]. They have also been implicated in regulating DNA replication in diverse organisms including mammalians [246] .

- **Slipped motifs:** Slipped motifs, are formed by direct repeat DNA sequences that present a certain symmetry [234]. These transient structures typically form during DNA replication and transcription. As the DNA strand unwinds, these direct repeat sequences on the single stranded DNA have the opportunity to fold back due to mispairing of the repeat units on the same strand [234].

Figure 1.6

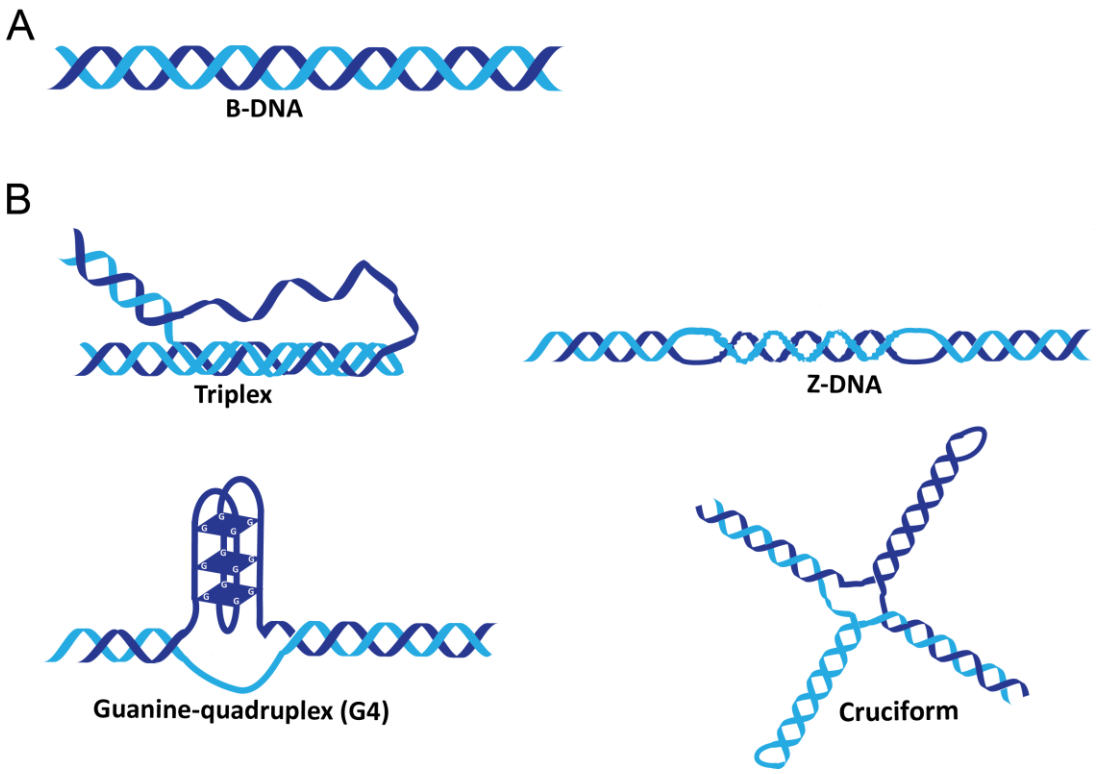


Figure 1.6: Canonical B-DNA structure and non-B DNA structures (adapted from reference [236]).

(A) Structure of canonical right handed double helix B-DNA. **(B)** Examples of non-B DNA structures. Triplex DNA is formed from (R.Y) mirror repeat sequence. Z-DNA is formed from alternating pyrimidine-purine sequences (YR_YR) *n*. Guanine-quadruplexes are formed from oligo $(G)_n$ tracts and cruciform structures are formed from inverted sequences.

The presence of these motifs has been associated with neurodegenerative and neuromuscular diseases [247,248].

- **Mirror repeats:** Mirror repeats are DNA sequences that are separated with a center of symmetry [249]. Mirror repeats are known to cause replication fork stalling due to their propensity of folding into triplex [250].

- **Short tandem repeats:** Short tandem repeats (STR) or microsatellites are short simple repeats of DNA sequences. Each repeat unit is about 1-6 bases long and are A-rich in the human genome [251,252]. STR can have a length of up to 100 nucleotides. STRs are present in 3% of the human genome and occur every 2000 bp in the human genome. They are mostly found in non-coding regions [252]. Some short tandem repeats are thought to act as transcription regulator and can affect certain genes expression. They may also be involved in recombination and could be associated with certain neurodegenerative diseases [252].

- **Triplex repeats:** Triplex DNA can form due to the presence of long stretches of purine-pyrimidine (R_Y) mirror repeat sequences [234]. There are two different types of triplex DNA: intermolecular and intramolecular. Intermolecular triplexes are triplexes that are formed between a duplex DNA and a triplex forming oligonucleotide (TFO) via Hoogsteen pairing [253]. Intramolecular triplexes are formed from a duplex DNA with homopurine and homopyrimidine at sites of DNA supercoiling [253]. An example of intramolecular triplex DNA are H-DNAs. Triplex sequences are predicted to be present in promoter and intergenic region near introns [254]. Triplex DNA has been implicated in regulating gene expression and genetic recombination [255,256].

- **Guanine-quadruplex (G4) motifs:** G4 motifs form secondary structures formed by guanine rich nucleic acids. G4 can assemble in G4-tetrads (G-G-G-G). Within a quadruplex, two to three tetrads stack together to form a compact, four-stranded DNA. The tetrad structure is stabilized by cations (Na^+, K^+) and hydrogen bonds in the center of the plane [257]. G4s have been found to be located in promoter region of oncogenes [258] and telomeres of chromosomes [224,257]. They can induce genomic instability in mammalians and bacteria [259].

- **Z-DNA motifs:** Z-DNA is a left-handed DNA helix structure. Z-DNA is formed based on alternating pyrimidine-purine sequences ((YR_YR) n)[235]. Compared to the canonical B DNA that have a major and a minor groove, Z-DNA only has a deep groove [235]. These motifs are more stable due to the energy release from negative supercoiling and they are usually present at transcription start sites, promoter region of genes and the 5'ends of genes [260,261]. Z-DNA can induce genomic instability. In fact, Z-DNA has been shown to cause double-strand breaks in mammals and bacteria, resulting in large scale deletions and chromosomal rearrangements [259].

1.7 Research overview and rationale

cART helps suppress HIV-1 replication in infected individuals, but it fails to eradicate virus from latently-infected reservoirs, such as memory $CD4^+$ T cells and macrophages. These latent proviruses are long-lived and undetectable by the immune system. Notably, proliferation of $CD4^+$ T cells through clonal expansion can further expand the latent reservoir. The potential for even a single virus to reinitiate infection despite successful antiviral therapy implies that it is necessary to eliminate all replication competent latent proviruses in order to eradicate HIV-1 from an infected individual. HIV-1 integration within active transcriptional units might promote viral gene expression and maximize viral production during the short lifespan of infected cells. In contrast, HIV-1 integration in low/inactive regions of genomes, such as satellite DNA, gene deserts and centromeric heterochromatin, has also been previously described [262–264]. It is possible that these regions might be involved in the establishment of latently-infected cells and HIV-1 persistence. Mechanisms underlying integration sites choice are not fully understood.

Furthermore, most of the studies available examine only the integration profile of HIV-1 subtype B, which is mostly prevalent in the Americas, Europe, Australian, Japan and Thailand. However, multiple subtypes and recombinants dominate other regions of the world, with 95% of HIV-1 infections occurring in developing countries. Subtype C contributes to most infection worldwide. Therefore, performing a comparative study of integration site profiles involving different subtypes will help determine if the site preferences seen with HIV-1 subtype B are common to all HIV-1 subtypes, and how this

may contribute to disease persistence. Additionally, the integration site selection of HIV-1 in compartmentalized sanctuary sites from infected individuals have not been defined. Integration sites studies will also provide more insights on the integration site profile seen in different latent reservoirs.

Given our previous findings that non-B DNA influences HIV-1 integration site selection and that certain non-B DNA motifs such as the G4 motifs can regulate gene expression, I proposed to further characterize the role of non-B DNA motifs in HIV-1 integration site selection and their contribution to HIV-1 persistence. Furthermore, I proposed to further investigate the integration site profile in evolutionary diverse retroviruses (e.g. HIV-1, SIV, MLV, and MMTV) with respect to non-B DNA.

1.8 Hypothesis

I hypothesized that host genomic non-B DNA motifs are favored in retroviral integration target site selection and that integration near non-B DNA contributes to HIV-1 persistence in latently infected cells. To address my hypothesis, I first characterized the integration site profile in productively and latently infected cells. I further assessed the integration profile among different HIV-1 subtypes (A, B, C and D) and other retroviruses in order to determine if any variation exists between their integration site preferences. Lastly, I delineated the integration site profile of different HIV-1 latent tissue reservoirs such as the brain/CNS and the gastrointestinal tract (GIT).

1.9 Thesis Chapters Overview

1.9.1 Specific host DNA structures are genomic beacons for integrated, quiescent/latent HIV-1 in patients receiving treatment

In chapter 2, I present analyses of the distribution of HIV-1 integration sites in productively infected cells, latently infected cells as well as clonally expanded cells, and non-clonally expanded cells. We have demonstrated a distinct integration profile in latently infected cells and clonally infected cells in different genomic regions. We have shown that HIV-1 favored integration in or near specific non-B DNA motifs in productively infected cells,

latently infected and clonally expanded cells and that these genomic features may attract HIV-1 for integration. Analysis of integration site placement in LEDGF/p75 and CPSF6 depleted cells showed that integration in or near specific non-B DNA was influenced by LEDGF/p75 and CPSF6. Additionally, we showed a strong bias toward integration into guanine-quadruplex (G4) structures that are generally associated with transcriptional silencing.

1.9.2 Non-B DNA structures are universally targeted by evolutionarily diverse retroviruses for integration

In chapter 3, I extended our analysis by assessing the integration distribution in evolutionary diverse retroviruses and in different HIV-1 subtypes. Throughout our analysis we were able to demonstrate that non-B DNA motifs are also targeted by other retroviruses and exhibit distinct integration profiles. Additionally, we showed via next generation sequencing of subtypes A, C and D that they exhibit an integration site profile that differs from subtype B, the most studied subtype. Importantly, integration site analysis from patient datasets correlated with integration into non-B DNA motifs know to suppress gene expression (e.g. G4 motifs and Z-DNA). Lastly, antiretroviral therapy altered HIV-1 integration site targeting, with enriched integration near G4 motifs observed in patients receiving antiretroviral therapy for subtypes A and C.

1.9.3 The quiescent/latent HIV-1 integration site landscape from different anatomical tissues reveals unique differences

In chapter 4, I have characterized the integration site profile in various anatomical sites (brain, PBLs/PBMCs and parts of the GIT) that harbor latent viruses during HIV-1 infection. We found that HIV-1 integration was enriched in genes in PBLs/PBMCs and the GIT. Integration into genes was strongly disfavored in the brain. Most importantly, integration was strongly enriched in or near non-B DNA motifs, which can play substantial role in regulating adjacent genes expression.

1.10 References

1. P.Goff, S. Retroviridae: The retroviruses and their replication. in *Fields Virology* (ed. David M. Knipe, P. M. H.) 1999–2069 (Lippincott Williams & Wilkins, 2007).

2. John M. Coffin, Stephen H. Hughes, H. E. V. Historical introduction to the general properties of retroviruses. in *Retroviruses* 1–25 (Cold Spring Harbor Laboratory Press, 1997).

3. Jay A. Levy. HIV and the pathogenesis of AIDS. in 1–26 (ASM Press, 2007).

4. Emanuele Fanales-Belasio, Mariangela Raimondo, Barbara Suligoi, and S. B. HIV virology and pathigenetic mechanism of infection, a brief overview. *Ann Ist Super Sanità* **46,** 5–14 (2010).

5. UNAIDS. Available at: http://www.unaids.org/en. (Accessed: 14th March 2018)

6. Barré-Sinoussi F, Chermann JC, Rey F, Nugeyre MT, Chamaret S, Gruest J, Dauguet C, Axler-Blin C, Vézinet-Brun F, Rouzioux C, Rozenbaum W, M. L. Isolation of a T-lymphotropic retrovirus from a patient at risk for acquired immune deficiency syndrome (AIDS). *Science* **220,** 868–71 (1983).

7. Arts, E. J. & Hazuda, D. J. HIV-1 antiretroviral drug therapy. *Cold Spring Harb. Perspect. Med.* **2,** a007161 (2012).

8. Chun, T. W. *et al.* Early establishment of a pool of latently infected, resting $CD4^{(+)}$ T cells during primary HIV-1 infection. *Proc. Natl. Acad. Sci. U. S. A.* **95,** 8869–8873 (1998).

9. Whitney, J. B. *et al.* Rapid seeding of the viral reservoir prior to SIV viraemia in rhesus monkeys. *Nature* **512,** 74–77 (2014).

10. Blankson, J. N., Persaud, D., Siliciano, R. F., In, P. R. & Ecycling, P. A. R. The challenge of viral reservoirs in HIV-1 infection. *Annu. Rev. Med.* **53,** 557–593 (2002).

11. Kamori, D. & Ueno, T. HIV-1 Tat and Viral Latency : What we can L'learn from naturally occurring sequence variations. *Front. Microbiol.* **8,** (2017).

12. Chavez, L., Calvanese, V. & Verdin, E. HIV latency is established directly and early in both resting and activated primary CD4 T. *PLoS Pathog.* **11,** e1004955 (2015).

13. Vanuitert, B. *et al.* Residual HIV-1 RNA in blood plasma of patients taking suppressive. **282,** 1627–1632 (1999).

14. Davey Jr., R. T. *et al.* HIV-1 and T cell dynamics after interruption of highly active antiretroviral therapy (HAART) in patients with a history of sustained viral suppression. *Proc Natl Acad Sci U S A* **96,** 15109–15114 (1999).

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

16. Eric O. Freed, M. A. M. HIVs and their replication. in *Fields Virology* (ed. David M Knipe, P. M. H.) 2108–2185 (Lippincott Williams & Wilkins, 2007).

17. Hernandez-Vargas, E. A. & Middleton, R. H. Modeling the three stages in HIV infection. *J. Theor. Biol.* **320,** 33–40 (2013).

18. Poiesz, B. J. *et al.* Detection and isolation of type C retrovirus particles from fresh and cultured lymphocytes of a patient with cutaneous T-cell lymphoma. *Proc. Natl. Acad. Sci.* **77,** 7415–7419 (1980).

19. Mikulas Popovic, M. G. Sarngadharan, E. R. and R. C. G. Detection , isolation , and continuous production of cytopathic retroviruses (HTLV-III) from patients with AIDS and pre-AIDS. *Science* **224,** 497–500 (1984).

20. Robert C. Gallo, Syed Z. Salahuddin, Mikulas Popovic, Gene M. Shearer, Mark Kaplan, Barton F. Haynes, Thomas J. Palker, Robert Redfield, James Oleske, Bijan Safai, Gilbert White, P. F. and P. D. M. Frequent detection and isolation of cytopathic retroviruses (HTLV-III) from patients with AIDS and at risk for AIDS.

Science **224,** 500–503 (1984).

21. Coffin, J., Haase, A., Levy, J. A., Montagnier, L. & Oroszlan, S. Human immunodeficiency viruses. *Sci. New Ser.* **232,** 697 (1986).

22. Robertson DL, Anderson JP, Bradac JA, Carr JK, Foley B, Funkhouser RK, Gao F, Hahn BH, Kalish ML, Kuiken C, Learn GH, Leitner T, McCutchan F, Osmanov S, Peeters M, Pieniazek D, Salminen M, Sharp PM, Wolinsky S, K. B. HIV-1 nomenclature proposal. *Science* **288,** 55 (2000).

23. Seitz, R. Human immunodeficiency virus (HIV). *Transfus. Med. Hemotherapy* **43,** 203–222 (2016).

24. Plantier, J. & Leoz, M. A new human immunodeficiency virus derived from gorillas. *Nat. Med.* **15,** 871–872 (2009).

25. Ward, M. J., Lycett, S. J., Kalish, M. L., Rambaut, A. & Leigh Brown, A. J. Estimating the rate of intersubtype recombination in early HIV-1 group M strains. *J. Virol.* **87,** 1967–1973 (2013).

26. Laboratories, L. A. N. HIV sequence database: nomenclature overview. Available at: www.hiv.lanl.gov/content/sequence/HelpDocs/subtypes-more.html. (Accessed: 14th March 2018)

27. Teixeira, C., Gomes, J. R. B., Gomes, P. & Maurel, F. Viral surface glycoproteins , gp120 and gp41 , as potential drug targets against HIV-1 : Brief overview one quarter of a century past the approval of zidovudine , the first anti-retroviral drug. *Eur. J. Med. Chem.* **46,** 979–992 (2011).

28. Frankel, A. D. & Young, J. A. T. HIV-1: Fifteen proteins and an RNA. *Annu. Rev. Biochem.* **67,** 1–25 (1998).

29. Bushman, F. D., Fujiwara, T. & Craigie, R. Retroviral DNA integration directed by HIV integration protein in vitro. *Science* **249,** 1555–8 (1990).

30. Kao, S.-Y., Calman, A. F., Luciw, P. A. & Peterlin, B. M. Anti-termination of transcription within the long terminal repeat of HIV-1 by tat gene product. *Nature* **330,** 489–493 (1987).

31. Feinberg, M. B., Baltimore, D. & Frankel, A. D. The role of Tat in the human immunodeficiency virus life cycle indicates a primary effect on transcriptional elongation. *Proc Natl Acad Sci U S A* **88,** 4045–4049 (1991).

32. Suhasini, M. & Reddy, T. R. Cellular proteins and HIV-1 Rev function. *Curr. HIV Res.* **7,** 91–100 (2009).

33. Miller, J. H., Presnyak, V. & Smith, H. C. The dimerization domain of HIV-1 viral infectivity factor Vif is required to block virion incorporation of APOBEC3G. *Retrovirology* **4,** (2007).

34. Kogan, M. & Rappaport, J. HIV-1 Accessory Protein Vpr: Relevance in the pathogenesis of HIV and potential for therapeutic intervention. *Retrovirology* **8,** 25 (2011).

35. Heinzinger, N. K. *et al.* The Vpr protein of human immunodeficiency virus type 1 influences nuclear localization of viral nucleic acids in nondividing host cells. *Biochemistry* **91,** 7311–7315 (1994).

36. Di Marzio, P., Choe, S., Ebright, M., Knoblauch, R. & Landau, N. R. Mutational analysis of cell cycle arrest, nuclear localization and virion packaging of human immunodeficiency virus type 1 Vpr. *J. Virol.* **69,** 7909–7916 (1995).

37. Jowett, J. B. M. *et al.* The human immunodeficiency virus type 1 vpr gene arrests infected T cells in the G2 + M phase of the cell cycle. *J. Virol.* **69,** 6304–6313 (1995).

38. Schubert, U. *et al.* The two biological activities of human immunodeficiency virus type 1 Vpu protein involve two separable structural domains. *J. Virol.* **70,** 809–19 (1996).

39. Willey, R. L., Maldarelli, F., Martin, M. A. & Strebel, K. Human immunodeficiency

virus type 1 Vpu protein induces rapid degradation of CD4. *J. Virol.* **66,** 7193–200 (1992).

40. Aiken, C., Konner, J., Landau, N. R., Lenburg, M. E. & Trono, D. Nef induces CD4 endocytosis: Requirement for a critical dileucine motif in the membrane-proximal CD4 cytoplasmic domain. *Cell* **76,** 853–864 (1994).

41. Schwartz, O., Maréchal, V., Le Gall, S., Lemonnier, F. & Heard, J. M. Endocytosis of major histocompatibility complex class I molecules is induced by the HIV-1 Nef protein. *Nat. Med.* **2,** 338–42 (1996).

42. Chowers, M. Y. *et al.* Optimal infectivity in vitro of human immunodeficiency virus type 1 requires an intact nef gene. *J. Virol.* **68,** 2906–14 (1994).

43. Pereira, L. A., Bentley, K., Peeters, A., Churchill, M. J. & Deacon, N. J. A compilation of cellular transcription factor interactions with the HIV-1 LTR promoter. *Nucleic Acids Res.* **28,** 663–8 (2000).

44. Johnson, S. F. & Telesnitsky, A. Retroviral RNA dimerization and packaging: The what, how, when, where, and why. *PLoS Pathog.* **6,** 10–13 (2010).

45. Valsamakis, A., Zeichner, S., Carswell, S. & Alwine, J. C. The human immunodeficiency virus type 1 polyadenylylation signal: a 3’ long terminal repeat element upstream of the AAUAAA necessary for efficient polyadenylylation. *Proc. Natl. Acad. Sci. U. S. A.* **88,** 2108–12 (1991).

46. DeZazzo, J. D., Kilpatrick, J. E. & Imperiale, M. J. Involvement of long terminal repeat U3 sequences overlapping the transcription control region in human immunodeficiency virus type 1 mRNA 3’ end formation. *Mol. Cell. Biol.* **11,** 1624–1630 (1991).

47. Greene, W. C. & Peterlin, B. M. Charting HIV’s remarkable voyage through the cell: Basic science as a passport to future therapy. *Nature Medicine* **8,** 673–680 (2002).

48. Engelman, A. & Cherepanov, P. The structural biology of HIV-1: Mechanistic and therapeutic insights. *Nat. Rev. Microbiol.* **10,** 279–290 (2012).

49. Klatzmann, D. *et al.* T-lymphocyte T4 molecule behaves as the receptor for human retrovirus LAV. *Nature* **312,** 767–768 (1984).

50. Dalgleish, A. G. *et al.* The CD4 (T4) antigen is an essential component of the receptor for the AIDS retrovirus. *Nature* **312,** 763–767 (1984).

51. Lasky, L. A. *et al.* Delineation of a region of the human immunodeficiency virus type 1 gp120 glycoprotein critical for interaction with the CD4 receptor. *Cell* **50,** 975–985 (1987).

52. Rizzuto, C. D. *et al.* A conserved HIV gp120 glycoprotein structure involved in chemokine receptor binding a conserved HIV gp120 glycoprotein structure involved in chemokine receptor binding. *Science* **280,** 1949–1953 (1998).

53. Naif, H. M. Pathogenesis of HIV infection. *Infect. Dis. Rep.* **5,** 26–30 (2013).

54. Freed, E. O., Myers, D. J. & Risser, R. Characterization of the fusion domain of the human immunodeficiency virus type 1 envelope glycoprotein gp41. *Proc. Natl. Acad. Sci.* **87,** 4650–4654 (1990).

55. Suzuki, Y., Chew, M. L. & Suzuki, Y. Role of host-encoded proteins in restriction of retroviral integration. *Front. Microbiol.* **227,** 1–13 (2012).

56. Miller, M. D., Farnet, C. M. & Bushman, F. D. Human immunodeficiency virus type 1 preintegration complexes: studies of organization and composition. *J. Virol.* **71,** 5382–5390 (1997).

57. McDonald, D. *et al.* Visualization of the intracellular behavior of HIV in living cells. *J. Cell Biol.* **159,** 441–452 (2002).

58. Bukrinskaya, a, Brichacek, B., Mann, a & Stevenson, M. Establishment of a functional human immunodeficiency virus type 1 (HIV-1) reverse transcription

complex involves the cytoskeleton. *J. Exp. Med.* **188,** 2113–2125 (1998).

59. Jonathan Karn and C. Martin Stoltzfus. Transcriptional and Posttranscriptional Regulation of HIV-1 Gene Expression. *Cold Spring Harb. Perspect. Med.* **2,** a006916 (2012).

60. Langford, S. E., Ananworanich, J. & Cooper, D. A. Predictors of disease progression in HIV infection: A review. *AIDS Res. Ther.* **4,** 1–14 (2007).

61. Epstein, F. H., Pantaleo, G., Graziosi, C. & Fauci, A. S. The immunopathogenesis of human immunodeficiency virus infection. *N. Engl. J. Med.* **328,** 327–335 (1993).

62. Piatak, A. M. *et al.* High levels of HIV-1 in plasma during all stages of infection determined by competitive PCR. *Science* **259,** 1749–1754 (1993).

63. Eric S. Daar, M.D., Tarsem Moudgil, M.S., Richard D Meyer, M.D., and Davis D. Ho, M. D. Transient high levels of virema in patients with primary human immunodefiency virus type 1 infection. *N. Engl. J. Med.* **324,** 961–964 (1991).

64. Grivel, J., Malkevitch, N. & Margolis, L. Human immunodeficiency virus type 1 induces apoptosis in $CD4^+$ but not in $CD8^+$ T cells in ex vivo-infected human lymphoid tissue. *J. Virol.* **74,** 8077–8084 (2000).

65. Jekle, A. *et al.* In vivo evolution of human immunodeficiency virus type 1 toward increased pathogenicity through CXCR4-mediated killing of uninfected CD4 T cells. *J. Virol.* **77,** 5846–5854 (2003).

66. M L Gougeon, H Lecoeur, A Dulioust, M G Enouf, M Crouvoiser, C Goujard, T. D. and L. M. Programmed cell death in peripheral lymphocytes from HIV-infected persons: increased susceptibility to apoptosis of CD4 and CD8 T cells correlates with lymphocyte activation and with disease progression. *J. Immunol.* **156,** 3509–3520 (1996).

67. Galloway, N. L. K. *et al.* Cell-to-cell transmission of HIV-1 is required to trigger pyroptotic death of lymphoid tissue-derived CD4 T cells. *Cell Rep.* **12,** 1555–1563

(2015).

68. Doitsh, G. *et al.* Pyroptosis drives CD4 T-cell depletion in HIV-1 infection. *Nature* **505,** 509–514 (2014).

69. Cuevas, J. M., Geller, R., Garijo, R., López-Aldeguer, J. & Sanjuán, R. Extremely high mutation rate of HIV-1 in vivo. *PLoS Biol.* **13,** 1–19 (2015).

70. Rooke, R. *et al.* Isolation of drug-resistant variants of HIV-1 from patients on long-term zidovudine therapy. *AIDS (London, England)* **3,** 411–415 (1989).

71. Li, J. W. & Vederas, J. C. Multiple mutations in HIV-1 reverse transcriptase confer high-level resistance to zidovudine. *Science* **246,** 1155–1158 (1989).

72. Harrigan, P. R., Whaley, M. & Montaner, J. S. G. Rate of HIV-1 RNA rebound upon stopping antiretroviral therapy. *Aids* **13,** F59-62 (1999).

73. Finzi, D. *et al.* Identification of a reservoir for HIV-1 in patients on highly active antiretroviral therapy. *Science* **278,** 1295–1300 (1997).

74. Igarashi, T. Macrophage are the principal reservoir and sustain high virus loads in rhesus macaques after the depletion of CD4+ T cells by a highly pathogenic simian immunodeficiency virus/HIV type 1 chimera (SHIV): Implications for HIV-1 infections of humans. *Proc. Natl. Acad. Sci.* **98,** 658–663 (2001).

75. Richman, D. D. *et al.* The Challenge of Finding a Cure for HIV Infection. *Science* **323,** 1304–1307 (2009).

76. McCune, J. M. Viral latency in HIV disease. *Cell* **82,** 183–188 (1995).

77. Pierson, T. C. *et al.* Molecular characterization of preintegration latency in human immunodeficiency virus type 1 infection. *J. Virol.* **76,** 8518–8531 (2002).

78. Zack, J. A. *et al.* HIV-1 entry into quiescent primary lymphocytes: Molecular analysis reveals a labile, latent viral structure. *Cell* **61,** 213–222 (1990).

79. Koelsch, K. K. *et al.* Dynamics of total, linear nonintegrated, and integrated HIV-1 DNA in vivo and in vitro. *J. Infect. Dis.* **197,** 411–419 (2008).

80. Zhou, Y., Zhang, H., Siliciano, J. D. & Siliciano, R. F. Kinetics of human immunodeficiency virus type 1 decay following entry into resting CD4. *Microbiology* **79,** 2199–2210 (2005).

81. Siliciano, J. D. & Siliciano, R. F. A long-term latent reservoir for HIV-1: Discovery and clinical implications. *J. Antimicrob. Chemother.* **54,** 6–9 (2004).

82. Siliciano, R. F. & Greene, W. C. HIV latency. *Cold Spring Harb. Perspect. Med.* **1,** 1–20 (2011).

83. Chun, T.-W. *et al.* Gene expression and viral prodution in latently infected, resting CD4+ T cells in viremic versus aviremic HIV-infected individuals. *Proc. Natl. Acad. Sci.* **100,** 1908–1913 (2003).

84. Strain, M. C. *et al.* Heterogeneous clearance rates of long-lived lymphocytes infected with HIV: intrinsic stability predicts lifelong persistence. *Proc. Natl. Acad. Sci.* **100,** 4819–24 (2003).

85. Siliciano, J. D. *et al.* Long-term follow-up studies confirm the stability of the latent reservoir for HIV-1 in resting CD4+T cells. *Nat. Med.* **9,** 727–728 (2003).

86. Chomont, N. *et al.* HIV reservoir size and persistence are driven by T cell survival and homeostatic proliferation. *Nat. Genet.* **15,** 893–900 (2009).

87. Maldarelli F1, Wu X2, Su L2, Simonetti FR3, Shao W2, Hill S1, Spindler J1, Ferris AL1, Mellors JW4, Kearney MF1, Coffin JM5, H. S. Specific HIV integration sites are linked to clonal expansion and persistence of infected cells. *Science* **345,** 179–83 (2014).

88. Wagner, T. A. *et al.* HIV latency. Proliferation of cells with HIV integrated into cancer genes contributes to persistent infection. *Science* **345,** 570–573 (2014).

89. Chomonta, N., DaFonsecaa, S., Vandergeetena, C., Ancuta, P. & Sékaly, R. P. Maintenance of $CD4^+$ T-cell memory and HIV persistence: Keeping memory, keeping HIV. *Curr. Opin. HIV AIDS* **6,** 30–36 (2011).

90. Kwon, K. J. & Siliciano, R. F. HIV persistence : clonal expansion of cells in the latent reservoir. *J. Clin. Invest.* **127,** 2536–2538 (2017).

91. Anderson, E. M. & Maldarelli, F. The role of integration and clonal expansion in HIV infection : live long and prosper. *Retrovirology* **15,** (2018).

92. Cohn, L. B. *et al.* HIV-1 integration landscape during latent and active infection. *Cell* **160,** 420–432 (2015).

93. Simonetti, F. R. *et al.* Clonally expanded $CD4^+$ T cells can produce infectious HIV-1 in vivo. *Proc. Natl. Acad. Sci.* **113,** 1883–1888 (2016).

94. Han, Y. *et al.* Orientation-dependent Regulation of Integrated HIV-1 Expression by Host Gene Transcriptional Readthrough. *Cell Host Microbe* **4,** 134–146 (2008).

95. Tina Lenasi1, Xavier Contreras1, and B. M. P. Transcriptional interference antagonizes proviral gene expression to promote HIV latency. *Cell Host Microbe* **4,** 123–133 (2008).

96. Ganesh, L. *et al.* The gene product Murr1 restricts HIV-1 replication in resting CD4+lymphocytes. *Nature* **426,** 853–857 (2003).

97. Yang, Z., Zhu, Q., Luo, K. & Zhou, Q. The 7SK small nuclear RNA inhibits the CDK9/cyclin T1 kinase to control transcription. *Nature* **414,** 317–322 (2001).

98. Huang, J. *et al.* Cellular microRNAs contribute to HIV-1 latency in resting primary CD4+T lymphocytes. *Nat. Med.* **13,** 1241–1247 (2007).

99. Giuseppe Pantaleo, Cecilia Graziosi, James F. Demarest, Luca Butini, Maria Montroni, Cecil H. Fox, Jan M. Orenstein, D. P. K. & A. S. F. HIV infection is active ‘and’ progressive in lymphoid tissue during the clinically latent stage of

disease. *Nature* **362,** 355–358 (1993).

100. Günthard, H. F. *et al.* Residual human immunodeficiency virus (HIV) type 1 RNA and DNA in lymph nodes and HIV RNA in genital secretions and in cerebrospinal fluid after suppression of viremia for 2 Years. *J. Infect. Dis.* **183,** 1318–1327 (2001).

101. Chun, T. *et al.* Persistence of HIV in Gut-associated lymphoid tissue despite long-term antiretroviral therapy. *J. Infect. Dis.* **197,** 714–720 (2008).

102. Wiley, C. A., Schrier, R. D., Nelson, J. A., Lampert, P. W. & Oldstone, M. B. Cellular localization of human immunodeficiency virus infection within the brains of acquired immune deficiency syndrome patients. *Proc. Natl. Acad. Sci. U. S. A.* **83,** 7089–93 (1986).

103. Williams, K. C. & Hickey, W. F. Central nervous system damage, monocytes and macrophages, and neurological disorders in AIDS. *Annu. Rev. Neurosci.* **25,** 537–562 (2002).

104. Maria M. Bednar, Christa Buckheit Sturdevant, Lauren A. Tompkins, Kathryn Twigg Arrildt, Elena Dukhovlinova, Laura P. Kincer, and R. S. author. Compartmentalization, viral evolution, and viral latency of HIV in the CNS. *Curr. HIV/AIDS Rep.* **12,** 262–271 (2015).

105. Subjects, S. Human Immunodeficiency Virus Type 1 in the Semen of Men Receiving Highly Active Antiretroviral Therapy. *Blood* **339,** 1803–1809 (1998).

106. Quayle, A. J., Xu, C., Mayer, K. H. & Anderson, D. J. T lymphocytes and macrophages, but not motile spermatozoa, are a significant source of human immunodeficiency virus in semen. *J Infect Dis* **176,** 960–968 (1997).

107. Bagasra, O. *et al.* Detection of HIV-1 proviral DNA in sperm from HIV-1-infected men. *Aids* **8,** 1669–1674 (1994).

108. Prins, J. M. *et al.* Immuno-activation with anti-CD3 and recombinant human IL-2 in HIV-1-infected patients on potent antiretroviral therapy. *AIDS* **13,** 2405–10

(1999).

109. Contreras, X. *et al.* Suberoylanilide Hydroxamic Acid Reactivates HIV from Latently Infected Cells. *J. Biol. Chem.* **284,** 6782–6789 (2009).

110. Archin, N. M. *et al.* Expression of Latent HIV Induced by the Potent HDAC Inhibitor Suberoylanilide Hydroxamic Acid. *AIDS Res. Hum. Retroviruses* **25,** 207–212 (2009).

111. Savarino, A. *et al.* 'Shock and kill' effects of class I-selective histone deacetylase inhibitors in combination with the glutathione synthesis inhibitor buthionine sulfoximine in cell line models for HIV-1 quiescence. *Retrovirology* **6,** (2009).

112. Elliott, J. H. *et al.* Activation of HIV transcription with short-course vorinostat in HIV-infected patients on suppressive antiretroviral therapy. *PLoS Pathog.* **10,** e1004473 (2014).

113. Rasmussen TA, Tolstrup M, Brinkmann CR, Olesen R, Erikstrup C, Solomon A, Winckelmann A, Palmer S, Dinarello C, Buzon M, Lichterfeld M, Lewin SR, Østergaard L, S. O. Panobinostat, a histone deacetylase inhibitor, for latent-virus reactivation in HIV-infected patients on suppressive antiretroviral therapy: a phase 1/2, single group, clinical trial. *Lancet HIV* **1,** e13-21 (2014).

114. Wei, D. G. *et al.* Histone Deacetylase Inhibitor Romidepsin Induces HIV Expression in CD4 T Cells from Patients on Suppressive Antiretroviral Therapy at Concentrations Achieved by Clinical Dosing. *PLoS Pathog.* **10,** e1004071 (2014).

115. Prete, G. Q. Del *et al.* Elevated Plasma Viral Loads in Romidepsin-Treated Simian Immunodeficiency Virus-Infected Rhesus Macaques on Suppressive Combination Antiretroviral Therapy. *Antimicrob. Agents Chemother.* **60,** 1560–1572 (2016).

116. Mehla, R. *et al.* Bryostatin modulates latent HIV-1 infection via PKC and AMPK signaling but inhibits acute infection in a receptor independent manner. *PLoS One* **5,** e11160 (2010).

117. Korin, Y. D., Brooks, D. G., Brown, S., Korotzer, A. & Zack, J. a. Effects of prostratin on T-cell activation and human immunodeficiency virus latency. *J. Virol.* **76,** 8118–23 (2002).

118. Kulkosky, J. *et al.* Prostratin : activation of latent HIV-1 expression suggests a potential inductive adjuvant therapy for HAART. *Gene* **98,** 3006–3015 (2011).

119. Zerbato, J., Mellors, J. W. & Sluis-cremer, N. Disulfiram reactivates latent HIV-1 expression through depletion of the phosphatase and tensin homolog. *AIDS* **27,** F7–F11 (2013).

120. Cyktor, J. C. & Mellors, J. W. Toll-like receptor agonists : Can they exact a toll on human immunodeficiency virus persistence ? *Clin. Infect. Dis.* **64,** 1696–1698 (2017).

121. Angela Tsai, Alivelu Irrinki, Jasmine Kaur, Tomas Cihlar, George Kukolj, D. D. S. and J. P. M. Toll-like receptor 7 agonist GS-9620 induces HIV expression and HIV-specific immunity in cells from HIV-infected individuals on suppressive antiretroviral therapy. *J. Virol.* **91,** e02166-16 (2017).

122. Rasmus Offersen, Sara Konstantin Nissen, Thomas A. Rasmussen, Lars Østergaard, Paul W. Denton, Ole Schmeltz Søgaard, M. T. A novel toll-like receptor 9 agonist, MGN1703, enhances HIV-1 transcription and NK cell-mediated inhibition of HIV-1-infected autologous CD4+ T cells. *J. Virol.* **90,** 4441–4453 (2016).

123. Winckelmann, A. A. *et al.* Administration of a toll-like receptor 9 agonist decreases the proviral reservoir in virologically suppressed HIV-infected patients. *PLoS One* **8,** e62074 (2013).

124. Martínez-bonet, M. *et al.* Synergistic activation of latent HIV-1 expression by novel histone deacetylase inhibitors and bryostatin-1. *Sci. Rep.* **13,** 16445 (2015).

125. Laird, G. M. *et al.* Ex vivo analysis identifies effective HIV-1 latency – reversing drug combinations. *J. Clin. Invest.* **125,** 1901–1912 (2015).

126. Van Maele, B., Busschots, K., Vandekerckhove, L., Christ, F. & Debyser, Z. Cellular co-factors of HIV-1 integration. *Trends Biochem. Sci.* **31,** 98–105 (2006).

127. Duane P Grandgenett, Krishan K Pandey, Sibes Bera, H. A. Multifunctional facets of retrovirus integrase. *World J. Biol. Chem.* **6,** 83–94 (2015).

128. Vink, C., van Gent, D. C., Elgersma, Y. & Plasterk, R. H. Human immunodeficiency virus integrase protein requires a subterminal position of its viral DNA recognition sequence for efficient cleavage. *J. Virol.* **65,** 4636–4644 (1991).

129. Masuda, T., Kuroda, M. J. & Harada, S. Specific and independent recognition of U3 and U5 *att* sites by human immunodeficiency virus type 1 integrase in vivo. *J. Virol.* **72,** 8396–8402 (1998).

130. Sherman, P. A. & Fyfe, J. A. Human immunodeficiency virus integration protein expressed in Escherichia coli possesses selective DNA cleaving activity. *Proc. Natl. Acad. Sci.* **87,** 5119–5123 (1990).

131. Vink, C., Groeneger, A. A. M. oude & Plasterk, R. H. A. Identification of the catalytic and DNA-binding region of the human immunodeficiency virus type I integrase protein. *Nucleic Acids Res.* **21,** 1419–1425 (1993).

132. Engelman, A., Mizuuchi, K. & Craigie, R. HIV-1 DNA integration: Mechanism of viral DNA cleavage and DNA strand transfer. *Cell* **67,** 1211–1221 (1991).

133. Bushman, F. D. & Craigie, R. Activities of human immunodeficiency virus (HIV) integration protein in vitro: specific cleavage and integration of HIV DNA. *Proc. Natl. Acad. Sci. U. S. A.* **88,** 1339–43 (1991).

134. Smith, J. A. & Daniel, R. Following the path of the virus: the exploitation of host DNA repair mechanisms by retroviruses. *ACS Chem. Biol.* **1,** 217–226 (2006).

135. Yoder, K. E. & Bushman, F. D. Repair of gaps in retroviral DNA integration intermediates. *J. Virol.* **74,** 11191–11200 (2000).

136. Delelis, O., Carayon, K., Saïb, A., Deprez, E. & Mouscadet, J. F. Integrase and integration: Biochemical activities of HIV-1 integrase. *Retrovirology* **5,** 1–13 (2008).

137. Eijkelenboom, A. P. A. M. *et al.* The DNA-binding domain of HIV-1 integrase has an SH3-like fold. *Nat. Struct. Mol. Biol.* **2,** 807–810 (1995).

138. Mengli Cai, Ronglan Zheng, Michael Caffrey, Robert Craigie, G. M. C. & A. M. G. Solution structure of the N-terminal zinc binding domain of HIV-1 integrase. *Nat. Struct. Biol.* **4,** 567–577 (1997).

139. Lee SP, Xiao J, Knutson JR, Lewis MS, H. M. Zn_2^+ promotes the self-association of human immunodeficiency virus type-1 integrase in vitro. *Biochemistry* **36,** 173–180 (1997).

140. Zheng, R., Jenkins, T. M. & Craigie, R. Zinc folds the N-terminal domain of HIV-1 integrase, promotes multimerization, and enhances catalytic activity. *Proc. Natl. Acad. Sci. U. S. A.* **93,** 13659–64 (1996).

141. Dyda, F. *et al.* Crystal structure of the catalytic domain of HIV-1 integrase: similarity to other polynucleotidyl transferases. *Science* **266,** 1981–6 (1994).

142. Maignan, S., Guilloteau, J. P., Zhou-Liu, Q., Clément-Mella, C. & Mikol, V. Crystal structures of the catalytic domain of HIV-1 integrase free and complexed with its metal cofactor: High level of similarity of the active site with other viral integrases. *J. Mol. Biol.* **282,** 359–368 (1998).

143. Goldgur, Y. *et al.* Three new structures of the core domain of HIV-1 integrase: an active site that binds magnesium. *Proc. Natl. Acad. Sci. U. S. A.* **95,** 9150–9154 (1998).

144. Johnson, A. A. *et al.* Integration requires a specific interaction of the donor DNA terminal 5′-cytosine with glutamine 148 of the HIV-1 integrase flexible loop. *J. Biol. Chem.* **281,** 461–467 (2006).

145. Esposito, D. & Craigie, R. Sequence specificity of viral end DNA binding by HIV-1 integrase reveals critical regions for protein-DNA interaction. *EMBO J.* **17,** 5832–5843 (1998).

146. Lutzke, R. a & Plasterk, R. H. Structure-based mutational analysis of the C-terminal DNA-binding domain of human immunodeficiency virus type 1 integrase: critical residues for protein oligomerization and DNA binding. *J. Virol.* **72,** 4841–4848 (1998).

147. Bowerman, B., Brown, P. O., Bishop, J. M. & Varmus, H. E. A nucleoprotein complex mediates the integration of retroviral DNA. *Genes Dev.* **3,** 469–478 (1989).

148. Brown, P. O., Bowerman, B., Varmus, H. E. & Bishop, J. M. Correct integration of retroviral DNA in vitro. *Cell* **49,** 347–356 (1987).

149. Farnet, C. M. & Haseltine, W. A. Integration of human immunodeficiency virus type 1 DNA in vitro. *Proc. Natl. Acad. Sci. U. S. A.* **87,** 4164–8 (1990).

150. Popov, S. *et al.* Viral protein R regulates nuclear import of the HIV-1 pre-integration complex. *EMBO J.* **17,** 909–917 (1998).

151. Emiliani, S. *et al.* Integrase mutants defective for interaction with LEDGF/p75 are impaired in chromosome tethering and HIV-1 replication. *J. Biol. Chem.* **280,** 25517–25523 (2005).

152. Kalpana, G. V, Marmon, S., Wang, W., Crabtree, G. R. & Goff, S. P. Binding and stimulation of HIV-1 integrase by a human homolog of yeast transcription factor SNF5. *Science* **266,** 2002–2006 (1994).

153. Lin, C. & Engelman, A. The Barrier-to-Autointegration Factor Is a Component of Functional Human Immunodeficiency Virus Type 1 Preintegration Complexes. *J. Virol.* **77,** 5030–6 (2003).

154. Cherepanov, P. *et al.* HIV-1 integrase forms stable tetramers and associates with LEDGF/p75 protein in human cells. *J. Biol. Chem.* **278,** 372–381 (2003).

155. Lee, M. S. & Craigie, R. Protection of retroviral DNA from autointegration: involvement of a cellular factor. *Proc. Natl. Acad. Sci. U. S. A.* **91,** 9823–9827 (1994).

156. Farnet, C. M. & Bushman, F. D. HIV-1 cDNA integration: Requirement of HMG I(Y) protein for function of preintegration complexes in vitro. *Cell* **88,** 483–492 (1997).

157. Carlson, M. & Laurent, B. C. The SNF/SWI family of global transcriptional activators. *Curr. Opin. Cell Biol.* **6,** 396–402 (1994).

158. Wang, W. *et al.* Purification and biochemical heterogeneity of the mammalian SWI-SNF complex. *EMBO J.* **15,** 5370–5382 (1996).

159. Morozov, a, Yung, E. & Kalpana, G. V. Structure-function analysis of integrase interactor 1/hSNF5L1 reveals differential properties of two repeat motifs present in the highly conserved region. *Proc. Natl. Acad. Sci. U. S. A.* **95,** 1120–1125 (1998).

160. Cheng, S. W. G. *et al.* c-MYC interacts with INI1/hSNF5 and requires the SWI/SNF complex for transactivation function. *Nat. Genet.* **22,** 102–105 (1999).

161. Lee, D. *et al.* SWI/SNF complex interacts with tumor suppressor p53 and is necessary for the activation of p53-mediated transcription. *J. Biol. Chem.* **277,** 22330–22337 (2002).

162. Lee, D., Sohn, H., Kalpana, G. V. & Choe, J. Interaction of E1 and hSNF5 proteins stimulates replication of human papillomavirus DNA. *Nature* **399,** 487–491 (1999).

163. Wu, D. Y., Kalpana, G. V, Goff, S. P. & Schubach, W. H. Epstein-Barr virus nuclear protein 2 (EBNA2) binds to a component of the human SNF-SWI complex, hSNF5/Ini1. *J. Virol.* **70,** 6020–8 (1996).

164. Yung, E. *et al.* Inhibition of HIV-1 virion production by a transdominant mutant of integrase interactor 1. *Nat. Med.* **7,** 920–926 (2001).

165. Yung, E. *et al.* Specificity of interaction of INI1/hSNF5 with retroviral integrases and its functional significance. *J. Virol.* **78,** 2222–31 (2004).

166. Craig, E., Zhang, Z. K., Davies, K. P. & Kalpana, G. V. A masked NES in INI1/hSNF5 mediates hCRM1-dependent nuclear export: Implications for tumorigenesis. *EMBO J.* **21,** 31–42 (2002).

167. Cai, M. *et al.* Solution structure of the cellular factor baf responsible for protecting retroviral dna from autointegration. *Nat. Struct. Biol.* **5,** 903–909 (1998).

168. Segura-Totten, M. & Wilson, K. L. BAF: Roles in chromatin, nuclear structure and retrovirus integration. *Trends Cell Biol.* **14,** 261–266 (2004).

169. Skoko, D. *et al.* Barrier-to-autointegration factor (BAF) condenses DNA by looping. *Proc. Natl. Acad. Sci.* **106,** 16610–16615 (2009).

170. Chen, H. & Engelman, A. The barrier-to-autointegration protein is a host factor for HIV type 1 integration. *Proc. Natl. Acad. Sci. U. S. A.* **95,** 15270–4 (1998).

171. Furukawa, K. LAP2 binding protein 1 (L2BP1/BAF) is a candidate mediator of LAP2-chromatin interaction. *J. Cell Sci.* **112,** 2485–2492 (1999).

172. Suzuki, Y., Yang, H. & Craigie, R. LAP2α and BAF collaborate to organize the Moloney murine leukemia virus preintegration complex. *EMBO J.* **23,** 4670–4678 (2004).

173. Li, L. *et al.* Retroviral cDNA integration: stimulation by HMG I family proteins. *J. Virol.* **74,** 10965–10974 (2000).

174. Hindmarsh, P. *et al.* HMG protein family members stimulate human immunodeficiency virus type 1 and avian sarcoma virus concerted DNA integration in vitro. *J. Virol.* **73,** 2994–3003 (1999).

175. Beitzel, B. & Bushman, F. Construction and analysis of cells lacking the HMGA gene family. *Nucleic Acids Res.* **31,** 5025–5032 (2003).

176. Sharma, P., Singh, D. P., Fatma, N., Chylack, L. T. & Shinohara, T. Activation of LEDGF gene by thermal- and oxidative-stresses. *Biochem. Biophys. Res. Commun.* **276,** 1320–1324 (2000).

177. Singh, D. P., Fatma, N., Kimura, A., Chylack, L. T. & Shinohara, T. LEDGF binds to heat shock and stress-related element to activate the expression of stress-related genes. *Biochem. Biophys. Res. Commun.* **283,** 943–955 (2001).

178. Cherepanov, P., Devroe, E., Silver, P. A. & Engelman, A. Identification of an evolutionarily conserved domain in human lens epithelium-derived growth factor/transcriptional co-activator p75 (LEDGF/p75) that binds HIV-1 integrase. *J. Biol. Chem.* **279,** 48883–48892 (2004).

179. Busschots, K. *et al.* The interaction of LEDGF/p75 with integrase is lentivirus-specific and promotes DNA binding. *J. Biol. Chem.* **280,** 17841–17847 (2005).

180. Vanegas, M. *et al.* Identification of the LEDGF/p75 HIV-1 integrase-interaction domain and NLS reveals NLS-independent chromatin tethering. *J. Cell Sci.* **118,** 1733–43 (2005).

181. Maertens, G. *et al.* LEDGF/p75 is essential for nuclear and chromosomal targeting of HIV-1 integrase in human cells. *J. Biol. Chem.* **278,** 33528–33539 (2003).

182. Cherepanov, P., Ambrosio, A. L. B., Rahman, S., Ellenberger, T. & Engelman, A. Structural basis for the recognition between HIV-1 integrase and transcriptional coactivator p75. *Proc. Natl. Acad. Sci. U. S. A.* **102,** 17308–13 (2005).

183. Stec, I., Nagl, S. B., Van Ommen, G. J. B. & Den Dunnen, J. T. The PWWP domain: A potential protein-protein interaction domain in nuclear proteins influencing differentiation? *FEBS Lett.* **473,** 1–5 (2000).

184. Maertens, G., Cherepanov, P., Debyser, Z., Engelborghs, Y. & Engelman, A. Identification and characterization of a functional nuclear localization signal in the HIV-1 integrase interactor LEDGF/p75. *J. Biol. Chem.* **279,** 33421–33429 (2004).

185. Turlure, F., Maertens, G., Rahman, S., Cherepanov, P. & Engelman, A. A tripartite DNA-binding element, comprised of the nuclear localization signal and two AT-hook motifs, mediates the association of LEDGF/p75 with chromatin in vivo. *Nucleic Acids Res.* **34,** 1653–1665 (2006).

186. Christ, F. *et al.* Transportin-SR2 imports HIV into the nucleus. *Curr. Biol.* **18,** 1192–1202 (2008).

187. Krishnan, L. *et al.* The Requirement for cellular Transportin 3 (TNPO3 or TRN-SR2) during infection maps to human immunodeficiency virus type 1 capsid and not integrase. *J. Virol.* **84,** 397–406 (2010).

188. De Iaco, A. & Luban, J. Inhibition of HIV-1 infection by TNPO3 depletion is determined by capsid and detectable after viral cDNA enters the nucleus. *Retrovirology* **8,** 98 (2011).

189. Anisenko, A. N. *et al.* Characterization of HIV-1 integrase interaction with human Ku70 protein and initial implications for drug targeting. *Sci. Rep.* **7,** 1–14 (2017).

190. Zheng, Y., Ao, Z., Wang, B., Jayappa, K. D. & Yao, X. Host protein Ku70 binds and protects HIV-1 integrase from proteasomal degradation and is required for HIV replication. *J. Biol. Chem.* **286,** 17722–17735 (2011).

191. Rüegsegger, U., Blank, D. & Keller, W. Human pre-mRNA cleavage factor Im Is related to spliceosomal SR proteins and can be reconstituted in vitro from recombinant subunits. *Mol. Cell* **1,** 243–253 (1998).

192. Bhattacharya, A. *et al.* Structural basis of HIV-1 capsid recognition by PF74 and CPSF6. *Proc. Natl. Acad. Sci.* **111,** 18625–18630 (2014).

193. Chin, C. R. *et al.* Direct visualization of HIV-1 replication intermediates shows that capsid and CPSF6 modulate HIV-1 intra-nuclear invasion and integration. *Cell Rep.* **13,** 1717–1731 (2015).

194. Lee, K. *et al.* HIV-1 capsid-targeting domain of cleavage and polyadenylation

specificity factor 6. *J. Virol.* **86,** 3851–3860 (2012).

195. Ciuffi, A. *et al.* Integration site selection by HIV-based vectors in dividing and growth-arrested IMR-90 lung fibroblasts. *Mol. Ther.* **13,** 366–373 (2006).

196. Shinn, P. *et al.* HIV-1 integration in the human genome favors active genes and local hotspots. *Cell Press* **110,** 521–529 (2002).

197. Wang, G. P., Ciuffi, A., Leipzig, J., Berry, C. C. & Bushman, F. D. HIV integration site selection: Analysis by massively parallel pyrosequencing reveals association with epigenetic modifications. *Genome Res.* **17,** 1186–1194 (2007).

198. Bushman, F. *et al.* Genome-wide analysis of retroviral DNA integration. *Nat. Rev. Microbiol.* **3,** 848–58 (2005).

199. Wu, X., Li, Y., Crise, B. & Burgess, S. M. Transcription start regions in the human genome are favored targets for MLV integration. *Science* **300,** 1749–1751 (2003).

200. Kim, S. *et al.* Integration site preference of xenotropic murine leukemia virus-related virus, a new human retrovirus associated with prostate cancer. *J. Virol.* **82,** 9964–9977 (2008).

201. Barr, S. D., Leipzig, J., Shinn, P., Ecker, J. R. & Bushman, F. D. Integration targeting by avian sarcoma-leukosis virus and human immunodeficiency virus in the chicken genome. *J. Virol.* **79,** 12035–44 (2005).

202. Narezkina, A. *et al.* Genome-wide analyses of avian sarcoma virus integration sites. *J. Virol.* **78,** 11656–63 (2004).

203. Derse, D. *et al.* Human T-cell leukemia virus type 1 integration target sites in the human genome: Comparison with those of other retroviruses. *J. Virol.* **81,** 6731–6741 (2007).

204. Faschinger, A. *et al.* Mouse mammary tumor virus integration site selection in human and mouse genomes. *J. Virol.* **82,** 1360–1367 (2008).

205. Bor, Y. C., Bushman, F. D. & Orgel, L. E. In vitro integration of human immunodeficiency virus type 1 cDNA into targets containing protein-induced bends. *Proc. Natl. Acad. Sci. U. S. A.* **92,** 10334–8 (1995).

206. Müller, H. P. & Varmus, H. E. DNA bending creates favored sites for retroviral integration: an explanation for preferred insertion sites in nucleosomes. *EMBO J.* **13,** 4704–14 (1994).

207. Pruss, D., Reeves, R., Bushman, F. D. & Wolffe, A. P. The influence of DNA and nucleosome structure on integration events directed by HIV integrase. *J. Biol. Chem.* **269,** 25031–25041 (1994).

208. Pruss, D., Bushman, F. D. & Wolffe, A. P. Human immunodeficiency virus integrase directs integration to sites of severe DNA distortion within the nucleosome core. *Proc. Natl. Acad. Sci.* **91,** 5913–5917 (1994).

209. Gallay, P. A., Hope, T. J., Chin, D. & Trono, D. HIV-1 infection of nondividing cells through the recognition of integrase by the importin/karyopherin pathway. *Proc. Natl. Acad. Sci. U. S. A.* **94,** 9825–30 (1997).

210. Weinberg, J. B., Matthews, T. J., Cullen, B. R. & Malim, M. H. Productive human immunodeficiency virus type 1 (HIV-1) infection of nonproliferating human monocytes. *J. Exp. Med.* **174,** 1477–82 (1991).

211. Lewis, P., Hensel, M. & Emerman, M. Human immunodeficiency virus infection of cells arrested in the cell cycle. *EMBO J.* **11,** 3053–8 (1992).

212. Roe, T., Reynolds, T. C., Yu, G. & Brown, P. O. Integration of murine leukemia virus DNA depends on mitosis. *EMBO J.* **12,** 2099–2108 (1993).

213. Barr, S. D. *et al.* HIV Integration Site Selection: Targeting in Macrophages and the Effects of Different Routes of Viral Entry. *Mol. Ther.* **14,** 218–225 (2006).

214. Ferris, A. L. *et al.* Lens epithelium-derived growth factor fusion proteins redirect HIV-1 DNA integration. *Proc. Natl. Acad. Sci.* **107,** 3135–3140 (2010).

215. Marshall, H. M. *et al.* Role of PSIP 1/LEDGF/p75 in lentiviral infectivity and integration targeting. *PLoS One* **2,** (2007).

216. Ciuffi, A. *et al.* A role for LEDGF/p75 in targeting HIV DNA integration. *Nat. Med.* **11,** 1287–9 (2005).

217. Vandekerckhove, L. *et al.* Transient and stable knockdown of the integrase cofactor LEDGF/p75 reveals its role in the replication cycle of human immunodeficiency virus. *J. Virol.* **80,** 1886–96 (2006).

218. Wu, X., Li, Y., Crise, B., Burgess, S. M. & Munroe, D. J. Weak palindromic consensus sequences are a common feature found at the integration target sites of many retroviruses. *J. Virol.* **79,** 5211–5214 (2005).

219. Yeou-Cherng, B., Miller, M. D., Bushman, F. D. & Orgel, L. E. Target-sequence preferences of HIV-1 integration complexes in vitro. *Virology* **222,** 283–288 (1996).

220. Carteau, S., Hoffmann, C. & Bushman, F. Chromosome structure and human immunodeficiency virus type 1 cDNA integration: centromeric alphoid repeats are a disfavored target. *J. Virol.* **72,** 4005–14 (1998).

221. McAllister, R. G. *et al.* Lentivector integration sites in ependymal cells from a model of metachromatic leukodystrophy: non-B DNA as a new factor influencing integration. *Mol. Ther. Nucleic Acids* **3,** e187 (2014).

222. Bacolla, A. & Wells, R. D. Non-B DNA conformations, genomic rearrangements, and human disease. *J. Biol. Chem.* **279,** 47411–47414 (2004).

223. Rhodes, D. & Lipps, H. J. G-quadruplexes and their regulatory roles in biology. *Nucleic Acids Res.* **43,** 8627–8637 (2015).

224. Bochman, M. L., Paeschke, K. & Zakian, V. a. DNA secondary structures: stability and function of G-quadruplex structures. *Nat. Rev. Genet.* **13,** 770–80 (2012).

225. Dahm, R. Friedrich Miescher and the discovery of DNA. *Dev. Biol.* **278,** 274–288

(2005).

226. Franklin, R. & Gosling, R. G. Molecular configuration in sodium thymonucleate. *Nature* **171,** 740–741 (1953).

227. Wilkins, M. H. F., Stokes, a R., Wilson, H. R., Strokes, A. R. & Wilson, H. R. Molecular structure of deoxypentose nucleic acids. 1953. *Nature* **171,** 738–740 (1953).

228. Watson, J. D. & Crick, F. H. C. Molecular structure of nucleic aids: A structure for deoxyribose nucleic acid. *Nature* **171,** 737–738 (1953).

229. Alberts B, Johnson A, Lewis J, et al. The structure and function of DNA. in *Molecular Biology of the Cell* (Garland Science, 2002).

230. Wells, R. D. Unusual DNA structures. *J. Biol. Chem.* **263,** 1095–1098 (1988).

231. Mirkin, S. M. Discovery of alternative DNA structures: a heroic decade (1979-1989). *Front. Biosci.* **13,** 1064–1071 (2008).

232. Shlyakhtenko, L. S., Potaman, V. N., Sinden, R. R. & Lyubchenko, Y. L. Structure and dynamics of supercoil-stabilized DNA cruciforms. *J. Mol. Biol.* **280,** 61–72 (1998).

233. Wells, A. R. R. and R. D. W. Stabilization of Z DNA in vivo by Localized Supercoiling. *Science* **246,** 358–363 (1989).

234. Wells, R. D., Dere, R., Hebert, M. L., Napierala, M. & Son, L. S. Advances in mechanisms of genetic instability related to hereditary neurological diseases. *Nucleic Acids Res.* **33,** 3785–98 (2005).

235. Choi, J. & Majima, T. Conformational changes of non-B DNA. *Chem. Soc. Rev.* **40,** 5893 (2011).

236. Kouzine, F. *et al.* Permanganate/S1 nuclease footprinting reveals non-B DNA structures with regulatory potential across a mammalian genome. *Cell Syst.* **4,** 344–

356.e7 (2017).

237. Dai, X. & Rothman-Denes, L. B. DNA structure and transcription. *Curr. Opin. Microbiol.* **2,** 126–130 (1999).

238. Bacolla, A. & Wells, R. D. Non-B DNA conformations, genomic rearrangements, and human disease. *J. Biol. Chem.* **279,** 47411–4 (2004).

239. Zhao, J., Bacolla, A., Wang, G. & Vasquez, K. M. Non-B DNA structure-induced genetic instability and evolution. *Cell. Mol. Life Sci.* **67,** 43–62 (2010).

240. Cer, R. Z. *et al.* Non-B DB v2.0: A database of predicted non-B DNA-forming motifs and its associated tools. *Nucleic Acids Res.* **41,** 94–100 (2013).

241. Gordenin, D. A. *et al.* Inverted DNA repeats: A source of eukaryotic genomic instability. *Mol. Cell. Biol.* **13,** 5315–5322 (1993).

242. Tanaka, H., Tapscott, S. J., Trask, B. J. & Yao, M.-C. Short inverted repeats initiate gene amplification through the formation of a large DNA palindrome in mammalian cells. *Proc. Natl. Acad. Sci.* **99,** 8772–8777 (2002).

243. Mita, S. *et al.* Recombination via flanking direct repeats is a major cause of large-scale deletions of human mitochondrial DNA. *Nucleic Acids Res.* **18,** 561–567 (1990).

244. Samuels, D. C., Schon, E. A. & Chinnery, P. F. Two direct repeats cause most human mtDNA deletions. *Cell Press* **20,** 393–398 (2004).

245. van Holde, K. & Zlatanova, J. Unusual DNA structures, chromatin and transcription. *BioEssays* **16,** 59–68 (1994).

246. Pearson, C. E., Zorbas, H., Price, G. B. & Zannis-Hadjopoulos, M. Inverted repeats, stem-loops, and cruciforms: Significance for initiation of DNA replication. *J. Cell. Biochem.* **63,** 1–22 (1996).

247. Pearson, C. E., Wang, Y. H., Griffith, J. D. & Sinden, R. R. Structural analysis of

slipped strand DNA (s-DNA) formed in CTG (N)nCAG(N)n from the myotonic dystrophy locus. *Nucleic Acids Res.* **26,** 816–823 (1998).

248. Axford, M. M. *et al.* Detection of slipped-DNAs at the trinucleotide repeats of the myotonic dystrophy type I disease locus in patient tissues. *PLoS Genet.* **9,** (2013).

249. Lang, D. M. Imperfect DNA mirror repeats in the gag gene of HIV-1 (HXB2) identify key functional domains and coincide with protein structural elements in each of the mature proteins. *Virol. J.* **4,** 1–13 (2007).

250. Liu, G. *et al.* Replication fork stalling and checkpoint activation by a PKD1 locus mirror repeat polypurine-polypyrimidine (Pu-Py) tract. *J. Biol. Chem.* **287,** 33412–33423 (2012).

251. Nadir, E., Margalit, H., Gallily, T. & Ben-Sasson, S. a. Microsatellite spreading in the human genome: evolutionary mechanisms and structural implications. *Proc. Natl. Acad. Sci. U. S. A.* **93,** 6470–6475 (1996).

252. Fan, H. & Chu, J. Y. A brief review of short tandem repeat mutation. *Genomics, Proteomics Bioinforma.* **5,** 7–14 (2007).

253. Frank-Kamenetskii, M. D. & Mirkin, S. M. Triplex DNA structures. *Annu. Rev. Biochem.* **64,** 65–95 (1995).

254. Lexa, M., Martinek, T. & Brazdova, M. Uneven distribution of potential triplex sequences in the human genome in silico study using the R/Bioconductor package triplex. *Bioinforma. 2014 Proc. Int. Conf. Bioinforma. Model. Methods Algorithms* 80–88 (2014).

255. Napierala, M., Dere, R., Vetcher, A. & Wells, R. D. Structure-dependent recombination hot spot activity of GAA·TTC sequences from intron 1 of the Friedreich's Ataxia gene. *J. Biol. Chem.* **279,** 6444–6454 (2004).

256. Grabczyk, E. The GAATTC triplet repeat expanded in Friedreich's ataxia impedes transcription elongation by T7 RNA polymerase in a length and supercoil dependent

manner. *Nucleic Acids Res.* **28,** 2815–2822 (2000).

257. Patel, D. J., Phan, A. T. & Kuryavyi, V. Human telomere, oncogenic promoter and 5′-UTR G-quadruplexes: Diverse higher order DNA and RNA targets for cancer therapeutics. *Nucleic Acids Res.* **35,** 7429–7455 (2007).

258. Tracy A. Brooks, Samantha Kendrick, and L. H. Making sense of G-quadruplex and i-motif functions in oncogene promoters. *FEBS J.* **277,** 3459–3469 (2010).

259. Wang, G., Christensen, L. A. & Vasquez, K. M. Z-DNA-forming sequences generate large-scale deletions in mammalian cells. *Proc. Natl. Acad. Sci. U. S. A.* **103,** 2677–82 (2006).

260. Shin, S. I. *et al.* Z-DNA-forming sites identified by ChIP-Seq are associated with actively transcribed regions in the human genome. *DNA Res.* **23,** 477–486 (2016).

261. Schroth, G. P., Chou, P. J. & Ho, P. S. Mapping Z-DNA in the human genome. Computer-aided mapping reveals a nonrandom distribution of potential Z-DNA-forming sequences in human genes. *J. Biol. Chem.* **267,** 11846–11855 (1992).

262. Lewinski, M. K. *et al.* Genome-wide analysis of chromosomal features repressing human immunodeficiency virus transcription. *J. Virol.* **79,** 6610–9 (2005).

263. Pace, M. J. *et al.* Directly infected resting CD4+T cells can produce HIV Gag without spreading infection in a model of HIV latency. *PLoS Pathog.* **8,** e1002818 (2012).

264. Bartholomae, C. C. *et al.* Lentiviral vector integration profiles differ in rodent postmitotic tissues. *Mol. Ther.* **19,** 703–10 (2011).

Chapter 2

2 Specific host DNA structures are genomic beacons for integrated, quiescent/latent HIV-1 in patients receiving treatment

Elimination of the latent reservoir is essential for curing HIV-1 infection. Integration sites of latent proviruses play a critical role in the clonal expansion, persistence and reactivation of HIV-1 expression. To better understand the local genomic environment surrounding integrated proviruses and its contribution to latency, we characterized integration site datasets from productively and latently infected cells. We showed that integration sites are enriched in and/or near non-B DNA motifs/structures, and that lens epithelium-derived growth factor (LEDGF)/p75 and cleavage and polyadenylation specificity factor 6 (CPSF6) influence this integration site targeting. Non-B DNA integration site profiles from productively and latently infected cells, including clonally and non-clonally expanded cells, are distinct. Importantly, we demonstrated a strong correlation between integration sites, guanine-quadruplex (G4) motifs and reactivation of latent proviruses with latency reversal agents. Our findings implicate non-B DNA as a key factor in HIV-1 integration site targeting and the establishment and maintenance of latency.

2.1 Introduction

An essential step in the life cycle of HIV-1 is the integration of its viral genome into the human genome. This event is permanent and leads to life-long persistence of the virus within its host. Combination antiretroviral therapy (cART) suppresses productive HIV-1 replication in infected individuals, thereby reducing circulating virus to undetectable levels [1]. Despite this, resting memory $CD4^+$ T-cells harbor integrated virus that persists in a transcriptionally silent state referred to as latency [2,3]. HIV-1 remains latent indefinitely until reactivated by means that are not fully understood, but include cessation of antiretroviral therapy, development of antiretroviral resistance or clinically-directed 'shock and kill' therapy [4]. Latency presents a major obstacle in curing an individual of HIV-1 infection. This is in part due to the slow decay rate of the latent reservoir after cART initiation, which has an estimated half-life of 44 months and an eradication timeline of >70

years in a patient [5,6]. Although the size of the latent pool is under much debate, modeling studies have suggested that expansion and contraction of latently infected cells can generate low-level persistent viremia and intermittent viral blips that can replenish the latent reservoir [7–9]. Therefore, elimination of the latent reservoir is essential for eradication of the virus from the body.

Multiple mechanisms have been attributed to establishing and maintaining proviruses in a latent state and are likely not mutually exclusive. For example, the site and orientation of integration, availability of cellular transcription factors and viral proteins, epigenetic regulation of the HIV-1 promoter, and microRNA regulation of chromatin remodeling and targeting of messenger RNAs have been shown to contribute to HIV-1 latency [10,11]. 'Shock and kill' strategies have been proposed to flush out latent HIV-1 reservoirs to induce depletion of the virus for a cure. The main objective of these strategies is to facilitate the reactivation of HIV-1 expression from latent reservoirs, which are then destroyed through either natural means (e.g. immune response and viral cytopathogenicity) or artificial means (e.g. drugs and antibodies) [12]. Many latency reversing agents have been used for reactivation including physiological stimuli, chemical compounds (phorbol esters), histone deacetylase (HDAC) inhibitors, p-TEFb activators, and antibodies (e.g. anti-CD3); however, these agents fail to reactivate the entire pool of latently infected cells [13–16]. Although somewhat controversial, several reports suggest that this failure is due to genomic location-driven differences in HIV-1 expression [14,16–21]. Specific integration sites are also associated with clonal expansion of latently infected cells [22]. Clonally expanded cells have been shown to produce infectious HIV-1 *in vivo*; however, this may not be the case in all infected individuals [23,24]. A better understanding of the molecular mechanisms contributing to the establishment and maintenance of latency will help current eradication strategies.

Much of the early retroviral integration site analyses focused on the most frequent integration events in an attempt to better understand the genomic environment surrounding sites that result in productive infection. Comparatively, there are fewer integration site analyses with respect to latent infection. This is likely attributed to the rarity of latent integration events and the small number of cells comprising the latent reservoir. Several

groups have proposed that the features affecting latency are highly local and heterogeneous [16,19,20,25]. HIV-1 LTR promoter activity was also shown to be sensitive to the local chromatin environment in such a way that it is not directly controlled by DNA methylation or histone acetylation in cell lines [18,26]. Importantly, the insertion site affects the response to latency reversal agents (LRAs), where different LRAs can activate different subsets of proviruses in the whole latent population [16,19]. These different subsets were distinguishable in terms of chromatin functional states and only represented <5% of cells carrying a latent provirus [16]. Furthermore, host cellular proteins such as LEDGF/p75 and CPSF6 have been shown to promote integration into actively transcribed genes residing in gene-dense regions, thereby reducing integration into other genomic regions conducive to latency such as heterochromatin [27–36].

We previously identified non-B DNA as a novel factor that influences HIV-1 integration site targeting in acute infection [37]. Non-B DNA motifs are abundant in the human genome and form secondary structures using non-canonical Watson-Crick base pairing. At least 10 non-B DNA conformations exist including G4 motifs, A-phased repeats, inverted repeats, direct repeats, cruciform motifs, slipped motifs, mirror repeats, short-tandem repeats, triplex repeats and Z-DNA motifs [38]. Several non-B DNA motifs preferentially act as the recipient of genetic information, stimulating homologous recombination >20-fold in human cells [39]. Non-B DNA structures (e.g. G4, Z-DNA, cruciform and triplex motifs) have also been shown to potently silence expression of adjacent genes [40–50].

In this study, we analyzed HIV-1 integration site profiles of productively and latently infected cells and identified a strong correlation between non-B DNA motifs and latent proviral integration sites. In particular, we showed that G4 motifs significantly influence integration site targeting and proviral reactivation potential by certain LRAs. Moreover, we showed that LEDGF/p75 and CPSF6 impact integration targeting of non-B DNA motifs.

2.2 Materials and methods

2.2.1 Cell lines

Human embryonic kidney 293T (HEK 293T) cells were obtained from the American Type Culture Collection. HEK 293T cells were maintained in standard Dulbecco's Modified Eagles Medium (DMEM) (Wisent, cat#:319-005-CL) or phenol red free DMEM (Wisent, cat#:319-051-CL) at 37°C with 5% CO_2 in a humidified incubator. All media were supplemented with 10% heat-inactivated fetal bovine serum ([FBS], [Wisent, cat#:080-450, lot#: 115690]), 100U/ml penicillin and 100µg/ml streptomycin.

2.2.2 Virus production

Pseudotyped HIV-1/VSV-G was generated by co-transfecting HEK 293T cells (plated at $3.5x10^6$ cells in 10cm dish) with plasmids p156RRLsinPPTCMVGFPWPRE (encoding the HIV vector segment), pCMVdeltaR9 (the packaging construct), and pMD.G (encoding the VSV-G envelope) as previously described [51]. The three plasmids were kindly provided by Dr. F. Bushman (University of Pennsylvania, USA). Co-transfection was performed using 5µg of each plasmid with LipoD293 (FroggaBio, cat#: SL1006680.1) following the manufacturer's instructions. 24 hours post co-transfection, the cells and culture medium were harvested and centrifuged at 1500 rpm (239x g) for 5 min at room temperature. The supernatants were collected and stored at -80°C and were used to infect cells as described in section 2.2.3.

2.2.3 Drug treatment and genomic DNA extraction

BRACO19 hydrochloride (BRACO19) was purchased from Sigma Aldrich (cat. #SML0560-5MG). TMPyP4 was purchased from Calbiochem-EMD Millipore (cat. #613560-25MG). $0.5x10^6$ HEK 293T cells were plated in 6 well plates for 24 hours. The cells were then left untreated (control) or treated for 24 hours with either BRACO19 (0, 1, 3, and 32 µM) or TMPyP4 (0, 0.5, 1 and 8 µM). BRACO19 and TMPyP4 concentrations were established from previously used concentrations [52,53,54]. Cells were then infected for 5 hours with pseudotyped HIV-1/VSV-G (800µl of DMEM, 200µl of HIV-1/VSV-G virus) in the presence of 1µl of 10µg/ml polybrene [(Sigma, cat#: H9268-5g]). Medium was

changed 5 hours post-infection and the cells were treated anew with either BRACO19 or TMPyP4 for 24 hours. The genomic DNA was then extracted from the cells as per manufacturer's instructions using the DNeasy Blood and Tissue Kit (Qiagen, cat#: 69504) and processed for integration site profile analyses (see section 2.2.5).

2.2.4 MTT assay

Cell metabolic activity was measured using the MTT (3-(4,5-Dimethylthiazol-2-yl)-2,5-Diphenyltetrazolium Bromide) kit (Thermo Fisher Scientific, cat#: M6494). HEK 293T cells were plated ($0.2x10^6$) in 24 well plates with phenol red free DMEM for 24 hours. Cells were treated with either BRACO19 (0, 1, 3, and 32 µM) or TMPyP4 (0, 0.5, 1 and 8 µM). 24 hours post-treatment, the medium was changed and the cells were incubated for 5 hours at 37°C without treatment. This step was performed to reproduce conditions during infection as described in section 2.2.3. The cells were later treated with BRACO19 and TMPyP4 for an additional 24 hours. Following the second treatment, MTT solution was added at a final concentration of 0.5mg/ml. The samples were incubated for 1 hour at 37°C. 100µl of DMSO (EMD, cat#: MX1456-6) was added to solubilized the purple formazan crystals for 10 min on a shaker. Absorbance of the plates were read at 540 nm using the Epoch microplate spectrophotometer (BioTek) plate reader and the Gen5.2.06 analysis software. To determine the percent (%) cell viability, the average values of the blank wells (medium + MTT solution) were subtracted from each sample read (BRACO19, TMPyP4 treated and untreated samples). Untreated value was used as positive control from which the % cell viability was determined.

2.2.5 HIV-1 integration library

Genomic DNA extracted from infected treated and untreated HEK 293T cells (see section 2.2.3) was processed for integration site analysis and sequenced using the Illumina MiSeq platform. First, 1µg of extracted genomic DNA was restriction enzyme digested with MseI overnight at 37°C. Digested DNA was column purified with the Gel/PCR DNA Fragments Kit (Geneaid, cat#: DF100) according to manufacturer's instructions. Next, compatible double-stranded linkers to the MseI sites were prepared as follows: MseI Linker (+) 5'GTAATACGACTCACTATAGGGCTCCGCTTAAGGGAC 3'and MseI Linker (-): 5'

[Phos]-TAGTCCCTTAAGCG GAG-[AmC7-Q] 3’ were mixed (20µl MseI Linker (+) [40 µM] and 20 µl MseI Linker (-) [40 µM]). The linker mixture was denatured for 5 min at 90°C and cooled 1°C every 3 min until the temperature reached 20°C using the T100™ Thermal Cycler (Bio-Rad). The prepared linkers are now referred to as the “adapter mix”.

Purified DNA was linker ligated with the adapter mix at 21°C for about 14 hours with 13.5µl of MseI digested samples, 3.5µl of adapter mix, 1µl of T4 DNA Ligase (400U/µl, [NEB, cat#: M0202S), and 2 µl of 10x ligase buffer. Subsequently, 20µl of the ligated sample was digested at 37°C for 4 hours with 2µl of DpnI (20U/µl), 2µl of NarI (5U/µl), 5µl of 10x buffer and water to a total volume of 50µl. Following digestion, the samples were column purified. The junctions between the integrated HIV-1 LTR sequence and adjacent genomic sequence were amplified in two separate rounds of PCR amplification. The HIV-1 NL4-3 LTR sequence were used to design primers that amplify through the HIV-1 LTR. The RuparLTR (Forward) 5’-TGCTTCAAGTAGTGTGTGC-3’ primer that anneals to the HIV-1 LTRs and the Linker1 (Reverse) 5’-GTAATACGACTCACTATAG GGC-3’primer specific to the MseI linker sequences were used for the first round of PCR amplification. Each PCR reaction mixture consisted of 15.5µl sterile water, 5µl of NarI/DnpI digested sample, 2.5µl of 10x Advantage 2 PCR Buffer, 0.5 µl of 15µM of Linker1 primer , 0.5 µl of 15µM RuparLTR primer, 0.5µl of 10mM dNTPs and 0.5 µl of 50X Advantage 2 PCR polymerase mix (Takara Bio Inc., cat#:639201). PCR was run on T100™ Thermal Cycler (Bio-Rad) under the following cycling conditions: 1 min at 94°C, 5 cycles of 2 sec at 94°C, 1 min at 72°C with an additional 20 cycles of 2 sec at 94°C, 1 min at 67°C and a final extension cycle for 1 min at 72°C and a 4°C hold. The second round of nested PCR amplification was performed using sample from the first round of PCR amplification. The PCR reaction mixture and cycling condition were as described for the first round of PRC amplification. The following primer set was used for nested PCR: Rupar-LTR2nested (Forward) 5’-CTCTGGTAACTAGAGATCCCTCAGACC-3’, Linker2nested (Reverse) 5’-AGGGCTCCGCTTAAGGGAC-3’ and Next, Illumina adapter overhang nucleotide sequences were added to the HIV-1 LTR sequence and the MseI linker sequence. Illutag-Forward 5’-GTCTCGTGGGCTCGG AGATGTGTATAA GAGACAGCTCTGGTAACTAGAGATCCCTCAGACC-3’ and Illutag-Reverse 5’-TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGAGGGCTCCGCTTAAGGGA

C-3'. Underlined section of the two Illutag primers represent the overhang section. Illumina adapters were utilized in a PCR reaction mixture contained 15.5µl sterile water, 5µl of nested PCR samples, 2.5µl of Advantage 2 PCR Buffer (10x), 0.5 µl of Forward adapter (10µM), 0.5 µl of Reverse adapter (10µM), 0.5µl of dNTPs (10mM) and 0.5 µl of 50X Advantage 2 PCR polymerase mix. PCR was run on T100^{TM} Thermal Cycler (Bio-Rad). Cycling conditions were as described for the first round of PRC amplification.

The PCR product were purified with AmPure XP beads (Beckman Coulter, cat#: A63881) and the DNA samples were processed using Nextera XT Index Kit (Illumina). The Nextera XT Indexes technology utilizes a single tagmentation reaction that fragments and tags input DNA with unique adapter and index (barcodes) sequences on both ends of the DNA as previously described [37]. The DNA samples were purified using AmPure XP beads following addition of the barcodes. The barcoded samples were quantified using the Quant-it PicoGreen dsDNA Assay Kit (Invitrogen, cat#: P7589). The absorbance of the plates were read (excitation 480nm for 10 sec and emission 540 nm for 10 sec) with the Cytation5 Imaging Reader (BioTek) and the Gen5 3.02.1 analysis software. Sample concentration was determined using a standard concentration curve. The barcoded samples were sequenced through Illumina MiSeq using 2×150 bp chemistry at the London Regional Genomics Centre at the Robarts Research Institute (Western University, Canada).

2.2.6 Computational analysis

Fastq sequencing reads were quality trimmed and unique integration sites identified using our in-house bioinformatics pipeline [37], which is called the Barr Lab Integration Site Identification Pipeline (BLISIP version 2.9). BLISIP version 2.9 includes the following updates: bedtools (v2.25.0) which is used to compute distances between integration sites and genomic features, bioawk (awk version 20110810) a programming language for biological data manipulation, bowtie2 (version 2.3.4.1) is used for aligning sequence reads to the human genome, and restrSiteUtils (v1.2.9) is used to generate *in silico* matched random control integration sites based on restriction enzyme used or DNA shearing methods. HIV-1 LTR-containing Fastq sequences were identified and filtered by allowing up to a maximum of five mismatches with the reference NL4-3 LTR sequence and if the LTR sequence had no match with any region of the human genome (GRCh37/hg19).

Integration site profile heatmaps were generated using our in-house python program BHmap (BHmap version 1.0). Sites that could not be unambiguously mapped to a single region in the genome were excluded from the study. Mapping of integration sites to non-B DNA motifs was performed using the Non-B DB for the human genome (GRCh37/hg19) [55, 56] as previously described [37]. Lamina associated domains (LADs) were retrieved from http://dx.doi.org/10.1038/ nature06947 [57].

2.2.7 Datasets analysis

All integration site datasets used in this study were independently analyzed using BLISIP version 2.9. The Cohn dataset was obtained from the National Center for Biotechnology Information Sequence Read Archive (NCBI SRA) using the accession number SRP045822 as described [24]. The Battivelli dataset was obtained from the "Integration Sites – Source Data" as described [16]. The Maldarelli/Wu dataset was obtained from the supplemental material as described [22]. The Achuthan dataset was obtained from the NCBI SRA using the accession number SRP132583 as described [34]. All genomic sites in each dataset that hosted two or more sites (i.e. identical sites) were collapsed into one unique site for the analysis.

2.2.8 Statistical analysis

The Fisher's exact test was used for all comparisons of integration site distributions in Figures 2.1B, 2.1C, 2.1D, 2.2B, 2.2C, 2.2D, 2.3, 2.4B, 2.4C, 2.4D and 2.5D. A single factor ANOVA test was used to confirm significant changes within the experiment for Figure 2.5C. For all Figures, * $P < 0.05$, ** $P < 0.01$, *** $P < 0.001$, **** $P < 0.0001$.

2.2.9 Data and software availability

The sequences reported in this paper have been deposited in the National Center for Biotechnology Information Sequence Read Archive (NCBI SRA) (SRP164286: SRR7975450-SRR7975468).

2.3 Results

2.3.1 HIV-1 integration sites in quiescent/latently infected cells are enriched in and near non-B DNA motifs

To determine if there is a correlation between HIV-1 integration sites and non-B DNA motifs during latent infection, we analyzed the integration site profile of a previously published HIV-1 integration dataset by Cohn and colleagues [24]. This dataset contained integration sites that were obtained from primary $CD4^+$ T cells from 13 HIV-1 infected individuals, categorized as untreated viremic (3,210-71,857 viral RNA copies/ml before therapy), untreated controller (<50-880 viral RNA copies/ml before therapy) and treated (<40 viral RNA copies/ml after therapy) groups **(Figure 2.1A, Table 2.1 and Supplemental Table 2.1**). To generate integration site profiles, we used an in-house bioinformatics pipeline designed to maximize the number of unique integration sites as similarly described [37,58]. The integration site profiles were compared with matched random control (MRC) datasets generated *in silico*.

In agreement with the work of Cohn et al. (2015) [24] and others [59], integration sites in all three patient groups were enriched in genes and short interspersed nuclear elements (SINEs) (e.g. *Alu* elements) and disfavored in long interspersed nuclear elements (LINEs) **(Figure 2.1B)**. Integration sites in the treated group were also enriched within 500 base pairs (bp) of satellite DNA, which is abundant in heterochromatin ($P < 0.0001$). This is in contrast with the untreated viremic and controller groups where integration sites were significantly depleted in and near satellite DNA ($P < 0.001$). In contrast with the untreated controller and treated groups, integration sites in the viremic group were enriched in CpG islands and disfavored in endogenous retroviral elements (ERVs) ($P < 0.0001$).

Although all groups exhibited enriched integration in genes, the untreated controller group also exhibited enrichment near genes (1-499 bp) compared to the untreated viremic and treated groups ($P < 0.0001$). Together, these data confirm previous findings by Cohn et al. (2015) and others and show that integration sites in treated patients are enriched near heterochromatin.

Table 2.1: List of integration site datasets used in chapter 2.

Dataset	Group	# of unique integration sites	Cell Type	Reference
Cohn	Untreated Viremic	26,342	Primary $CD4^+$ T-cells	Cohn et al. (2015)
	Untreated Controllers	26,034	Primary $CD4^+$ T-cells	
	Treated	101,881	Primary $CD4^+$ T-cells	
Maldarelli/Wu	Clonal	216	Primary $CD4^+$ T-cells	Maldarelli /Wu et al. (2014)
	Non-clonal	1507	Primary $CD4^+$ T-cells	
Battivelli	Productively Infected (PIC)	950	Primary $CD4^+$ T-cells	Battivelli et al. (2018)
	Reactivated Latently Infected (RLIC)	153	Primary $CD4^+$ T-cells	
	Non-Reactivated Latently Infected (NRLIC)	669	Primary $CD4^+$ T-cells	
Achuthan	Wild type (WT)	277	293T cells	

	LEDGF/p75 depletion (BID)	2949	293T cells	Achuthan et al. (2018)
	CPSF6 depletion (A77V)	4431	293T cells	
This study	Untreated	2017	293T cells	This study
	BRACO19 (1 µM)	797	293T cells	
	BRACO19 (3 µM)	1073	293T cells	
	BRACO19 (32 µM)	759	293T cells	
	TMPyP4 (0.5 µM)	1223	293T cells	
	TMPyP4 (1 µM)	1286	293T cells	
	TMPyP4 (8 µM)	1302	293T cells	

Figure 2.1: HIV-1 integration sites in quiescent/latently infected cells are enriched in and near non-B DNA motifs. (A) Table showing the range of viral load and $CD4^+$ T cells counts from the untreated viremic, untreated controller and treated groups. (**B**) Heatmap depicting the fold enrichment or depletion of integration sites in common genomic features compared to the matched random control (MRC). (**C**) Bar graphs representing the proportion of unique HIV-1 integration sites in various non-B DNA motifs. Asterisks denote significant differences from MRC. (**D**) Heatmap depicting the fold enrichment or depletion of integration sites in non-B DNA motifs compared to the MRC. Within each heatmap, numbers represent the fold-change in the ratio of integration sites compared to MRC sites. Darker shades represent higher fold-changes. Fisher's exact test was used for all comparisons. * $P < 0.05$, ** $P < 0.01$, *** $P < 0.001$, **** $P < 0.0001$.

Figure 2.1

A

Group	Viral Load (copies/ml)	CD4 Count
Untreated Viremic	3210 - 71857	165 - 674
Untreated Controller	<50 - 880	430 - 1070
Treated	<40	280 - 965

B

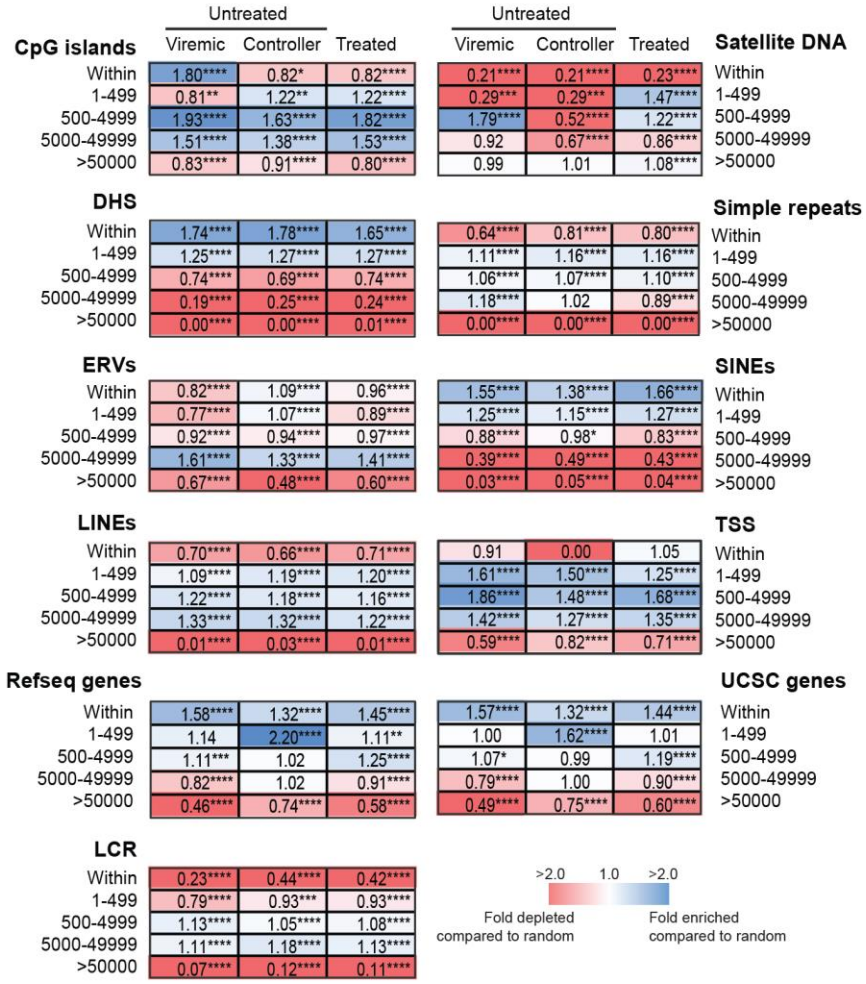


Figure 2.1: HIV-1 integration sites in quiescent/latently infected cells are enriched in and near non-B DNA motifs.

Figure 2.1

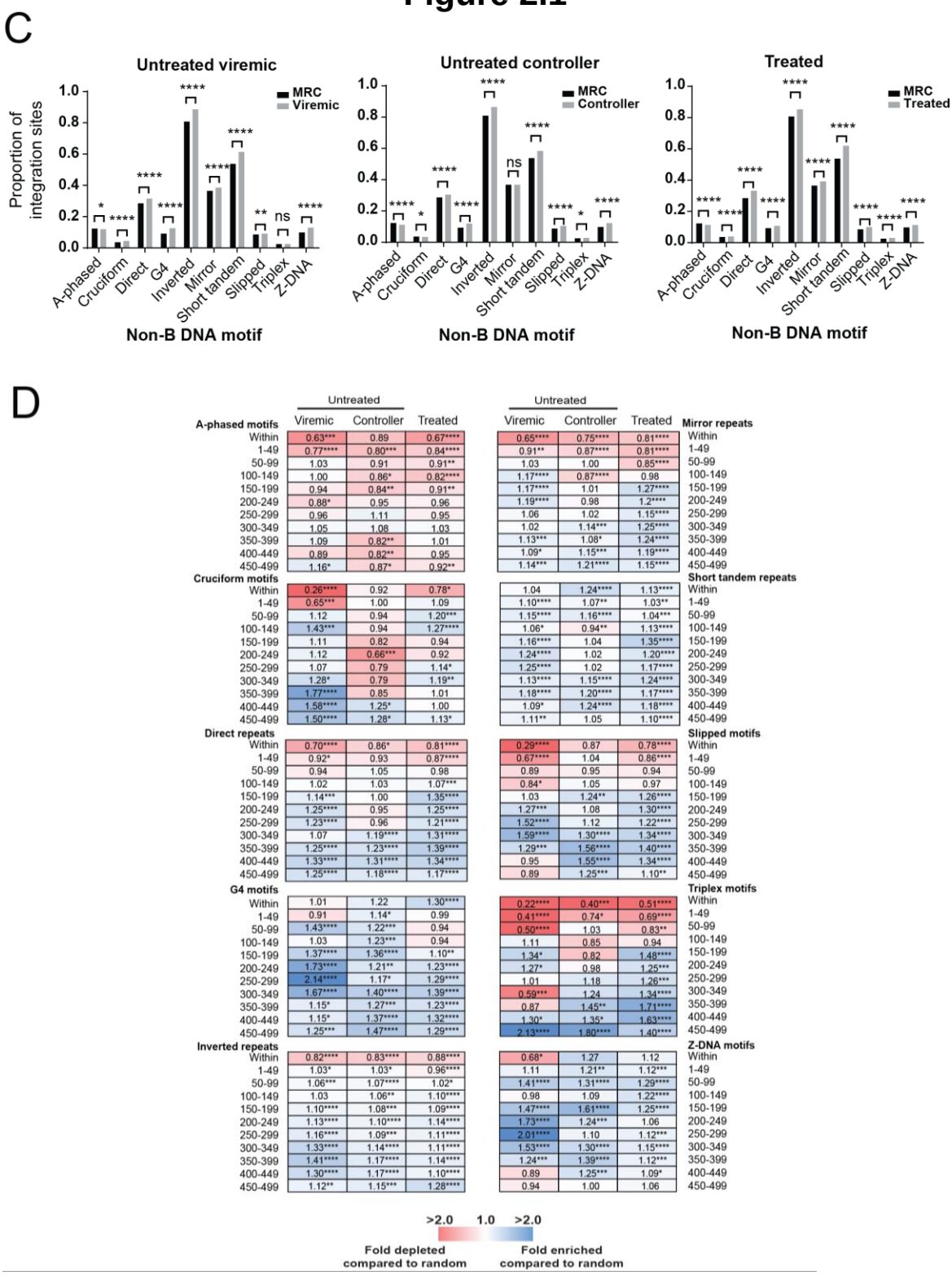


Figure 2.1: HIV-1 integration sites in quiescent/latently infected cells are enriched in and near non-B DNA motifs.

To assess the correlation between the same integration sites from the Cohn dataset and integration in or near non-B DNA motifs, we quantified integration sites within 500 bp of several different non-B DNA motifs from each of the treatment groups. As shown in **Figure 2.1C**, integration sites from each group were significantly enriched within 500 bp of all non-B DNA motifs examined, except for A-phased motifs.

To examine integration site placement more local to the non-B DNA motifs, we quantified integration sites directly within the non-B DNA motif itself or in distance bins of 50 bp up to 500 bp away from the feature. As shown in **Figure 2.1D**, the integration site profile differed among the different treatment groups, especially between the viremic and controller/treated groups. Notably, integration sites were enriched directly in G4 motifs, Z-DNA motifs and short tandem repeats of the controller and treated groups (compared to the MRC), whereas integration sites in the viremic group were enriched 150-399 bp away from these features. Integration directly in cruciform motifs, Z-DNA motifs and slipped motifs was strongly disfavored in the viremic group compared to both controller and treated groups. Furthermore, integration sites in the controller and treated groups were notably enriched 150-399 bp away from triplex motifs compared to the viremic group. Together, these data show that HIV-1 favors integration in and/or near non-B DNA motifs in infected individuals, and that a distinct integration site bias for specific non-B DNA motifs exists based on treatment status.

2.3.2 Integration near G4 motifs is associated with proviral reactivation in quiescent/latently infected cells

Recently, it was shown by Battivelli and colleagues that HIV-1 integration sites were distinguishable with respect to chromatin functional states and that these locations correlated with latency reactivation [16]. This integration site dataset was generated by infecting primary $CD4^+$ T cells with a novel dual-fluorescence HIV-1 reporter virus (HIV_{GKO}) designed for the accurate quantification and purification of a large number of latently infected cells as previously described [16]. Briefly, the HIV_{GKO} reporter is a dual-color reporter that carries the HIV-1 promoter in the 5'LTR driving a codon-switch eGFP (csGFP) expression and the cellular elongation factor one alpha (EF1α) promoter that drives expression of the monomeric kusabira-orange2 (mKO2) fluorescent protein. The

HIV_{GKO} reporter virus allows for quantification and discrimination of productively infected cells (csGFP positive, mKO2 positive), latently infected cells (csGFP negative, mKO2 positive) and uninfected cells (csGFP negative, mKO2 negative) **(Figure 2.2A)**. Sorted latently infected cells were then subjected to reactivation with the αCD3/CD28 LRA [16]. Overall, the Battivelli study involved three populations of HIV-1infected cells. The first population contained productively-infected cells (PIC), whereas the second population contained latently infected cells that could be reactivated with the αCD3/CD28 LRA which are the reactivated latently-infected cells (RLIC) and the third population contained cells that could not be reactivated with the αCD3/CD28 LRA which are the non-reactivated latently-infected cells (NRLIC) **(Figure 2.2A)**. We first assessed the integration site selection with respect to the most common genomic features. In agreement with the Battivelli study, integration sites in each of the PIC, RLIC and NRLIC populations were enriched in genes (compared to the MRC), with the majority of sites located in genes (82%, 70% and 59% respectively) **(Table 2.1, Figure 2.2 B and Supplemental Table 2.2)**.

We also observed that the frequency of integration directly in DNaseI hypersensitivity sites in the NRLIC population was not significantly different from the frequency expected for random placement. Also in agreement, integration sites in the NRLIC population were enriched in regions of heterochromatin (compared to the MRC), such as those containing satellite DNA ($P < 0.0001$). This is in contrast with the PIC population where integration sites were depleted in satellite DNA. The proportion of integration sites in heterochromatic lamin associated domains (LADs) were also enriched in the RLIC and NRLIC samples (25% and 29%) compared to the PIC samples (14%) **(Supplemental Table 2.2)**. Together, these data show that our bioinformatic analyses agree with the findings of the Battivelli study and further supports an integration site bias towards regions of heterochromatin in latently infected cells.

We then analyzed the Battivelli integration site dataset to determine if proviral reactivation in latently infected cells correlated with a distinct non-B DNA integration site profile. Integration sites in the PIC populations were enriched within 500 bp of direct repeats, inverted repeats, mirror repeats, short tandem repeats, and slipped motifs **(Figure 2.2 C**

Figure 2.2: Integration near G4 motifs is associated with proviral reactivation in quiescent/latently infected cells. (A) Schematic depicting isolation of the productively-infected cells (PIC), the reactivated latently-infected cells (RLIC) and the non-reactivated latently-infected cells (NRLIC) population from the Battivelli dataset. The αCD3/CD28 was used for reactivation of latently infected cells in the Battivelli dataset (**B**) Heatmap depicting the fold enrichment or depletion of integration sites in the most common genomic features compared to the matched random control (MRC). The 'RLIC+NRLIC' population was compared to the PIC population. **(C)** Bar graphs showing the proportion of unique HIV-1 integration sites in non-B DNA motifs. Asterisks denote significant differences from the MRC **(D)** Heatmap depicting the fold enrichment or depletion of integration sites in non-B DNA motifs compared to the MRC. The 'RLIC+NRLIC' population was compared to the PIC population. Within each heatmap, numbers represent the fold-change in the ratio of integration sites compared to MRC sites. Darker shades within each hetmap represent higher fold-changes. Within each heatmap, infinite number (inf) indicates that 1 or more integrations were observed when 0 integrations were expected by chance. Not a number (nan) indicates that 0 integrations were observed and 0 were expected by chance. Fisher's exact test was used for all comparisons. * $P < 0.05$, ** $P < 0.01$, *** $P < 0.001$, **** $P < 0.0001$.

Figure 2.2

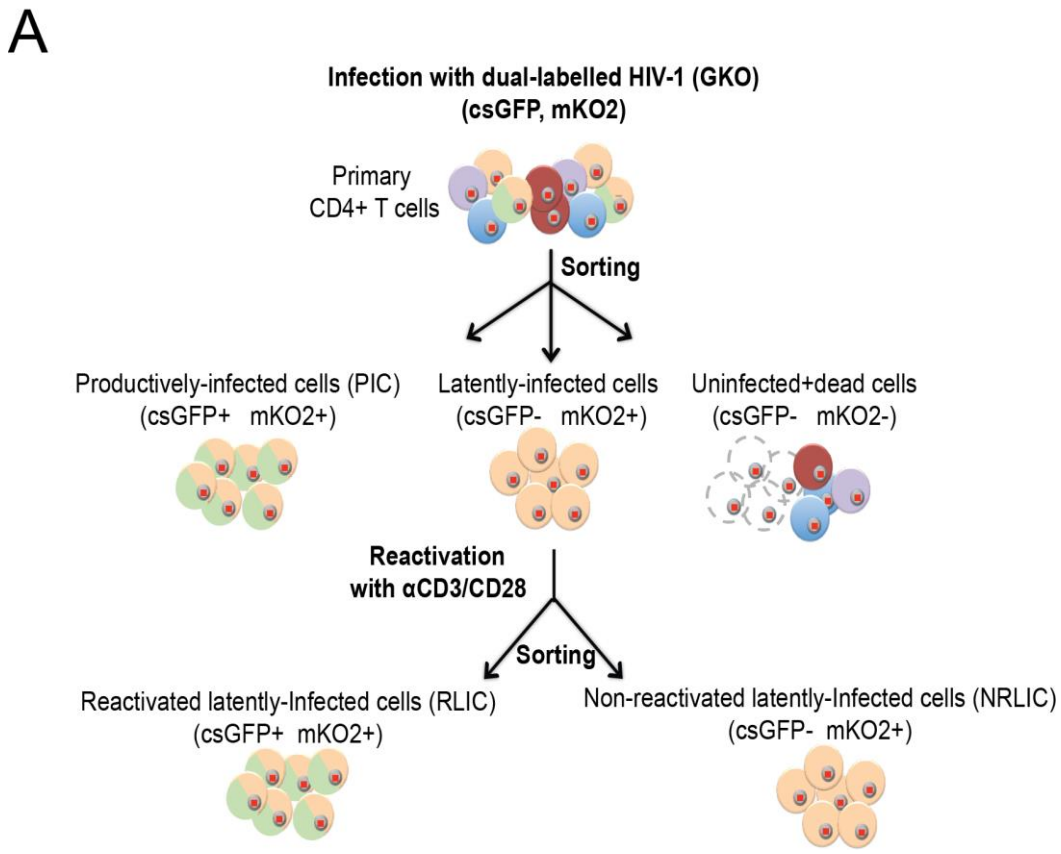


Figure 2.2: Integration near G4 motifs is associated with proviral reactivation in quiescent/latently infected cells.

Figure 2.2

B

Figure 2.2: Integration near G4 motifs is associated with proviral reactivation in quiescent/latently infected cells.

Figure 2.2

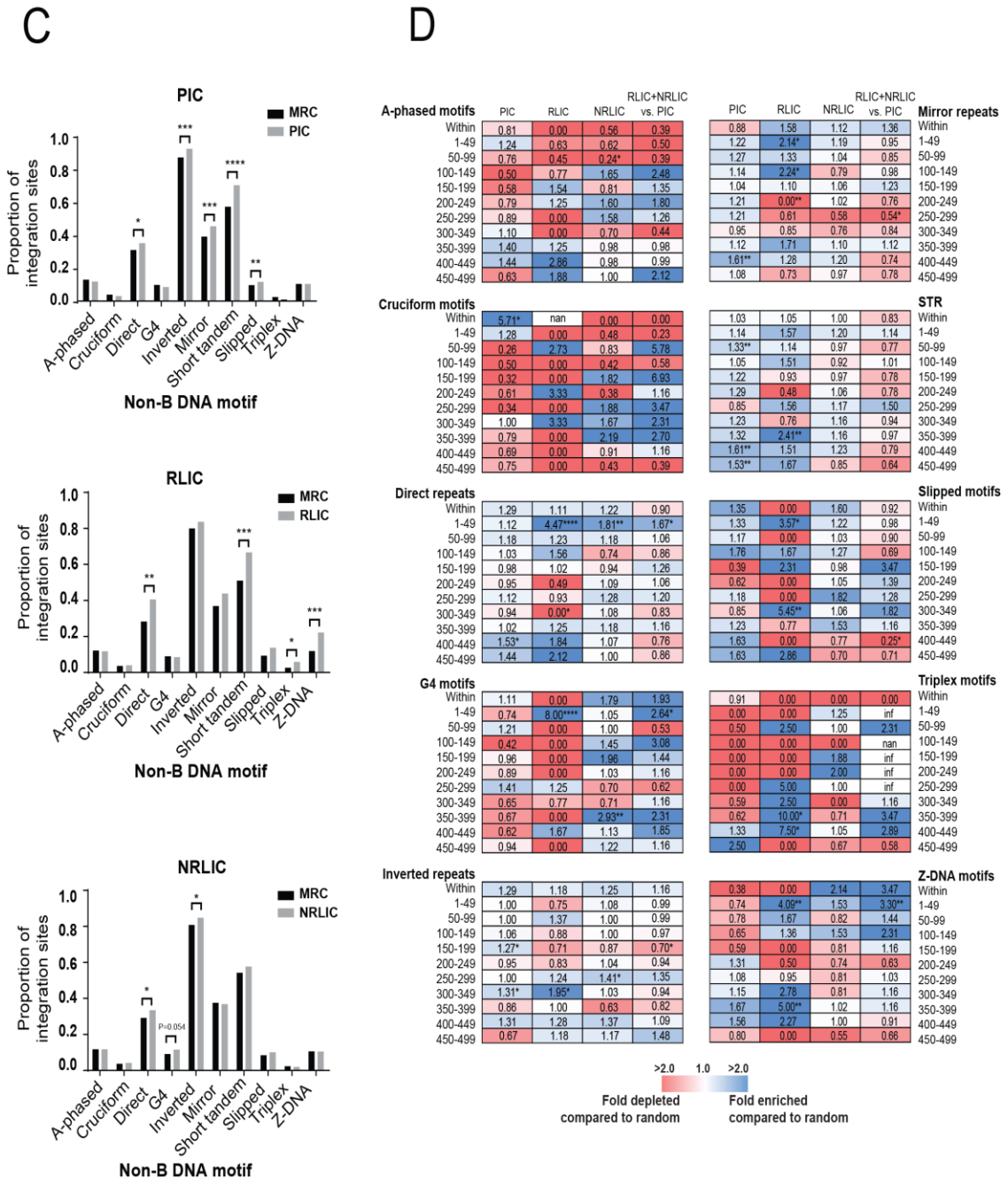


Figure 2.2: Integration near G4 motifs is associated with proviral reactivation in quiescent/latently infected cells.

and Supplemental Table 2.2). Integration sites in the RLIC populations were enriched within 500 bp of all non-B DNA motifs except for A-phased and G4 motifs. Integration sites in the NRLIC populations were enriched within 500 bp of all non-B DNA motifs except for A-phased motifs, mirror repeats, triplex motifs and Z-DNA motifs. To examine integration site placement more local to the non-B DNA motifs, we quantified integration sites directly in the non-B DNA motif itself or in bins of 50 bp up to 500 bp away from the feature. Distinct differences in the integration site profiles were observed for each of the different populations **(Figure 2.2D and Supplemental Table 2.2)**. Notable differences were identified for cruciform, G4, slipped, triplex and Z-DNA motifs, where integration was highly enriched in and/or near these features in the RLIC and NRLIC populations compared to the MRC. In contrast with the RLIC and NRLIC populations, integration sites in the PIC population were enriched in, but not near, cruciform motifs.

After comparing the frequency of integration sites in the RLIC+NRLIC populations with the PIC population, we observed strong enrichment of integration sites in and/or near most non-B DNA motifs, indicating that the latently infected populations were more enriched in these motifs compared to the productively infected population **(Figure 2.2D and Supplemental Table 2.2).** Distinct differences in integration site profiles were also observed between the RLIC and NRLIC populations, particularly with respect to G4, slipped and Z-DNA motifs. Integration sites in the NRLIC population were enriched in these motifs. In striking contrast, integration sites in the RLIC population were depleted in these features, but highly enriched within 1-49 bp of these features. Integration sites in the RLIC population were also highly enriched within a region 250-450 bp away from inverted repeats, short tandem repeats, triplex motifs and Z-DNA motifs compared to the NRLIC population. Together, these data show that integration sites in latently infected cells are more enriched in and/or near non-B DNA motifs compared to productively infected cells. Furthermore, reactivation of latent proviral expression correlates with integration site placement adjacent to, but not within, G4, slipped and Z-DNA motifs.

2.3.3 CSPF6 and LEDGF/p75 promote integration into specific non-B DNA

The interaction of CPSF6 with HIV-1 capsid protein licenses HIV-1 pre-integration complexes to bypass peripheral heterochromatin in the nucleus and penetrate the nuclear interior to locate gene-dense euchromatin for integration. Lens epithelium-derived growth factor (LEDGF/p75) is a host factor that tethers the HIV-1 pre-integration complex to euchromatin where it promotes integration into transcriptionally active genes [28,29,31–33,60–63,34]. To determine if CPSF6 and LEDGF/p75 influence the targeting of non-B DNA motifs for integration, we analyzed the recently published integration site dataset by Achuthan and colleagues (2018) who studied the impact of CPSF6 and LEDGF/p75 on integration site targeting [34]. In that study, CPSF6 function was depleted by using the HIV-1 capsid mutant A77V, which impairs CPSF6 binding efficiency without severely decreasing infectivity. LEDGF/p75 function was depleted by treating cells at the time of infection with the allosteric integrase inhibitor BI-D, which competes with integrase-LEDGF/p75 binding and inhibits HIV-1 integration.

Consistent with the findings of Achuthan and colleagues, independent CPSF6 depletion and LEDGF/p75 depletion resulted in decreased integration within genes and increased integration into heterochromatin (e.g. LADs and satellite DNA) compared to the wild type control **(Table 2.1 and Supplemental Table 2.3).** CPSF6 and LEDGF/p75 depletion also correlated with a strong reduction in integration sites within CpG islands compared to the control. With respect to non-B DNA, CPSF6 and LEDGF/p75 depletion resulted in substantial reductions in the percentage of integration sites falling within 500 bp of G4 motifs, mirror repeats, short tandem repeats and Z-DNA compared to the wild type control **(Figure 4, Table 2.1 and Supplemental Table 2.3).** Although not achieving statistical significance, increases in the percentage of integration sites falling near A-phased motifs, slipped motifs and triplex motifs were observed with CPSF6 and LEDGF/p75 depletion. Analysis of the distribution of integration sites around the non-B DNA motifs revealed that integration was generally disfavored at several 50 bp distance intervals from most non-B DNA features compared to the control cells.

Figure 2.3: CPSF6 and LEDGF/p75 promote integration into specific non-B DNA. Graphs show the percentage of total unique HIV-1 integration sites located in or within 500 bp of various non-B DNA motifs (distributed in 50 bp bins) from wild type (control), CPSF6 depleted or LEDGF/p75 depleted cells. Inset numbers show the percentage of total unique integration sites falling within 500 bp of the non-B DNA motif. Heatmaps show the fold enrichment (blue) or depletion (red) of integration sites at each distance interval from the non-B DNA motif compared to the matched random control (MRC). Fisher's exact test was used for all comparisons. * $P < 0.05$, ** $P < 0.01$, *** $P < 0.001$, **** $P < 0.0001$.

Figure 2.3

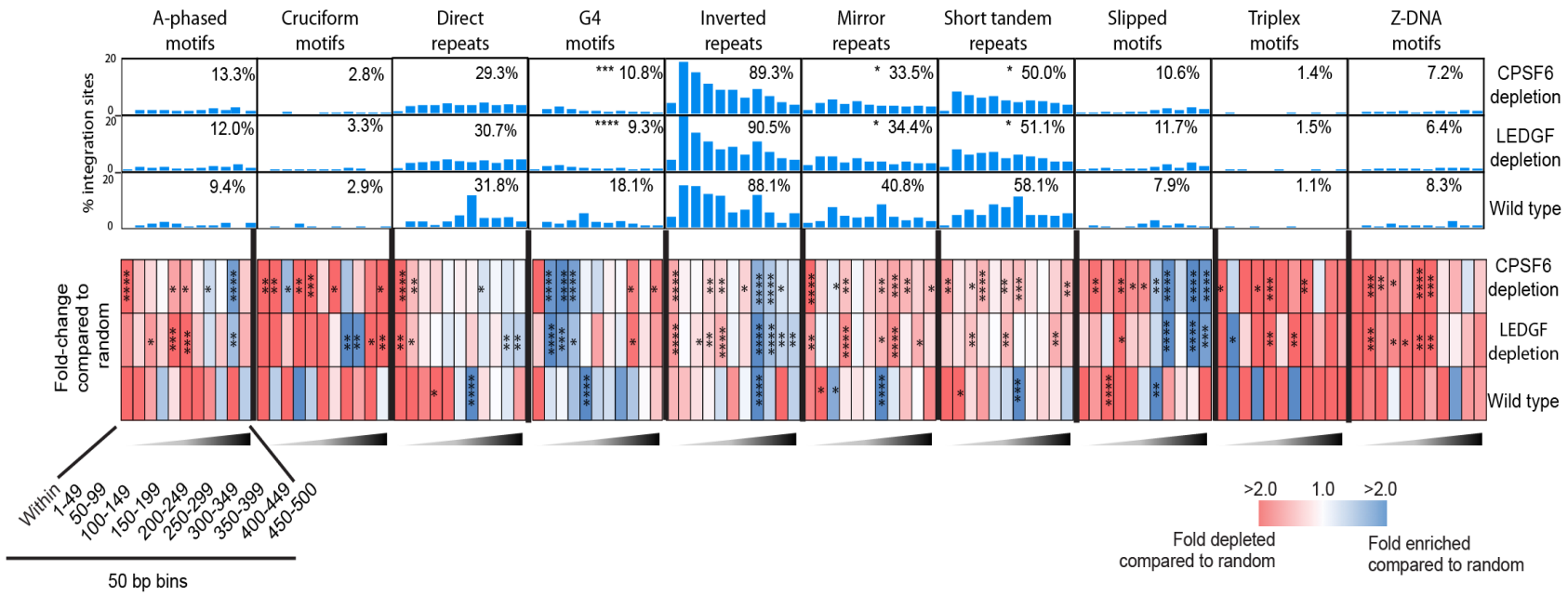


Figure 2.3: CPSF6 and LEDGF/p75 promote integration into specific non-B DNA.

However, integration sites were highly enriched 1-150 bp away from G4 motifs and 300-500 bp away from slipped motifs in cells depleted of CPSF6 compared to the control cells. Together, these data confirm that CPSF6 and LEDGF/p75 promote integration in genes and gene-dense euchromatin and show that they also promote integration near the non-B DNA features G4 motifs, mirror repeats, short tandem repeats and Z-DNA.

2.3.4 Clonally-expanded quiescent/latently infected cells exhibit a distinct non-B DNA integration site profile

Previous studies have demonstrated that a large fraction of HIV-1 infected cells in patients can arise from expansion of a single cellular clone **(Figure 2.4A)** [22–24]. Maldarelli and Wu and colleagues identified specific HIV-1 integration sites linked to this clonal expansion, particularly integration into genes involved in cellular growth, development and persistence [22]. We analyzed the Maldarelli/Wu dataset to determine if there are distinct non-B DNA integration site profiles for clonal and non-clonal populations of latently infected cells. Our analysis showed that the integration site profiles between these two populations of cells were highly similar with respect to several common genomic features with integration strongly favoring genes and DNaseI hypersensitivity sites **(Figure 2.4B, Table 2.1 and Supplemental Table 2.4)** when compared to the matched random control (MRC).

With respect to non-B DNA motifs, our analysis revealed that both clonal and non-clonal populations favored integration near many non-B DNA motifs, with the non-clonal population exhibiting significantly more enrichment near non-B DNA motifs compared to the MRC **(Figure 2.4C and Supplemental Table 2.4).** Integration sites in the clonal population were enriched near cruciform motifs, inverted repeats, mirror repeats and short tandem repeats, although significance was only achieved for inverted repeats. Integration sites in the non-clonal population were significantly enriched near direct repeats, inverted repeats, mirror repeats, short tandem repeats and slipped motifs. Analysis of integration site placement more local to the motifs revealed distinct integration site profiles between the clonal and non-clonal populations **(Figure 2.4D and Supplemental Table 2. 4).**

Figure 2.4: Clonally-expanded quiescent/latently infected cells exhibit a distinct non-B DNA integration site profile. (A) Schematic depicting the isolation of clonally-expanded and non-clonally-expanded populations in the Maldarelli/Wu study. **(B)** Heatmap depicting the fold enrichment or depletion of integration sites in common genomic features compared to the matched random control (MRC). **(C)** Proportion of unique HIV-1 integration sites in various non-B DNA motifs. Asterisks denote significant difference from MRC **(D)** Heatmap depicting the fold enrichment or depletion of integration sites in non-B DNA motifs of the clonal or non-clonal populations compared to the MRC or to each other. Darker shades represent higher fold-changes. Within each heatmap, infinite number (inf) indicates that 1 or more integrations were observed when 0 integrations were expected by chance. Not a number (nan) indicates that 0 integrations were observed and 0 were expected by chance. Fisher's exact test was used for all comparisons. * $P < 0.05$, ** $P < 0.01$, *** $P < 0.001$, **** $P < 0.0001$.

Figure 2.4

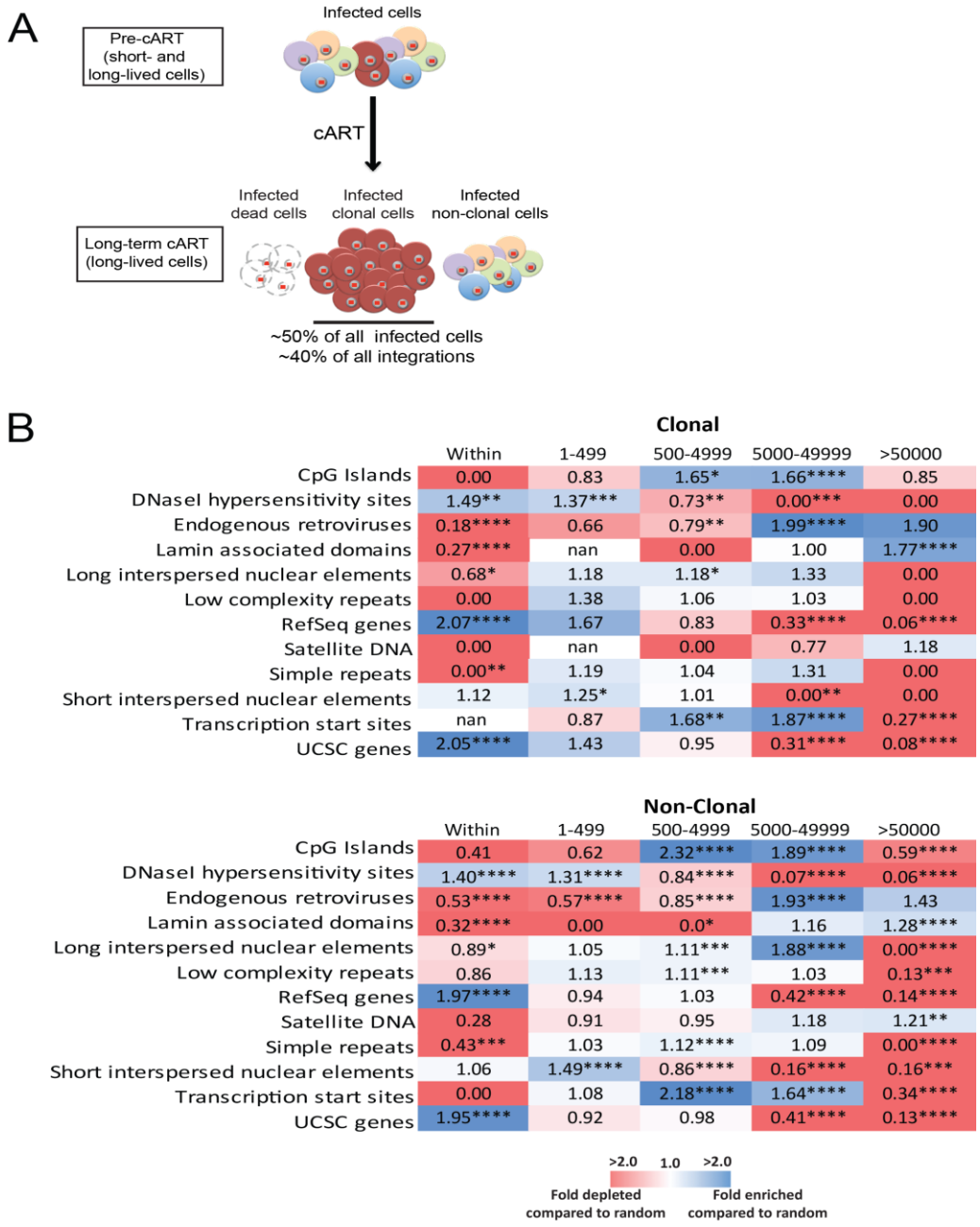
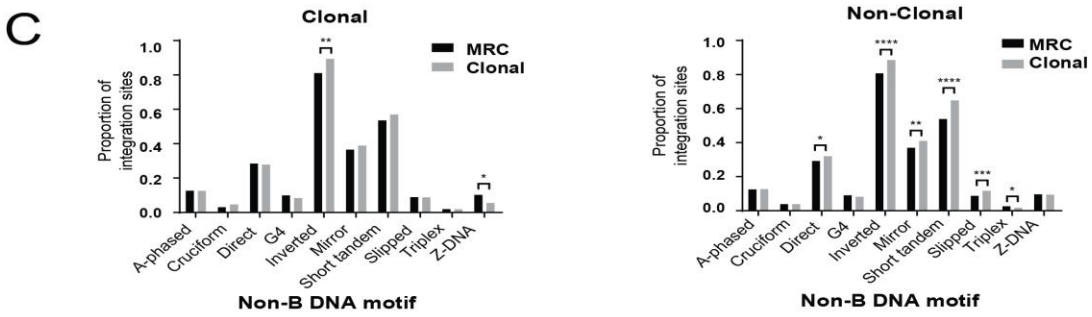


Figure 2.4: Clonally-expanded quiescent/latently infected cells exhibit a distinct non-B DNA integration site profile.

Figure 2.4



D

A-phased motifs	Clonal	Non-Clonal	Clonal vs. Non-Clonal
Within	0.00	0.68	0.00
1-49	1.11	1.22	0.91
50-99	0.37	0.90	0.44
100-149	1.32	0.95	1.84
150-199	1.90	1.04	1.40
200-249	1.07	1.09	1.05
250-299	0.29	0.90	0.41
300-349	1.87	1.05	1.10
350-399	0.00	0.86	0.00
400-449	1.67	0.92	0.87
450-499	1.79	1.33	1.59

Mirror repeats	Clonal	Non-Clonal	Clonal vs. Non-Clonal
Within	0.65	0.86	0.78
1-49	1.06	1.00	1.01
50-99	1.36	1.04	1.31
100-149	0.85	0.88	0.81
150-199	1.37	1.16	1.06
200-249	1.64	1.13	1.44
250-299	0.34	1.18	0.23*
300-349	0.80	1.43**	0.62
350-399	0.99	1.56**	0.75
400-449	1.33	1.01	1.52
450-499	1.03	1.05	1.00

Cruciform motifs	Clonal	Non-Clonal	Clonal vs. Non-Clonal
Within	nan	0.45	0.00
1-49	0.00	0.32	0.00
50-99	5.00	0.96	2.79
100-149	3.75	1.32	2.99
150-199	0.00	0.68	0.00
200-249	0.00	1.20	0.00
250-299	0.00	0.77	0.00
300-349	0.00	1.07	0.00
350-399	0.00	1.15	0.00
400-449	7.14**	1.51	4.36*
450-499	0.00	1.29	0.00

Short tandem repeats	Clonal	Non-Clonal	Clonal vs. Non-Clonal
Within	0.00	0.70	0.00
1-49	0.76	1.13	0.69
50-99	1.49	1.10	1.30
100-149	1.09	1.03	0.95
150-199	1.02	1.39***	0.76
200-249	1.14	1.24*	0.98
250-299	0.90	1.42**	0.54
300-349	0.47	1.23	0.43*
350-399	2.43**	1.40**	1.50
400-449	1.07	1.25	0.92
450-499	1.13	1.12	1.03

Direct repeats	Clonal	Non-Clonal	Clonal vs. Non-Clonal
Within	0.74	0.56	1.07
1-49	0.94	0.95	0.91
50-99	1.55	1.17	1.40
100-149	1.35	0.66*	1.58
150-199	0.95	1.02	0.93
200-249	1.13	1.32*	0.83
250-299	0.46	1.23	0.45
300-349	0.62	1.28	0.61
350-399	1.25	1.47**	0.56
400-449	1.15	0.94	1.40
450-499	0.43	1.34	0.32

Slipped motifs	Clonal	Non-Clonal	Clonal vs. Non-Clonal
Within	0.00	0.37	0.00
1-49	1.15	0.75	1.74
50-99	2.78	1.05	2.33
100-149	1.50	0.51	3.49
150-199	0.67	5.37****	0.11**
200-249	1.05	1.36	0.82
250-299	0.00	0.59	0.00
300-349	0.42	1.35	0.58
350-399	1.25	0.98	0.50
400-449	0.87	0.62	1.74
450-499	0.71	1.64	0.39

G4 motifs	Clonal	Non-Clonal	Clonal vs. Non-Clonal
Within	0.00	2.35*	0.00
1-49	1.74	0.79	2.15
50-99	1.92	0.81	3.49*
100-149	0.43	0.77	0.63
150-199	0.45	0.98	0.54
200-249	1.05	1.70*	0.73
250-299	1.43	0.70	2.09
300-349	0.48	0.93	0.54
350-399	0.40	0.45	1.40
400-449	0.00	0.36*	0.00
450-499	0.00	1.07	0.00

Triplex motifs	Clonal	Non-Clonal	Clonal vs. Non-Clonal
Within	0.00	0.62	0.00
1-49	nan	1.00	0.00
50-99	3.33	0.26	6.98
100-149	0.00	0.67	0.00
150-199	0.00	0.00*	nan
200-249	6.67	1.29	3.49
250-299	0.00	0.65	0.00
300-349	0.00	1.67	0.00
350-399	0.00	0.59	0.00
400-449	1.43	0.0*	inf
450-499	0.00	0.75	0.00

Inverted repeats	Clonal	Non-Clonal	Clonal vs. Non-Clonal
Within	1.40	1.15	1.19
1-49	1.01	1.07	1.00
50-99	0.73	1.12	0.66*
100-149	1.36	1.03	1.27
150-199	1.40	1.04	1.36
200-249	1.29	1.24*	1.00
250-299	0.89	1.09	0.88
300-349	0.78	1.19	0.64
350-399	1.33	1.06	1.31
400-449	1.25	0.92	1.14
450-499	1.07	1.25	0.89

Z-DNA motifs	Clonal	Non-Clonal	Clonal vs. Non-Clonal
Within	0.00	0.29	0.00
1-49	1.20	0.76	1.90
50-99	0.67	1.04	0.47
100-149	0.00	0.65	0.00
150-199	1.11	0.94	1.07
200-249	0.00	1.19	0.00
250-299	0.40	1.38	0.35
300-349	0.00	1.39	0.00
350-399	1.05	0.89	1.27
400-449	0.00	0.54	0.00
450-499	1.50	1.05	1.50

Figure 2.4: Clonally-expanded quiescent/latently infected cells exhibit a distinct non-B DNA integration site profile.

Integration was highly disfavored in all non-B DNA motifs except inverted repeats in the clonal population compared to the non-clonal population. However, integration sites in the clonal population were enriched in regions spanning 1-149 bp and/or 350-450 bp away from most motifs. Notably, integration sites in the non-clonal population were significantly enriched directly in G4 motifs. Together, these data identify distinct non-B DNA integration site profiles for clonally- and non-clonally-expanded cells.

2.3.5 G4 structure influences integration site targeting in the genome

The only non-B DNA motif where integration sites were consistently and significantly enriched in or near the motif in the latently infected populations from each of the Cohn, Battivelli and Maldarelli/Wu datasets was G4 motifs. To determine the influence of G4 structures on integration site targeting in the human genome during HIV-1 infection, we utilized G4 structure-stabilizing and -destabilizing ligands. G4 structures consist of four guanine bases which are stabilized by hydrogen bonds forming a G-tetrad in a planar arrangement **(Figure 2.5A)** [64]. BRACO19 is a 3,6,9-trisubstituted acridine derivative that interacts with and stabilizes G4 structures **(Figure 2.5B)** [52,65–69]. We asked if stabilization of G4 structures increased the frequency of integration in and/or near G4 motifs. HEK 293T cells treated with increasing concentrations of BRACO19 for 24 hours were infected with HIV-1 pseudotyped with the vesicular stomatitis virus G envelope glycoprotein (HIV/VSV-G). Twenty-four hours after infection, the integration site profile was determined for each drug concentration and compared to the infected untreated control cells. No significant reduction in cell viability was detected after treatment with BRACO19 using the MTT assay **(Figure 2.5C).** Treatment of cells with increasing concentrations of BRACO19 resulted in a substantial increase in the proportion of integration sites located 250-499 nucleotides away from G4 motifs **(Figure 2.5D, Table 2.1 and Supplemental Table 2.5).** In contrast, integration was strongly disfavored in and adjacent to (1-149 bp) G4 motifs. Interestingly, integration sites were highly enriched 150- 199 bp from G4 motifs at low concentrations of BRACO19 and became less enriched at higher concentrations.

Figure 2.5: G4 structure influences integration site targeting in the genome. (A) Depiction of the nucleoside arrangement of a guanine-quartet and the G4 structure. **(B)** Structure of the TMPyP4 and BRACO19 compounds. **(C)** HEK 293T cells were treated with increasing concentrations of BRACO19 or TMPyP4 for 48 hours and the percent cell viability was determined using the MTT assay. Data are represented as mean ± SEM of at least 3 independent experiments. **(D)** Heatmap depicting the fold enrichment or depletion of integration sites in G4 motifs compared to untreated infected cells. Significance was determined by Fisher's exact test. * $P < 0.05$, ** $P < 0.01$, *** $P < 0.001$, **** $P < 0.0001$. **(E)** Data from D shown spatially with respect to the G4 structure and flanking nucleosomes.

Figure 2.5

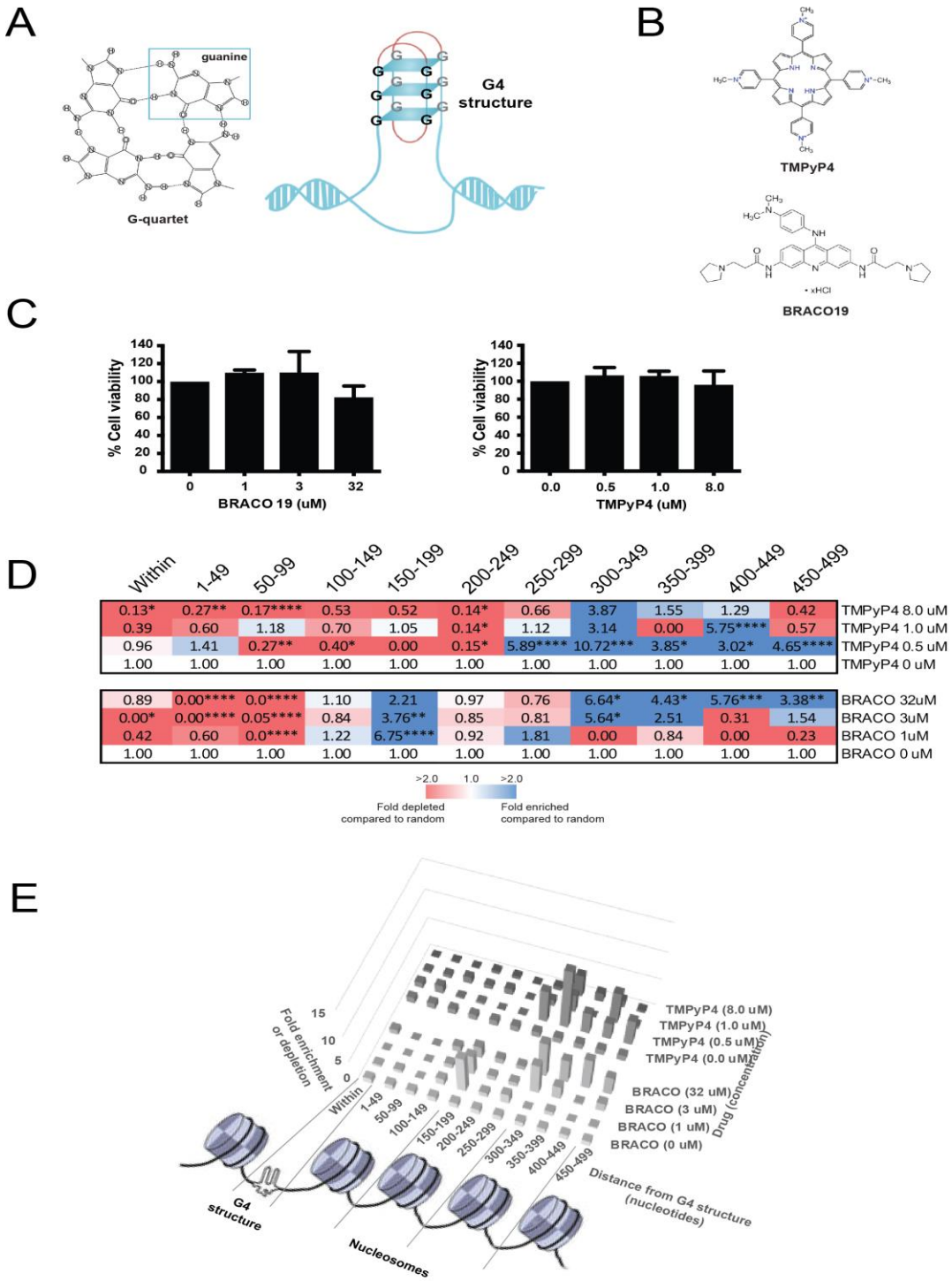


Figure 2.5: G4 structure influences integration site targeting in the genome.

The cationic porphyrin 5,10,15,20-tetra (N-methyl-4-pyridyl) porphin (TMPyP4) interacts with and destabilizes non-telomeric G4 structures while, paradoxically, stabilizing G4 structures located in telomeric DNA **(Figure 2.5B)** [53,54,69–72]. We asked if destabilization of non-telomeric G4 structures reduced integration in and/or near G4 motifs. HEK 293T control cells or cells treated with increasing concentrations of TMPyP4 were infected with HIV/VSV-G. No significant reduction in cell viability was detected after treatment with TMPyP4 using the MTT assay **(Figure 2.5C).**

Twenty-four hours after infection, the integration site profile was determined for each drug concentration and compared to the integration site profile from untreated cells. Treatment of cells with increasing concentrations of TMPyP4 resulted in a substantial reduction in the proportion of integration sites directly in, and 250-499 nucleotides away from, G4 motifs **(Figure 2.5D, Table 2.1 and Supplemental Table 2.5)**. Unexpectedly, low concentrations of TMPyP4 caused an enrichment in integration sites located 250-499 bp away from G4 motifs. Intriguingly, when integration site placement between the BRACO19 and TMPyP4 datasets were compared, the largest changes in integration site enrichment occurred between ~300-450 bp away from the G4 motif, consistent with a region of DNA located approximately three nucleosomes away **(Figure 2.5E).** Together, these data show that modulating G4 structure stability in the host genome significantly influences HIV-1 integration site targeting in and near G4 motifs.

2.3.6 Integration in or near G4 motifs favors G4 structures with long loops

Putative G4 structures are identified using the motif $G_xN_{y1}G_xN_{y2}G_xN_{y3}G_x$ [73]. The motif consists of four guanine tracts with three intervening loops **(Figure 2.6 A).** In this expression, x represents the number of guanine nucleotides, N_y1-N_y3 represent the 3 intervening loops and can be categorized as short-loop G4 structures (1-7 nucleotides) or long-loop G4 structures ($\geq$7 nucleotides) based on the number of nucleotides (N) in the loop. Loop-length has been shown to play an important role in G4 structure stability and protein-binding specificity [74,75]. We asked if HIV-1 integration sites in latently infected cells are biased towards G4 motifs with short or long loops and if loop-length is associated with clonal expansion or reactivation of latently infected cells.

Figure 2.6: Integration in or near G4 motifs favors G4 structures with long-loops. (A) Schematic depicting a G4 motif (left) and G4 structure (right) consisting of four adjacent runs of two or more guanines, with three loop regions of nucleotide subsequences (L1, L2 and L3) connecting the G-runs. Loops containing <7 nucleotides are considered short-loop (SL) G4s (solid red line), whereas ≥7 are considered long-loop (LL) G4s (dashed red line). **(B)** G4 motifs hosting integration sites or located in or within 500 bp upstream or downstream of an integration site were identified from the Cohn, Maldarelli/Wu and Bativelli datasets. The percentage of G4 motifs in each dataset classified as short-loops or long-loops were compared in the bar graphs. The average loop lengths for each of the three loops were calculated and are shown below the bar graphs.

Figure 2.6

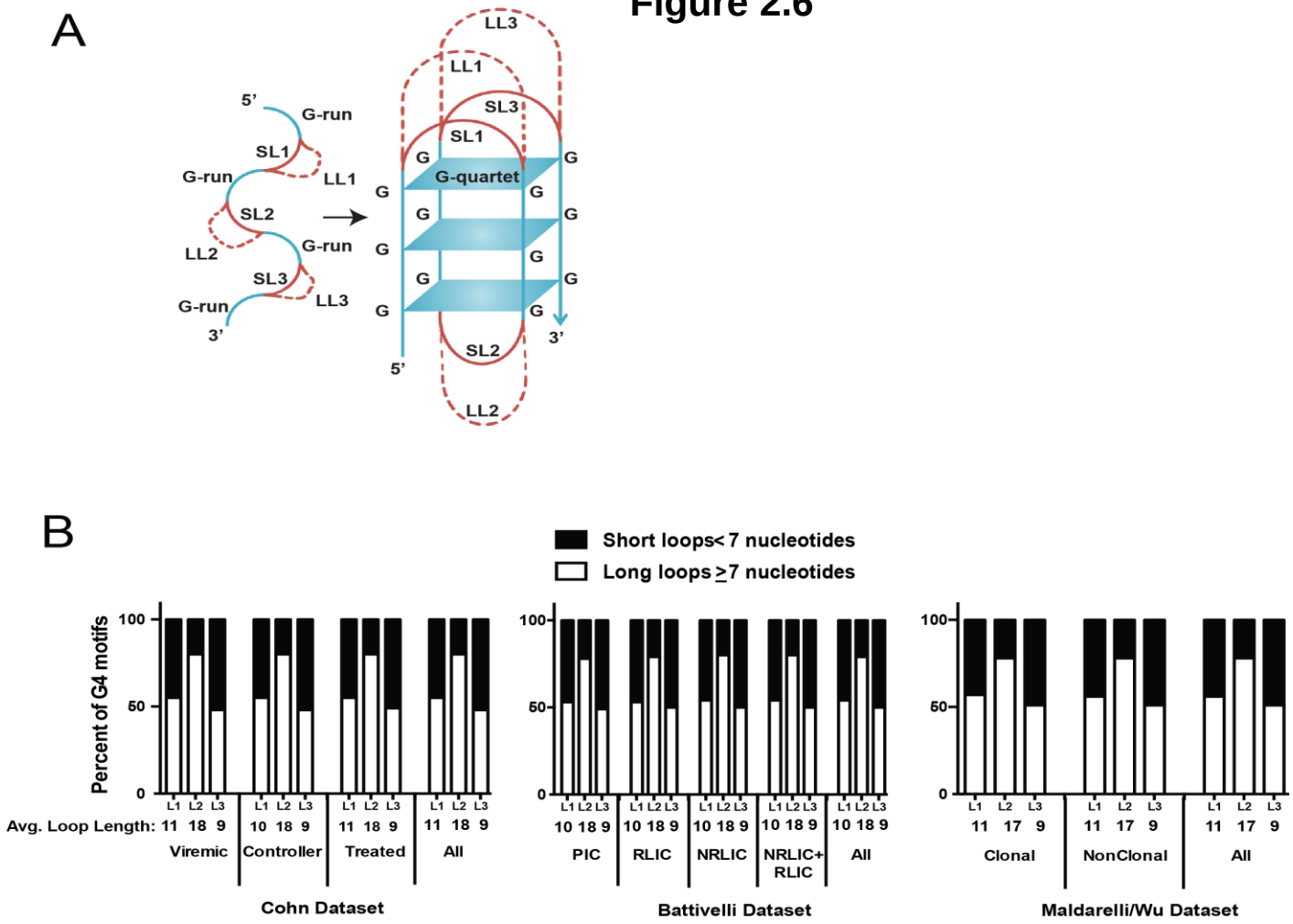


Figure 2.6: Integration in or near G4 motifs favors G4 structures with long-loops.

G4 motif sequences were extracted from the Cohn, Battivelli and Maldarelli/Wu datasets that hosted integration sites directly in the motif or within 500 bp upstream or downstream from the motif. The average loop-lengths were calculated and compared for each of the three loops from each dataset. The average loop-lengths were 10-11 nucleotides for loop 1, 17-18 nucleotides for loop 2 and 9 nucleotides for loop 3 **(Figure 2.6B)**. No substantial differences in the average loop-lengths were observed between any of the datasets. Notably, the majority of G4 motifs in each dataset contained a longer loop 2 than loop 1 or loop 3. Together, these data indicate that HIV-1 integration is biased towards long-loop G4 structures and that loop-length does not correlate with clonal expansion or reactivation potential of latently infected cells.

2.4 Discussion

The data presented herein show that non-B DNA motifs are novel features that influence HIV-1 integration site targeting in HIV-1-infected cells. Importantly, we identified non-B DNA as a genomic feature that correlates with the establishment and maintenance of HIV-1 latency. We showed that the locations of integration sites that predominate in latently infected cells are enriched in or near non-B DNA motifs, some of which are well-known to inhibit gene expression such as G4, cruciform, Z-DNA and triplex structures [40–50]. Latently infected cells, including those that underwent clonal expansion, also demonstrated a distinct non-B DNA integration site profile compared to non-clonally-expanded cells, with a bias for integration near specific types of non-B DNA, especially long-loop G4 motifs. Treating cells with G4 ligands that stabilize or destabilize G4 structures altered integration site preference for G4 motifs. Remarkably, integration adjacent to G4 motifs correlated with the ability of latent proviruses to be reactivated by LRAs.

The ability of HIV-1 to target non-B DNA structures for integration has several important implications for productive and latent infection. Numerous non-B DNA structures are associated with active genes *in vivo* and contribute to a dynamic interplay between DNA structure, chromatin organization and transcriptional activities [76]. In fact, non-B DNA structures are recognized by non-B DNA-specific transcription factors, leading to transcriptional activation [77–80]. Conversely, the unusual non-B DNA structure can block

binding of B-DNA-specific transcription factors, resulting in constitutive repression of adjacent genes [81–84]. The non-B DNA sequence itself might alter the intrinsic sequence preference of nucleosomes, thereby affecting nucleosome occupancy [76,85]. Similarly, the non-B DNA structure might sterically exclude nucleosomes, thereby affecting nucleosome positioning [76,86]. Indeed, G4 motifs form in nucleosome-free regions in the genome [87]. As such, this ability to locally and dynamically organize flanking nucleosomes may contribute to transcriptional regulation of adjacent genes and integrated proviruses. Intriguingly, our analyses revealed a notable enrichment of integration sites at intervals of ~150 bp away from non-B DNA motifs, which may be a result of this non-B DNA-induced repositioning of nucleosomes, which are comprised of ~147 bp of DNA wrapped around a histone octamer core [88].

Our data analyses are consistent with recent studies showing that integration does occur at low levels in heterochromatin, and that the frequency of integration in heterochromatin is much higher in latently infected cells compared to productively infected cells [16,18]. The factors that attract integration into heterochromatin are not fully understood. G4 structures, which are known to repress transcription, are highly localized to heterochromatin [82]. These structures are typically not found within DNA wrapped around a histone octamer, which could create a partially open state in heterochromatin for integration directly into or adjacent to these structures [87,89]. Notably, HIV integrase is known to bind directly to G4 motif-containing DNA [90–99]. Alternatively, it is also possible that G4-structure-binding proteins serve as tethers for the pre-integration complex as observed for LEDGF/p75.

It has been previously suggested that an optimally tuned bias for integrating into transcriptionally active (euchromatin) versus inactive (heterochromatin) regions of the genome may help establish a diverse latent viral reservoir [100,101,102]. CPSF6 and LEDGF/p75 are two host proteins that promote integration into euchromatin. Specifically, CPSF6 traffics the pre-integration complex away from peripheral heterochromatin in the nucleus towards gene-dense regions in the interior, whereas LEDGF/p75 promotes integration within transcriptionally active genes. We showed that both LEDGF/p75 and CPSF6 expression resulted in increased integration near non-B DNA motifs, especially those known to influence gene expression (e.g. G4 and Z-DNA motifs). Although the

mechanism by which LEDGF/p75 and CPSF6 increase integration near non-B DNA is unknown, it is possible that LEDGF/p75 and CPSF6 recognize certain non-B DNA structures (directly or indirectly via non-B DNA-binding proteins) and promote interactions between the pre-integration complex and the genomic DNA leading to integration.

Research efforts have attempted to purge the latent HIV-1 reservoir via LRAs to force expression of proviruses so that the infected cells can be cleared via the immune system or cytopathic effects [4,103]. Unfortunately, currently available agents have proven ineffective, reactivating only a small proportion (<5%) of cells carrying a latent provirus [16,18,19,104–106]. We showed that the ability of latent proviruses to become reactivated by αCD3/CD28 LRA correlated with integration sites situated adjacent to, but not in, G4, Z-DNA and slipped motifs. This is in contrast with non-reactivatable latent proviruses whose integration sites are enriched directly in the motifs and, in the case of G4, more distal (250-400 bp) to the motif. Given that G4 and Z-DNA have been shown to interfere with the assembly of transcription pre-initiation complexes and/or polymerase elongation, it is possible that expression of proviruses integrated near these non-B DNA motifs can be silenced by these structures, thereby contributing to latency [41,48,49]. In addition, the proximity of integration sites to these motifs may play an important role in the ability of LRAs to reactivate proviral expression.

The importance of clonal expansion of latently infected cells is not fully understood, but it is thought to be important for persistence of HIV-infected cells. The mechanism driving this clonal expansion is also unknown; however, there is a correlation of increased integration events in genes involved in the growth and development of cells [107]. Additionally, it is also thought that antigen stimulation and homeostatic cytokine-driven proliferation may contribute to clonal expansion thus helping maintain the latent reservoir [108,109]. In the present study, we observed enrichment of integration sites adjacent to several non-B DNA motifs, particularly G4 motifs, in clonally expanded cells. Interestingly, G4 structures and several other non-B DNA structures, are highly enriched in genes involved in growth and development or their promoters [76,110–114]. Abnormal expression of genes involved in developmental regulation can be detrimental (e.g. oncogenic transformation)

and are repressed most of the time. This is likely attributed to their localization in facultative heterochromatin. Non-B DNA structures located in the promoter of these genes may be incompatible with the assembly of the transcription complex, explaining the paucity of RNA polymerase II at these sites. It is possible that G4 and other non-B DNA motifs, play an important role in attracting integration into regions of the genome critical for prolonged persistence of expanded clones. This is biologically important given that a single integration event can generate a latently infected cell that could undergo clonal expansion, thereby seeding and maintaining the latent reservoir. While the current data identifies non-B DNA as a target for integration in clonally expanded cells, it is unclear how many clonal cells harbor replication competent virus. Therefore, it is possible that the number of integration sites could be inflated from cells containing defective virus.

Previous searches to identify a consensus sequence for integration site targeting have only revealed an apparent weak palindromic sequence at the site of insertion of several retroviruses. Recent work by Kirk and colleagues challenged this notion by showing that the palindromic consensus sequence arises in the population average as a consequence of non-palindromic motifs existing in equal proportions on the plus and minus strand of the target sequence [115]. Our study not only supports the notion that there is not likely a single palindromic consensus sequence at the integration site but shows that the integration sites are heterogeneous in nature, many of which fall into or near different types of non-B DNA motifs. In fact, we inspected each of the non-palindromic sequences identified by Kirk and colleagues in subpopulations of target integration sites from HTLV-1, HIV-1, MLV, ASLV and PFV (IV) and found that they all represent different non-B DNA sequences that are all predicted to form slipped-strand DNA structures. This shows that despite having distinct nucleotide consensus sequences (palindromic or non-palindromic), the nature of the sequence is such that it is predicted to form non-B DNA slipped-strand structures. Our findings using G4-stabilizing and -destabilizing ligands further highlights the likelihood that it may not be the primary DNA sequence itself that plays an important role in attracting the HIV-1 pre-integration complex, but rather the secondary structure formed by the non-B DNA motif itself. Additionally, we found that treating cells with G4 stabilizing and destabilizing ligands altered integration site preference for G4 motifs. It is important to note that in our study, infection was performed with pseudotyped HIV-1/VSV-G leading

to a single round of infection. As a result, it is possible that variation in integration site preference might occurs when compared to the integration site preference during prolonged, long-termed infection *in vivo.* We also showed that structural variations in these non-B DNA structures may also be important for attracting integration since HIV-1 demonstrated a strong bias for integration in or near long-loop G4 structures instead of short-loop G4 structures. Short-loop G4 motifs ($(TTAGGG)_n$) are highly enriched in the telomeres of chromosomes, for which no bias for HIV-1 integration has been observed.

In conclusion, our findings that HIV-1 integration sites in latently infected cells are enriched in and/or near non-B DNA motifs indicates that non-B DNA structures, particularly G4 structures, contribute to the establishment and maintenance of HIV-1 latency. Manipulation of these structures could be a novel approach for improving 'shock and kill' and/or 'block and lock' therapies aimed at HIV-1 cure.

2.5 References

1. Arts, E. J. & Hazuda, D. J. HIV-1 antiretroviral drug therapy. *Cold Spring Harb. Perspect. Med.* **2,** a007161 (2012).

2. Chun, T. W. *et al.* In vivo fate of HIV-1-infected T cells: quantitative analysis of the transition to stable latency. *Nat. Med.* **1,** 1284–1290 (1995).

3. Chun, T. W. *et al.* Early establishment of a pool of latently infected, resting $CD4^{(+)}$ T cells during primary HIV-1 infection. *Proc. Natl. Acad. Sci. U. S. A.* **95,** 8869–8873 (1998).

4. Deeks, S. G. HIV: Shock and kill. *Nature* **487,** 439–440 (2012).

5. Finzi, D. *et al.* Latent infection of CD4+ T cells provides a mechanism for lifelong persistence of HIV-1, even in patients on effective combination therapy. *Nat. Med.* **5,** 512–7 (1999).

6. Siliciano, J. D. *et al.* Long-term follow-up studies confirm the stability of the latent reservoir for HIV-1 in resting CD4+ T cells. *Nat. Med.* **9,** 727–8 (2003).

7. Archin, N. M. *et al.* Immediate antiviral therapy appears to restrict resting CD4+ cell HIV-1 infection without accelerating the decay of latent infection. *Proc. Natl. Acad. Sci. U. S. A.* **109,** 9523–8 (2012).

8. Rong, L. & Perelson, A. S. Modeling latently infected cell activation: Viral and latent reservoir persistence, and viral blips in HIV-infected patients on potent therapy. *PLoS Comput. Biol.* **5,** (2009).

9. Palmer, S. *et al.* 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 (2003).

10. Debbie S. Ruelas and Warner C. Greene. An integrated overview of HIV-1 latency. *Cell* **155,** 519–529 (2013).

11. Dahabieh, M., Battivelli, E. & Verdin, E. Understanding HIV latency: The road to an HIV cure. *Annu. Rev. Med.* **66,** 407–421 (2015).

12. Hamer., D. H. Can HIV be cured? Mechanisms of HIV persistence and strategies to combat it. *Curr. HIV Res.* **2,** 99–111 (2004).

13. Geeraert, L., Kraus, G. & Pomerantz, R. J. Hide-and-seek: The challenge of viral persistence in HIV-1 infection. *Annu. Rev. Med.* **59,** 487–501 (2008).

14. Savarino, A. *et al.* 'Shock and kill' effects of class I-selective histone deacetylase inhibitors in combination with the glutathione synthesis inhibitor buthionine sulfoximine in cell line models for HIV-1 quiescence. *Retrovirology* **6,** (2009).

15. Rasmussen, T. A. & Lewin, S. R. Shocking HIV out of hiding: Where are we with clinical trials of latency reversing agents? *Curr. Opin. HIV AIDS* **11,** 394–401 (2016).

16. Battivelli, E. *et al.* Distinct chromatin functional states correlate with HIV latency reactivation in infected primary CD4+ T cells. *Elife* **7,** e34655 (2018).

17. Archin, N. M. *et al.* Administration of vorinostat disrupts HIV-1 latency in patients on antiretroviral therapy. *Nature* **487,** 482–5 (2012).

18. Jordan, A., Defechereux, P. & Verdin, E. The site of HIV-1 integration in the human genome determines basal transcriptional activity and response to Tat transactivation. *EMBO J.* **20,** 1726–38 (2001).

19. Chen, H.-C., Martinez, J. P., Zorita, E., Meyerhans, A. & Filion, G. J. Position effects influence HIV latency reversal. *Nat. Struct. Mol. Biol.* **24,** 47–54 (2017).

20. Sherrill-Mix, S. *et al.* HIV latency and integration site placement in five cell-based models. *Retrovirology* **10,** 90 (2013).

21. Dahabieh, M. S. *et al.* Direct non-productive HIV-1 infection in a T-cell line is driven by cellular activation state and NFκB. *Retrovirology* **11,** 17 (2014).

22. Maldarelli, F. *et al.* Specific HIV integration sites are linked to clonal expansion and persistence of infected cells. *Science* **345,** 179–183 (2014).

23. Simonetti, F. R. *et al.* Clonally expanded $CD4^+$ T cells can produce infectious HIV-1 in vivo. *Proc. Natl. Acad. Sci.* **113,** 1883–1888 (2016).

24. Cohn, L. B. *et al.* HIV-1 integration landscape during latent and active infection. *Cell* **160,** 420–432 (2015).

25. Ho, Y.-C. *et al.* Replication-competent noninduced proviruses in the latent reservoir increase barrier to HIV-1 cure. *Cell* **155,** 540–551 (2013).

26. Jordan, A., Bisgrove, D. & Verdin, E. HIV reporducibly establishes a latent infection after acute infection of T cells ni vitro. *EMBO J.* **22,** 1868–1877 (2003).

27. Schroder, A. *et al.* HIV-1 Integration in the human genome favors active genes and local hotspots. *Cell* **110,** 521–529 (2002).

28. Cherepanov, P. *et al.* HIV-1 integrase forms stable tetramers and associates with LEDGF/p75 protein in human cells. *J.Biol.Chem.* **278,** 372–381 (2003).

29. Maertens, G. *et al.* LEDGF/p75 is essential for nuclear and chromosomal targeting of HIV-1 integrase in human cells. *J.Biol.Chem.* **278,** 33528–33539 (2003).

30. Marshall, H. M. *et al.* Role of PSIP1/LEDGF/p75 in lentiviral infectivity and integration targeting. *PLoS One* **2,** e1340 (2007).

31. Ciuffi, A. *et al.* A role for LEDGF/p75 in targeting HIV DNA integration. *Nat. Med.* **11,** 1287–9 (2005).

32. Shun, M. C. *et al.* LEDGF/p75 functions downstream from preintegration complex formation to effect gene-specific HIV-1 integration. *Genes Dev.* **21,** 1767–1778 (2007).

33. Singh, P. K. *et al.* LEDGF/p75 interacts with mRNA splicing factors and targets HIV-1 integration to highly spliced genes. *Genes Dev.* **29,** 2287–2297 (2015).

34. Achuthan, V. *et al.* Capsid-CPSF6 interaction licenses nuclear HIV-1 trafficking to sites of viral DNA integration. *Cell Host Microbe* **24,** 392–404 (2018).

35. Sowd, G. A. *et al.* A critical role for alternative polyadenylation factor CPSF6 in targeting HIV-1 integration to transcriptionally active chromatin. *Proc. Natl. Acad. Sci.* **113,** E1054–E1063 (2016).

36. Lee, K. *et al.* Flexible use of nuclear import pathways by HIV-1. *Cell Host Microbe* **7,** 221–233 (2010).

37. McAllister, R. G. *et al.* Lentivector integration sites in ependymal cells from a model of metachromatic leukodystrophy: non-B DNA as a new factor influencing integration. *Mol. Ther. Nucleic Acids* **3,** e187 (2014).

38. Bacolla, A. & Wells, R. D. Non-B DNA conformations, genomic rearrangements, and human disease. *J. Biol. Chem.* **279,** 47411–4 (2004).

39. Wahls, W. P., Wallace, L. J. & Moore, P. D. The Z-DNA motif d(TG)30 promotes reception of information during gene conversion events while stimulating

homologous recombination in human cells in culture. *Mol. Cell. Biol.* **10,** 785–93 (1990).

40. Siddiqui-Jain, A., Grand, C. L., Bearss, D. J. & Hurley, L. H. Direct evidence for a G-quadruplex in a promoter region and its targeting with a small molecule to repress c-MYC transcription. *Proc. Natl. Acad. Sci. U. S. A.* **99,** 11593–8 (2002).

41. Verma, A., Yadav, V. K., Basundra, R., Kumar, A. & Chowdhury, S. Evidence of genome-wide G4 DNA-mediated gene expression in human cancer cells. *Nucleic Acids Res.* **37,** 4194–4204 (2009).

42. Waga, S., Mizuno, S. & Yoshida, M. Chromosomal protein HMG1 removes the transcriptional block caused by the cruciform in supercoiled DNA. *J. Biol. Chem.* **265,** 19424–8 (1990).

43. Waga, S., Mizuno, S. & Yoshida, M. Nonhistone protein HMG1 removes the transcriptional block caused by left-handed Z-form segment in a supercoiled DNA. *Biochem. Biophys. Res. Commun.* **153,** 334–9 (1988).

44. Jain, A., Magistri, M., Napoli, S., Carbone, G. M. & Catapano, C. V. Mechanisms of triplex DNA-mediated inhibition of transcription initiation in cells. *Biochimie* **92,** 317–320 (2010).

45. Maher, L. J., Dervan, P. B. & Wold, B. Analysis of promoter-specific repression by triple-helical DNA complexes in a eukaryotic cell-free transcription system. *Biochemistry* **31,** 70–81 (1992).

46. Bochman, M. L., Paeschke, K. & Zakian, V. a. DNA secondary structures: stability and function of G-quadruplex structures. *Nat. Rev. Genet.* **13,** 770–80 (2012).

47. Brázda, V., Laister, R. C., Jagelská, E. B. & Arrowsmith, C. Cruciform structures are a common DNA feature important for regulating biological processes. *BMC Mol. Biol.* **12,** 33 (2011).

48. Delic J, Onclercq R, M.-C. M. Inhibition and enhancement of eukaryotic gene

expression by potential non-B DNA sequences. *Biochem Biophys Res Commun* **180,** 1273–83 (1991).

49. Tornaletti, S., Park-Snyder, S. & Hanawalt, P. C. G4-forming sequences in the non-transcribed DNA strand pose blocks to T7 RNA polymerase and mammalian RNA polymerase II. *J. Biol. Chem.* **283,** 12756–62 (2008).

50. Belotserkovskii, B. P. *et al.* A triplex-forming sequence from the human c-MYC promoter interferes with DNA transcription. *J. Biol. Chem.* **282,** 32433–32441 (2007).

51. Barr, S. D. *et al.* HIV integration site selection: targeting in macrophages and the effects of different routes of viral entry. *Mol. Ther.* **14,** 218–225 (2006).

52. Perrone, R. *et al.* Anti-HIV-1 activity of the G-quadruplex ligand BRACO-19. *J. Antimicrob. Chemother.* **69,** 3248–3258 (2014).

53. Ofer, N., Weisman-Shomer, P., Shklover, J. & Fry, M. The quadruplex r(CGG)n destabilizing cationic porphyrin TMPyP4 cooperates with hnRNPs to increase the translation efficiency of fragile X premutation mRNA. *Nucleic Acids Res.* **37,** 2712–2722 (2009).

54. Morris, M. J., Wingate, K. L., Silwal, J., Leeper, T. C. & Basu, S. The porphyrin TmPyP4 unfolds the extremely stable G-quadruplex in MT3-MMP mRNA and alleviates its repressive effect to enhance translation in eukaryotic cells. *Nucleic Acids Res.* **40,** 4137–45 (2012).

55. Cer, R. Z. *et al.* Non-B DB v2.0: A database of predicted non-B DNA-forming motifs and its associated tools. *Nucleic Acids Res.* **41,** 94–100 (2013).

56. Cer, R. Z. *et al.* Non-B DB: a database of predicted non-B DNA-forming motifs in mammalian genomes. *Nucleic Acids Res.* **39,** D383-91 (2011).

57. Guelen, L. *et al.* Domain organization of human chromosomes revealed by mapping of nuclear lamina interactions. *Nature* **453,** 948–951 (2008).

58. Ciuffi, A. & Barr, S. D. Identification of HIV integration sites in infected host genomic DNA. *Methods* **53,** 39–46 (2011).

59. Bushman, F. *et al.* Genome-wide analysis of retroviral DNA integration. *Nat. Rev. Microbiol.* **3,** 848–58 (2005).

60. Sowd, G. A. *et al.* A critical role for alternative polyadenylation factor CPSF6 in targeting HIV-1 integration to transcriptionally active chromatin. *Pnas* **113,** E1054-63 (2016).

61. KyeongEun Lee, Zandrea Ambrose, Thomas D. Martin, Ilker Oztop, A., Mulky, John G. Julias, Nick Vandegraaff, Joerg G. Baumann, Rui Wang, W., Yuen, Taichiro Takemura1 Kenneth Shelton, Ichiro Taniuchi, Yuan Li, J., Sodroski, Dan R. Littman, John M. Coffin, Stephen H. Hughes, Derya Unutmaz, A. & Engelman, and V. N. K. Flexible Use of Nuclear Import Pathways by HIV-1. *Cell Host Microbe* **7,** 221–233 (2006).

62. Marshall, H. M. *et al.* Role of PSIP 1/LEDGF/p75 in lentiviral infectivity and integration targeting. *PLoS One* **2,** e1340 (2007).

63. Shinn, P. *et al.* HIV-1 integration in the human genome favors active genes and local hotspots. *Cell Press* **110,** 521–529 (2002).

64. Huppert, J. L. Structure , location and interactions of G-quadruplexes. *FEBS J.* **277,** 3452–3458 (2010).

65. White, E. W. *et al.* Structure-specific recognition of quadruplex DNA by organic cations: Influence of shape, substituents and charge. *Biophys. Chem.* **126,** 140–153 (2007).

66. Read, M. *et al.* Structure-based design of selective and potent G quadruplex-mediated telomerase inhibitors. *Proc. Natl. Acad. Sci.* **98,** 4844–4849 (2001).

67. Burger, A. M. *et al.* The G-quadruplex-interactive molecule BRACO-19 inhibits tumor growth, consistent with telomere targeting and interference with telomerase

function. *Cancer Res.* **65,** 1489–1496 (2005).

68. Tippana, R., Hwang, H., Opresko, P. L., Bohr, V. A. & Myong, S. Single-molecule imaging reveals a common mechanism shared by G-quadruplex-resolving helicases. *Proc. Natl. Acad. Sci. U. S. A.* **113,** 8448–53 (2016).

69. Hu, M.-H. *et al.* Specific targeting of telomeric multimeric G-quadruplexes by a new triaryl-substituted imidazole. *Nucleic Acids Res.* **45,** 1606–1618 (2017).

70. Weisman-Shomer, P. *et al.* The cationic porphyrin TMPyP4 destabilizes the tetraplex form of the fragile X syndrome expanded sequence d(CGG)n. *Nucleic Acids Res.* **31,** 3963–3970 (2003).

71. Han, H., Langley, D. R., Rangan, A. & Hurley, L. H. Selective interactions of cationic porphyrins with G-quadruplex structures. *J. Am. Chem. Soc.* **123,** 8902–8913 (2001).

72. Grand, C. L. *et al.* The cationic porphyrin TMPyP4 down-regulates c-MYC and human telomerase reverse transcriptase expression and inhibits tumor growth in vivo. *Mol. Cancer Ther.* **1,** 565–73 (2002).

73. Frees, S., Menendez, C., Crum, M. & Bagga, P. S. QGRS-Conserve: A computational method for discovering evolutionarily conserved G-quadruplex motifs. *Hum. Genomics* **8,** 1–13 (2014).

74. Lago, S., Tosoni, E., Nadai, M., Palumbo, M. & Richter, S. N. The cellular protein nucleolin preferentially binds long-looped G-quadruplex nucleic acids. *Biochim. Biophys. Acta - Gen. Subj.* **1861,** 1371–1381 (2017).

75. Bugaut, A. & Balasubramanian, S. A sequence-independent study of the influence of short loop lengths on the stability and topology of intramolecular DNA G-quadraplexes. *Biochemistry* **47,** 689–697 (2008).

76. Kouzine, F. *et al.* Permanganate/S1 nuclease footprinting reveals non-B DNA structures with regulatory potential across a mammalian genome. *Cell Syst.* **4,** 344–

356.e7 (2017).

77. Brooks, T. A. & Hurley, L. H. The role of supercoiling in transcriptional control of MYC and its importance in molecular therapeutics. *Nat. Rev. Cancer* **9,** 849–861 (2009).

78. Cogoi, S., Shchekotikhin, A. E. & Xodo, L. E. HRAS is silenced by two neighboring G-quadruplexes and activated by MAZ, a zinc-finger transcription factor with DNA unfolding property. *Nucleic Acids Res.* **42,** 8379–8388 (2014).

79. Kang, H.-J. *et al.* Novel interaction of the Z-DNA binding domain of human ADAR1 with the oncogenic c-Myc promoter G-quadruplex. *J. Mol. Biol.* **426,** 2594–2604 (2014).

80. Murat, P. & Balasubramanian, S. Existence and consequences of G-quadruplex structures in DNA. *Curr. Opin. Genet. Dev.* **25,** 22–29 (2014).

81. Michelotti, G. A. *et al.* Multiple single-stranded cis elements are associated with activated chromatin of the human c-myc gene in vivo. *Mol. Cell. Biol.* **16,** 2656–69 (1996).

82. Hoffmann, R. F. *et al.* Guanine quadruplex structures localize to heterochromatin. *Nucleic Acids Res.* **44,** 152–163 (2016).

83. Lam, E. Y. N., Beraldi, D., Tannahill, D. & Balasubramanian, S. G-quadruplex structures are stable and detectable in human genomic DNA. *Nat. Commun.* **4,** 1796 (2013).

84. Ray, B. K., Dhar, S., Henry, C., Rich, A. & Ray, A. Epigenetic regulation by Z-DNA silencer function controls cancer-associated ADAM-12 expression in breast cancer: Cross-talk between MeCP2 and NF1 transcription factor family. *Cancer Res.* **73,** 736–744 (2013).

85. Struhl, K. & Segal, E. Determinants of nucleosome positioning. *Nat. Struct. Mol. Biol.* **20,** 267–273 (2013).

86. Sadeh, R. & Allis, C. D. Genome-wide 'Re'-Modeling of Nucleosome Positions. *Cell* **147,** 263–266 (2011).

87. Wong, H. M. & Huppert, J. L. Stable G-quadruplexes are found outside nucleosome-bound regions. *Mol. Biosyst.* **5,** 1713–1719 (2009).

88. Timothy J. Richmond & Curt A. Davey. The structure of DNA in the nucleosome core. *Nature* **423,** 145–150 (2003).

89. van Holde, K. & Zlatanova, J. Unusual DNA structures, chromatin and transcription. *Bioessays* **16,** 59–68 (1994).

90. Ojwang, J. O. *et al.* T30177, an oligonucleotide stabilized by an intramolecular guanosine octet, is a potent inhibitor of laboratory strains and clinical isolates of human immunodeficiency virus type 1. *Antimicrob. Agents Chemother.* **39,** 2426–2435 (1995).

91. Rando, R. F. *et al.* Suppression of human immunodeficiency virus type 1 activity in vitro by oligonucleotides which form intramolecular tetrads. *Journal of Biological Chemistry* **270,** 1754–1760 (1995).

92. Mazumder, A. N. *et al.* Inhibition of human immunodeficiency virus type 1 integrase by guanosine quartet structures. *Biochemistry* **35,** 13762–13771 (1996).

93. Jing, N., Rando, R. F., Pommier, Y. & Hogan, M. E. Ion selective folding of loop domains in a potent anti-HIV oligonucleotide. *Biochemistry* **36,** 12498–12505 (1997).

94. Jing, N. *et al.* Mechanism of inhibition of HIV-1 integrase by G-tetrad-forming oligonucleotides in vitro. *J. Biol. Chem.* **275,** 21460–21467 (2000).

95. De Soultrait, V. R. *et al.* DNA aptamers derived from HIV-1 RNase H inhibitors are strong anti-integrase agents. *J. Mol. Biol.* **324,** 195–203 (2002).

96. Phan, A. T. *et al.* From The Cover: An interlocked dimeric parallel-stranded DNA

quadruplex: A potent inhibitor of HIV-1 integrase. *Proc. Natl. Acad. Sci.* **102,** 634–639 (2005).

97. Koizumi, M. *et al.* Biologically active oligodeoxyribonucleotides-IX. Synthesis and anti-HIV-1 activity of hexadeoxyribonucleotides, TGGGAG, bearing 3'- and 5'-end-modification. *Bioorg. Med. Chem.* **5,** 2235–43 (1997).

98. Urata, H., Kumashiro, T., Kawahata, T., Otake, T. & Akagi, M. Anti-HIV-1 activity and mode of action of mirror image oligodeoxynucleotide analogue of zintevir. *Biochem. Biophys. Res. Commun.* **313,** 55–61 (2004).

99. Pedersen, E. B., Nielsen, J. T., Nielsen, C. & Filichev, V. V. Enhanced anti-HIV-1 activity of G-quadruplexes comprising locked nucleic acids and intercalating nucleic acids. *Nucleic Acids Res.* **39,** 2470–2481 (2011).

100. Poeschla, E. M. Integrase, LEDGF/p75 and HIV replication. *cell Mol Life Sci* **65,** 1403–1424 (2014).

101. Han, Y., Wind-Rotolo, M., Yang, H. C., Siliciano, J. D. & Siliciano, R. F. Experimental approaches to the study of HIV-1 latency. *Nat. Rev. Microbiol.* **5,** 95–106 (2007).

102. Dwayne Bisgrove, Mary Lewinski, Frederic Bushman, E. V. Molecular mechanisms of HIV-1 proviral latency. *Expert Rev. Anti. Infect. Ther.* **3,** 805–814 (2005).

103. Churchill, M. J., Deeks, S. G., Margolis, D. M., Siliciano, R. F. & Swanstrom, R. HIV reservoirs: What, where and how to target them. *Nat. Rev. Microbiol.* **14,** 55–60 (2015).

104. Sanyal, A. *et al.* Novel assay reveals a large, inducible, replication-competent HIV-1 reservoir in resting CD4+ T cells. *Nat. Med.* **23,** 885–889 (2017).

105. Yucha, R. W. *et al.* High-throughput characterization of HIV-1 reactivation using a single-cell-in-droplet PCR assay. *EBioMedicine* **20,** 217–229 (2017).

106. Cillo, A. R. *et al.* Quantification of HIV-1 latency reversal in resting $CD4^+$ T cells from patients on suppressive antiretroviral therapy. *Proc. Natl. Acad. Sci.* **111,** 7078–7083 (2014).

107. Maldarelli, F. The role of HIV integration in viral persistence: No more whistling past the proviral graveyard. *J. Clin. Invest.* **126,** 438–447 (2016).

108. Kwon, K. J. & Siliciano, R. F. HIV persistence : clonal expansion of cells in the latent reservoir. *J. Clin. Invest.* **127,** 2536–2538 (2017).

109. Anderson, E. M. & Maldarelli, F. The role of integration and clonal expansion in HIV infection : live long and prosper. *Retrovirology* **15,** (2018).

110. Eddy, J. & Maizels, N. Gene function correlates with potential for G4 DNA formation in the human genome. *Nucleic Acids Res.* **34,** 3887–3896 (2006).

111. Simonsson, T., Pecinka, P. & Kubista, M. DNA tetraplex formation in the control region of c-myc. *Nucleic Acids Res.* **26,** 1167–72 (1998).

112. Sun, D., Guo, K., Rusche, J. J. & Hurley, L. H. Facilitation of a structural transition in the polypurine/polypyrimidine tract within the proximal promoter region of the human VEGF gene by the presence of potassium and G-quadruplex-interactive agents. *Nucleic Acids Res.* **33,** 6070–6080 (2005).

113. Rankin, S. *et al.* Putative DNA Quadruplex Formation within the Human *c-kit* Oncogene. *J. Am. Chem. Soc.* **127,** 10584–10589 (2005).

114. Dai, J. *et al.* An intramolecular G-quadruplex structure with mixed parallel/antiparallel G-strands formed in the human BCL-2 promoter region in solution. *J Am Chem Soc.* **128,** 1096–1098 (2006).

115. Kirk, P. D. W., Huvet, M., Melamed, A., Maertens, G. N. & Bangham, C. R. M. Retroviruses integrate into a shared, non-palindromic DNA motif. *Nat. Microbiol.* **2,** 1–6 (2016).

Chapter 3

3 Non-B DNA structures are universally targeted by evolutionarily diverse retroviruses for integration

In an effort to characterize retroviral integration sites, in this study we present an analysis of the integration sites of different HIV-1 subtypes and other retroviruses. Using several published integration site datasets, and samples from HIV-1 infected population from a Uganda and Zimbabwe cohort, we showed differences in integration sites between evolutionary diverse retroviruses, HIV-1 infected *in vitro* and patient datasets and between HIV-1 subtypes A, B, C and D. Integration was highly enriched in and/or near non-B DNA motifs for all subtypes and in other retroviruses. Particularly, integration targeting in and near G4 motifs were strongly enriched in patients compared to *in vitro* datasets and this was also associated with integration of latent proviruses during antiretroviral therapy. Moreover, we showed that antiretroviral therapy significantly alters the integration site profile of different HIV-1 subtypes.

3.1 Introduction

Two defining features of the replication cycle of retroviruses such as human immunodeficiency virus type 1 (HIV-1) is reverse transcription of the single-stranded viral RNA genome into double-stranded complementary DNA (cDNA) and the subsequent integration of the cDNA into the chromosomal host DNA [1]. Reverse transcription is performed by the reverse transcriptase (RT) enzyme. The integration reaction is mediated by the viral integrase (IN) protein that interacts with the viral cDNA and other host and viral proteins forming the pre-integration complex [2,3]. Integration is an essential step in the retrovirus life cycle. In the case of HIV-1 infection, stable insertion of the viral cDNA into the host genome results in a persistent life-long infection. Although combination antiretroviral therapy (cART) substantially suppresses HIV-1 viral replication in infected individuals and improves their quality of life, HIV-1 still remains an incurable infection [4]. The emergence of drug resistant viruses during prolonged cART treatment [5,6], increase in the genetic diversity of the virus [7], high rate of viral replication and residual viremia that can replenish the reservoir to maintain an on-going replication [8–10] all contribute to the

difficulty in curing HIV-1 infection. The existence of a small pool of transcriptionally quiescent/latent cells that are reactivated when treatment is interrupted is a major obstacle for an effective cure [11,12]. Latently infected cells are established early during infection. The site of integration in the genome has been proposed to contribute to a transcriptional block in HIV-1 gene expression[13,14]. Integration site profiles from evolutionarily diverse retroviruses have also been performed and each exhibits a unique integration site profile with some common genomic features[15–23] . In one of the first pioneering studies, integration of murine leukemia virus (MLV) was shown to have a strong preference for gene promoters and was linked to the activation of oncogenes [24,25]. Therefore, a better understanding of the genomic environment surrounding integration sites in evolutionarily diverse retroviruses will provide more insight into pathogenesis, as well as the mechanisms that retroviruses use to integrate into their target host genomes.

Retroviral integration site selection is not a random process. For instance, early studies investigating HIV-1 integration site preferences showed that HIV-1 favors active transcriptional units and genes for insertion [26,27]. These gene rich regions are usually associated with several chromosomal features, including a high GC and CpG islands content, high Alu repeat elements, low long interspersed nuclear elements (LINEs) and DNaseI hypersensitive sites. Moreover, a number of studies also reported HIV-1 integration to occur within transcriptionally silent regions of the genome such as gene deserts, centromeric heterochromatin, satellite DNA, introns and alphoid repeats [28,29,30,31,32]. Comparable to HIV-1, other lentiviruses like the simian immunodeficiency virus (SIV) and feline immunodeficiency virus (FIV) have been shown to also integrate within transcription units of actively transcribed genes, while disfavoring integration into transcription start sites [18,17,19]. In contrast to these complex retroviruses, MLV integrates near transcription start sites, in gene promoters and near CpG islands [15]. Similar to MLV, integration of the foamy virus (FV) showed preference for integration in the vicinity of CpG islands and transcription start sites [20]. While certain retroviruses showed preference for integration into specific sites, other retroviruses such as avian sarcoma leukosis virus (ASLV) and human T-lymphotrophic virus 1 (HTLV-1) showed the most random distributions, with weaker preferences for transcription units, transcription start sites, genes

and CpG islands [23,22,19]. Mouse mammary tumor virus (MMTV) also exhibited no significant preference for transcription start sites or CpG islands [16].

Several factors have been implicated in influencing integration site selection. During HIV-1 infection, it has been reported that the condensed or relaxed structure of the chromatin influences access of the pre-integration complex to the target DNA. For example, wrapping of the DNA around the nucleosome creates sites of DNA distortion that facilitate integration especially within the major groove of the DNA [33,34,35,36]. Host proteins are also critical in influencing integration site selection. One of the best described host tethering protein is lens epithelium derived growth factor and co-factor p75 (LEDGF/p75) [37,38]. LEDGF/p75 interacts with HIV-1 integrase (IN) and targets the pre-integration complex to transcriptionally active genes [39,40]. Another host factor called cleavage and polyadenylation specificity factor 6 (CPSF6) interacts with the viral capsid protein and allows HIV-1 to bypass integration into nuclear peripheral heterochromatin, thereby promoting integration into gene-rich regions of the nucleus [41]. Furthermore, the host cellular protein bromodomain and extraterminal domain (BET) proteins (Brd2, -3, -4) has been demonstrated to interact with MLV IN enzyme promoting integration near transcription start sites [42]. We previously assessed the primary sequence flanking HIV-1 integration in the genome and discovered that HIV-1 integration sites were highly enriched near specialized genomic features called non-B DNA motifs [30]. Non-B DNA motifs are secondary structures in our genome formed by specific nucleotide sequences that exhibit non-canonical DNA base pairing. At least 10 non-B DNA conformations have been identified, including guanine-quadruplex (G4)/tetraplex, A-phased repeats, inverted repeats, direct repeats, cruciform, slipped motifs, mirror repeats, short-tandem repeats, triplex repeats and Z-DNA [43,44]. Moreover, we found integration to be enriched in or near certain non-B DNA motifs that are known to regulate gene expression[45,46,47,48]. Notably, integration sites in latently infected cells that were enriched near G4 and Z-DNA motifs could not be reactivated by the αCD3/CD28 latency reversing agent (see chapter 2). These findings suggest that targeting of non-B DNA motifs for integration can influence the establishment and maintenance of HIV-1 latency. So far, the integration site selection preferences of other retroviruses belonging to different genera has not been analyzed with respect to non-B DNA motifs.

Thus far, HIV-1 integration site analyses have only been conducted on subtype B infections. It is unknown if the integration site profiles of HIV-1 non-subtype B viruses differ from subtype B virus. Based on phylogenetic analyses of full-length genomic sequences, HIV-1 isolates are classified into four distinct groups: group M, N, O and P [49]. HIV-1 M group accounts for the majority of the global pandemic and is subdivided into nine subtypes or clades (A, B, C, D, F, G, H, J and K) [49]. Subtype A and F are further divided into sub-subtypes. Additionally, several circulating recombinant forms (CRFs) and unique recombinant forms (URFs) have been identified. CRF and URF are the result of a recombination event between two different subtypes [50]. HIV-1 geographical prevalence is extremely diverse. Most group M subtypes are present in Sub-Saharan Africa. Subtype C which represents more than 50% of the infection worldwide is prevalent in Africa and Asia [51]. Subtype B infection is common in the Americas, Europe, Australia, and part of South Asia, Northern Africa and the Middle East. Subtypes A, D, F, G, H, J and K occur mostly in Sub-Saharan Africa. Infection with CRFs and URFs occurs in most geographical areas around the world. Infections with groups N, O and P have been found in confined regions of West-Central Africa [52].

In this study, we present a comprehensive and comparative analysis of the integration site profiles of diverse retroviruses and from individuals infected with different HIV-1 subtypes. Our analyses revealed significant differences in the integration site profiles among different retroviruses and different HIV-1 subtypes, particularly with respect to non-B DNA motifs. Moreover, antiretroviral treatment strongly altered the integration site profile in HIV-1 infected individuals. Since non-B DNA structures are known to significantly influence gene expression, the targeting of non-B DNA structures by HIV-1 may play an important role in the establishment and maintenance of latency and/or disease progression among different HIV-1 subtypes.

3.2 Materials and methods

3.2.1 Ethics statement and participants samples

Details pertaining to the Uganda study population have been reported previously [53–56]. Briefly, women who became HIV-1 infected while participating in the Hormonal

Contraception and Risk of HIV Acquisition Study in Uganda were enrolled upon primary infection with HIV-1 into a subsequent study, the Hormonal Contraception and HIV-1 Genital Shedding and Disease Progression among Women with Primary HIV Infection (GS) Study. Ethical approval was obtained from the Institutional Review boards (IRBs) from the Joint Clinical Research Centre and UNST in Uganda, from University of Zimbabwe, from the University Hospitals of Cleveland, and recently, from Western University. All adult subjects provided written informed consent and no child participants were included in the study. Protocol numbers and documentation of these approvals/renewals are available upon request. Blood and cervical samples were collected every month for the first six months, then every three months for the first two years, and then every six months up to 9.5 years. Women who had CD4 lymphocyte counts of 200 cells/ml and/or who developed severe symptoms of HIV infection (WHO clinical stage IV or advanced stage III disease) were offered combination antiretroviral therapy (cART) and trimethoprim-sulfamethoxazole (for prophylaxis against bacterial infections and *Pneumocystis jeroveci* pneumonia).

3.2.2 DNA isolation and HIV-1 integration library

Total genomic DNA from the Uganda and Zimbabwe cohort was extracted from peripheral blood mononuclear cells (PBMCs) using the QIAmp DNA mini kit (Qiagen, cat#: 51306). All genomic DNA was processed for integration site analysis and sequenced using the Illumina MiSeq platform. Genomic DNA was digested with the restriction enzyme MseI overnight at 37°C. Digested DNA was column purified with the Gel/PCR DNA Fragments Kit (Geneaid, cat#: DF100) according to manufacturer's instructions. Next, compatible double-stranded linkers to the MseI sites were prepared as follows: MseI Linker (+) 5'GTAATACGACTCACTATAGG GCTCCGCTTAAGGG AC 3'and MseI Linker (-): 5' [Phos]-TAGTCCCTTAAGCG GAG-[AmC7-Q] 3' were mixed (20µl MseI Linker (+) [40 µM] and 20 µl MseI Linker (-) [40 µM]). The linker mixture was denatured for 5 min at 90°C and cooled 1°C every 3 min until the temperature reached 20°C using the T100^{TM} Thermal Cycler (Bio-Rad). The prepared linkers are now referred to as the "adapter mix". Purified DNA was combined with the adapter mix at 21°C for about 14 hours with 13.5µl of MseI digested samples, 3.5µl of adapter mix, 1µl of T4 DNA Ligase (400U/µl, [NEB,

cat#: M0202S), and 2 µl of 10x ligase buffer. Subsequently, 20µl of the ligated sample was digested at 37°C for 4 hours with 2µl of DpnI (20U/µl), 2µl of NarI (5U/µl), 5µl of 10x buffer and water to a total volume of 50µl. Following digestion, the samples were column purified. The junctions between the integrated HIV-1 LTR sequence and adjacent genomic sequence were amplified in two separate rounds of PCR amplification.

The HIV-1 NL4-3 LTR sequence was used to design primers that amplify through the HIV-1 LTR. The RuparLTR (Forward) 5'-TGCTTCAAGTAGTGTGTGC-3' primer that anneals to the HIV-1 LTRs and the Linker1 (Reverse) 5'-GTAATACGACTCACTATAG GGC-3'primer specific to the MseI linker sequences were used for the first round of PCR amplification. Each PCR reaction mixture consisted of 15.5µl sterile water, 5µl of NarI/DpnI digested sample, 2.5µl of 10x Advantage 2 PCR Buffer, 0.5 µl of 15µM of Linker1 primer , 0.5 µl of 15µM RuparLTR primer, 0.5µl of 10mM dNTPs and 0.5 µl of 50X Advantage 2 PCR polymerase mix (Takara Bio Inc., cat#:639201). PCR was run on T100^{TM} Thermal Cycler (Bio-Rad) under the following cycling conditions: 1 min at 94°C, 5 cycles of 2 sec at 94°C, 1 min at 72°C with an additional 20 cycles of 2 sec at 94°C, 1 min at 67°C and a final extension cycle for 1 min at 72°C and a 4°C hold. The second round of nested PCR amplification was performed using the amplified sample from the first round of PCR amplification. The PCR reaction mixture and cycling conditions were as described for the first round of PRC amplification. The following primer set was used for nested PCR: Rupar-LTR2nested (Forward) 5'-CTCTGGTAACTAGAGATCCCTCA GAC C-3' and Linker2nested (Reverse) 5'-AGGGCTCCGCTTAAGGGAC-3'. Next, Illumina adapter overhang nucleotide sequences were added to the HIV-1 LTR sequence and the MseI linker sequence by PCR amplification using the following primers: Illutag-Forward 5'-<u>GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAG</u>CTCTGGTAACT AGAGATCCCT CAGAC C-3'and Illutag-Reverse 5'-<u>TCGT CGGCAGCGTCAGATGT GTATAAGAGACAG</u>A GGGCTCCGC TTAAGGGAC-3'. Underlined section of the two Illutag primers represent the overhang section. Illumina adapters were utilized in a PCR reaction mixture containing 15.5µl sterile water, 5µl of nested PCR samples, 2.5µl of Advantage 2 PCR Buffer (10x), 0.5 µl of Forward adapter (10µM), 0.5 µl of Reverse adapter (10µM), 0.5µl of dNTPs (10mM) and 0.5 µl of 50X Advantage 2 PCR polymerase

mix. PCR was run on T100^{TM} Thermal Cycler (Bio-Rad). Cycling conditions were as described for the first round of PRC amplification.

The PCR products were purified with AmPure XP beads (Beckman Coulter, cat#: A63881) and the DNA samples were processed using the Nextera XT Index Kit (Illumina). The Nextera XT Indexes technology utilizes a single tagmentation reaction that fragments and tags input DNA with unique adapter and index (barcodes) sequences on both ends of the DNA as previously described [30]. The DNA samples were purified using AmPure XP beads following addition of the barcodes. The barcoded samples were quantified using the Quant-it PicoGreen dsDNA Assay Kit (Invitrogen, cat#: P7589). The absorbance of the plates were read (excitation 480nm for 10 sec and emission 540 nm for 10 sec) with the Cytation5 Imaging Reader (BioTek) and the Gen5 3.02.1 analysis software. Sample concentration was determined by standard curve assessment. The barcoded samples was sequenced through Illumina MiSeq using 2×150 bp chemistry at the London Regional Genomics Centre /Robarts Research Institute from Western University (Canada) and at Case Western Reserve University (USA).

3.2.3 HIV-1 integration site library and computational analysis

Genomic DNA was processed for integration site analysis and sequenced using the Illumina MiSeq platform as described [30,57]. Fastq sequencing reads were quality trimmed and unique integration sites identified using our in-house bioinformatics pipeline [30], which is now called the Barr Lab Integration Site Identification Pipeline (BLISIP version 2.9). BLISIP version 2.9 includes the following updates: bedtools (v2.25.0) which is used to compute distances between integration sites and genomic features, bioawk (awk version 20110810) a programming language for biological data manipulation, bowtie2 (version 2.3.4.1) is used for aligning sequence reads to the human genome, and restrSiteUtils (v1.2.9) is used to generate *in silico* matched random control integration sites based on restriction enzyme used or DNA shearing method. HIV-1 LTR-containing fastq sequences were identified and filtered by allowing up to a maximum of five mismatches with the reference NL4-3 LTR sequence and if the LTR sequence had no match with any region of the human genome (GRCh37/hg19). Integration site profile heatmaps were generated using our in-house python program BHmap (BHmap version 1.0). Sites that could not be

unambiguously mapped to a single region in the genome were excluded from the study. Mapping of integration sites to non-B DNA motifs was performed using the Non-B DB for the human genome (GRCh37/hg19) as previously described [58, 59]. LADs were retrieved from http://dx.doi.org/10.1038/nature06947 [60].

3.2.4 Datasets

All integration site datasets used in this study were independently analyzed using BLISIP version 2.9 and BHmap version 1.0. Integration site datasets from patients infected with HIV-1 subtype B, were obtained from the Cohn dataset and Maldarelli/Wu dataset. Cohn dataset was obtained within the National Center for Biotechnology Information Sequence Read Archive (NCBI SRA) using the accession number SRP045822 as described [61]. The Maldarelli/Wu dataset was obtained from the supplemental material as described [62]. Identical integration sites in each dataset were collapsed into one unique site for the analysis. Integration site datasets from *in vitro* infection of HIV-1 subtype B were obtained from published datasets **(Supplemental Table 3.1)** [15,20,23,26,63,64]. Integration site datasets for SIV, FIV, HTLV-1, MLV, MMTV, FV and endogenous retroviruses (ERVs) were obtained from several published datasets **(Supplemental Table 3.1)** [15,17–20,23,65–69].

3.2.5 Statistical analysis

Fisher's exact test was used for all comparisons of integration site distributions and using Graphpad Prism v6.0. * $P < 0.05$, ** $P < 0.01$, *** $P < 0.001$, **** $P < 0.0001$.

3.3 Results

3.3.1 Evolutionarily divergent retroviruses exhibit distinct preferences for integration into the genome.

Previous analyses of integration sites for multiple retroviruses have shown that integration is more or less likely to occur in or near certain genomic features. However, the integration site profiles of these evolutionarily diverse retroviruses with respect to non-B DNA is unknown. To determine if different retroviruses display a distinct integration profile with respect to non-B DNA motifs, we analyzed the integration sites from previously published datasets using viruses from the lentivirus (e.g. HIV-1, SIV and FIV), deltaretrovirus (e.g.

HTLV-1), the gammaretrovirus (e.g. MLV), the spumavirus (e.g. FV), the alpharetrovirus (e.g. ASLV) and the betaretrovirus (e.g. MMTV) genera **(Supplemental Table 3.1).** We also analyzed the integration site profiles of endogenous retroviruses (ERVs) naturally found within the human genome. Integration sites from all datasets were obtained from various human cell lines and primary cells, such as HEK 293 T, HeLa, Jurkat and primary $CD4^+$ T cells as shown in **(Supplemental Table 3.1)**. To ensure a comprehensive comparative analysis of integration site profiles using the same human genome annotation, we utilized hg19 for our analyses. We also used our previously developed in-house bioinformatics pipeline called the Barr Lab Integration Site Pipeline (BLISIP) to generate integration site profiles from the different retroviral integration site datasets [57,30]. Our analyses involved unique integration site events and excluded sites arising from clonal expansion, sites falling in repeat regions or regions that cannot be confidently placed on a specific chromosome (e.g. ChrUn). Enrichment of integration sites within genomic features was determined by comparing the proportion of sites with either a matched random control (MRC) to account for restriction site bias in the cloning procedure during library construction and for comparison of datasets that used DNA shearing/fragmentation during library construction **(Supplemental Table 3.1).** To further validate our in-house bioinformatics analysis pipeline (BLISIP), and to provide a direct comparison of the integration site profiles of evolutionarily diverse retroviruses, we analyzed previously published integration site datasets using BLISIP. Integration sites in several common genomic features were quantified and placed in four distance bins starting from within each genomic feature (Bin 0) to > 50,000 base pairs (bp) away from the feature (Bin 4). Heatmaps from each retrovirus showing the fold enrichment and fold depletion of sites in each bin compared to MRC are shown in **Figure 3.1A**. Heatmaps are superimposed on a phylogenetic tree constructed using reverse transcriptase sequences to show the evolutionary relatedness of the different retrovirus genera [70].

Consistent with previous studies, HIV-1, SIV and FIV integration sites are highly enriched within genes (64%, 84%, 90% respectively) and are present at levels significantly more than that expected by chance ($P < 0.0001$, Fisher's exact test) **(Figure 3.1A, 3.1B and Supplemental Table 3.2)** [18,71,17].

Figure 3.1: Evolutionarily divergent retroviruses exhibit distinct preferences for integration into the genome. (A) Heatmap depicting the fold enrichment or depletion of integration sites in common genomic features compared to the matched random control (MRC). Darker shades represent higher fold-changes in the ratio of integration sites to MRC sites. Asterisks within each heatmap (*P < 0.05; **P < 0.01; ***P < 0.001; ****P < 0.0001, Fisher's exact test) represent significant differences in the number of integration sites in different retroviruses compared to MRC. Bins represent the distance of the integration sites from the genomic feature. Bin 0 = within the feature, Bin 1= 1 - 499 bp; Bin 2 = 500 -4,999 bp; Bin 3 = 5,000-49,999 bp; Bin 4 = < 49,999 bp. Infinite number (inf), 1 or more integrations were observed when 0 integrations were expected by chance. Not a number (nan), 0 integrations were observed and 0 were expected by chance. **(B)** Proportion of unique HIV-1 integration sites in common genomic features. Blue lines represent MRC values. Significant differences are with respect to the paired MRC (blue lines) and are denoted by asterisks (Fisher's exact test; *P < 0.05; **P < 0.01; ***P < 0.001; ****P < 0.0001). HIV-1 = human immunodeficiency virus, SIV = simian immunodeficiency virus, FIV= feline immunodeficiency virus, HTLV-1= human T-lymphotrophic virus 1, MLV = murine leukemia virus, FV = foamy virus, ASLV = avian sarcoma leukosis virus, MMTV = mouse mammary tumor virus and ERVs = endogenous retroviruses. The number of unique integration sites for each retrovirus are as follow: HIV-1 = 180699 sites SIV = 160 sites, FIV= 226 sites, HTLV-1= 624 sites, MLV = 1484 sites, FV = 3423 sites, ASLV = 680 sites, MMTV = 268 sites and ERVs = 325 sites.

Figure 3.1

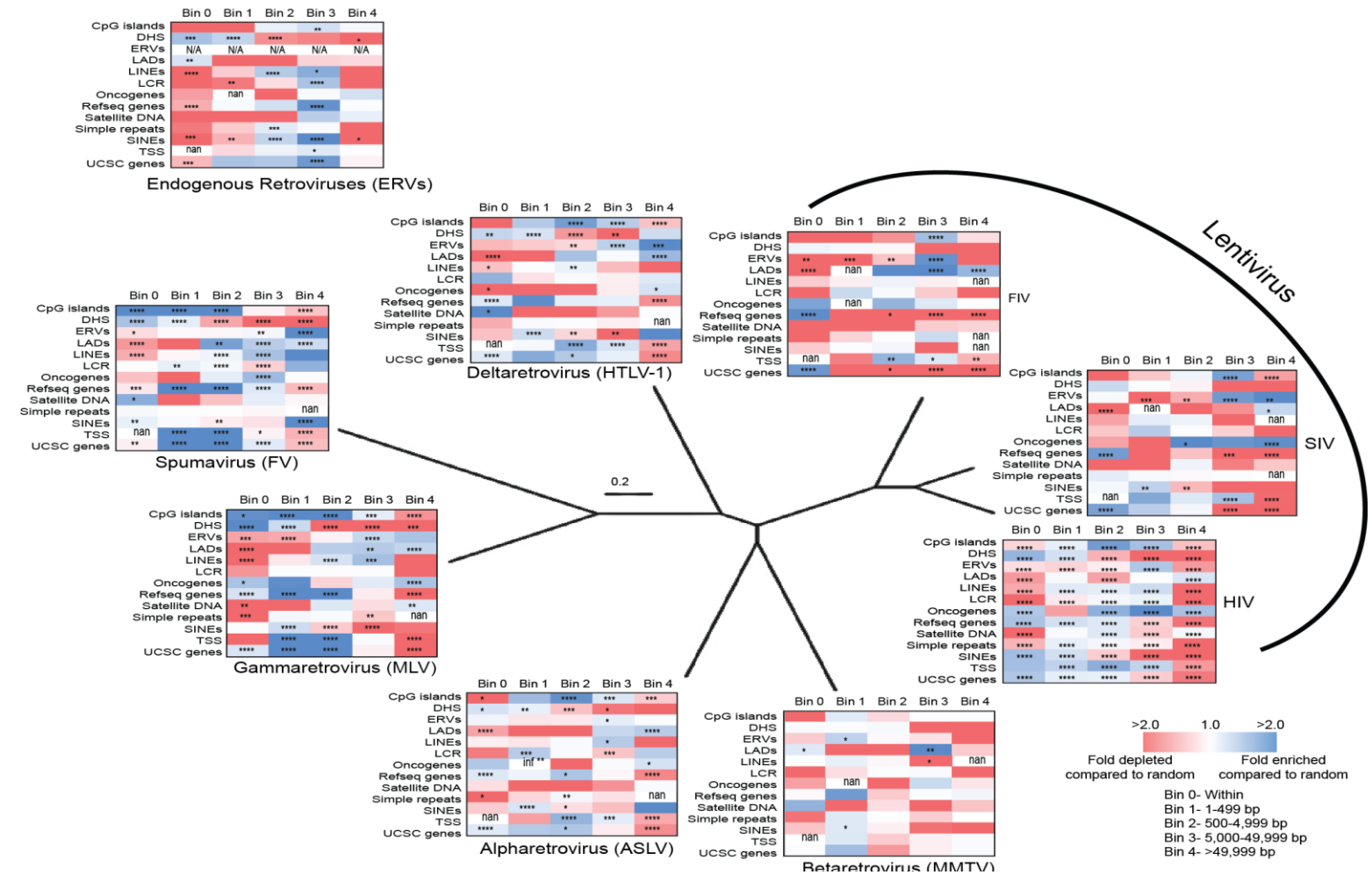


Figure 3.1: Evolutionarily divergent retroviruses exhibit distinct preferences for integration into the genome.

Figure 3.1

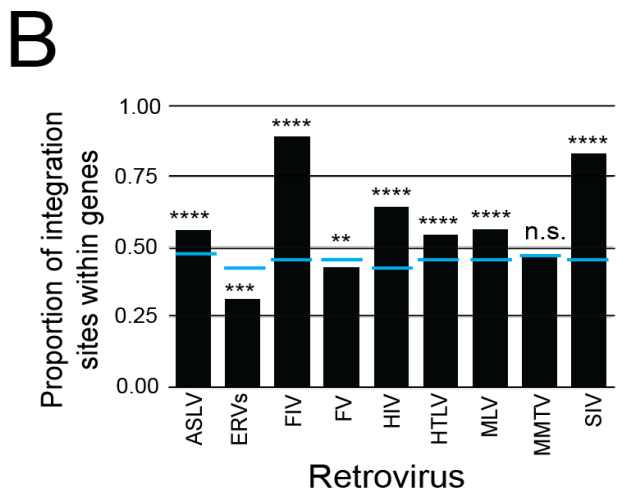


Figure 3.1: Evolutionarily divergent retroviruses exhibit distinct preferences for integration into the genome.

Our analysis of HTLV-1, ASLV and MLV also agreed with previous reports, confirming that these viruses exhibit only modest preferences for integration within genes (54%, 56% and 56% of integration sites found within genes respectively) **(Figure 3.1A and 3.1B, Supplemental Table 3.2)** [16, 72].

In contrast, MMTV, FV and ERVs showed no preference for integration into genes (47%, 43% and 32% respectively), with FV and ERVs showing a significant disfavoring for integration into genes (P < 0.001, Fisher's exact test). All retroviruses showed no preference for integration directly into transcription start sites (TSS). Notably, all retroviruses showed enriched integration near TSS. This was particularly noteworthy for FV and MLV, which showed a strong preference for integration near (<5,000 bp) TSS (25% and 34% respectively), which was 2- to 3-fold (respectively) more than that expected by chance (P <0.0001, Fisher's exact test) **(Figure 3.1A, Supplemental Table 3.2)**.

We further observed that all exogenous retroviruses and ERVs integrated near CpG islands except for MMTV, which showed no preference for integration in or near CpG islands. Additionally, FV and MLV also showed a strong preference for integration in and near (<500 bp) CpG islands (7% and 10%), which was 2- to 11 fold (respectively) more than that expected by chance (P <0.0001, Fisher's exact test) **(Figure 3.1A, Supplemental Table 3.2)**. Repetitive elements, such as long interspersed nuclear elements (LINEs), short interspersed nuclear elements (SINEs), ERVs (retrotransposons), satellite DNA, simple repeats (microsatellites), and low-complexity repeats (LCR) account for nearly half of the human genome sequence. However, no preference for integration into these regions was observed for any of the retroviruses except for HIV-1 and FV which targeted SINEs, and FV and HTLV-1 which targeted satellite DNA **(Figure 3.1A).**

The nuclear architecture is known to influence HIV-1 integration site selection and proviral expression [60,73]. HIV-1 strongly disfavors integration into heterochromatic condensed regions positioned in lamin-associated domains (LADs) at the nuclear periphery, although some integration does occur in these regions and significantly influences expression of latent proviruses [74]. We asked whether other retroviruses also disfavor integration into LADs. Like HIV-1, most retroviruses strongly disfavored integration into LADs with only

11-28% of sites falling within LADs ($P < 0.0001$, Fisher's exact test) **(Figure 3.1A, Supplemental Table 3.2).** In contrast, 48% of MMTV and 44% of ERV integration sites were located in LADs and were significantly more than that expected by chance ($P < 0.016$, Fisher's exact test). Together, these data confirm and extend previous findings that retroviruses of diverse genera have different preferences for integrating into common genomic features. Furthermore, these data indicate that most retroviruses avoid integrating into transcriptionally inactive heterochromatin, except MMTV and ERVs, which exhibit a strong preference for heterochromatin.

3.3.2 Evolutionarily diverse retroviruses target non-B DNA for integration.

Non-B DNA motifs are new host factors we previously identified that influence lentiviral integration and serve as important integration site targets for HIV-1 [30] (also see chapter 2). However, their influence on integration site targeting for other retroviruses was previously unknown. We analyzed several published integration site datasets from evolutionarily diverse retroviruses **(Supplemental Table 3.1)** to obtain and compare their non-B DNA integration site profiles. All retroviral integration sites were located in or within <500 bp of non-B DNA, with strong preference for certain non-B DNA motifs (**Figure 3.2**). Notably, HTLV-1 and FV were the only two retroviruses that exhibited an enrichment of sites directly in the majority of non-B DNA motifs compared to MRC (**Figure 3.2 and Supplemental Table 3.3**). Some other notable preferences for non-B DNA include HTLV-1, which showed a significant enrichment of integration sites directly in and/or near A-phased motifs ($P < 0.05$, Fisher's exact test), cruciform motifs ($P < 0.05$, Fisher's exact test) and inverted repeats ($P < 0.01$, Fisher's exact test). FV showed significant enrichment in and adjacent (<50 bp) to G4 motifs ($P < 0.001$, Fisher's exact test) and Z-DNA motifs ($P < 0.05$, Fisher's exact test). MLV showed strong enrichment of sites near G4 motifs, triplex motifs and Z-DNA motifs. ASLV strongly favored integration near slipped and triplex motifs. MMTV showed a strong preference for short tandem repeats, slipped motifs and triplex motifs.

Figure 3.2: Evolutionarily diverse retroviruses target non-B DNA for integration. Distribution of unique retroviral integration sites in non-B DNA-forming motifs. Darker shades represent higher fold-changes in the ratio of integration sites to matched random control (MRC) sites. Asterisks within each heatmap (*P < 0.05; **P < 0.01; ***P < 0.001; ****P < 0.0001, Fisher's exact test) represent significant differences in the number of integration sites in different retroviruses compared to MRC. HIV-1 = human immunodeficiency virus, SIV = simian immunodeficiency virus, FIV= feline immunodeficiency virus, HTLV-1= human T-lymphotrophic virus 1, MLV = murine leukemia virus, FV = foamy virus, ASLV = avian sarcoma leukosis virus, MMTV = mouse mammary tumor virus and ERVs = endogenous retroviruses. The number of unique integration sites for each retrovirus are as follow: HIV-1 = 180699 sites SIV = 160 sites, FIV= 226 sites, HTLV-1= 624 sites, MLV = 1484 sites, FV = 3423 sites, ASLV = 680 sites, MMTV = 268 sites and ERVs = 325 sites.

Figure 3.2

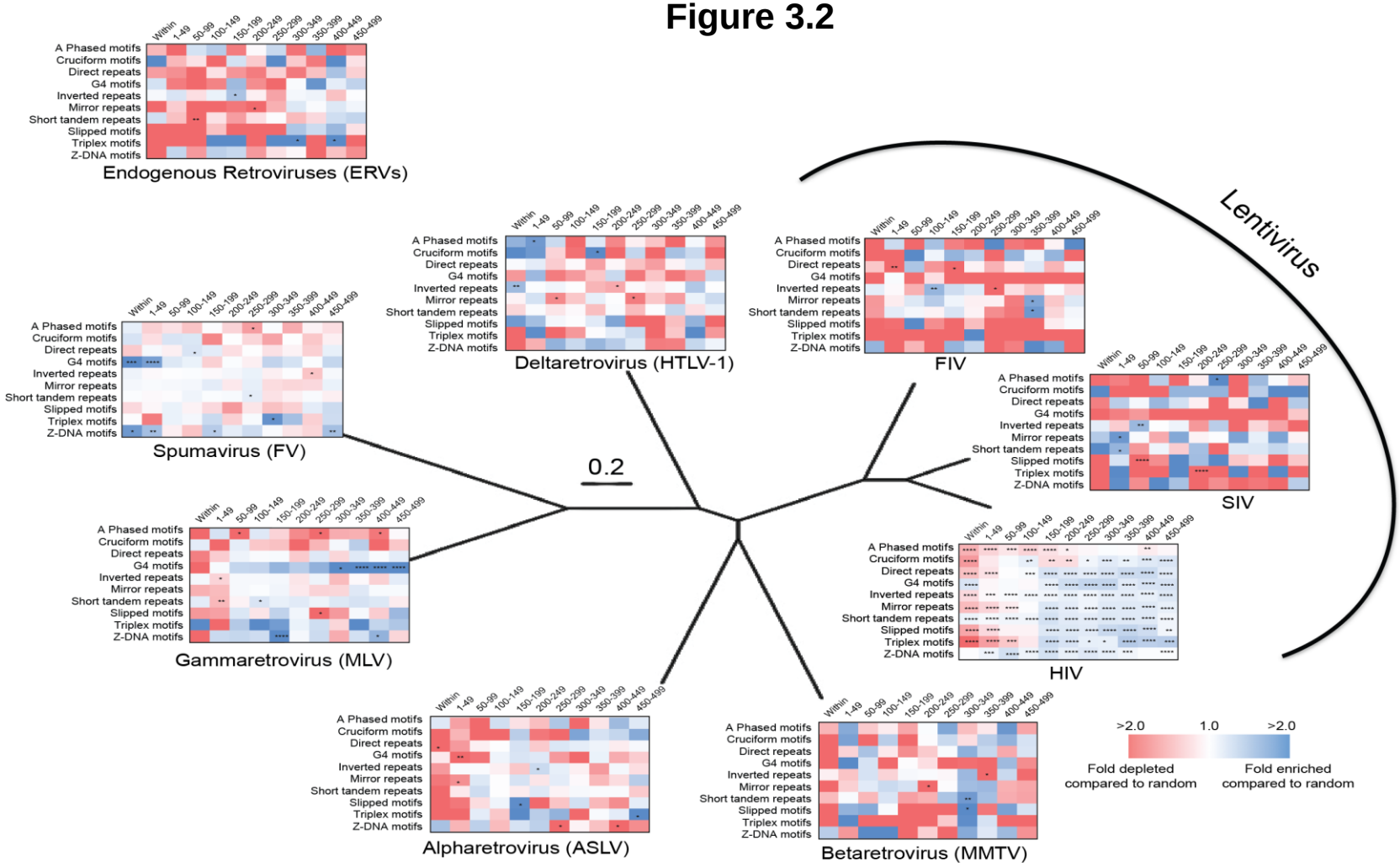


Figure 3.2: Evolutionarily diverse retroviruses target non-B DNA for integration.

SIV and FIV both exhibited strong preferences for integration near inverted repeats, mirror repeats and short tandem repeats, although SIV preferred to integrate closer to these non-B DNA than FIV. FIV and SIV also strongly disfavored integration in or near G4 motifs. ERVs exhibited a strong enrichment of sites near triplex motifs. HIV-1 showed significant enrichment of sites directly in G4 motifs and short tandem repeats, and 150-500 bp away from all non-B DNA motifs except A-phased motifs. Taken together, these data show that all retroviruses exhibit distinct non-B DNA integration site profiles, but also share similar strong preferences for integration into specific non-B DNA motifs.

3.3.3 Integration site profiles differ between *in vitro*-derived and patient-derived datasets

HIV-1 infection is the most clinically prevalent retroviral infection in the human population, yet most integration site analyses have been performed using HIV-1/vector infections performed *in vitro* using cell lines [27]. To determine if the HIV-1 integration site profiles from acute HIV-1 infections of cell lines (*in vitro*-derived infections) differ from those from chronically infected individuals (patient-derived) , we analyzed and compared the integration sites from previously published datasets **(Supplemental Table 3.1, Figure 3.3)**. We first analyzed the integration site profiles with respect to several commonly studied genomic features. A total of 9 previously published datasets were used to evaluate the *in vitro*-derived (13,601 sites) and patient-derived sites (167,098 sites). We first investigated the integration site profile with respect to several commonly studied genomic features. In general, the integration site profiles were similar with some notable differences. Both *in vitro* and patient datasets strongly favored integration in genes (83% *in vitro*, 62% in patients, $P < 0.0001$, Fisher's exact test) **(Figure 3.3A, 3.3B and Supplemental Table 3.4)**. Compared to the *in vitro* dataset, integration sites in the patient dataset were significantly enriched in CpG islands, DNaseI hypersensitivity sites, ERVs, LADs, satellite DNA, simple repeats and SINEs ($P < 0.0001$, Fisher's exact test). In contrast, integration sites in the patient dataset were significantly depleted in LINEs and low complexity repeats compared to the *in vitro* dataset ($P < 0.0001$, Fisher's exact test). We also identified differences in integration site targeting preferences for non-B DNA motifs between the two datasets.

Figure 3.3: Integration site profiles differ between *in vitro*-derived and patient-derived datasets. (A) Heatmaps illustrating the distribution of unique integration sites in common genomic features for HIV-1 *in vitro* (n= 13601sites) and infected patients datasets (n= 167098 sites). Darker shades represent higher fold-changes in the ratio of integration sites to matched random control (MRC) sites. Numbers within each heatmap represent the fold-increase or decrease in the number of unique integration sites in patient-derived dataset or *in vitro*-derived dataset compared to the MRC. Integration sites of patient-derived dataset were also compared to *in vitro*-derived dataset. (**B**) Percentage of unique HIV-1 integration sites directly within common genomic features in *in vitro*-derived dataset or patient-derived dataset. (**C**) Heatmaps depicting the fold enrichment or depletion of integration sites in non-B DNA motifs compared to the MRC. (**D**) Percentage of unique HIV-1 integration sites directly within non-B DNA motifs (G4, triplex and Z-DNA motifs) in *in vitro*-derived dataset or patient-derived dataset. All significant differences are denoted by asterisks (*P < 0.05; **P < 0.01; ***P < 0.001; ****P < 0.0001) and were determined by Fisher's exact test.

Figure 3.3

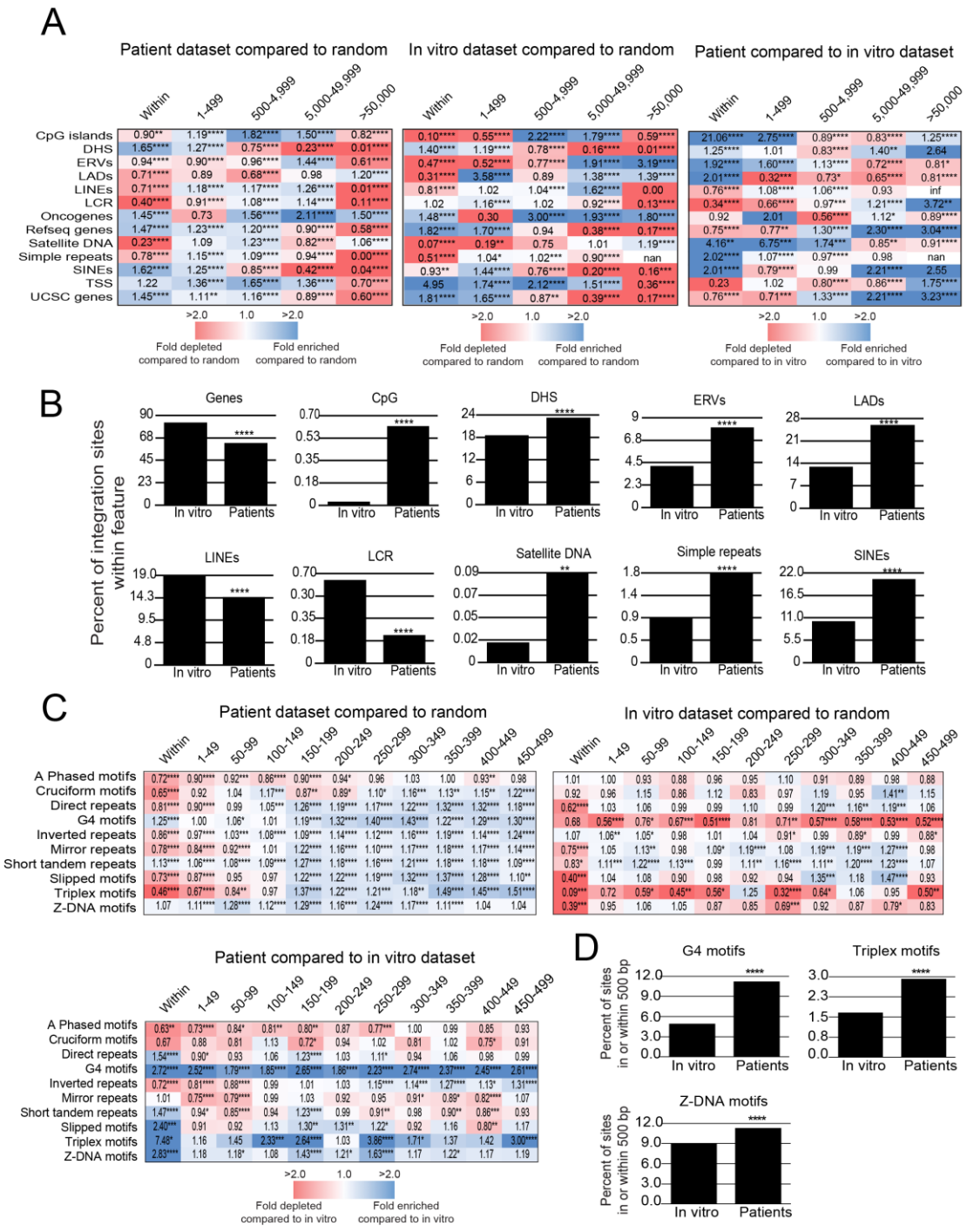


Figure 3.3: Integration site profiles differ between in vitro-derived and patient-derived datasets.

When compared to the MRCs, integration sites in the patient dataset were enriched directly in G4 motifs and short tandem repeats, whereas these features were disfavored in the *in vitro* dataset **(Figure 3.3C and Supplemental Table 3.4)**. Integration sites in the patient dataset were also highly enriched in a region 150 to 500 bp away from all non-B DNA motifs except A-phased motifs compared to the *in vitro* dataset. Notably, when the patient dataset was compared to the *in vitro* dataset, significant enrichment was observed within 500 bp of G4, triplex and Z-DNA motifs **(Figure 3.3C, 3.3D and Supplemental Table 3.4)**. Together, these data show striking differences in integration site targeting preferences between *in vitro*-derived and patient-derived datasets.

3.3.4 HIV-1 subtypes A, B, C and D have different integration site preferences.

To our knowledge available integration site studies have only been conducted with HIV-1 subtype B, which only represents ~10% of the infections worldwide. As of today, much less is known on the integration site profile of other HIV-1 subtypes. We asked if the integration site profiles from individuals infected with HIV-1 non-subtype B virus are similar to those infected with subtype B virus.

Genomic DNA was isolated from peripheral blood mononuclear cells (PBMCs) from a cohort of women in Uganda and Zimbabwe infected with HIV-1 subtypes A, C or D and used to generate integration site libraries **(Supplemental Table 3.5 A)**. Integration site profiles were generated from a total of 48 infected females (16 subtype A, 19 subtype C and 13 subtype D) and compared to the integration site profile from 14 men and women infected with subtype B virus generated from previously published datasets **(Supplemental Table 3.5 B and C).** The number of integration sites analyzed in this study were: subtype A, 429 sites; subtype B, 139480 sites; subtype C, 484 sites; and subtype D, 323 sites. Integration sites from all HIV-1 subtype viruses were highly enriched in genes **(Figure 3.4 A, B and Supplemental Table 3.6**). Notably, subtypes A, C and D exhibited a significantly stronger preference for integrating into genes compared to subtype B (A: 82%, C: 71%, D: 78% and B: 63%) (**Figure 3.4 B and Supplemental Table 3.6**). Integration was disfavored in CpG islands for each subtype; however, sites were enriched in a region 500-5,000 bp away from CpG islands for each subtype. Integration sites for subtypes B, C and

D were enriched in and near DHS, whereas subtype A sites were only enriched near (1-500 bp) this feature. Subtypes A and C exhibited enriched integration in low complexity repeats, whereas these regions were disfavored by subtypes B and D. Subtypes A, C and D disfavored integration in or near satellite DNA, whereas subtype B favored integration near (1-5,000 bp) satellite DNA. Integration in or near LADs was strongly disfavored in all subtypes.

Next, we analyzed the integration site profiles from the same datasets with respect to non-B DNA motifs. Integration sites falling directly within non-B DNA motifs or in distance bins of 50 bp up to 500 bp away from each motif were quantified. As shown in **Figure 3.4 C and Supplemental Table 3.7**, the integration site profiles differed substantially among the different subtypes, especially when compared to subtype B. The most notable difference in integration site preference among the different subtypes was towards G4 motifs. Subtype B virus favored integration directly in and near (150-500 bp) G4 motifs. Subtypes A, C and D generally disfavored integration in or near G4 motifs, except in a region 100-150 bp away from the motif where integration was favored. Like subtype B, subtype A also favored integration directly in the G4 motif itself. Some other notable differences were as follows. Subtypes A and C exhibited strong preferences for integration near A-phased motifs, whereas subtypes D and B showed little to no preference for these motifs.

All subtypes favored integration near cruciform motifs, with non-subtype B viruses exhibiting a stronger preference for these motifs than subtype B virus. All subtypes favored integration 150-500 bp away from direct repeats, with subtypes A, C and D also favoring integration adjacent (1-50 bp) to direct repeats All subtypes favored integration near slipped motifs, with subtype B only favoring a region 150-500 bp away from these motifs. Similar to its preference for slipped motifs, subtype B virus favored integration in triplex motifs and Z-DNA motifs in a region 150-500 bp away from these features. Non-subtype B viruses showed little preference for slipped or Z-DNA motifs; however, subtypes A and D favored integration in defined regions within 500 bp from these motifs (**Figure 3.4 C**). In contrast, subtype C virus showed very little preference for triplex or Z-DNA motifs.

Figure 3.4: HIV-1 subtypes A, B, C and D have different integration site preferences. **(A)** Heatmaps showing unique integration sites distribution of HIV-1 subtypes A, B, C and D in common genomic features compared to the matched random control (MRC). Darker shades represent higher fold-changes in the ratio of integration sites to matched random MRC sites. Numbers within each heatmap show the fold-increase or decrease in the number of unique integration sites in HIV-1 subtypes A, B, C, and D compared to MRC. Not a number (nan) indicates that 0 integrations were observed and 0 were expected by chance. **(B)** Proportion of integration directly within genes for subtypes A, B, C and D. **(C)** Distribution of unique integration sites for of HIV-1 subtypes A, B, C and D in non-B DNA motifs compared to MRC. Darker shades represent higher fold-changes in the ratio of integration sites to matched random MRC sites. Numbers within each heatmap show the fold-increase or decrease in the number of unique integration sites in HIV-1 subtypes A, B, C, and D compared to the MRC. All significant differences were determined by Fisher's exact test and are denoted by asterisks (*P < 0.05; **P < 0.01; ***P < 0.001; ****P < 0.0001). Number of integration sites are as follow: HIV-1 subtype A = 429 sites; HIV-1 subtype B = 139480 sites; HIV-1 subtype C = 484 sites; and HIV-1 subtype D = 323 sites.

Figure 3.4

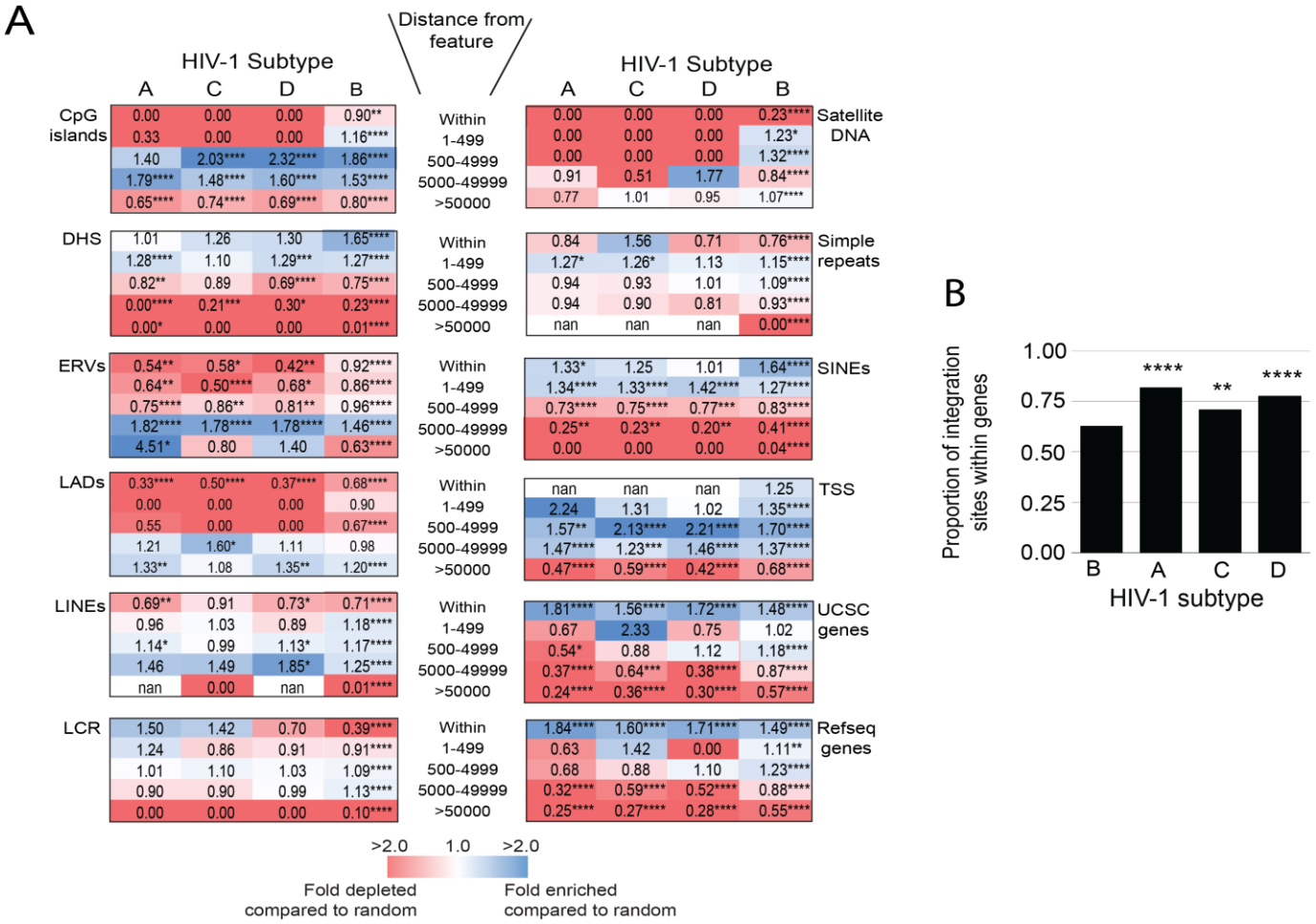


Figure 3.4: HIV-1 subtypes A, B, C and D have different integration site preferences.

Figure 3.4

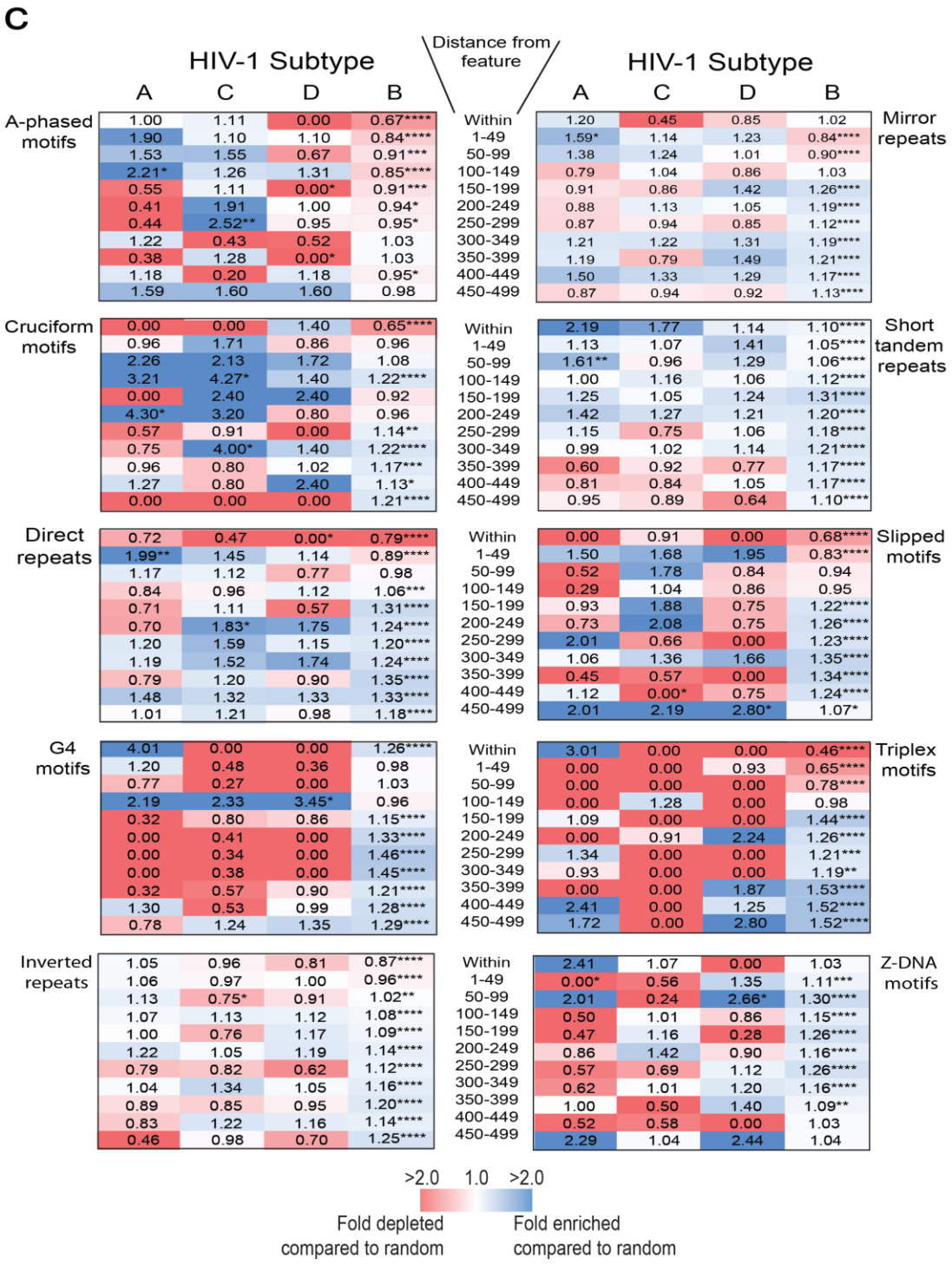


Figure 3.4: HIV-1 subtypes A, B, C and D have different integration site preferences.

Overall, integration preferences were most similar between non-subtype B and subtype B viruses for inverted repeats, mirror repeats and short tandem repeats, which all showed enrichment near these features. Together, these data show that the integration site profiles from HIV-1 subtype A, C and D infections differ substantially from subtype B infections. Moreover, each subtype exhibits a distinct integration site bias for specific genomic features, particularly non-B DNA motifs.

3.3.5 Combination antiretroviral therapy (cART) alters HIV-1 integration site selection in common genomic features

Once infected individuals start receiving cART to control HIV-1 replication, infected cells bearing silent proviruses becomes highly selected for with treatment. Given that cART helps select for integration into silent regions such as intergenic regions that can maintain latency [61], we further investigated the association between treatment and HIV-1 integration site selection among different HIV-1 subtypes, including with respect to non-B DNA. Unique integration sites from individuals infected with either HIV-1 subtype A, C, D and B were grouped into untreated (not receiving cART) and treated populations/samples (receiving cART) **(Supplemental Table 3.5 A, B, and C)**. All integration sites were compared to MRC. Interestingly, cART treatment lead to a decrease in frequency into genes for subtypes B (67% untreated, 62% treated) and C (73% untreated, 55% treated) **(Figure 3.5, Supplemental Table 3.8)** compared to MRC. cART did not induce a significant change in the frequency of integration into genes for subtype A (82% untreated, 84% treated), D (79% untreated , 77% treated). Integration within endogenous retroviruses (ERVs), satellite DNA and lamina associated domains (LADs) remained strongly depleted in the untreated and treated groups of each subtype. Integration directly into CpG islands was highly enriched in the untreated subtype B population compared to the treated population, which showed a reduction in CpG island integration ($P < 0.001$). In contrast to other subtypes, subtype B showed a strong preference for integration directly within DNaseI hypersensitivity sites and SINEs for both untreated and treated samples. However, subtypes C and D showed enriched integration within DnaseI hypersensitive sites in the untreated group compared to the treated group **(Figure 3.5)**.

Figure 3.5: Combination antiretroviral therapy (cART) alters HIV-1 integration site selection in common genomic features. Heatmaps illustrating the distribution of unique integration sites in common genomic features for HIV-1 subtype A, B, C and D untreated and treated patients' samples. Integration sites were compared to matched random control (MRC). Darker shades represent higher fold-changes in the ratio of integration sites to MRC sites. Numbers within each heatmap show the fold-increase or decrease in the number of unique integration sites in untreated and treated HIV-1 subtypes A, B, C, and D compared to the MRC. Not a number (nan), 0 integrations were observed and 0 were expected by chance. Significant differences were determined by Fisher's exact test and are denoted by asterisks (*P < 0.05; **P < 0.01; ***P < 0.001; ****P < 0.0001). Number of sites are as follow: Subtype A untreated = 259 sites, Subtype A treated = 170 sites, Subtype B untreated = 27077, Subtype B treated = 112403 sites, Subtype C untreated = 448 sites, Subtype C treated = 36 sites, Subtype D untreated = 156 sites and Subtype D treated = 167 sites.

Figure 3.5

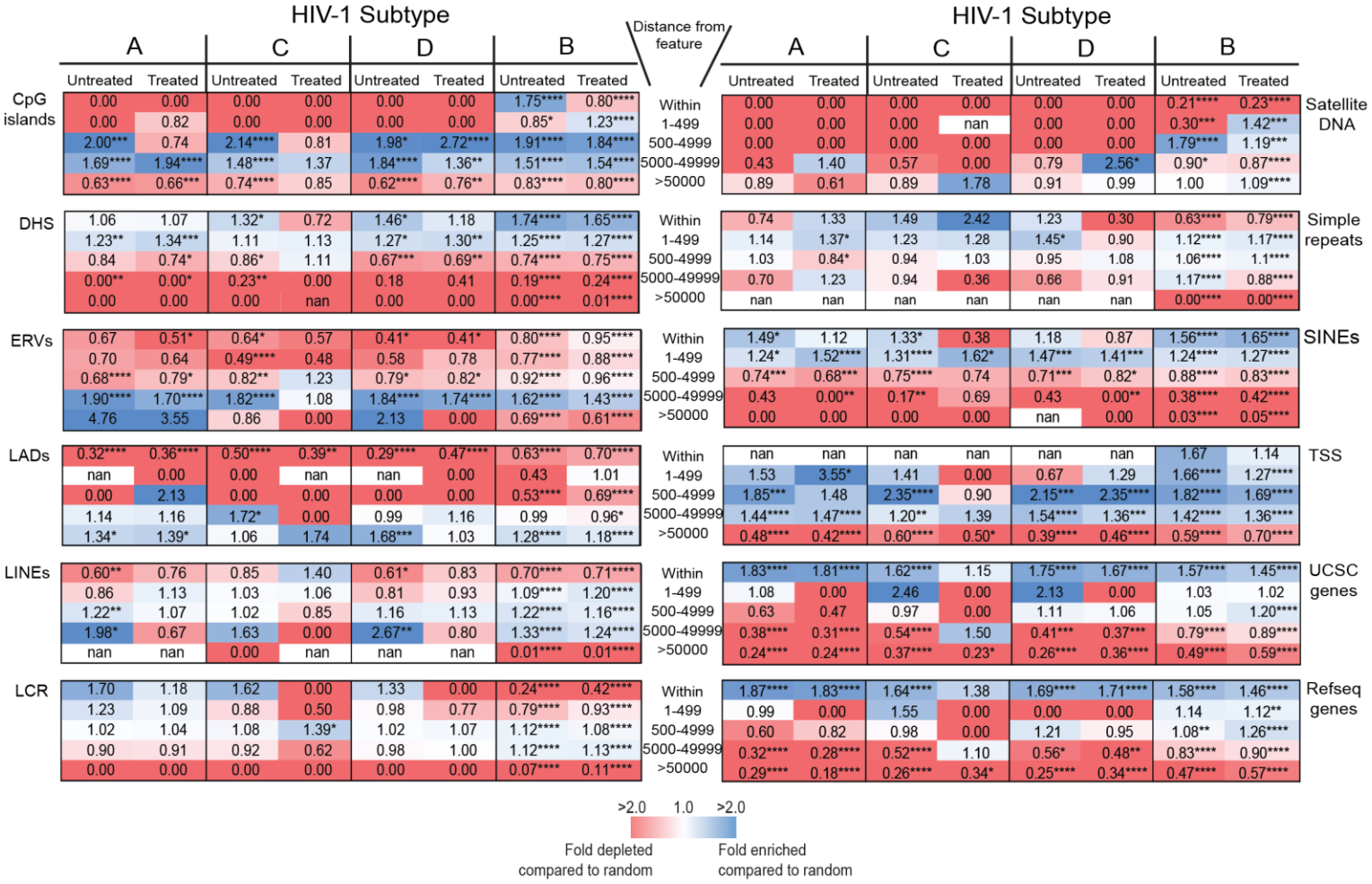


Figure 3.5: Combination antiretroviral therapy (cART) alters HIV-1 integration site selection in common genomic features.

Overall, these data strongly suggest that cART alters HIV-1 integration site target selection by favoring or disfavoring certain features among the different subtypes.

3.3.6 cART and HIV-1 integration site selection in non-B DNA motifs

Next, we further evaluated the frequency of integration sites in untreated and treated samples from subtypes A, B, C and D with respect to non-B DNA motifs using the same datasets as in **Supplemental Table 3.5 A, B and C**. Integration sites were quantified either directly within non-B DNA motifs or in distance bins of 50 bp up to 500 bp away from each motif. In the case of non-B DNA motifs, subtype B showed enriched integration 50-499bp from the motifs for both untreated and treated patient samples. A similar enrichment was observed in the other subtypes **(Figure 3.6, Supplemental Table 3.9).** Notably, integration in or near A-phased motif was significantly depleted in subtype B and enriched in subtypes A, C and D.

Since certain non-B DNA motifs such as G4 and Z-DNA can inhibit gene expression, we further assessed whether patients receiving cART had enriched integration into those motifs. We observed that all subtypes targeted G4 motifs at a distance of 100-149 bp for both untreated and treated group except for subtype B, which showed no preference for integration (untreated and treated samples) at that distance compared to other subtypes. Most important, we observed a 3 fold and 12 fold enrichment in the treated groups for subtype A and C respectively **(Figure 3.6).** In contrast, no substantial change in integration near G4 (100-149 bp from G4) was observed for subtype D between the untreated and treated group. These results again indicate that cART alters integration site targeting of non-B DNA motifs.

3.4 Discussion

The analysis presented here showed that non-B DNA motifs, which are novel features that influence HIV-1 integration, are also targeted by other retroviruses. Specifically, evolutionarily diverse retroviruses exhibit distinct non-B DNA integration site profiles, but also share similar strong preferences for integration into several non-B DNA motifs.

Figure 3.6: cART and HIV-1 integration site selection in non-B DNA motifs. Heatmaps showing the distribution of unique integration sites for HIV-1 subtypes A, B, C, D untreated and subtype A, B, C and D treated patients samples in non-B DNA motifs. Integration sites were compared to matched random control (MRC). Darker shades represent higher fold-changes in the ratio of integration sites to MRC sites. Numbers within each heatmap show the fold-increase or decrease in the number of unique integration sites in untreated and treated HIV-1 subtypes A, B, C, and D compared to the MRC. Not a number (nan), 0 integrations were observed and 0 were expected by chance. Significant differences were determined by Fisher's exact test and are denoted by asterisks (*P < 0.05; **P < 0.01; ***P < 0.001; ****P < 0.0001). Subtype A untreated = 259 sites, Subtype A treated = 170 sites, 27077, Subtype B treated = 112403 , Subtype C untreated = 448 sites, Subtype C treated = 36 sites, Subtype D untreated = 156 sites and Subtype D treated = 167 sites.

Figure 3.6

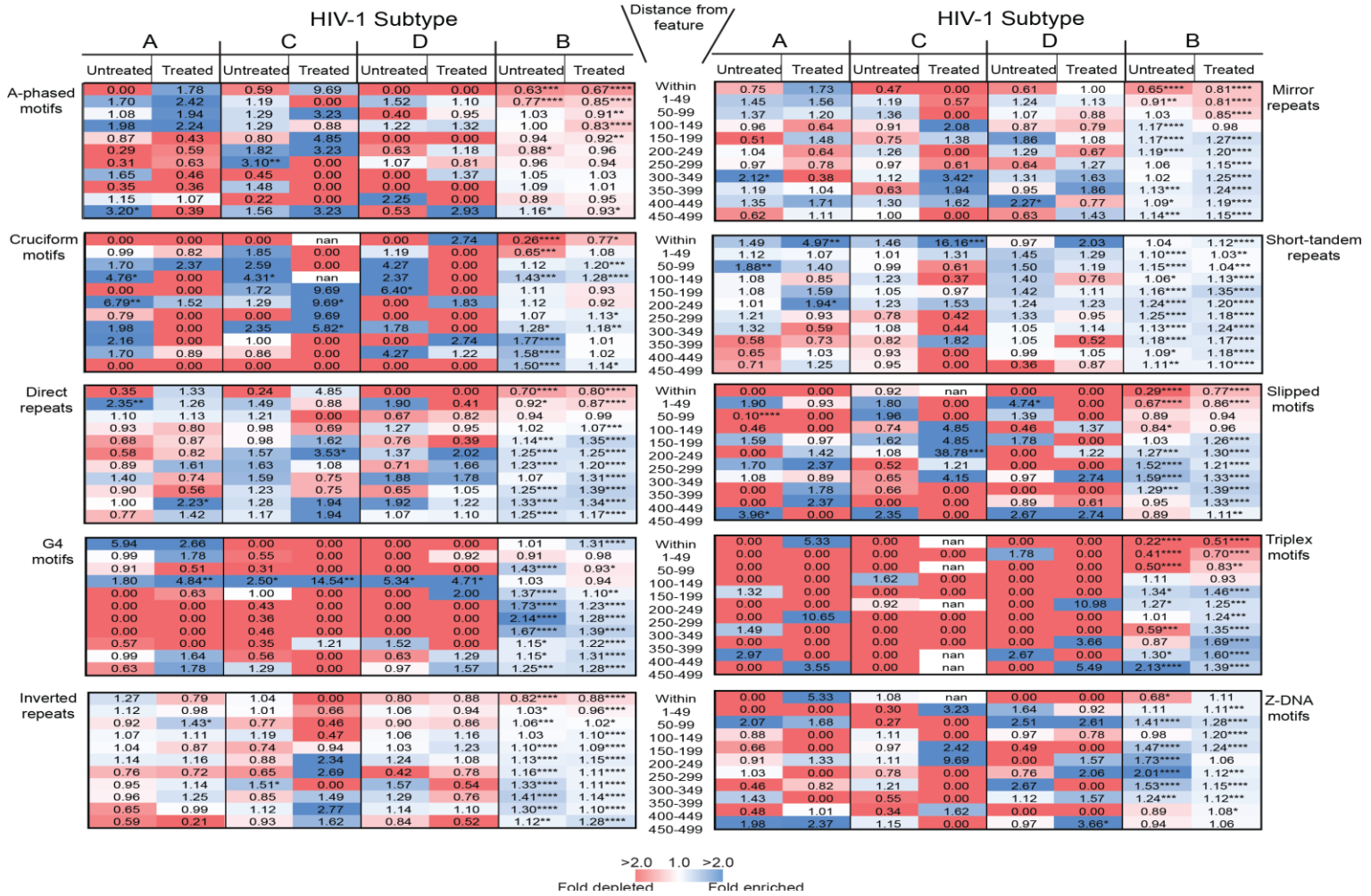


Figure 3.6: cART and HIV-1 integration site selection in non-B DNA motifs.

We also showed that most retroviruses disfavored integration into transcriptionally inactive heterochromatin, except MMTV and ERVs which exhibit a strong preference for heterochromatin. We further demonstrated striking differences in integration site targeting preferences between *in vitro*-derived and patient-derived datasets and showed that enrichment of HIV-1 integrations in and/or near non-B DNA motifs that regulate gene expression is specific to patient datasets. Our data also shows that HIV-1 subtype A, C and D infections differ substantially from subtype B infections and that cART alters integration site targeting of non-B DNA.

Our analyses are consistent with previous studies and extend this work to show that additional diverse retroviruses genera have different preferences for integrating into common genomic features, such as genes and near transcription start sites and CpG islands. Specifically, we were able to confirm that HIV-1, SIV and FIV strongly favored integration in genes, while HTLV-1, ASLV and MLV, viruses exhibited only modest preferences for integration into genes [16 ,72,]. However, MMTV, FV and ERVs disfavored integration into genes. It is important to note that our analysis showed a slightly higher percentage of integration into genes by HTLV-1, ASLV and MLV (54%, 56% and 56% respectively) than previously reported (HTLV 46.8% , ASLV 46.4% or 40%, and MLV 45.7% or 40.2%) [19, 22]. In our study, we used an earlier version of the UCSC genome database and RefSeq genes human gene annotation, which might be larger than the assembly previously used by others. This may have correlated to the slightly higher values seen in our current study. Overall, our results are consistent with previous findings. Factors that may influence integration site selection might include variation in the properties of their integrase proteins. In fact, previous phylogenetic analysis that looked at the integrase sequences of different retroviruses showed clustering among certain retroviruses, where HIV-1, SIV and FIV were found in the same cluster and MLV and FV in the same cluster [19]. This may contribute to why certain retroviruses showed similar integration site preferences with respect to specific genomic features. Nonetheless, properties of the pre-integration complex and virus interaction with cellular host factors and chromosomal DNA might also be critical in determining proviral integration site selection.

Our analysis of integration site selection also showed that non-B DNA motifs are not targeted only by HIV-1. Other retroviruses targeted non-B DNA motifs for integration and exhibited distinct integration site preference with respect to non-B DNA. While most integration occurred within the majority of the motifs for HTLV-1 and FV, HIV-1 showed integration spanning 150-500 bp away from each motif. Thus, integration targeting for other retroviruses seems to not occur at specific distances for the most part. The integration events seen in the other retroviruses (e.g. SIV, FIV, MLV, ASLV and MMTV) might suggest the lack of cellular factors capable of interacting with their viral integrase and/or pre-integration complex that could help direct their integration into specific distances from non-B DNA motifs. Importantly, HIV-1 integrase is known to bind directly to certain non-B DNA motifs [75–84]. Therefore, this could facilitate and consistently direct integration within a specific distance from non-B DNA motifs. Like HIV-1 SIV, FIV, MLV, ASLV and MMTV, also targeted certain non-B DNA motifs such as G4 motifs and Z-DNA, which had substantial implications in regulating nearby gene expression and potentially proviral gene expression. Notably, on one hand, interaction of non-B DNA structures with their specific transcription factors can induce transcriptional activation [85–88]. On the other hand, the three dimensional structure of non-B DNA motifs can prevent binding of B-DNA-specific transcription factors, leading to suppression of adjacent genes [89–92]. Thus, non-B DNA motifs can play a substantial role in regulating gene expression of all exogenous retroviruses, including HIV-1.

Historically, most of the conclusions regarding HIV-1 integration site targeting preferences were made *in vitro* using HIV-1 vectors and cell lines. We showed here that there are profound differences in integration site targeting preferences between *in vitro*-derived and patient-derived integration site datasets. Notably, sites in the patient dataset were highly enriched in a region spanning 150 to 500 bp away from most non-B DNA motifs compared to the *in vitro* dataset. It is unclear what causes these differences in integration as related specifically to non-B DNA. However, it could be suggested that integration sites studies from *in vitro* experiments are usually associated with acute short term infection, therefore, representing integration sites during the early stages of infection. In contrast, integration sites from patients are associated with chronic and persistent infection where some cells are productively infected while many cells also undergo latency, especially when receiving

antiretroviral treatment. As a result, integrations from patient data may be biased towards genomic sites that are more favorable for both HIV-1 expression (e.g. genes) and latency (e.g. non-B DNA motifs). Most importantly, HIV-1 integration in patient-derived datasets was significantly enriched near G4 motifs, as opposed to the cell line dataset which did not target G4 motifs. However, integration near other non-B DNA motifs also occurred, including integration near G4 motifs, which is relevant due to its role in gene regulation. From these results, it is clear that distinct integration site biases exist between *in vitro*-derived and patient-derived datasets. As such, generalizations regarding HIV-1 integration site preferences cannot be made solely from either *in vitro*-derived or patient-derived datasets. Additionally, the majority of *in vitro* studies are conducted in cancer derived cell lines that do not reflect the normal cell types infected by HIV-1. Importantly, these cells may lack of the expression of integrase cofactors as well as having abnormal chromatin structure.

Genetic variation between subtypes can range between 25 to 35% [49]. Importantly, the genetic differences between subtypes may influence their interaction with the host, therefore also influencing disease transmission, progression and notably integration site selection. Here, we have shown similar integration preference with respect to commonly studied genomic features such as genes, CpG islands, satellite DNA and LAD among subtypes A, B, C and D. However, substantial differences were observed among the different subtypes with respect to non-B DNA motifs. Notably, it is unclear why subtype B is the only subtype that showed enriched integration within specific distances from non-B DNA compared to other subtypes. Variation in integrase structure could contribute to this difference in integration. In fact, natural polymorphisms in HIV-1 integrase have been observed among different subtypes, which might affect their integration site selection [93,94]. Additionally, specific polymorphisms in HIV-1 integrase have been reported to retarget integration away from gene dense regions, which correlated with increase disease progression and virulence [95]. This further suggests that viral integrase can substantially contribute to disease progression as it relates to the virus integration site targeting. Of importance, we have also shown enriched integration between 100 -149 bp away from non-B DNA motifs, particularly G4 motifs, among subtypes A, C and D but not subtype B. Integration at that distance may play an essential role in integration site selection and may

be a result of G4 structure-induced repositioning nucleosomes, which are comprised of ~147 bp of DNA wrapped around a histone octamer core [96]. In fact, G4 motifs form in nucleosome-free regions in the genome [97]. As such, this ability to locally and dynamically organize flanking nucleosomes may contribute to transcriptional regulation of adjacent genes.

Upon cART initiation, infected cells bearing silent proviruses become highly selected for. As cART helps select for integration into silent regions such as intergenic regions [31], it was expected that integration site selection will also retain a strong bias toward heterochromatin rich regions (e.g. satellite DNA and LADs) in cART treatment patients. Our results showed enrichment in LADs in treated samples compared to untreated samples in subtype A, B and D infections. This further supports the idea that cART selects for integration sites located in heterochromatin among different subtypes. Notably, our observations indicated that cART changed the integration profile for all subtypes. We also observed that subtypes A, C and D shared higher similarity in integration site preferences compared to subtype B. Additionally, cART lead to an enriched integration near G4 motifs, which are also found in heterochromatin rich regions [90].

In conclusion, we identified distinct non-B DNA structures surrounding integration sites that are targeted differentially by evolutionarily diverse retroviruses, including HIV-1 viral subtypes. Additionally, different HIV-1 subtypes have distinct integration profiles that differ before and after cART.

3.5 References

1. Craigie, R. & Bushman, F. D. HIV DNA integration. *Cold Spring Harb. Perspect. Med.* **2,** 1–18 (2012).

2. Suzuki, Y., Chew, M. L. & Suzuki, Y. Role of host-encoded proteins in restriction of retroviral integration. *Front. Microbiol.* **227,** 1–13 (2012).

3. Miller, M. D., Farnet, C. M. & Bushman, F. D. Human immunodeficiency virus type 1 preintegration complexes: studies of organization and composition. *J. Virol.* **71,** 5382–5390 (1997).

4. Arts, E. J. & Hazuda, D. J. HIV-1 antiretroviral drug therapy. *Cold Spring Harb. Perspect. Med.* **2,** a007161 (2012).

5. Miller, V. *et al.* Dual resistance to zidovudine and lamivudine in patients treated with zidovudine-lamivudine combination therapy: association with therapy failure. *J Infect Dis* **177,** 1521–1532 (1998).

6. Karmochkine, M. *et al.* The cumulative occurrence of resistance mutations in the HIV-1 protease gene is associated with failure of salvage therapy with ritonavir and saquinavir in protease inhibitor-experienced patients. *Antiviral Res.* **47,** 179–188 (2000).

7. Santos, A. F. & Soares, M. A. HIV genetic diversity and drug resistance. *Viruses* **2,** 503–531 (2010).

8. Vanuitert, B. *et al.* Residual HIV-1 RNA in blood plasma of patients taking suppressive highly active antiretroviral thereapy. *JAMA* **282,** 1627–1632 (1999).

9. Furtado, M. R. *et al.* Persistence of HIV-1 transcription in peripheral-blood mononuclear cells in patients receiving potent antiretroviral therapy. *N. Engl. J. Med.* **340,** 1614–1622 (1999).

10. Chun, T. W. *et al.* HIV-infected individuals receiving effective antiviral therapy for extended periods of time continually replenish their viral reservoir. *J. Clin. Invest.* **115,** 3250–3255 (2005).

11. Chun, T.-W. *et al.* Early establishment of a pool of latently infected, resting $CD4^+$ T cells during primary HIV-1 infection. *Proc. Natl. Acad. Sci. U. S. A.* **95,** 8869–8873 (1998).

12. Davey Jr., R. T. *et al.* HIV-1 and T cell dynamics after interruption of highly active antiretroviral therapy (HAART) in patients with a history of sustained viral suppression. *Proc Natl Acad Sci U S A* **96,** 15109–15114 (1999).

13. Jordan, A., Defechereux, P. & Verdin, E. The site of HIV-1 integration in the human

genome determines basal transcriptional activity and response to Tat transactivation. *EMBO J.* **20,** 1726–1738 (2001).

14. Hughes, S. H. & Coffin, J. M. What integration sites tell us about HIV persistence. *Cell Host Microbe* **19,** 588–598 (2016).

15. Wu, X., Li, Y., Crise, B. & Burgess, S. M. Transcription start regions in the human genome are favored targets for MLV integration. *Science* **300,** 1749–1751 (2003).

16. Faschinger, A. *et al.* Mouse mammary tumor virus integration site selection in human and mouse genomes. *J. Virol.* **82,** 1360–1367 (2008).

17. Kang, Y. *et al.* Integration site choice of a feline immunodeficiency virus vector. *J. Virol.* **80,** 8820–8823 (2006).

18. Crise, B. *et al.* Simian immunodeficiency virus integration preference is similar to that of human immunodeficiency virus type 1. *J. Virol.* **79,** 12199–12204 (2005).

19. Derse, D. *et al.* Human T-cell leukemia virus type 1 integration target sites in the human genome: comparison with those of other retroviruses. *J. Virol.* **81,** 6731–6741 (2007).

20. Trobridge, G. D. *et al.* Foamy virus vector integration sites in normal human cells. *Pnas* **103,** 1498–1503 (2016).

21. Doi, K. *et al.* Preferential selection of human T-cell leukemia virus type I provirus integration sites in leukemic versus carrier states. *Blood* **106,** 1048–1053 (2005).

22. Narezkina, A. *et al.* Genome-wide analyses of avian sarcoma virus integration sites. *J. Virol.* **78,** 11656–63 (2004).

23. Mitchell, R. S. *et al.* Retroviral DNA integration: ASLV, HIV, and MLV show distinct target site preferences. *PLoS Biol.* **2,** E234 (2004).

24. Tsuruyama, T., Hiratsuka, T. & Yamada, N. Hotspots of MLV integration in the hematopoietic tumor genome. *Oncogene* **36,** 1169–1175 (2017).

25. Hacein-Bey-Abina, S., Cavazzana-Calvo, M. & et. al. Insertional Oncogenesis in 4 patients after retroviral-mediated gene therapy of SCID-X1. *J. Clin. Invest.* **118,** 3132–3142 (2008).

26. Shinn, P. *et al.* HIV-1 integration in the human genome favors active genes and local hotspots. *Cell Press* **110,** 521–529 (2002).

27. Bushman, F. *et al.* Genome-wide analysis of retroviral DNA integration. *Nat. Rev. Microbiol.* **3,** 848–58 (2005).

28. Lewinski, M. K. *et al.* Genome-wide analysis of chromosomal features repressing human immunodeficiency virus transcription. *J. Virol.* **79,** 6610–6619 (2005).

29. Jordan, A., Bisgrove, D. & Verdin, E. HIV reproducibly establishes a latent infection after acute infection of T cells in vitro. *EMBO J.* **22,** 1868–1877 (2003).

30. McAllister, R. G. *et al.* Lentivector integration sites in ependymal cells from a model of metachromatic leukodystrophy: non-B DNA as a new factor influencing integration. *Mol. Ther. Nucleic Acids* **3,** e187 (2014).

31. Cohn, L. B. *et al.* HIV-1 integration landscape during latent and active infection. *Cell* **160,** 420–432 (2015).

32. F. Maldarelli, 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. H. Specific HIV integration sites are linked to clonal expansion and persistence of infected cells. *Science* **345,** 179–183 (2014).

33. Bor, Y. C., Bushman, F. D. & Orgel, L. E. In vitro integration of human immunodeficiency virus type 1 cDNA into targets containing protein-induced bends. *Proc. Natl. Acad. Sci. U. S. A.* **92,** 10334–8 (1995).

34. Müller, H. P. & Varmus, H. E. DNA bending creates favored sites for retroviral integration: an explanation for preferred insertion sites in nucleosomes. *EMBO J.* **13,** 4704–14 (1994).

35. Pruss, D., Reeves, R., Bushman, F. D. & Wolffe, A. P. The influence of DNA and nucleosome structure on integration events directed by HIV integrase. *J. Biol. Chem.* **269,** 25031–25041 (1994).

36. Pruss, D., Bushman, F. D. & Wolffe, A. P. Human immunodeficiency virus integrase directs integration to sites of severe DNA distortion within the nucleosome core. *Proc. Natl. Acad. Sci.* **91,** 5913–5917 (1994).

37. Engelman, A. & Cherepanov, P. The lentiviral integrase binding protein LEDGF/p75 and HIV-1 replication. *PLoS Pathog.* **4,** e1000046 (2008).

38. Poeschla, E. M. Integrase, LEDGF/p75 and HIV replication. *cell Mol Life Sci* **65,** 1403–1424 (2014).

39. Shun, M. C. *et al.* LEDGF/p75 functions downstream from preintegration complex formation to effect gene-specific HIV-1 integration. *Genes Dev.* **21,** 1767–1778 (2007).

40. Ciuffi, A. *et al.* A role for LEDGF/p75 in targeting HIV DNA integration. *Nat. Med.* **11,** 1287–1289 (2005).

41. Achuthan, V. *et al.* Capsid-CPSF6 interaction licenses nuclear HIV-1 trafficking to sites of viral DNA integration. *Cell Host Microbe* **24,** 392–404 (2018).

42. Sharma, A. *et al.* BET proteins promote efficient murine leukemia virus integration at transcription start sites. *Proc. Natl. Acad. Sci.* **110,** 12036–12041 (2013).

43. Jungkweon Choi and Tetsuro Majima. Conformational changes of non-BDNA. *Chem. Soc. Rev.* **40,** 5893–5909 (2011).

44. Bacolla, A. & Wells, R. D. Non-B DNA conformations, genomic rearrangements, and human disease. *J. Biol. Chem.* **279,** 47411–47414 (2004).

45. Siddiqui-Jain, A., Grand, C. L., Bearss, D. J. & Hurley, L. H. Direct evidence for a G-quadruplex in a promoter region and its targeting with a small molecule to repress

c-MYC transcription. *Proc. Natl. Acad. Sci.* **99,** 11593–11598 (2002).

46. Verma, A., Yadav, V. K., Basundra, R., Kumar, A. & Chowdhury, S. Evidence of genome-wide G4 DNA-mediated gene expression in human cancer cells. *Nucleic Acids Res.* **37,** 4194–4204 (2009).

47. Nelson, L. D. *et al.* Triplex DNA-binding proteins are associated with clinical outcomes revealed by proteomic measurements in patients with colorectal cancer. *Mol. Cancer* **11,** 1–17 (2012).

48. Waga, S. Chromosomal Protein HMGl Removes the Transcriptional Caused by the Cruciform in Supercoiled DNA *. *J. Biol. Chem.* **265,** 19424–19428 (1990).

49. Taylor, B. *et al.* The challenge of HIV-1 subtype diversity. *N. Engl. J. Med.* **358,** 1590–1602 (2008).

50. Joris HemelaarHemelaar, J. The origin and diversity of the HIV-1 pandemic. *Cell Press* **18,** 182–92 (2012).

51. Patiño-Galindo, J. Á. & González-Candelas, F. The substitution rate of HIV-1 subtypes: a genomic approach. *Virus Evol.* **3,** 1–7 (2017).

52. Hahn, B. H., Shaw, G. M., Cock, K. M. De & Sharp, P. M. AIDS as a zoonosis: scientic and public health implications. *Science* **287,** 607–614 (2000).

53. Venner, C. M. *et al.* Infecting HIV-1 subtype predicts disease progression in women of Sub-Saharan Africa. *EBioMedicine* **13,** 305–314 (2016).

54. Morrison, C. S. *et al.* Hormonal contraceptive use and HIV disease progression among women in Uganda and Zimbabwe. *J Acquir Immune Defic Syndr.* **57,** 157–164 (2011).

55. Morrisona, C. S. *et al.* Hormonal contraception and the risk of HIV acquisition. *AIDS* **21,** 85–95 (2007).

56. Lemonovich, T. L. *et al.* Differences in clinical manifestations of acute and early

HIV-1 infection between HIV-1 subtypes in African women. *J. Int. Assoc. Provid. AIDS Care* **14,** 415–422 (2015).

57. Ciuffi, A. & Barr, S. D. Identification of HIV integration sites in infected host genomic DNA. *Methods* **53,** 39–46 (2011).

58. Cer, R. Z. *et al.* Non-B DB v2.0: A database of predicted non-B DNA-forming motifs and its associated tools. *Nucleic Acids Res.* **41,** 94–100 (2013).

59. Cer, R. Z. *et al.* Non-B DB: a database of predicted non-B DNA-forming motifs in mammalian genomes. *Nucleic Acids Res.* **39,** D383-91 (2011).

60. Guelen, L. *et al.* Domain organization of human chromosomes revealed by mapping of nuclear lamina interactions. *Nature* **453,** 948–951 (2008).

61. Cohn, L. B. *et al.* HIV-1 Integration Landscape during Latent and Active Infection. *Cell* **160,** 420–432 (2015).

62. Maldarelli, F. *et al.* Specific HIV integration sites are linked to clonal expansion and persistence of infected cells. *Science* **345,** 179–183 (2014).

63. Lewinski, M. K. *et al.* Genome-wide analysis of chromosomal features repressing human immunodeficiency virus transcription. *J. Vriology* **79,** 6610–6619 (2005).

64. Sherrill-Mix, S. *et al.* HIV latency and integration site placement in five cell-based models. *Retrovirology* **10,** (2013).

65. Jurka, J. Repbase. Available at: http://www.girinst.org/repbase/update/index.html.

66. Lewinski, M. K. *et al.* Retroviral DNA integration : Viral and cellular determinants of target-site selection. *PLoS Pathog.* **2,** e60 (2006).

67. Nowrouzi, A. *et al.* Genome-wide mapping of foamy virus vector integrations into a human cell line. *J. Gen. Virol.* **87,** 1339–1347 (2006).

68. Narezkina, A. *et al.* Genome-wide analyses of avian sarcoma virus integration sites.

J. Vriology **78,** 11656–11663 (2004).

69. Faschinger, A. *et al.* Mouse mammary tumor virus integration site selection in human and mouse genomes. *J. Virol.* **82,** 1360–1367 (2008).

70. Dereeper, A. *et al.* Phylogeny.fr: robust phylogenetic analysis for the non-specialist. *Nucleic Acids Res.* **36,** W465–W469 (2008).

71. Schroder, A. *et al.* HIV-1 Integration in the Human Genome Favors Active Genes and Local Hotspots. *Cell* **110,** 521–529 (2002).

72. Derse, D. *et al.* Human T-cell leukemia virus type 1 integration target sites in the human genome: comparison with those of other retroviruses. *J. Virol.* **81,** 6731–41 (2007).

73. Marini, B. *et al.* Nuclear architecture dictates HIV-1 integration site selection. *Nature* **521,** 227–31 (2015).

74. Battivelli, E. *et al.* Distinct chromatin functional states correlate with HIV latency reactivation in infected primary CD4+ T cells. *Elife* **7,** e34655 (2018).

75. Ojwang, J. O. *et al.* T30177, an oligonucleotide stabilized by an intramolecular guanosine octet, is a potent inhibitor of laboratory strains and clinical isolates of human immunodeficiency virus type 1. *Antimicrob. Agents Chemother.* **39,** 2426–2435 (1995).

76. Rando, R. F. *et al.* Suppression of human immunodeficiency virus type 1 activity in vitro by oligonucleotides which form intramolecular tetrads. *Journal of Biological Chemistry* **270,** 1754–1760 (1995).

77. Mazumder, A. N. *et al.* Inhibition of human immunodeficiency virus type 1 integrase by guanosine quartet structures. *Biochemistry* **35,** 13762–13771 (1996).

78. Jing, N., Rando, R. F., Pommier, Y. & Hogan, M. E. Ion selective Folding of loop domains in a potent anti-HIV oligonucleotide. *Biochemistry* **36,** 12498–12505

(1997).

79. Jing, N. *et al.* Mechanism of inhibition of HIV-1 integrase by G-tetrad-forming oligonucleotides in vitro. *J. Biol. Chem.* **275,** 21460–21467 (2000).

80. De Soultrait, V. R. *et al.* DNA aptamers derived from HIV-1 RNase H inhibitors are strong anti-integrase agents. *J. Mol. Biol.* **324,** 195–203 (2002).

81. Phan, A. T. *et al.* From The Cover: An interlocked dimeric parallel-stranded DNA quadruplex: A potent inhibitor of HIV-1 integrase. *Proc. Natl. Acad. Sci.* **102,** 634–639 (2005).

82. Koizumi, M. *et al.* Biologically active oligodeoxyribonucleotides--IX. Synthesis and anti-HIV-1 activity of hexadeoxyribonucleotides, TGGGAG, bearing 3'- and 5'-end-modification. *Bioorg. Med. Chem.* **5,** 2235–43 (1997).

83. Urata, H., Kumashiro, T., Kawahata, T., Otake, T. & Akagi, M. Anti-HIV-1 activity and mode of action of mirror image oligodeoxynucleotide analogue of zintevir. *Biochem. Biophys. Res. Commun.* **313,** 55–61 (2004).

84. Pedersen, E. B., Nielsen, J. T., Nielsen, C. & Filichev, V. V. Enhanced anti-HIV-1 activity of G-quadruplexes comprising locked nucleic acids and intercalating nucleic acids. *Nucleic Acids Res.* **39,** 2470–2481 (2011).

85. Brooks, T. A. & Hurley, L. H. The role of supercoiling in transcriptional control of MYC and its importance in molecular therapeutics. *Nat. Rev. Cancer* **9,** 849–861 (2009).

86. Cogoi, S., Shchekotikhin, A. E. & Xodo, L. E. HRAS is silenced by two neighboring G-quadruplexes and activated by MAZ, a zinc-finger transcription factor with DNA unfolding property. *Nucleic Acids Res.* **42,** 8379–8388 (2014).

87. Kang, H.-J. *et al.* Novel Interaction of the Z-DNA Binding Domain of Human ADAR1 with the Oncogenic c-Myc Promoter G-Quadruplex. *J. Mol. Biol.* **426,** 2594–2604 (2014).

88. Murat, P. & Balasubramanian, S. Existence and consequences of G-quadruplex structures in DNA. *Curr. Opin. Genet. Dev.* **25,** 22–29 (2014).

89. Michelotti, G. A. *et al.* Multiple single-stranded cis elements are associated with activated chromatin of the human c-myc gene in vivo. *Mol. Cell. Biol.* **16,** 2656–69 (1996).

90. Hoffmann, R. F. *et al.* Guanine quadruplex structures localize to heterochromatin. *Nucleic Acids Res.* **44,** 152–163 (2016).

91. Lam, E. Y. N., Beraldi, D., Tannahill, D. & Balasubramanian, S. G-quadruplex structures are stable and detectable in human genomic DNA. *Nat. Commun.* **4,** 1796 (2013).

92. Ray, B. K., Dhar, S., Henry, C., Rich, A. & Ray, A. Epigenetic regulation by Z-DNA silencer function controls cancer-associated ADAM-12 expression in breast cancer: Cross-talk between MeCP2 and NF1 transcription factor family. *Cancer Res.* **73,** 736–744 (2013).

93. Myers, R. E. & Pillay, D. Analysis of natural sequence variation and covariation in human immunodeficiency virus type 1 integrase. *J. Virol.* **82,** 9228–9235 (2008).

94. Rhee, S. Y. *et al.* Natural variation of HIV-1 group M integrase: Implications for a new class of antiretroviral inhibitors. *Retrovirology* **5,** 1–11 (2008).

95. Demeulemeester, J. *et al.* HIV-1 integrase variants retarget viral integration and are associated with disease progression in a chronic infection cohort. *Cell Host Microbe* **16,** 651–662 (2014).

96. Timothy J. Richmond & Curt A. Davey. The structure of DNA in the nucleosome core. *Nature* **423,** 145–150 (2003).

97. Wong, H. M. & Huppert, J. L. Stable G-quadruplexes are found outside nucleosome-bound regions. *Mol. Biosyst.* **5,** 1713–9 (2009).

Chapter 4

4 The quiescent/latent HIV-1 integration site landscape from different anatomical tissues reveals unique differences

HIV-1 infection persists in latently infected $CD4^+$ T cells from peripheral blood despite suppressive antiretroviral therapy. Additional anatomical sites harboring HIV-1 are also key reservoirs that are established early during infection. Together, these sites present a major barrier for HIV-1 eradication, where replication competent virus can persist, thus restoring the latent reservoir upon cessation of antiretroviral therapy. Additionally, these anatomical sites may help maintain a continuous low-level of viral replication despite antiretroviral therapy. While the integration site selection of HIV-1 has been extensively studied in peripheral blood from infected individuals, little is known about the integration site distribution in other sanctuary/anatomical sites. Here, we compared the distribution of integration sites of HIV-1 in the peripheral blood, the brain, and the gastrointestinal tract. All anatomical sites exhibited a significant preference for integration in genes, as previously observed in peripheral blood studies, with a somewhat lower frequency in the brain. We also showed distinct integration profiles from the peripheral blood, the brain, and the gastrointestinal tract with respect to non-B DNA motifs. Importantly, integration site were strongly enriched in and/or near guanine-quadruplex (G4) motifs, a type of non-B DNA motif, which are known to suppress expression of adjacent genes. Our findings demonstrate clinically favorable integration site profiles in anatomical sites of HIV-1 infected individuals and implicate non-B DNA motifs as an essential factor in HIV-1 integration within various latent reservoirs.

4.1 Introduction

Since the discovery of the human immunodeficiency virus type 1 (HIV-1) in the early 1980's as the key ethological agent of acquire immunodeficiency syndrome (AIDS), more than 35 million individuals have died from HIV-1/AIDS related illness [1,2,3,4]. The Joint United Nations Programme on HIV and AIDS (UNAIDS) reported that in the year 2017 over 36 million people were currently living with HIV-1, with approximately 2 million

new infections occurring annually [4]. The use of combination antiretroviral therapy (cART) has significantly improved the quality of life of infected individuals and has led to a substantial reduction in morbidity and mortality related to HIV-1/AIDS infection [5]. Despite the use of cART, which helps reduce HIV-1 plasma viremia (viral load) below the detectable limit (<50 copies of viral ribonucleic acid (RNA)/ml), no functional cure has been achieved with cART [5]. Upon cessation of cART, there is a rapid rebound in viremia [6]. Notably, the existence of a subset of transcriptionally silent/latent cells that can be reactivated when treatment is discontinued has been detected in individuals on effective cART [7,6,8]. Multiple factors can contribute to cART failure such as the occurrence of drug resistant virus during prolonged treatment [9] . However, the presence of latently infected cells is a major obstacle for HIV-1 eradication [10,11]. Latent reservoirs of infected cells can be found in infected cells in the blood. However, HIV-1 can also establish reservoirs of latently infected cells in several anatomical sites in the body, which are commonly known as sanctuary sites.

During infection, suboptimal drug penetration at certain anatomical sites may contribute to persistent HIV-1 replication and the replenishment of the latent reservoir [12]. A number of anatomical sites have been reported including: the gastrointestinal tract (GIT), the genital tract, semen, lymphoid tissues (e.g. lymph nodes, spleen), and the brain/central nervous system (CNS) [13–17]. The CNS is an important anatomical reservoir of HIV-1 [18]. HIV-1 is known to primarily infect macrophages and microglia cells in the CNS [12]. The cerebrospinal fluid (CSF) also represents a separate compartment for HIV-1 replication [12]. The blood brain barrier separates the brain from the peripheral blood, and the blood-cerebrospinal fluid barrier restricts the movement of free molecules and cells into the CSF. Both barriers provide obstacles for the passage of cART agents [12]. Although cART reduces HIV-1 in the CSF, viral genomes have been identified in the CSF, as well as in brain tissues of infected individuals that were on suppressive cART [19,20]. This further suggest the persistence of HIV-1 in the CNS. The GIT contains the largest amount of lymphoid tissues and lymphocytes in the body [21]. Gut associated lymphoid tissues (GALT) are highly targeted for infection and can maintain an elevated level of HIV-1 replication, which may be related with the large proportion of activated T cells and high predominance of cells

expressing the HIV-1 co-receptor CCR5 [22]. Furthermore, the GIT harbors persistent infections in individuals on long-term treatment, also making the GIT an important anatomical site for HIV-1 persistence [23]. HIV-1 has also been reported to infect cells of the female and male genital tract [24]. In the male genital tract, infections have been found in T lymphocytes and macrophages isolated from semen. Lymphocytes and macrophages infiltrating the testes, as well as spermatocytes, spermatids and residual germ cells can be targeted by HIV-1 [25,26] . Genital shedding of the virus has been observed in the semen despite undetectable levels of viral RNA in the blood [26]. Similarly, HIV-1 was detected in multiple cells and tissues of the female genital tract, such as epithelial and stromal cells of the uterus [27]. Viral shedding from the genital tract has also been demonstrated in women on cART [28]. Lymphoid tissues such as the lymph nodes and the spleen are important sites for viral infection and contain an abundance of infected cells. Despite a decrease in viral RNA in the lymph nodes and spleen following cART, HIV-1 still persists in the lymph nodes of infected people and in non-human primates on prolonged treatment [29–31].

cART distribution studies have shown variability in drug concentrations in certain anatomical sites such as the CNS and lymph nodes compared to the peripheral blood. These suboptimal concentrations contribute to the emergence of drug-resistant variants within the anatomical reservoirs of individuals on treatment [32,33,34,31]. Previous studies also reported a significant difference in composition of drug-resistant variants within different parts of GIT, such as the colon and the large intestine [35]. Overall, HIV-1 may be under selective pressure leading to the evolution and emergence of distinct viral sequences in sanctuary sites compared to those present in blood compartments. While viral variability significantly affects pathogenesis and disease progression, little is known about the impact of viral selective pressure on HIV-1 integration site distribution within anatomical sites as other than peripheral blood. Integration is meditated by the viral integrase enzyme and is an essential event in the life cycle of HIV-1 in which the viral genome is permanently incorporated into the host genome. HIV-1 integration site distribution from peripheral blood has been previously described [36,37]. Indeed, HIV-1 integration site selection from peripheral blood studies is known to occur in transcriptionally active genes, which are rich in GC and CpG islands content, high density of Alu repeat elements, low density in long interspersed nuclear element (LINE) and DNaseI hypersensitive sites [38]. Moreover,

integration also occurred in transcriptionally silent regions of the genome, such as gene deserts, centromeric heterochromatin, satellite DNA, introns and alphoid repeats [39,40,41,36,37].

Of importance, we previously identified non-B DNA structures as a novel factor that influences HIV-1 integration during infection [41]. Non-B DNA motifs are secondary structures that are abundant in our genome. They are formed by specific nucleotide sequences that exhibit non-canonical DNA base pairing. Several of those motifs have been identified. These include guanine-quadruplex (G4) motifs, A-phased repeats, inverted repeats, direct repeats, cruciform, slipped motifs, mirror repeats, short-tandem repeats, triplex repeats and Z-DNA. Latently infected cells with enriched integration near G4 and Z-DNA motifs could not be reactivated by the αCD3/CD28 latency reversal agent (see chapter 2). Since G4 and Z-DNA are known to suppress expression of adjacent genes, it is possible that such structures can impact expression of adjacent proviruses [42, 43,44, 45]. Furthermore, we have shown that different HIV-1 subtypes exhibit different integration preferences for non-B DNA motifs, which are further altered by cART (see chapter 3).

We also previously determined the integration site selection in an HIV-1 vector based system in murine brain cells [41]. However, to our knowledge the integration site selection of HIV-1 in compartmentalized sites from infected individuals has not been defined. Moreover, most studies have assessed HIV-1 integration sites in peripheral blood [36,37]. Findings from peripheral blood might not be observed in other body compartments, particularly in lymphoid tissues of the GIT, where the frequency of infected cells is higher [46] , and the CNS and lymph nodes, where drug concentrations are lower compared to the blood [18,33]. Since distinct HIV-1 strains can be found in compartmentalized sites in comparison to those present in the peripheral blood during cART [35] , we analyzed HIV-1 integration site profile in different anatomical sites in the body. Our data showed that the integration site profiles in the different anatomical sites are distinct, with striking differences in preferences for G4 motifs.

4.2 Materials and methods

4.2.1 Ethical statement and study participants' information for gastrointestinal tract biopsies samples and brain samples

Gastrointestinal tissue samples, peripheral blood mononuclear cells (PBMCs) and peripheral blood lymphocytes (PBLs) (PBMCs/PBLs) for this study had been collected from 5 infected individuals as previously described [35,47,48]. Briefly, patients were enrolled from a cohort of HIV-1 seropositive men who have sex with men (MSM). The cohort was followed at the Southern Alberta Clinic (SAC), Calgary, Alberta from the year 1993 to 1996. Ethical approval for all protocols and procedures were obtained from the Conjoint Health Ethics Research Board (CHREB, protocol approval #: REB15-1941) at the University of Calgary and Alberta Health Services (Calgary). All patients signed an informed consent upon enrollment. Patients were prospectively followed and testing for plasma viral load and $CD4^+$ T counts were performed for each individual at each visit. Additionally, upper and lower gastrointestinal endoscopies were performed in order to collect biopsies of tissues from the esophagus, stomach, duodenum, and colon. Samples were cryopreserved during shipment and stored at -70°C within 1 hour of collection [47]. PBMCs/PBLs were isolated from blood and stored in liquid nitrogen [35]. This cohort was recruited prior to the introduction of Highly Active Antiretroviral therapy (HAART)/combination antiretroviral therapy (cART) at the SAC in late 1997. Samples analyzed in this study were from individuals who received monotherapy, or dual therapy with the nucleoside reverse transcriptase inhibitors (NRTIs) such as azidothymidine (AZT/zidovudine), dideoxyinosine (ddI) or cART prior to the study and during the study as previously described [35]. Brain tissue samples used in this study have been collected from the frontal lobe of 8 HIV-1 infected individuals as previously described [49]. Samples were collected at autopsy with appropriate consent and frozen at -80°C. Patients were not receiving antiretroviral therapy at the time of samples collection.

4.2.2 DNA isolation and HIV-1 integration library

Total genomic DNA was extracted from gastrointestinal tract tissue biopsies and (PBMCs/PBLs) using Trizol Reagent (Invitrogen) as described [35,47]. DNA was also extracted from brain tissues with Trizol Reagent.

All genomic DNAs were processed for integration site analysis and sequenced using the Illumina MiSeq platform. Extracted genomic DNA was restriction enzyme digested with MseI overnight at 37°C. Digested DNA was column purified with the Gel/PCR DNA Fragments Kit (Geneaid, cat#: DF100) according to manufacturer's instructions. Next, compatible double-stranded linkers to the MseI sites were prepared as follows: MseI Linker (+) 5'GTAATACGACTCACTATAGGGCTCCG CTTAAGGGAC 3'and MseI Linker (-): 5' [Phos]-TAGTCCCTTAAGCG GAG-[AmC7-Q] 3' were mixed (20µl MseI Linker (+) [40 µM] and 20 µl MseI Linker (-) [40 µM]). The linker mixture was denatured for 5 min at 90°C and cooled 1°C every 3 min until the temperature reached 20°C using the T100^{TM} Thermal Cycler (Bio-Rad). The prepared linkers from here on forward are referred to as the "adapter mix".

Purified DNA was linker ligated with the adapter mix at 21°C for about 14 hours with 13.5µl of MseI digested samples, 3.5µl of adapter mix, 1µl of T4 DNA Ligase (400U/µl, [NEB, cat#: M0202S), and 2 µl of 10x ligase buffer. Subsequently, 20µl of the ligated sample was digested at 37°C for 4 hours with 2µl of DpnI (20U/µl), 2µl of NarI (5U/µl), 5µl of 10x buffer and water to a total volume of 50µl. Following digestion, the samples were column purified. The junctions between the integrated HIV-1 LTR sequence and adjacent genomic sequence were amplified in two separate rounds of PCR amplification.

The HIV-1 NL4-3 LTR sequence were used to design primers that amplify through the HIV-1 LTR. The RuparLTR (Forward) 5'-TGCTTCAAGTAGTGTGTGC-3' primer that anneals to the HIV-1 LTRs and the Linker1 (Reverse) 5'-GTAATACGACTCACTATAG GGC-3'primer specific to the MseI linker sequences were used for the first round of PCR amplification. Each PCR reaction mixture consisted of 15.5µl sterile water, 5µl of NarI/DnpI digested sample, 2.5µl of 10x Advantage 2 PCR Buffer, 0.5 µl of 15µM of Linker1 primer, 0.5 µl of 15µM RuparLTR primer, 0.5µl of 10mM dNTPs and 0.5 µl of 50X Advantage 2 PCR polymerase mix (Takara Bio Inc., cat#:639201). PCR was run on T100^{TM} Thermal Cycler (Bio-Rad) under the following cycling conditions: 1 min at 94°C, 5 cycles of 2 sec at 94°C, 1 min at 72°C with an additional 20 cycles of 2 sec at 94°C, 1 min at 67°C and a final extension cycle for 1 min at 72°C and a 4°C hold. The second round of nested PCR amplification was performed using sample from the first round of

PCR amplification. The PCR reaction mixture and cycling condition were as described for the first round of PRC amplification. The following primer set was used for nested PCR: Rupar-LTR2nested (Forward) 5'-CTCTGGTAACTAGAGATCCCTCAGAC C-3', Linker2nested (Reverse) 5'-AGGGCTCCGCTTAAGGGAC-3' and. Next, Illumina adapter overhang nucleotide sequences were added to the HIV-1 LTR sequence and the MseI linker sequence. Illutag-Forward 5'-GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAGCTCTGGTAACTAGAGATCCCTCAGAC C-3'and Illutag-Reverse 5'-TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGAGGGCTCCGCTTAAGGGAC -3'. Underlined section of the two Illutag primers represent the overhang section. Illumina adapters were utilized in a PCR reaction mixture contained 15.5µl sterile water, 5µl of nested PCR samples, 2.5µl of Advantage 2 PCR Buffer (10x), 0.5 µl of Forward adapter (10µM), 0.5 µl of Reverse adapter (10µM), 0.5µl of dNTPs (10mM) and 0.5 µl of 50X Advantage 2 PCR polymerase mix. PCR was run on T100^{TM} Thermal Cycler (Bio-Rad). Cycling conditions were as described for the first round of PRC amplification.

The PCR products were purified using AmPure XP beads (Beckman Coulter, cat#: A63881) and the DNA samples were processed using the Nextera XT Index Kit (Illumina). The Nextera XT Indexes technology utilizes a single tagmentation reaction that fragments and tags input DNA with unique adapter and index (barcodes) sequences on both ends of the DNA as previously described [41]. . The DNA samples were purified using AmPure XP beads following addition of the barcodes. The barcoded samples were quantified using the Quant-it PicoGreen dsDNA Assay Kit (Invitrogen, cat#: P7589). The absorbance of the samples were read (excitation 480nm for 10 sec and emission 540 nm for 10 sec) with the Cytation5 Imaging Reader (BioTek) and the Gen5 3.02.1 analysis software. Sample concentration was determined using a standard concentration curve. The barcoded samples were sequenced through Illumina MiSeq using 2×150 bp chemistry at the London Regional Genomics Centre at the Robarts Research Institute (Western University, Canada) and at Case Western Reserve University (USA).

4.2.3 Integration site analysis

Fastq sequencing reads were quality trimmed and unique integration sites identified using our in-house bioinformatics pipeline Barr Lab Integration Site Identification Pipeline

(BLISIP) (version 2.9) [41]. BLISIP version 2.9 includes the following updates: bedtools (v2.25.0) which is used to compute distances between integration sites and genomic features, bioawk (awk version 20110810) a programming language for biological data manipulation, bowtie2 (version 2.3.4.1) is used for aligning sequence reads to the human genome, and restrSiteUtils (v1.2.9) is used to generate *in silico* matched random control integration sites based on restriction enzyme used or DNA shearing method. HIV-1 LTR-containing Fastq sequences were identified and filtered by allowing up to a maximum of five mismatches with the reference NL4-3 LTR sequence and if the LTR sequence had no match with any region of the human genome (GRCh37/hg19). Integration site profile heatmaps were generated using our in-house python program BHmap (BHmap version 1.0). Sites that could not be unambiguously mapped to a single region in the genome were excluded from the study. Mapping of integration sites to non-B DNA motifs was performed using the Non-B DB for the human genome (GRCh37/hg19) as previously described [50, 51]. Lamina associated domains (LADs) were retrieved from http://dx.doi.org/10.1038/nature06947 [52].

4.2.4 Statistical analysis

The Fisher's exact test was used for all comparisons of integration site distributions in Figures 4.1, and 4.2. * $P < 0.05$, ** $P < 0.01$, *** $P < 0.001$, **** $P < 0.0001$.

4.3 Results

4.3.1 HIV-1 anatomical reservoirs exhibit distinct integration site preferences

To compare the integration site profiles in different anatomical sites in infected individuals, we generated integration site libraries from genomic DNA isolated from tissues from the blood (PBMCs/PBLs), esophagus, stomach, duodenum and colon of 5 HIV-1 infected individuals receiving antiretroviral therapy **(Supplemental Table 4.1)**. Integration sites from brain samples were determined from a separate cohort of 8 HIV-1 infected individuals who were not receiving antiretroviral therapy. To generate the integration site profile, we used an in-house bioinformatics pipeline called BLISIP, as used previously (see chapters 2 and 3) and [41]. Unique integration sites from each anatomical site were compared with

matched random control (MRC) datasets generated *in silico* as previously described [41]. The integration site distribution was divided into four bins starting from within each genomic feature to > 50,000 base pairs (bp) away from the feature **(Figure 4.1).** In agreement with the work presented in chapter 2 of this thesis and others [36], integration sites in cells from our peripheral blood (PMBCs/PBLs) samples were enriched in genes (81% of all integration sites) compared to MRC **(Figure 4.1, Supplemental Table 4.2)**. Integration sites in PBMCs/PBLs were also disfavored within CpG islands but highly enriched near these genomic features. In the esophagus, stomach, duodenum, colon and brain the majority of integration occurred within genes (79%, 71 %, 82%, 85% and 57% respectively) similarly to PBMCs/PBLs. Interestingly, integration within genes was drastically lower in brain samples compared to other tissues. Although all anatomical sites exhibited enriched integration near CpG islands (5000-49999 bp; $P < 0.001$ or $P < 0.0001$) including in PBMCs/PBLs, the duodenum also exhibited enrichment directly in CpG islands compared to MRC **(Figure 4.1)**. Additionally, integration in or near heterochromatic lamina-associated domains (LADs) and satellite DNA, which are abundant in heterochromatin, were depleted in the brain as opposed to other reservoirs. Together, these data show that our bioinformatics analyses agree with previous findings as it related to integration sites in peripheral blood and shows that tissue from different anatomical sites exhibit different preferences in their integration site selection.

4.3.2 Non-B DNA motifs are targeted for integration in different anatomical reservoirs of HIV-1 infected individuals

Our previous findings showed that HIV-1 favors integration in and/or near non-B DNA motifs in peripheral blood samples of infected individuals (see chapter 2 and 3). To assess integration site frequency with respect to non-B DNA motifs in other anatomical reservoirs **(Supplemental Table 4.1)** we quantified integration sites directly within each non-B DNA motif or in distance bins of 50 bp up to 500 bp away from the feature. Integration analyses were performed using BLISIP and unique integration sites were compared to the MRC. As shown in **Figure 4.2**, the integration site profile differed considerably among the different anatomical sites. The majority of integration sites were enriched directly in non-B DNA motifs for PBMCs/PBLs, stomach and duodenum.

Figure 4.1: HIV-1 anatomical reservoirs exhibit distinct integration site preferences. Heatmap depicting the fold enrichment or depletion of integration sites in common genomic features compared to the matched random control (MRC). Darker shades represent higher fold-changes in the ratio of integration sites to MRC. Numbers within each heatmap represent the fold-increase or decrease in the number of unique integration sites in different anatomical reservoirs compared to MRC. With each heatmap, not a number (nan) indicates that 0 integrations were observed and 0 were expected by chance. Significant differences were determined by Fisher's exact test and are denoted by asterisks (*P < 0.05; **P < 0.01; ***P < 0.001; ****P < 0.0001). Number of unique integration sites are as follow: PBMCs/PBLs = 232 sites, esophagus = 255 sites, stomach = 308 sites, duodenum = 195 sites, colon = 176 sites and the brain = 155 sites.

Figure 4.1

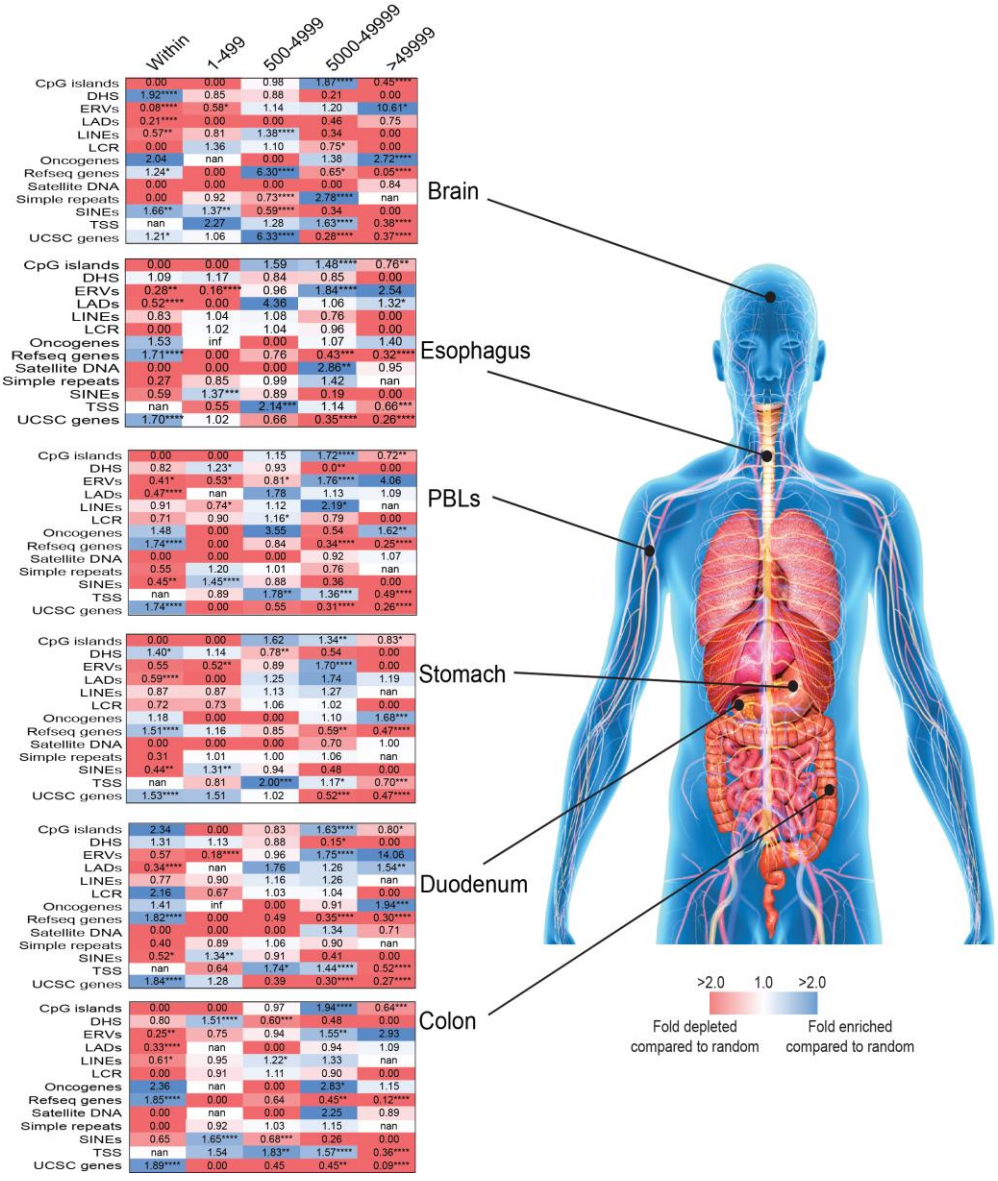


Figure 4.1: HIV-1 anatomical reservoirs exhibit distinct integration site preferences.

On the other hand, most integration occurred within 1-49 bp from non-B DNA in the colon. Notably, integration in and/or near G4, Z-DNA, cruciform and triplex motifs was highly enriched in the different anatomical reservoirs (**Figure 4.2, Supplemental Table 4.3**). Interestingly, enriched integration 100-150 bp away from G4 was observed in all anatomical reservoir as previous observed (see chapter 2 and 3). Additionally, integration within cruciform motifs was only targeted in PBMCs/PBLs. Integration was also favored near cruciform in PBMCs/PBLs. Other compartments (esophagus, stomach, colon, duodenum and brain) showed enriched integration only near cruciform motifs. Compared to other anatomical sites, the brain showed no preference for integration either in and/or near triplex motifs. Together, these data identified distinct non-B DNA integration site profiles for PBMCs/PBLs, colon, stomach, duodenum, esophagus and the brain.

4.4 Discussion

The data presented here show that different anatomical reservoirs of HIV-1 have different integration site preferences. Importantly, we also identified integration to be enriched in or near non-B DNA motifs in all anatomical reservoirs. Specifically, we determined that the integration sites in all latent reservoirs are strongly enriched in or near specific non-B DNA motifs that are known to inhibit gene expression, such as G4, cruciform, Z-DNA and triplex structures [42,44,53–61]. Previous HIV-1 integration sites analyses in infected individuals showed that active genes are preferred sites for integration *in vivo*. These datasets were obtained from cells originating from peripheral blood (e.g. PBMCs or $CD4^+$ T cells) [36,37]. In the current study, our analysis in PBMCs/PBLs of integration site distribution in genes are consistent with other findings.

Integration site distribution in other anatomical reservoirs showed similar results as in the peripheral blood except in the brain, which showed 15 to 28% lower frequency in genes compared to other compartments. The lens epithelium derived growth factor and co-factor p75 (LEDGF/p75) is a ubiquitously expressed protein which is well known to interact with the viral integrase enzyme and help target integration into active genes of the genome. Even though LEDGF/p75 is ubiquitously expressed, the expression of this protein is lower in the adult human brain and within specific regions of the brain [62].

Figure 4.2: Non-B DNA motifs are targeted for integration in different anatomical reservoirs of HIV-1 infected individuals. Heatmaps showing the distribution of unique integration sites in non-B DNA motifs from PBMCs/PBLs, esophagus, stomach, duodenum, colon and brain tissue samples compared to the matched random control (MRC). Darker shades represent higher fold-changes in the ratio of integration sites to MRC sites. Numbers within each heatmap represent the fold-increase or decrease in the number of unique integration sites in different anatomical reservoir compared to MRC. Not a number (nan) indicates that 0 integrations were observed and 0 were expected by chance. Significant differences were determined by Fisher's exact test and are denoted by asterisks (*P < 0.05; **P < 0.01; ***P < 0.001; ****P < 0.0001). Number of unique integration sites are as follow: PBMC/PBL = 232 sites, esophagus = 255 sites, stomach = 308 sites, duodenum = 195 sites, colon = 176 sites and the brain = 155 sites.

Figure 4.2

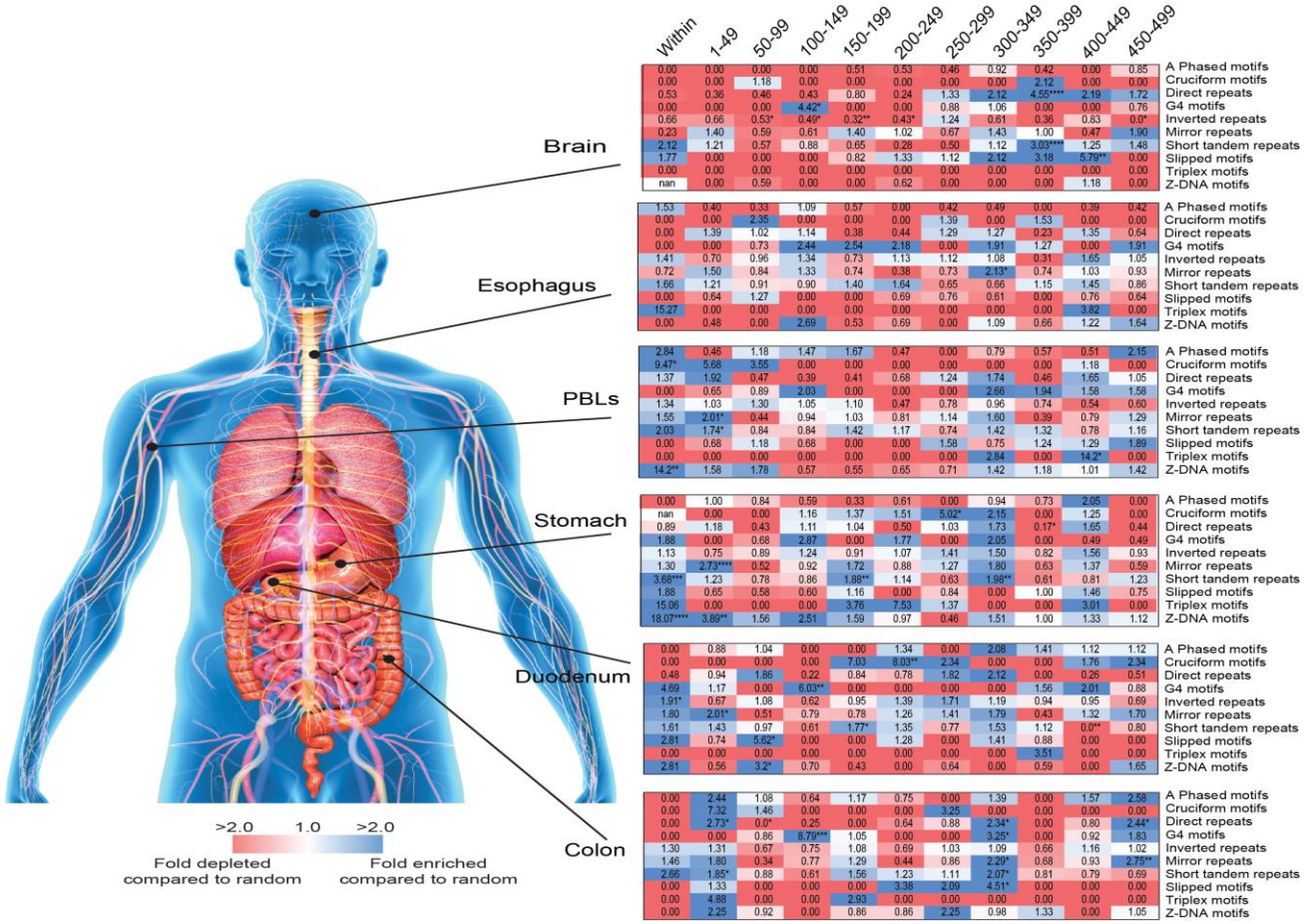


Figure 4.2: Non-B DNA motifs that are for integration in different anatomical reservoirs of HIV-1 infected individuals.

We and others also previously showed that depletion of LEDGF/p75 caused a substantial decrease in integration into genes **(chapter 2, Figure 2.3)** [63,64,65,66]. The same was true for the polyadenylation specificity factor 6 (CPSF6) protein which helps promote HIV-1 integration into actively transcribed genes residing in gene-dense regions, thereby reducing integration into other genomic regions conducive to latency such as heterochromatin. Therefore, a decreased expression of either LEDGF/p75 or CPSF6 may contribute to the lower frequency of integration in genes seen in the brain compared to other anatomical reservoirs. It is possible that the differences in integration are not only the result of selective pressure of cART on the variability of HIV-1 sequences in different compartments, but the result of the host cellular environment or the different phenotypes between the infected cells in each compartment.

A recent study also showed that the transcriptional initiation in $CD4^+$ T cells from the GIT (e.g. rectum) is much lower than that of $CD4^+$ T cells obtained from the blood [67]. However, it was unclear which cellular or viral factors contributed to the difference in transcriptional repression within the two compartments. Given the ability of HIV-1 to target non-B DNA motifs in the genome within different latent reservoirs, it is possible that the difference in the transcriptional repression seen in compartments other than in the blood could be due to their selective preferences in integration in or near specific non-B DNA motifs. Non-B DNA are abundant in the human genome and have been associated with chromatin remodeling and transcriptional activities [45]. It is possible that the expression of specific host proteins influence HIV-1 site targeting into or near specific non-B DNA structures within the different reservoirs.

In conclusion, our observations indicate that different anatomical reservoirs of HIV-1 infection are enriched in and/or near non-B DNA motifs and implicate non-B DNA motifs as a potential factor influencing HIV-1 integration site targeting within various latent reservoirs.

4.5 References

1. Barré-Sinoussi F, Chermann JC, Rey F, Nugeyre MT, Chamaret S, Gruest J,

Dauguet C, Axler-Blin C, Vézinet-Brun F, Rouzioux C, Rozenbaum W, M. L. Isolation of a T-lymphotropic retrovirus from a patient at risk for acquired immune deficiency syndrome (AIDS). *Science* **220,** 868–71 (1983).

2. Mikulas Popovic, M. G. Sarngadharan, E. R. and R. C. G. Detection , isolation , and continuous production of cytopathic retroviruses (HTLV-III) from patients with AIDS and pre-AIDS. *Science* **224,** 497–500 (1984).

3. Robert C. Gallo, Syed Z. Salahuddin, Mikulas Popovic, Gene M. Shearer, Mark Kaplan, Barton F. Haynes, Thomas J. Palker, Robert Redfield, James Oleske, Bijan Safai, Gilbert White, P. F. and P. D. M. Frequent detection and isolation of cytopathic retroviruses (HTLV-III) from patients with AIDS and at risk for AIDS. *Science* **224,** 500–503 (1984).

4. UNAIDS. Global HIV & AIDS statistics — 2018 fact sheet. (2018). Available at: http://www.unaids.org/en/resources/fact-sheet. (Accessed: 17th November 2018)

5. Arts, E. J. & Hazuda, D. J. HIV-1 antiretroviral drug therapy. *Cold Spring Harb. Perspect. Med.* **2,** a007161 (2012).

6. Davey Jr., R. T. *et al.* HIV-1 and T cell dynamics after interruption of highly active antiretroviral therapy (HAART) in patients with a history of sustained viral suppression. *Proc Natl Acad Sci U S A* **96,** 15109–15114 (1999).

7. Chun, T.-W. *et al.* Early establishment of a pool of latently infected, resting CD4+ T cells during primary HIV-1 infection. *Proc. Natl. Acad. Sci. U. S. A.* **95,** 8869–8873 (1998).

8. Chun, T.-W. *et al.* Presence of an inducible HIV-1 latent reservoir during highly active antiretroviral therapy. *Proc. Natl. Acad. Sci.* **94,** 13193–13197 (1997).

9. Karmochkine, M. *et al.* The cumulative occurrence of resistance mutations in the HIV-1 protease gene is associated with failure of salvage therapy with ritonavir and saquinavir in protease inhibitor-experienced patients. *Antiviral Res.* **47,** 179–188 (2000).

10. Wong, J. K. *et al.* Recovery of replication-competent HIV despite prolonged suppression of plasma viremia. *Science* **278,** 1291–1295 (1997).

11. Siliciano, J. D. *et al.* Long-term follow-up studies confirm the stability of the latent reservoir for HIV-1 in resting CD4+T cells. *Nat. Med.* **9,** 727–728 (2003).

12. Saksena, N. K. & Potter, S. J. Reservoirs of HIV-1 in vivo: Implications for antiretroviral therapy. *AIDS Reviews* **5,** 3–18 (2003).

13. Poss, M. *et al.* Diversity in virus populations from genital secretions and peripheral blood from women recently infected with human immunodeficiency virus type 1. *J. Virol.* **69,** 8118–8122 (1995).

14. Delwart, E. L. *et al.* Human immunodeficiency virus type 1 populations in blood and semen. *J. Virol.* **72,** 617–23 (1998).

15. van't Wout, A. B. *et al.* Analysis of the temporal relationship between human immunodeficiency virus type 1 quasispecies in sequential blood samples and various organs obtained at autopsy. *J Virol* **72,** 488–496 (1998).

16. Chun, T.-W. *et al.* Quantification of latent tissue reservoirs and total body viral load in HIV-1 infection. *Nature* **387,** 183–188 (1997).

17. Di Stefano, M. *et al.* Reverse transcriptase sequence of paired isolates of cerebrospinal fluid and blood from patients infected with human immunodeficiency virus type 1 during zidovudine treatment. *J. Clin. Microbiol.* **33,** 352–355 (1995).

18. Pomerantz, R. J. Reservoirs, sanctuaries, and residual disease: The hiding spots of HIV-1. *HIV Clin. Trials* **4,** 137–143 (2003).

19. Spudich, S., Lollo, N., Liegler, T., Deeks, S. G. & Price, R. W. Treatment benefit on cerebrospinal fluid HIV-1 levels in the setting of systemic virological suppression and failure. *J. Infect. Dis.* **194,** 1686–1696 (2006).

20. Kumar, A. M., Borodowsky, I., Fernandez, B., Gonzalez, L. & Kumar, M. Human

immunodeficiency virus type 1 RNA levels in different regions of human brain: Quantification using real-time reverse transcriptase-polymerase chain reaction. *J. Neurovirol.* **13,** 210–224 (2007).

21. Mowat, A. M. I. & Viney, J. L. The anatomical basis of intestinal immunity. *Immunol. Rev.* **156,** 145–166 (1997).

22. Poles, M., Elliott, J. & Taing, P. A preponderance of $CCR5^+$ $CXCR4^+$ mononuclear cells enhances gastrointestinal mucosal susceptibility to human immunodeficiency virus type 1 infection. *J. Virol.* **75,** 8390–8399 (2001).

23. Chun, T. *et al.* Persistence of HIV in Gut-associated lymphoid tissue despite long-term antiretroviral therapy. *J. Infect. Dis.* **197,** 714–720 (2008).

24. Coombs, R.W., Reichelderfer, P.S. & Landay, A. L. Recent observations on HIV type-1 infection in the genital tract of men and women. *AIDS* **17,** 455–480 (2003).

25. Muciaccia, B. *et al.* Testicular germ cells of HIV-seropositive asymptomatic men are infected by the virus. *J. Reprod. Immunol.* **41,** 81–93 (1998).

26. Nuovo, G. J. *et al.* Spermatogonia and their progeny a study by polymerase chain reaction in situ hybridization. *Am. J. Pathol.* **144,** 1142–1148 (1994).

27. Howell, A. *et al.* Human immunodeficiency virus type 1 infection of cells and tissues from the upper and lower human female reproductive tract. *J. Vriology* **71,** 3498–3506 (1997).

28. Kovacs, A. *et al.* Determinants of HIV-1 shedding in the genital tract of women. *Lancet* **358,** 1593–1601 (2001).

29. North, T. W. *et al.* Viral Sanctuaries during highly active antiretroviral therapy in a nonhuman primate model for AIDS. *J. Virol.* **84,** 2913–2922 (2010).

30. Günthard, H. F. *et al.* Residual human immunodeficiency virus (HIV) type 1 RNA and DNA in lymph nodes and HIV RNA in genital secretions and in cerebrospinal

fluid after suppression of viremia for 2 Years. *J. Infect. Dis.* **183,** 1318–1327 (2001).

31. Wong, J. K. *et al.* Reduction of HIV-1 in blood and lymph nodes following potent antiretroviral therapy and the virologic correlates of treatment failure. *Proc. Natl. Acad. Sci.* **94,** 12574–12579 (1997).

32. Ene L, Duiculescu D, Sm, R. & Victor. How much do antiretroviral drugs penetrate into the central nervous system? *J. Med. Life* **4,** 432–439 (2011).

33. Fletcher, C. V. *et al.* Persistent HIV-1 replication is associated with lower antiretroviral drug concentrations in lymphatic tissues. *Proc. Natl. Acad. Sci.* **111,** 2307–2312 (2014).

34. Letendre, S. *et al.* Validation of the CNS penetration-effectiveness rank for quantifying antiretroviral penetration into the central nervous system. *Arch Neurol* **65,** 65–70 (2008).

35. van Marle, G. *et al.* Higher levels of Zidovudine resistant HIV in the colon compared to blood and other gastrointestinal compartments in HIV infection. *Retrovirology* **7:74,** (2010).

36. Cohn, L. B. *et al.* HIV-1 integration landscape during latent and active infection. *Cell* **160,** 420–432 (2015).

37. F. Maldarelli, 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. H. Specific HIV integration sites are linked to clonal expansion and persistence of infected cells. *Science* **345,** 179–183 (2014).

38. Bushman, F. *et al.* Genome-wide analysis of retroviral DNA integration. *Nat. Rev. Microbiol.* **3,** 848–58 (2005).

39. Lewinski, M. K. *et al.* Genome-wide analysis of chromosomal features repressing human immunodeficiency virus transcription. *J. Virol.* **79,** 6610–6619 (2005).

40. Jordan, A., Bisgrove, D. & Verdin, E. HIV reproducibly establishes a latent infection after acute infection of T cells in vitro. *EMBO J.* **22,** 1868–1877 (2003).

41. McAllister, R. G. *et al.* Lentivector integration sites in ependymal cells from a model of metachromatic leukodystrophy: non-B DNA as a new factor influencing integration. *Mol. Ther. Nucleic Acids* **3,** e187 (2014).

42. Bochman, M. L., Paeschke, K. & Zakian, V. a. DNA secondary structures: stability and function of G-quadruplex structures. *Nat. Rev. Genet.* **13,** 770–80 (2012).

43. Tornaletti, S., Park-Snyder, S. & Hanawalt, P. C. G4-forming sequences in the non-transcribed DNA strand pose blocks to T7 RNA polymerase and mammalian RNA polymerase II. *J. Biol. Chem.* **283,** 12756–12762 (2008).

44. Delic J, Onclercq R, M.-C. M. Inhibition and enhancement of eukaryotic gene expression by potential non-B DNA sequences. *Biochem Biophys Res Commun* **180,** 1273–83 (1991).

45. Kouzine, F. *et al.* Permanganate/S1 nuclease footprinting reveals non-B DNA structures with regulatory potential across a mammalian genome. *Cell Syst.* **4,** 344–356.e7 (2017).

46. Yukl, S. A. *et al.* The distribution of HIV DNA and RNA in cell subsets differs in gut and blood of HIV-positive patients on ART: Implications for viral persistence. *J. Infect. Dis.* **208,** 1212–1220 (2013).

47. van Marle, G. *et al.* Compartmentalization of the gut viral reservoir in HIV-1 infected patients. *Retrovirology* **4,** 1–14 (2007).

48. van Marle, G., Sharkey, K. A., Gill, M. J. & Church, D. L. Gastrointestinal viral load and enteroendocrine cell number are associated with altered survival in HIV-1 infected individuals. *PLoS One* **8,** e75967 (2013).

49. Maingat, F. *et al.* Regulation of Lentivirus Neurovirulence by Lipopolysaccharide Conditioning: Suppression of CXCL10 in the Brain by IL-10. *J Immunol.* **184,**

1566–1574 (2010).

50. Cer, R. Z. *et al.* Non-B DB v2.0: A database of predicted non-B DNA-forming motifs and its associated tools. *Nucleic Acids Res.* **41,** 94–100 (2013).

51. Cer, R. Z. *et al.* Non-B DB: a database of predicted non-B DNA-forming motifs in mammalian genomes. *Nucleic Acids Res.* **39,** D383-91 (2011).

52. Guelen, L. *et al.* Domain organization of human chromosomes revealed by mapping of nuclear lamina interactions. *Nature* **453,** 948–951 (2008).

53. Siddiqui-Jain, A., Grand, C. L., Bearss, D. J. & Hurley, L. H. Direct evidence for a G-quadruplex in a promoter region and its targeting with a small molecule to repress c-MYC transcription. *Proc. Natl. Acad. Sci. U. S. A.* **99,** 11593–8 (2002).

54. Verma, A., Yadav, V. K., Basundra, R., Kumar, A. & Chowdhury, S. Evidence of genome-wide G4 DNA-mediated gene expression in human cancer cells. *Nucleic Acids Res.* **37,** 4194–4204 (2009).

55. Waga, S., Mizuno, S. & Yoshida, M. Chromosomal protein HMG1 removes the transcriptional block caused by the cruciform in supercoiled DNA. *J. Biol. Chem.* **265,** 19424–8 (1990).

56. Waga, S., Mizuno, S. & Yoshida, M. Nonhistone protein HMG1 removes the transcriptional block caused by left-handed Z-form segment in a supercoiled DNA. *Biochem. Biophys. Res. Commun.* **153,** 334–9 (1988).

57. Jain, A., Magistri, M., Napoli, S., Carbone, G. M. & Catapano, C. V. Mechanisms of triplex DNA-mediated inhibition of transcription initiation in cells. *Biochimie* **92,** 317–320 (2010).

58. Maher, L. J., Dervan, P. B. & Wold, B. Analysis of promoter-specific repression by triple-helical DNA complexes in a eukaryotic cell-free transcription system. *Biochemistry* **31,** 70–81 (1992).

59. Brázda, V., Laister, R. C., Jagelská, E. B. & Arrowsmith, C. Cruciform structures are a common DNA feature important for regulating biological processes. *BMC Mol. Biol.* **12,** 33 (2011).

60. Tornaletti, S., Park-Snyder, S. & Hanawalt, P. C. G4-forming sequences in the non-transcribed DNA strand pose blocks to T7 RNA polymerase and mammalian RNA polymerase II. *J. Biol. Chem.* **283,** 12756–62 (2008).

61. Belotserkovskii, B. P. *et al.* A triplex-forming sequence from the human c-MYC promoter interferes with DNA transcription. *J. Biol. Chem.* **282,** 32433–32441 (2007).

62. Chylack, L. T. *et al.* Lens epithelium-derived growth factor (LEDGF/p75) expression in fetal and adult human brain. *Exp. Eye Res.* **79,** 941–948 (2004).

63. Ciuffi, A. *et al.* A role for LEDGF/p75 in targeting HIV DNA integration. *Nat. Med.* **11,** 1287–9 (2005).

64. Shun, M. C. *et al.* LEDGF/p75 functions downstream from preintegration complex formation to effect gene-specific HIV-1 integration. *Genes Dev.* **21,** 1767–1778 (2007).

65. Marshall, H. M. *et al.* Role of PSIP 1/LEDGF/p75 in lentiviral infectivity and integration targeting. *PLoS One* **2,** e1340 (2007).

66. Achuthan, V. *et al.* Capsid-CPSF6 interaction licenses nuclear HIV-1 trafficking to sites of viral DNA integration. *Cell Host Microbe* **24,** 392–404 (2018).

67. Telwatte, S. *et al.* Gut and blood differ in constitutive blocks to HIV transcription, suggesting tissue-specific differences in the mechanisms that govern HIV latency. *PLOS Pathog.* **14,** e1007357 (2018).

Chapter 5

5 General discussion and future directions

5.1 Thesis summary

HIV-1 is the most clinically prevalent retrovirus in the human population that causes life-long infection. The latent reservoir of HIV-1, which harbors transcriptionally silent replication competent virus, represents the major obstacle for curing HIV-1 infection. Although several bodies of work suggest that the integration site of the virus in the human genome plays a critical role in disease persistence and reactivation of HIV-1 expression, better understanding of the local genomic environment surrounding integrated proviruses is needed to provide more insights into its contribution to latency. The work described in my thesis presents my characterization of retroviral integration site selection and genomic determinants that may impact latency establishment or reversal.

In Chapter 2 of this thesis, my work, and that of others in the laboratory, established a strong correlation between HIV-1 integration sites in and/or near non-B DNA motifs and latency. We were able to show that the location of integration sites that predominate in latently infected cells are enriched in or near non-B DNA motifs, which are well-known to inhibit gene expression. Specifically, we demonstrated that HIV-1 integration near guanine-quadruplex (G4) motifs, a type of non-B DNA motif may influence reactivation of the latent proviruses by latency reversal agents **(Figure 2.2)**. To further characterize the implication of G4 motifs in HIV-1 latency and integration site selection, we treated cells with G4 ligands that stabilize or destabilize G4 structures, which were found to significantly alter HIV-1 integration site preference for G4 motifs **(Figure 2.5).** Historically, much of our understanding of HIV-1 infection has been primarily modeled on HIV-1 subtype B infections. However, other subtypes exist around the world which are known to present different disease progression as opposed to subtypes B [1]. In Chapter 3, I characterized the HIV-1 integration profiles of different HIV-1 subtypes via next-generation sequencing to determine if any difference exists in their integration site selection preference. I also assessed the integration site profile among evolutionary diverse retroviruses such as SIV, FIV, HTLV-1, MLV, MMTV, FV, ASLV and ERVs.

Remarkably, our studies showed that non-B DNA motifs were highly targeted for integration not only by HIV-1 subtype A, B, C and D but by other retroviruses **(Figure 3.2 and 3.4 C)**. Antiretroviral treatment was shown to strongly alter the integration site profile in different HIV-1 subtypes **(Figure 3.6).** Finally, in Chapter 4 I explored HIV-1 integration site preference in anatomical sites which are also known to harbor latent viruses despite antiretroviral therapy. I utilized next generation sequencing to investigate the difference in HIV-1 integration site selection in the peripheral blood, central nervous system and gastrointestinal tract (GIT). Notably, we also showed here that non-B DNA motifs are strongly targeted in different anatomical sites during infection. Interestingly, all compartments also showed strong preferences for integration near G4 motifs **(Figure 4.2).**

Overall, we were able to expand our current understanding of HIV-1 integration site preferences and identified non-B DNA motifs as novel factors that influence HIV-1 integration site targeting. Furthermore, we demonstrated that integration sites in latently infected cells are enriched in and/or near non-B DNA motifs suggesting that non-B DNA structures, particularly G4 motifs, likely contribute to the establishment and maintenance of HIV-1 latency in subtypes B and in other subtypes.

5.1.1 Non-B DNA motifs are targeted in quiescent/latently infected cells

Our analysis of HIV-1 integration site preference began through the assessment of viral integration profile in latently infected cells from infected individuals. In our study, we presented analyses that are consistent with previous studies showing that integration in heterochromatin regions are more frequent in latently infected cells as opposed to productively infected cells [2,3] **(Figure 2.2).** We also observed a strong enrichment of integration near non-B DNA (e.g. G4 motifs) that strongly influence nearby gene expression **(Figure 2.1D and 2.2)** and latency reactivation. Previous studies have shown that G4 motifs are highly localized to heterochromatin [4]. The fact that integration sites in or near G4-motifs were strongly enriched in latently infected cells could further explain why heterochromatin regions are targeted in latently infected cells. Alternatively, G4 motif binding proteins might influence integration near these motifs. This could also be the case for other non-B DNA motifs. Interestingly, HIV-1 integrase is known to bind directly to

G4 motifs [5–14]. Other tethering factors such LEDGF/p75 and CPSF6 have been shown to promote integration into euchromatin regions, which are more permissive for gene expression. In our present study, we have showed that LEDGF/p75 and CPSF6 promote integration in and/or near non-B DNA motifs, especially those known to repress gene expression (e.g. G4, Z-DNA, cruciform and triplex motifs) **(Figure 2.3)**.

Latency reversal agents have been proposed to reactivate the latent reservoir and induce depletion of the virus [15]. However, only a small fraction of the latent reservoir can be reactivated [3,16]. It is therefore conceivable that the integration site placement within certain genomic features contribute to the maintenance and establishment of the latent reservoir. In this thesis we further demonstrated that failure to become reactivated by latency reversal agent (αCD3/CD28) correlated with HIV-1 integration near non-B DNA motifs particularly G4, and Z-DNA motifs **(Figure 2.2)**. Since G4 motifs and Z-DNA are known to both impede nearby gene expression though mechanisms that involve stalling of the RNA polymerase or interfering with the assembly of transcription pre-initiation complexes [17,18] [17–19] it is possible that integration near non-B DNA motifs can significantly silence proviral expression therefore contributing to latency. Our findings using G4-stabilizing and G4-destabilizing ligands to modulate G4 formation indicate that the secondary structure of non-B DNA motifs and not their primary sequences is more likely to play an important role in attracting HIV-1 pre-integration complex **(Figure 2.5).**

5.1.2 A comparative analysis of the integration site distribution of evolutionary diverse retroviruses

Different HIV-1 subtypes are known to present different disease progression. Integration which is an important event in the life cycle of retroviruses is essential for replication to occur. We therefore explored the differences in integration profile among different HIV-1 subtypes and evolutionary diverse retroviruses. Interestingly, we showed that non-B DNA motifs are also targeted by other retroviruses. However, they presented distinct integration profiles **(Figure 3.2).** Variations in the properties of their integrase proteins could greatly influence their integration site targeting. Furthermore, sequence differences in their integrase gene could also influence their integration site preference for specific genomic

features. Interaction of the virus with cellular host factors and chromosomal DNA might also be essential in determining proviral integration site selection.

In another series of experiments, we further dissected the differences in integration profile seen in *in vitro* infection models of HIV-1 and datasets from infected individuals. Here we showed profound differences in integration site targeting preferences between *in vitro*-derived and patient-derived integration site datasets. Notably, enrichment of HIV-1 integration in and/or near non-DNA motifs that potentially regulate gene expression, such as G4 motifs, is observed in patient derived data **(Figure 3.3).** Interestingly our data also showed similar integration preference with respect to commonly studied genomic features such as genes, CpG islands, satellite DNA and LADs among subtypes A, B, C and D **(Figure 3.4 A, and B).** However, substantial differences were observed among the different subtypes with respect to non-B DNA motifs **(Figure 3.4 C).** Notably, subtype B is the only subtype that showed enriched integration within specific distances from non-B DNA motifs (between 150-500 bp). It remains unclear why subtype B is the only subtype that showed enriched integration within specific distances from non-B DNA compared to other subtypes. Differences in the integrase structure of each subtype could contribute to this difference in integration. Natural polymorphisms in HIV-1 integrase have been observed among different subtypes, which might affect their integration site selection [20,21]. Additionally specific polymorphisms in HIV-1 integrase have been reported to retarget integration away from gene dense regions, which correlated with increase disease progression and virulence [22]. This further suggest that viral integrase can substantially contribute to disease progression, as it is related to the virus integration site.

5.1.3 Combination antiretroviral therapy (cART) alters HIV-1 integration site selection

cART helps select for integration into silent regions such as intergenic regions that can maintain latency [23]. We further investigated the association between treatment and HIV-1 integration site selection among different HIV-1subtypes. Our observations indicate that cART changed the integration profile for all subtypes with respect to common genomic features and non-B DNA motifs **(Figure 3.5 and 3.6).** Our data showed integration enrichment in heterochromatin rich region such as LAD regions, in treated individuals

compared to untreated individuals of subtype A, B and D infections **(Figure 3.5)**. This further confirms that cART selects for cells harboring integration sites in heterochromatin among different subtypes. Additionally, cART led to enriched integration near G4 motifs, which are also found in heterochromatin rich regions [4].

5.1.4 Non-B DNA motifs influence HIV-1 integration in anatomical reservoirs

Besides peripheral blood, which is the most studied anatomical site for HIV-1 infection, other anatomical sites are also known to harbor latent virus during HIV-1 infection [24,25, 26, 27, 28, 29–31]. In this final study, our integration site distribution analysis across anatomical sites such as in PBMCs/ PBLs, GIT (esophagus, colon, duodenum and stomach) and brain revealed that other anatomical sites showed similar integration level in genes as in the peripheral blood except in the brain which showed drastically reduced integration within genes **(Figure 4.1)**. This difference in integration into genes could be due to the lower expression of host cellular factors known to target integration into transcriptionally active genes such as LEDGF/p75 and CPSF6 the in the brain. In fact, LEDGF/p75 expression is slightly lower in the adult human brain and within specific regions of the brain [32]. Expression levels of CPSF6 is thought to be highly expressed in the brain and the stomach.

Additionally, integration in and/or near non-B DNA motifs (e.g. G4, Z-DNA, cruciform and triplex motifs) that are known to suppress gene expression were strongly targeted in the different anatomical reservoirs **(Figure 4.2)** [33,18,34,19,35–39,17,40]. It is possible that the expression of specific host proteins influence HIV-1 site targeting into or near specific non-B DNA structures within these different reservoirs.

Throughout this study, enriched HIV-1 integration at intervals of 100-149bp away from G4 motifs was consistently observed. The ability of HIV-1 to integrate at that specific distance could be the results of non-B DNA inducing repositioning of nucleosomes, which are comprised of ~147 bp of DNA wrapped around a histone octamer core [41]. In fact, G4 motifs are known to form in nucleosome-free regions in the genome and their ability to dynamically organize flanking nucleosomes may contribute to transcriptional regulation of

integrated proviruses and potentially cause a transcriptional block of proviral gene expression [42].

5.2 Future directions

Latency is a major obstacle to a functional cure. In this work we have shown that G4 motifs are strongly targeted for integration during latent infection, which is important as they are known to impede adjacent gene expression. These motifs were shown to correlate with the failure of latency reversal agents **(Figure 2.2B).** We have also demonstrated that compounds either stabilizing or destabilizing G4 alter HIV-1 integration profile with respect to G4 **(Figure 2.5).** It will be important to further assess whether these compounds (BRACO19 and TMPyP4) can enhance or lower the activity of latency reversal agents and proviral gene expression. This could be tested *in vitro* by using a dual color reporter as described in Chapter 2, section 2.3.2 to establish the percent of latently reactivated cells following treatment with these compounds. Additional, the synergetic effect of these G4 compounds and currently used latency reversal agents could be tested to determine whether this will induce a substantial reactivation of latently infected cells.

To further assess the implication of G4 on gene expression and how this might also affect proviral gene expression, G4 sequence motifs could be cloned either upstream or downstream of a promoter site in a luciferase construct containing HIV-1 LTR where luciferase expression will be determined.

The integrase protein is essential for HIV-1 infection. Another study that could be performed would be to determine the sequence variation among different HIV-1 to determine how specific changes might correlate to the differences seen in their integration profiles that we observed. Experiments involving switching the integrase sequence of HIV-1 and other retroviruses will further help assess the differences in integration profile seen among evolutionary diverse retroviruses.

5.3 Concluding remarks and significance

HIV-1 persistence from latency presents a major barrier for eradication and a functional cure. This is in part due to the slow decay of the latent reservoir (which has an estimated

half-life of 44 months [43,44]). Advances in the development of cART help control HIV-1 replication in infected individuals, but fail to eradicate this latent pool. In fact, it has been calculated that it would take more than 70 years to eradicate HIV-1 under cART treatment. The long-term goal of this study is to characterize and further understand the implication of HIV-1 integration site selection on latency. Our study has provided novel insight into the role of HIV-1 integration site selection and it potential contribution to latency. Specifically, we identified non-B DNA motifs to be a novel factor that could substantially contribute to HIV-1 persistence/latency and integration site selection. With this work, we hope to help inform the design of future experiments in HIV-1 eradication research and in designing better gene therapy vectors based on HIV-1 biology.

5.4 References

1. Taylor, B. *et al.* The challenge of HIV-1 subtype diversity. *N. Engl. J. Med.* **358,** 1590–1602 (2008).

2. Jordan, A., Defechereux, P. & Verdin, E. The site of HIV-1 integration in the human genome determines basal transcriptional activity and response to Tat transactivation. *EMBO J.* **20,** 1726–38 (2001).

3. Battivelli, E. *et al.* Distinct chromatin functional states correlate with HIV latency reactivation in infected primary CD4+ T cells. *Elife* **7,** e34655 (2018).

4. Hoffmann, R. F. *et al.* Guanine quadruplex structures localize to heterochromatin. *Nucleic Acids Res.* **44,** 152–163 (2016).

5. Ojwang, J. O. *et al.* T30177, an oligonucleotide stabilized by an intramolecular guanosine octet, is a potent inhibitor of laboratory strains and clinical isolates of human immunodeficiency virus type 1. *Antimicrob. Agents Chemother.* **39,** 2426–2435 (1995).

6. Rando, R. F. *et al.* Suppression of human immunodeficiency virus type 1 activity in vitro by oligonucleotides which form intramolecular tetrads. *Journal of Biological Chemistry* **270,** 1754–1760 (1995).

7. Mazumder, A. N. *et al.* Inhibition of human immunodeficiency virus type 1 integrase by guanosine quartet structures. *Biochemistry* **35,** 13762–13771 (1996).

8. Jing, N., Rando, R. F., Pommier, Y. & Hogan, M. E. Ion selective Folding of loop domains in a potent anti-HIV oligonucleotide. *Biochemistry* **36,** 12498–12505 (1997).

9. Jing, N. *et al.* Mechanism of inhibition of HIV-1 integrase by G-tetrad-forming oligonucleotides in vitro. *J. Biol. Chem.* **275,** 21460–21467 (2000).

10. De Soultrait, V. R. *et al.* DNA aptamers derived from HIV-1 RNase H inhibitors are strong anti-integrase agents. *J. Mol. Biol.* **324,** 195–203 (2002).

11. Phan, A. T. *et al.* From The Cover: An interlocked dimeric parallel-stranded DNA quadruplex: A potent inhibitor of HIV-1 integrase. *Proc. Natl. Acad. Sci.* **102,** 634–639 (2005).

12. Koizumi, M. *et al.* Biologically active oligodeoxyribonucleotides--IX. Synthesis and anti-HIV-1 activity of hexadeoxyribonucleotides, TGGGAG, bearing 3’- and 5’-end-modification. *Bioorg. Med. Chem.* **5,** 2235–43 (1997).

13. Urata, H., Kumashiro, T., Kawahata, T., Otake, T. & Akagi, M. Anti-HIV-1 activity and mode of action of mirror image oligodeoxynucleotide analogue of zintevir. *Biochem. Biophys. Res. Commun.* **313,** 55–61 (2004).

14. Pedersen, E. B., Nielsen, J. T., Nielsen, C. & Filichev, V. V. Enhanced anti-HIV-1 activity of G-quadruplexes comprising locked nucleic acids and intercalating nucleic acids. *Nucleic Acids Res.* **39,** 2470–2481 (2011).

15. Xing S, S. R. Targeting HIV latency: pharmacologic strategies toward eradication. *Drug Discov. Today* **18,** 541–551 (2013).

16. Cillo, A. R. *et al.* Quantification of HIV-1 latency reversal in resting CD4+ T cells from patients on suppressive antiretroviral therapy. *Proc. Natl. Acad. Sci.* **111,** 7078–7083 (2014).

17. Tornaletti, S., Park-Snyder, S. & Hanawalt, P. C. G4-forming sequences in the non-transcribed DNA strand pose blocks to T7 RNA polymerase and mammalian RNA polymerase II. *J. Biol. Chem.* **283,** 12756–62 (2008).

18. Delic J, Onclercq R, M.-C. M. Inhibition and enhancement of eukaryotic gene expression by potential non-B DNA sequences. *Biochem Biophys Res Commun* **180,** 1273–83 (1991).

19. Verma, A., Yadav, V. K., Basundra, R., Kumar, A. & Chowdhury, S. Evidence of genome-wide G4 DNA-mediated gene expression in human cancer cells. *Nucleic Acids Res.* **37,** 4194–4204 (2009).

20. Myers, R. E. & Pillay, D. Analysis of natural sequence variation and covariation in human immunodeficiency virus type 1 integrase. *J. Virol.* **82,** 9228–9235 (2008).

21. Rhee, S. Y. *et al.* Natural variation of HIV-1 group M integrase: Implications for a new class of antiretroviral inhibitors. *Retrovirology* **5,** 1–11 (2008).

22. Demeulemeester, J. *et al.* HIV-1 integrase variants retarget viral integration and are associated with disease progression in a chronic infection cohort. *Cell Host Microbe* **16,** 651–662 (2014).

23. Cohn, L. B. *et al.* HIV-1 Integration Landscape during Latent and Active Infection. *Cell* **160,** 420–432 (2015).

24. Spudich, S., Lollo, N., Liegler, T., Deeks, S. G. & Price, R. W. Treatment benefit on cerebrospinal fluid HIV-1 levels in the setting of systemic virological suppression and failure. *J. Infect. Dis.* **194,** 1686–1696 (2006).

25. Kumar, A. M., Borodowsky, I., Fernandez, B., Gonzalez, L. & Kumar, M. Human immunodeficiency virus type 1 RNA levels in different regions of human brain: Quantification using real-time reverse transcriptase-polymerase chain reaction. *J. Neurovirol.* **13,** 210–224 (2007).

26. Chun, T. *et al.* Persistence of HIV in Gut-associated lymphoid tissue despite long-

term antiretroviral therapy. *J. Infect. Dis.* **197,** 714–720 (2008).

27. Nuovo, G. J. *et al.* Spermatogonia and their progeny a study by Polymerase chain reaction in situ hybridization. *Am. J. Pathol.* **144,** 1142–1148 (1994).

28. Kovacs, A. *et al.* Determinants of HIV-1 shedding in the genital tract of women. *Lancet* **358,** 1593–1601 (2001).

29. North, T. W. *et al.* Viral Sanctuaries during highly active antiretroviral therapy in a nonhuman primate model for AIDS. *J. Virol.* **84,** 2913–2922 (2010).

30. Günthard, H. F. *et al.* Residual human immunodeficiency virus (HIV) type 1 RNA and DNA in lymph nodes and HIV RNA in genital secretions and in cerebrospinal fluid after suppression of viremia for 2 Years. *J. Infect. Dis.* **183,** 1318–1327 (2001).

31. Wong, J. K. *et al.* Reduction of HIV-1 in blood and lymph nodes following potent antiretroviral therapy and the virologic correlates of treatment failure. *Proc. Natl. Acad. Sci.* **94,** 12574–12579 (1997).

32. Chylack, L. T. *et al.* Lens epithelium-derived growth factor (LEDGF/p75) expression in fetal and adult human brain. *Exp. Eye Res.* **79,** 941–948 (2004).

33. Bochman, M. L., Paeschke, K. & Zakian, V. a. DNA secondary structures: stability and function of G-quadruplex structures. *Nat. Rev. Genet.* **13,** 770–80 (2012).

34. Siddiqui-Jain, A., Grand, C. L., Bearss, D. J. & Hurley, L. H. Direct evidence for a G-quadruplex in a promoter region and its targeting with a small molecule to repress c-MYC transcription. *Proc. Natl. Acad. Sci. U. S. A.* **99,** 11593–8 (2002).

35. Waga, S., Mizuno, S. & Yoshida, M. Chromosomal protein HMG1 removes the transcriptional block caused by the cruciform in supercoiled DNA. *J. Biol. Chem.* **265,** 19424–8 (1990).

36. Waga, S., Mizuno, S. & Yoshida, M. Nonhistone protein HMG1 removes the transcriptional block caused by left-handed Z-form segment in a supercoiled DNA.

Biochem. Biophys. Res. Commun. **153,** 334–9 (1988).

37. Jain, A., Magistri, M., Napoli, S., Carbone, G. M. & Catapano, C. V. Mechanisms of triplex DNA-mediated inhibition of transcription initiation in cells. *Biochimie* **92,** 317–320 (2010).

38. Maher, L. J., Dervan, P. B. & Wold, B. Analysis of promoter-specific repression by triple-helical DNA complexes in a eukaryotic cell-free transcription system. *Biochemistry* **31,** 70–81 (1992).

39. Brázda, V., Laister, R. C., Jagelská, E. B. & Arrowsmith, C. Cruciform structures are a common DNA feature important for regulating biological processes. *BMC Mol. Biol.* **12,** 33 (2011).

40. Belotserkovskii, B. P. *et al.* A triplex-forming sequence from the human c-MYC promoter interferes with DNA transcription. *J. Biol. Chem.* **282,** 32433–32441 (2007).

41. Timothy J. Richmond & Curt A. Davey. The structure of DNA in the nucleosome core. *Nature* **423,** 145–150 (2003).

42. Wong, H. M. & Huppert, J. L. Stable G-quadruplexes are found outside nucleosome-bound regions. *Mol. Biosyst.* **5,** 1713–1719 (2009).

43. Siliciano, J. D. *et al.* Long-term follow-up studies confirm the stability of the latent reservoir for HIV-1 in resting CD4+ T cells. *Nat. Med.* **9,** 727–8 (2003).

44. Diana Finzi, Joel Blankson, Janet D. Siliciano, Joseph B. Margolick, Karen Chadwick, Theodore Pierson, Kendall Smith, Julianna Lisziewicz, Franco Lori, Charles Flexner, Thomas C. Quinn, Richard E. Chaisson, Eric Rosenberg, Bruce Walker, Stephen Gange, J. G. & R. F. S. Latent infection of CD4 $^{+}$ T cells provides a mechanism for lifelong persistence of HIV-1 , even in patients on effective com bination therapy. *Nat. Med.* **5,** 512–517 (1999).

Curriculum Vitae

Name: **Hinissan Pascaline Kohio**

Education

Ph.D. candidate in Microbiology and Immunology **September 2014 - Present**
Western University, London, Ontario (Canada)
Concentration: Molecular Virology

Master of Science in Biology **August 2012**
The University of North Carolina at Greensboro, Greensboro, North Carolina (USA)
Thesis title: "Glycolytic control of Vacuolar ATPase Pump activity: A mechanism to regulate influenza viral infection"

Bachelor of Science in Biology **May 2009**
Armstrong Atlantic State University, Savannah, Georgia (USA)
Institutional Honors: *Cum Laude*

Work Experience

Teaching Assistant **January 2016 - April 2018**
MICROIMM 3620G (Immunology Lab)
Western University, London, Ontario (Canada)

Graduate Teaching Assistant **May 2012 - June 2012**
BIO 271 Lab (Human Anatomy Lab)
The University of North Carolina at Greensboro, Greensboro, North Carolina (USA)

Graduate Teaching Assistant **January 2011 - April 2012**
BIO 105 (Major Concepts of Biology Lecture)
The University of North Carolina at Greensboro, Greensboro, North Carolina (USA)

Graduate Teaching Assistant **August 2010 - April 2012**
BIO 105 Lab (Major Concepts of Biology Lab)
The University of North Carolina at Greensboro, Greensboro, North Carolina (USA)

Awards and recognitions

Western University, London, Ontario (Canada)
PSAC Local 610 Scholarship for Community Involvement **September 2018**
Dr. Frederick W. Luney Graduate Travel Award in M&I **July 2018**
Canadian Bioinformatics Workshop Registration Award **May 2018**
CIHR National Student Research Poster Competition invitation **June 3rd 2015**
London Health Research Day, Best Poster Presentation Award **April 1st 2015**
Graduate Teaching Assistantship **January 2015 - April 2018**

Western graduate Research Scholarship **September 2014 - August 2018**

Publications

Ermela Paparisto, Matthew Woods, Macon Coleman, Seyed Moghadasi, Divjyot Kochar, Sean Tom, **Hinissan Kohio**, Richard Gibson, Taryn Rohringer, Nina Hunt, Eric Di Gravio, Jonathan Zhang, Meijuan Tian, Yong Gao, Eric Arts, and Stephen Barr. "Evolution-guided structural and functional analyses of the HERC family reveals an ancient marine origin and determinants of antiviral activity". *Journal of Virology* 2018; 13; 92(13), e00528-18

Volk-Draper L, Hall K, Griggs C, Rajput S, **Kohio P**, DeNardo D and Ran S. "Paclitaxel therapy promotes breast cancer metastasis in a TLR4-dependent manner". *Cancer Research* 2014; 74: 5421-5434.

Hinissan P. Kohio and Amy L. Adamson. "Glycolytic control of vacuolar-type ATPase activity: A mechanism to regulate influenza viral Infection". *Virology* 2013; 444:301-309.

Selected oral and poster presentations

Hinissan P. Kohio. Hannah O. Ajoge, Sean K. Tom, Macon Coleman and Stephen D. Barr. Novel insights into the genomic integration site landscape of HIV, *American Society for Virology* at The University of Maryland at College Park, College Park, Maryland (USA), [July 17, 2018]. **Oral presentation**

Pascaline Kohio. "Human Immunodeficiency Virus (HIV) integration and latency: a hunt for a cure", *Western University 3 Minutes Thesis Competition,* Western University, London ON (Canada), [April 05 2017]. **Oral presentation.**

Hinissan P. Kohio. Hannah O. Ajoge, Sean K. Tom, Macon Coleman and Stephen D. Barr. "Host genomic non-B DNA structures significantly influence integration site selection of latent HIV-1 and potently inhibit gene expression: implications for cure-focused antiretrovirals", *American Society for Virology*, Western University London ON (Canada), [July 11, 2015]. **Oral presentation**

Hinissan P. Kohio, Hannah O. Ajoge, Sean K. Tom, Macon Coleman and Stephen D. Barr. "Genomic Non-B DNA motifs significantly influence integration site selection of latent HIV-1 and potently inhibit gene expression: implications for cure-focused antiretroviral". *Canadian Student Health Research Forum*, the University of Manitoba, Winnipeg MB (Canada), [June 03, 2015]**. Poster presentation**

Hinissan P. Kohio, Hannah O. Ajoge, Sean K. Tom, Macon Coleman and Stephen D. Barr. "Host genomic non-B DNA structures significantly Influence Integration Site Selection of Latent HIV-1 and potently inhibit gene expression: implications for cure-focused antiretrovirals", London Health Research Day, London ON (Canada), [April 01, 2015]. Poster presentation

Hinissan P. Kohio and Amy Adamson. *"*A novel mechanism to regulate influenza viral infection. Poster presentation", *Graduate Research and Creativity Expo Presentation,* Department of Biology at the University of North Carolina at Greensboro, Greensboro NC (USA), [April 03, 2012]. **Poster presentation**