Unique Molecular Properties of HIV-1C Reverse Transcriptase Conferring a Possible Replication Advantage

A thesis submitted in partial fulfilment
of the requirements for the degree of

Doctor of Philosophy

By

Arun Panchapakesan

JNCASR

HIV-AIDS Laboratory
Molecular Biology and Genetics Unit
Jawaharlal Nehru Centre for Advanced Scientific Research
Bangalore - 560064, Karnataka, India

April 2022

To my family - Pati, Shyamala Athai, Appa, Amma and
Shravan, the giants on whose shoulders I stood,
to be able to see this far.

To
those
illiterate millions
of
my country,
at
whose cost
I have had
this
rarest opportunity

(Udaykumar Ranga, PhD Thesis, 1990)

To the late Dr. Francis Crick, who has inspired and
influenced my thinking more than any other scientist
in history

DECLARATION

I hereby declare that the thesis entitled **‘Unique Molecular Properties of HIV-1C Reverse Transcriptase Conferring a possible Replication Advantage**’ is the result of research work carried out by myself under the supervision of Prof. Ranga Udaykumar at the HIV-AIDS Laboratory, Molecular Biology and Genetics Unit, Jawaharlal Nehru Centre for Advanced Scientific Research, Bangalore, India.

In keeping with the general practice in reporting the scientific observations, due acknowledgements have been made wherever the work described is based on the findings of other investigators. Any oversight due to an error of judgment is inadvertent and highly regretted.

Date: 11th April 2022
Place: Bangalore

Arun Panchapakesan

Jawaharlal Nehru Center for Advanced Scientific Research
(Autonomous Body under the Department of Science & Technology, Government of India)
Jakkur Campus, Jakkur Post
Bengaluru 560 064, INDIA
Office Tel: +91 (80) 2208 2830
Office Fax: +91 (80) 2208 2766
Email: udaykumar@jncasr.ac.in

Udaykumar Ranga, Ph.D.
Professor
HIV-AIDS Laboratory
Molecular Biology and Genetics Unit
Jawaharlal Nehru Centre for Advanced Scientific Research

CERTIFICATE

I hereby certify that the work described in the thesis entitled **'Unique Molecular Properties of HIV-1C Reverse Transcriptase Conferring a possible Replication Advantage'** has been carried out by **Mr. Arun Panchapakesan** at the HIV-AIDS Laboratory, Molecular Biology and Genetics Unit, Jawaharlal Nehru Centre for Advanced Scientific Research, Bangalore, India under my supervision.

This work has not been submitted in part or full, to any other institute or university for the award of any other degree or diploma.

R. Udaykumar

Place: Bangalore
Date: 11th April, 2022

Prof. Udaykumar Ranga

Acknowledgements

Through the course of my work, I’ve been at the receiving end of a great deal of help, love and support from a good many number of people. I take up this space to express my gratitude to each of them. However, my words here do not do justice to what each of these individuals has done for me.

I begin by thanking my supervisor, Prof. Udaykumar Ranga, for being so much more than a PhD thesis advisor to me. In my career, he has essayed the role of a mentor, a critic, a father figure, and most importantly, a friend. His constant support and encouragement have been the driving force behind my work. His ability to bring out the best in people is the prime reason for some of the deeper insights presented in the later sections of this work. Every page of this thesis speaks, in subtext, of the dedication and commitment he has shown towards his work over the last two decades.

I want to express my gratitude to Dr. Udupi Ramagopal and Prof. Vinayaka Prasad for their valuable insights and time to discuss my work. Their inputs were invaluable to me.

The models for sequence duplication would not have been possible without Chhavi’s prowess in digital animation, and I am grateful to her for her help.

I thank the faculty of the Molecular Biology and Genetics Unit, Prof. M R S Rao, Prof. Namita Surolia, Prof. Hemalatha Balaram, Prof. Anuranjan Anand, Prof. Tapas Kundu, Prof. Maneesha Inamdar, Prof. Kaustuv Sanyal, Prof. Ravi Manjithaya and Dr. Kushagra Bansal, as well as the faculty of the Neurosciences Unit Prof. James Chelliah and Prof. Sheeba Vasu for the useful discussions, and the feedback that they constantly provided during my work.

I shall remain eternally indebted to Lakkappa, Sridhar, Lavanya, Swathi and Bhuvana for making my life easy in the laboratory. Their assistance to our work is often overlooked and yet, remains at the core of every PhD thesis from the laboratory.

I thank the people who worked with me on this project – Haider, Neelam, Afzal and Kavita, without whom it would have been impossible to do some of the experiments. Each of them poured their heart out into this work in their own way, and I feel blessed to have been able to work with them.

My lab mates - Dr. Anjali, Dr. Shilpee, Dr. Malini, Dr. Prabhu, Dr. Sutanuka, Deepak, Disha, Sreshtha, Haider, Brahmaiah, Neha, Kavita, Yuvraj, Swati, Chhavi, Anish, Afzal, Vijeta, Jyotsna, Harshit, Swarnima, Jayendra, Hrimkar and Shobith have ensured that the laboratory remained a vibrant and fun place to be. Their inputs to my work during the laboratory meetings have been vital to its completion.

I would also like to thank Deepak, Disha, Neha, Brahmaiah, Kavita, Swati, Yuvraj, Afzal, Neelam, Chhavi, Jyotsna, Swarnima, Jayendra and Anish for another reason altogether. They have been so much more to me than merely valuable colleagues; they are invaluable friends.

I am grateful to Dr. Vijay, Dr. Amrutha, Dr. Santosh, Dr. Sanjeev, Dr. Lakshmeesha, Arpitha, Asutosh, for always treating me as a member of their own laboratories and the helpful discussions that arose as a result.

Mrs. Rama Vidyarthi, Dr. Narendra Nala and Dr. Thirunavakarasu Dharmalingam have been constant sources of guidance, and have been my well-wishers throughout. I am thankful to them for their support.

I want to thank Prof. C N R Rao and Dr. Indumati Rao for allowing me to teach at the C N R Rao Foundation, which was a truly enlightening experience.

Words fail me as I think of what Anirudha, Neelakshi, Srikant, Naveen and Chitral mean to me. They were with me at every step of this incredible journey. They have been constant sources of encouragement to me throughout, and I am incredibly fortunate to call them my closest friends.

Table of Contents

List of Abbreviations

AIDS	–	Acquired Immunodeficiency Syndrome
ATP	–	Adeosine Tri-Phosphate
cART	–	Combinatorial Anti-Retroviral Therapy
cDNA	–	complementary DNA
CRF	–	Circulating Recombinant Form
dATP	–	deoxy-Adenosine Tri-Phosphate
DEAE	–	Diethylaminoethyl
DMEM	–	Dulbecco’s Modified Eagle Medium
DNA	–	Deoxy-ribonucleic Acid
dNTP	–	deoxy-Nucleotide Tri-Phosphate
DTT	–	Dithiothreitol
EDTA	–	Ethylenediaminetetraacetic Acid
EGFP	–	Enhanced Green Fluorescent Protein
ELISA	–	Enzyme-Linked Immunosorbent Assay
FBS	–	Fetal Bovine Serum
HIV	–	Human Immunodeficiency Virus
HTLV	–	Human T-Lymphotropic Viruses
IL-2	–	Interleukin-2
LANL	–	Los Alamos National Laboratory
LTR	–	Long Terminal Repeat
MHD	–	Micro-Homology Domain
MLV	–	Murine Leukemia Virus
MR-ME	–	MHD-based Recombination and Matched-base Extension
MR-MME	–	MHD-based Recombination and Mis-Matched-base Extension
NIAID	–	National Institute of Allergy and Infectious Diseases
NIH	–	National Institutes of Health
Ni-NTA	–	Nickel-Nitriloacetic acid
NR-ME	–	Non-MHD-based recombination and Matched-base Extension
NR-MME	–	Non-MHD-based Recombination and Mis-Matched-base Extension
PBMCs	–	Peripheral Blood Mononuclear Cells
PBS	–	Primer Binding Site
PCR	–	Polymerase Chain Reaction
PDB	–	Protein Data Bank
PHA-P	–	Phytohaemagglutinin-P
PMA	–	Phorbol 12-Myristate 13-Acetate
PPT	–	Polypurine Tract
RNA	–	Ribonucleic Acid
RPMI	–	Roswell Park Memorial Institute
RSV	–	Rous Sarcoma Virus
RT	–	Reverse Transcriptase
SIV	–	Simian Immunodeficiency Virus
TFBS	–	Transcription Factor Binding Sites
TNF	–	Tumour Necrosis Factor
UNAIDS	–	The Joint United Nations Programme on HIV/AIDS
URF	–	Unique Recombinant Form
VESPA	–	Viral Epidemiology Signature Pattern Analysis
WHO	–	World Health Organization

Synopsis of the Thesis

Introduction: The Human Immunodeficiency Virus (HIV) is classified into HIV-1 and HIV-2, based on the zoonotic origin of the viruses. HIV-1 can be further divided into four groups, M, N, O, and P, to represent viral strains from four independent events of cross-species transmission(Sharp & Hahn, 2001, 2010). Group M, responsible for the global pandemic, is further divided into ten primary genetic subtypes labelled A – D, F – H, J – L, and many recombinant forms of the primary subtypes. The global distribution of the subtypes is non-uniform, with HIV-1C accounting for approximately half of the HIV-1 infections of the world and nearly all of those in India (Hemelaar et al., 2019).

HIV-1C possesses many unique molecular characteristics, including its ability to duplicate sequences at a high frequency. Previous work from our laboratory demonstrated the ability of HIV-1C to duplicate the sequence motifs of biological significance in two distinct regions of the viral genome – the Long Terminal Repeat (LTR) and the p6 protein in Gag (Bachu, Mukthey, et al., 2012; Sharma et al., 2017). Both these sequence duplication events appear to be associated with a replication advantage to the virus (Bachu, Yalla, et al., 2012; Sharma et al., 2018).

The LTR duplications consist of repeats of transcription factor binding sites, such as NF-κB, LEF-1, and/or RBE-III, that are crucial for regulating viral latency. Experimental evidence suggests that the duplication of the NF-κB site enhances the transcriptional strength of the LTR, while that of the RBE-III site hinders the reactivation of latent HIV-1. In contrast, a co-duplication of both motifs appears to enhance the transcriptional strength when activated and stabilise viral latency when the host cell is not activated (Bhange D et al., manuscript in preparation). Thus, when associated with the NF-κB site duplication, the RBE-III site duplication appears to provide the balance necessary to maintain the viral latency of a stronger viral promoter.

The p6 protein in Gag, the second hotspot for sequence duplication, is responsible for viral budding. Here, amino acid sequences of varying lengths, ranging from 3-14 aa, are duplicated. The core region of these duplications is the four amino acid PTAP motif that recruits the host factor TSG-101 to bud out of the cell. Previous work from our laboratory demonstrated the dominance of dual-PTAP motif variants over single-PTAP strains in natural infection and several experimental conditions (Martins et al., 2015; Sharma et al., 2018).

A detailed bioinformatic analysis of HIV-1 sequences downloaded from the LANL database revealed that approximately 17% of all deposited sequences contain duplications in these two regions – the LTR and Gag. Notably, the frequency of sequence duplications in the two regions is significantly higher in HIV-1C than other HIV-1 subtypes. For example, 14.1% and 19.1% of HIV-1B and HIV-1C sequences, respectively, harbour sequence duplications in the LTR. The difference is more pronounced in the PTAP region of p6, where 6.4% and 29.0% of HIV-1B and HIV-1C sequences contain duplications, respectively.

These sequence duplications are a result of the micro-homology-mediated recombination by the reverse transcriptase of HIV-1. Micro-homology-mediated recombination, previously known as non-homologous recombination, occurs at a frequency several orders of magnitude lower than that of homologous recombination. The Micro-homology domain-mediated recombination leads to the generation of sequence duplications and deletions that are then subjected to natural selection (Onafuwa-Nuga & Telesnitsky, 2009).

The significantly higher frequency of sequence motif duplications observed in HIV-1C alludes to subtype-associated differences in the function of the Reverse Transcriptase at the molecular level.

The present thesis attempts to understand subtype-specific differences in the functioning of RT and the associated cis and trans elements underlying the observed high-frequency of sequence motif duplication in HIV-1C.

Chapter – 1 introduces the reader to the Human Immunodeficiency Virus, its classification, molecular biology, and life cycle. The chapter also introduces HIV reverse transcriptase, the phenomenon of high-frequency sequence duplications in HIV-1C, and outlines the mechanism of HIV reverse transcription. The chapter concludes with the presentation of various models of genetic recombination and the process of sequence duplication, and the impact they may have on viral evolution.

Chapter – 2 reviews the various viral factors that may influence HIV-1 recombination. The chapter also depicts the publications that have compared the functioning of HIV-1B and C RTs, highlighting the inherent limitations associated with such experimental models, including the sporadic occurrence of such molecular events, and concludes with the presentation of the rationale underlying this study.

Chapter – 3 describes the experimental strategies used for the present analysis, explained here in brief,

- **Protein Expression and Purification:** The pNL4-3 and pIndie molecular clones were used as representatives for HIV-1B and C, respectively, through the experimental work. The subunits (p66 and p51) of all variant RTs were cloned and expressed independently in *E. coli* M15 cells. Cell lysates of appropriate subunits were pooled, and the assembled RT was purified using Ni-NTA affinity chromatography. Following a second round of purification by ion-exchange chromatography, the purified proteins were quantified by densitometry and stored in aliquots at -20°C until use.

- **Polymerase Stall Assay:** Homologous sequences of 120 bp representing the NF-κB and RBE-III duplication regions from the HIV-1B and C LTRs were generated using *in vitro* transcription. The two RNA molecules were used as templates for extending a 5' radio-labelled oligonucleotide primer using HIV-1B or C RT. The extension products were resolved on a 10% polyacrylamide-urea denaturing gel and visualised by autoradiography.

- **Sequence Analysis:** HIV-1 Reverse Transcriptase (p51) and p6-Gag sequences were downloaded from the HIV-1 Los Alamos National Laboratory Sequence database and aligned using the Clustal Omega software. The sequences were then analysed using the BioEdit software.

- **Structural Analysis of RT:** A high-resolution structure of HIV-1B RT in complex with DNA was downloaded from the Protein Data Bank (PDB ID – 5J2M) and analysed using the PyMOL software.

- **Specific Activity Measurement:** A homopolymeric polyU template was annealed with oligodA and, the incorporation of ^{32}P labelled dATP by each RT at varying concentrations was measured using liquid scintillation spectrometry, and the specific activity was calculated.

- **cDNA Synthesis Rate:** The cDNA synthesis rate was estimated at 5-minute intervals by measuring the incorporation of ^{32}P labelled dATP into a polyU template by various RTs at a constant dATP concentration of 250 µM.

- **Recombination Assay:** The recombination frequency of the RT was quantitated by flow cytometry using complementation between two EGFP coding RNA molecules harbouring debilitating frameshift mutations at two different locations (positions 4 and 204) in the protein. A recombination event occurring between these two mutations would generate functional GFP, the frequency of which must reflect the functional activity of the RT variant. We tested a panel of eight RT variants using this assay.

- **Strand Transfer Assay:** The ability of RT variants to switch strands and initiate acceptor strand synthesis was estimated using the strand transfer assay. Briefly, two different sense-strand template RNA molecules containing a 20 bp overlap were generated by *in vitro* transcription. A pair of primers were designed such that each primer annealed to one of the two template RNAs. Productive PCR would be possible only when the nascent cDNA molecule switches between the two strands - as a strand-switch mediated by the function of RT. The Ct value of the real-time PCR was used to compare the strand-switch efficiency mediated by the individual members of a panel of eight variant

RT proteins. For the purpose of normalisation, a different primer pair was also designed to amplify only the larger RNA template without the need for a template switch.

- **3'-terminus mismatch Extension Assay:** The propensity of HIV-1C RT variants to extend a mismatch at the 3'-end of a template DNA strand was optimised. A panel of 12 primers, each representing one of the 12 possible nucleotide mismatches, was designed. The incorporation of ^{32}P labelled dATP was compared using liquid scintillation spectrometry.

- **Infectivity Assay:** A panel of eight infectious viral strains harbouring RT mutations was constructed and used to infect the reporter TZM-bl cell line. The production of luciferase was compared at 48 h of infection.

- **Replication Kinetics:** Healthy donor PBMC were infected with the individual members of the infectious RT-variant viral strains of the panel, and the viral p24 protein secretion was measured from the spent medium at several time points post-infection.

Chapter – 4 presents experimental data, as summarised below:

- **The subtype identity of the template, not RT, determines polymerase stalling:** The extension of a 5' radio-labelled oligonucleotide primer on analogous HIV-1B and C RNA templates sequences using purified RTs from both subtypes identified several locations where both enzymes stalled during polymerisation. The stalling pattern is nearly comparable with minor differences for a specific template regardless of the identity of the polymerase. Thus, as previously reported by others, the polymerisation and the stalling of the polymerase activity appear to be majorly determined by the nature of the template than that of the polymerase.

- **The G359T substitution may create an additional hydrogen bond:** A comparative bioinformatic analysis of HIV-1 p51 RT sequences downloaded from the Los Alamos sequence database identified six amino acid residues unique to HIV-1C RT, - E39, T48, A173, T359, R530, and S534. Multiple crystal structures of HIV-1B RT are available in the Protein Data Bank, but that of HIV-1C has not been solved yet. We, therefore, used the PyMol software to simulate the effect of changing one amino acid residue at one time of the six identified signature residues of HIV-1C RT, using the structure of HIV-1B RT. Of the six signature residues, the location of T359 and the possibility of forming an additional hydrogen bond between the nascent negative-strand DNA and the threonine residue deserved more attention. The bond length of this hydrogen bond obtained after energy minimisation analysis was estimated to be 2.4 Å.

  The significance of the T359 residue to RT function became apparent when the specific activities of purified RT proteins were compared by measuring the incorporation of ^{32}P labelled dATP using a homo-polymeric template. While the activities of wild type RTs of HIV-1B and C were comparable, a 6.5-fold increase in specific activity was observed when the T359 residue was mutated to a Glycine in C-RT. A corresponding reduction in activity was not observed when the Glycine residue of HIV-1B RT was changed to a threonine, with the decrease being minimal. The presence of a serine at position 359, as is naturally seen in some C-RT sequences, appeared to marginally increase the activity of both the RTs, while the substitution of Alanine at this location demonstrated little to no effect on the activity of the RTs.

  Two panels of four infectious full-length viral strains representing each subtype - containing Glycine, Threonine, Serine, or Alanine at position 359 were constructed. The replication kinetics of each viral strain in PBMC was determined by monitoring the concentration of the p24 protein in the culture medium at various time points using ELISA. As expected, the wild type strain of the respective viral subtype (Glycine and Threonine representing subtypes B and C, respectively) outperformed the other members of the corresponding panel in p24 production.

- **HIV-1C RT displays superior ability for acceptor-strand synthesis:** To assess the ability of variant RTs to forcibly switch strands during cDNA synthesis, we used two RNA template molecules representing subtype B and C and the two panels of RTs described above. In the strand transfer assay, the wild type HIV-1C RT demonstrated a nearly 3-fold increase in the ability to initiate acceptor strand synthesis compared to the wild type HIV-1B RT. Both glycine and serine mutations reduced the strand-transfer ability of C-RT, while Alanine conferred no change. Interestingly, all HIV-1C RT variants displayed a superior magnitude of acceptor-strand synthesis compared to the corresponding HIV-1B RT variants. However, in a cell-culture-based recombination assay, these profiles and differences were not consistent. The recombination assay estimates frequencies of restoration of EGFP function by a strand transfer event occurring between two debilitating frameshift mutations in EGFP. In this assay, the wild type RT variants of both subtypes performed comparably, and the presence of Glycine/Threonine, Serine, or Alanine at position 359 reduced the recombination frequency significantly.

- **C-RT can extend a 3' mismatched terminus residue more efficiently:** An analysis of nucleotide sequences containing a PTAP motif duplication from the extant database representing HIV-1B and 1C identified a significantly higher duplication frequency when the terminal base pair is a correct match. Based on this observation, we could successfully explain why the length of PTAP motif duplication was variable between subtypes B and C and even within the same subtype. We designed a panel of 4 oligo-nucleotide primers annealing to a specific target sequence on an RNA template to test this model experimentally. Only one of the four primers will contain a correct match for the template at the 3' end while the other three a mismatch. We designed four different primer sets priming on four different locations on the template, each containing one of the four bases. The two RT viral panels representing subtype B and C were used in the assay to compare the ability of each RT variant to incorporate ^{32}P labelled dATP during polymerisation. C-RT appeared to be superior to B-RT in extending nine of the twelve mismatches tested, of which four combinations displayed a statistically significant difference. Glycine substitution in HIV-1C RT reduced the mismatch extension ability of nearly half the mismatched base pairs, and a corresponding increase was observed for the Threonine variant of HIV-1B RT. Interestingly, the Alanine variant of both RTs appeared to perform moderately in the 3'- mismatch extension assay.

Chapter 5 discusses the results presented in chapter 4, interprets the data in the light of the literature, and offers an experimental model for the differential activity of HIV-1C RT in the context of high-frequency sequence duplications in HIV-1C.

Recombination is an essential evolutionary strategy common to nearly all life forms, particularly viruses with an RNA genome. The generation of diverse viral quasi-species by the promiscuous nature of viral polymerases is an indispensable part of the evolution of these viruses (Domingo & Perales, 2019). The pseudo-diploid nature of lentiviruses and the inherent dependence on programmed strand transfers during replication ensure that recombination rates of lentiviral reverse transcriptases are typically many folds higher than those of other RNA viruses (Onafuwa-Nuga & Telesnitsky, 2009; Onafuwa et al., 2003). Even among lentiviruses, the recombination rates of HIV-1 are significantly higher, at least 10-fold according to one estimate, as compared to other viruses such as Murine Leukemia Virus or Spleen Necrosis Virus (Onafuwa et al., 2003). While this extraordinarily high frequency of recombination in HIV-1 has been primarily attributed to the differences in how HIV-1 packages its genomic RNA, the same cannot be said for differences in recombination rates that appear to exist within different genetic subtypes of HIV-1. In HIV-1C, the frequency of sequence duplication is significantly higher than that in other HIV-1 subtypes. Since recombination is a prerequisite to sequence duplication, it is reasonable to hypothesise that HIV-1C recombines at rates higher than other subtypes. However, previous work from other groups (Chin et al., 2005; Galli et al., 2010) and our experimental results confirm that the recombination rates of different HIV-1 subtypes are comparable, in direct contrast to our conjectures.

Of note, data presented here indicate that the difference between HIV-1C and other subtypes exists not in the rate of recombination itself but in initiating acceptor strand synthesis after RT has dissociated from the RNA template. To the best of our knowledge, previous publications did not directly compare the ability of RTs of different subtypes to switch between strands. The cell culture-based EGFP complementation assay measures the frequency of homologous recombination, where the subtle differences between the RTs of B and C subtypes to initiate acceptor strand synthesis may not be conspicuous enough to warrant further study. On the other hand, sequence duplications result from micro-homologous recombination, a process estimated to occur at frequencies between 10 and 100-fold lower than that of homologous recombination (Zhang & Temin, 1993). The strand transfer assay, while not a perfect indicator, provides a closer estimate of the ability of different RTs to perform micro-homologous recombination. Since the assay measures the frequency of acceptor strand synthesis after a forced transfer, it also mimics the conditions of a sequence duplication more closely. The relevance of a superior acceptor strand synthesis ability to micro-homologous recombination is exceptionally high since it is the initiation of polymerisation on a mispaired template that is critical for a successful sequence duplication.

Our work has bridged an important gap in the field of HIV-1C molecular biology. Multiple publications speculated on the crucial role HIV-1C RT may play in the high-frequency of sequence duplications observed in this subtype (Bachu, Yalla, et al., 2012; Martins et al., 2011; Sharma et al., 2018). Our work has evaluated these differences experimentally and reduced the myriad possibilities to two superior abilities of HIV-1C RT – acceptor strand synthesis and mismatch extension, both of which appear to be enhanced in this subtype, in part, due to the T359 residue. Importantly, we have also established an association between HIV-1C template polymorphism and the high frequency of PTAP motif duplication. Collectively, these observations enable us to explain why HIV-1C can duplicate sequences more efficiently than the other subtypes of HIV-1. Our model alludes to the formation of an additional hydrogen bond between the nascent DNA and the T359 residue, given the proximity between them. An additional hydrogen bond may stabilise the RT-Template-DNA complex formation during micro-homologous recombination, thereby enabling extensions of mismatches and ensuring superior acceptor strand synthesis, both of which lead to a higher magnitude of sequence duplications. The absence of a stabilising hydrogen bond here in the other subtypes of HIV-1 may make sequence motif duplications sporadic events in these subtypes.

Efforts are currently underway to experimentally determine the presence of a hydrogen bond due to T359. A major technical limitation is the non-availability of a crystal structure of HIV-1C RT. Despite this technical difficulty, two experimental observations support the possibility of a hydrogen bond formation by T359. First, a substantial increase in RT activity was observed when T359 residue in HIV-1C RT was substituted with Glycine that lacks a hydrogen bond-forming -OH group. Further, substitution with Serine, which is predicted to form a weak hydrogen bond, results only in a marginal increase in activity, strongly alluding to the formation of a hydrogen bond by Threonine. Second, given that the sequence identity between the RTs of HIV-1B and 1C ranges between 88 – 93%, the probability of the PyMol software to correctly predict the formation of a hydrogen bond at location 359 is significantly high.

The foundation of the present work lies in the identification of two hotspots of sequence-motif duplication in the genome of HIV-1C – p6-Gag and the LTR. Numerous previous publications reported sequence duplications or deletions in other regions spanning the viral genome of HIV-1 subtypes. The fixation of a molecular and biological phenotype of a virus depends on two crucial events – the generation of a variation and its selection. The p6-Gag and the LTR hotspots in HIV-1C appear to be subjected to strong positive selection, as sequence-motif duplications in these regions of the viral genome confer a significant replication fitness advantage on the variant viral strains in natural infection and experimental conditions (Bachu, Mukthey, et al., 2012; Bachu, Yalla, et al., 2012; Sharma et al., 2017, 2018). Thus, despite the natural propensity of the RT to create sequence duplications in various other regions of the viral genome, probably uniformly, these events may not have the same selection advantage as these two hotspots enjoy in HIV-1C. As a result, these two molecular events are selected at a significantly higher frequency in the population once they are generated. Therefore, while a higher propensity to create duplications will also be associated with a concomitant increase in the creation of defective viral variants, the shallow frequency of non-homologous recombination ensures that these defective viral strains form a tiny fraction of the reverse-transcribed viral genomes.

Our findings have significant implications for HIV-1 disease therapeutics and management. As mandated by the test and treat policy (WHO, 2016), ART initiation following a positive diagnosis may lead to the emergence of double-PTAP variant strains of HIV-1C and possibly of other HIV-1 subtypes. Several publications reported a significant association between ART initiation and PTAP duplication (Martins et al., 2011, 2015; Peters et al., 2001). While the molecular mechanisms underlying the sequence duplication following ART initiation remain enigmatic, preliminary data from our laboratory suggest the possibility of PTAP motif duplication playing a compensatory role in restoring replication fitness after a drug-resistance mutation has been selected. Further, the RBE-III motif duplication in the LTR appears to resist latency reactivation of the variant viral strains. Unpublished work from our laboratory demonstrates that the known latency-reversing agents fail to activate the dual RBE-III, but not canonical, viral strains (Bhange D et al., manuscript in preparation). Therefore, the ability of HIV-1C to duplicate sequences at a higher frequency is of major concern considering that HIV-1C is responsible for nearly half of the infections globally.

References:

Bachu, M., Mukthey, A. B., Murali, R. V., Cheedarla, N., Mahadevan, A., Shankar, S. K., Satish, K. S., Kundu, T. K., & Ranga, U. (2012). Sequence Insertions in the HIV Type 1 Subtype C Viral Promoter Predominantly Generate an Additional NF-κB Binding Site. AIDS Research and Human Retroviruses, 28(10), 1362–1368. https://doi.org/10.1089/aid.2011.0388

Bachu, M., Yalla, S., Asokan, M., Verma, A., Neogi, U., Sharma, S., Murali, R. V., Mukthey, A. B., Bhatt, R., Chatterjee, S., Rajan, R. E., Cheedarla, N., Yadavalli, V. S., Mahadevan, A., Shankar, S. K., Rajagopalan, N., Shet, A., Saravanan, S., Balakrishnan, P., … Ranga, U. (2012). Multiple NF-κB Sites in HIV-1 Subtype C Long Terminal Repeat Confer Superior Magnitude of Transcription and Thereby the Enhanced Viral Predominance. Journal of Biological Chemistry, 287(53), 44714–44735. https://doi.org/10.1074/jbc.M112.397158

Chin, M. P. S., Rhodes, T., Chen, J., Fu, W., & Hu, W.-S. (2005). Identification of a Major Restriction in HIV-1 Inter-subtype recombination. Proc Natl Acad Sci U S A, 102(25), 9002–9007. https://doi.org/10.1073/pnas.0502522102

Domingo, E., & Perales, C. (2019). Viral quasispecies. PLoS Genetics, 15(10), 1–20. https://doi.org/10.1371/journal.pgen.1008271

Galli, A., Kearney, M., Nikolaitchik, O. A., Yu, S., Chin, M. P. S., Maldarelli, F., Coffin, J. M., Pathak, V. K., & Hu, W.-S. (2010). Patterns of Human Immunodeficiency Virus Type 1 Recombination Ex Vivo Provide Evidence for Coadaptation of Distant Sites, Resulting in Purifying Selection for Intersubtype Recombinants during Replication. J. Virol., 84(15), 7651–7661. https://doi.org/10.1128/JVI.00276-10

Hemelaar, J., Elangovan, R., Yun, J., Dickson-Tetteh, L., Fleminger, I., Kirtley, S., Williams, B., Gouws-Williams, E., Ghys, P. D., Abimiku, A. G., Agwale, S., Archibald, C., Avidor, B., Barbás, M. G., Barre-Sinoussi, F., Barugahare, B., Belabbes, E. H., Bertagnolio, S., Birx, D., … Zhang, R. (2019). Global and regional molecular epidemiology of HIV-1, 1990–2015: a systematic review, global survey, and trend analysis. The Lancet Infectious Diseases, 19(2), 143–155. https://doi.org/10.1016/S1473-3099(18)30647-9

Martins, Angélica N., Arruda, M. B., Pires, A. F., Tanuri, A., & Brindeiro, R. M. (2011). Accumulation of P(T/S)AP Late Domain Duplications in HIV Type 1 Subtypes B, C, and F Derived from Individuals Failing ARV Therapy and ARV Drug-Naive Patients. AIDS Research and Human Retroviruses, 27(6), 687–692. https://doi.org/10.1089/aid.2010.0282

Martins, Angelica N., Waheed, A. A., Ablan, S. D., Huang, W., Newton, A., Petropoulos, C. J., Brindeiro, R. de M., & Freed, E. O. (2015). Elucidation of the Molecular Mechanism Driving Duplication of the HIV-1 PTAP Late Domain. Journal of Virology, 90(October), JVI.01640-15. https://doi.org/10.1128/JVI.01640-15

Onafuwa-Nuga, A., & Telesnitsky, A. (2009). The Remarkable Frequency of Human Immunodeficiency Virus Type 1 Genetic Recombination. Microbiology and Molecular Biology Reviews, 73(3), 451–480. https://doi.org/10.1128/MMBR.00012-09

Onafuwa, A., An, W., Robson, N. D., & Telesnitsky, A. (2003). Human Immunodeficiency Virus Type 1 Genetic Recombination Is More Frequent Than That of Moloney Murine Leukemia Virus despite Similar Template Switching Rates. Journal of Virology, 77(8), 4577–4587. https://doi.org/10.1128/JVI.77.8.4577

Peters, S., Muñoz, M., Yerly, S., Lopez-galindez, C., Perrin, L., Larder, B., Cmarko, D., Fakan, S., Noz, M. M. U., & Perrin, L. U. C. (2001). Resistance to Nucleoside Analog Reverse Transcriptase Inhibitors Mediated by Human Immunodeficiency Virus Type 1 p6 Protein Resistance to Nucleoside Analog Reverse Transcriptase Inhibitors Mediated by Human Immunodeficiency Virus Type 1 p6 Protein. Journal of Virology, 75(20), 9644–9653. https://doi.org/10.1128/JVI.75.20.9644

Sharma, S., Aralaguppe, S. G., Abrahams, M.-R., Williamson, C., Gray, C., Balakrishnan, P., Saravanan, S., Murugavel, K. G., Solomon, S., & Ranga, U. (2017). The PTAP sequence duplication in HIV-1 subtype C Gag p6 in drug-naive subjects of India and South Africa. BMC Infectious Diseases, 17(1), 95. https://doi.org/10.1186/s12879-017-2184-4

Sharma, S., Arunachalam, P. S., Menon, M., Ragupathy, V., Satya, R. V., Jebaraj, J., Aralaguppe, S. G., Rao, C., Pal, S., Saravanan, S., Murugavel, K. G., Balakrishnan, P., Solomon, S., Hewlett, I., & Ranga, U. (2018). PTAP motif duplication in the p6 Gag protein confers a replication advantage on HIV-1 subtype C. Journal of Biological Chemistry, 293(30), 11687–11708. https://doi.org/10.1074/jbc.M117.815829

Sharp, P. M., & Hahn, B. H. (2001). Origins of HIV and the AIDS Pandemic. Cold Spring Harbor Laboratory Press, 1–22. https://doi.org/10.1101/cshperspect.a006841

Sharp, P. M., & Hahn, B. H. (2010). The evolution of HIV-1 and the origin of AIDS. Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences, 365(1552), 2487–2494. https://doi.org/10.1098/rstb.2010.0031

Zhang, J., & Temin, H. (1993). Rate and mechanism of non-homologous recombination during a single cycle of retroviral replication. 727(1984).

Chapter – 1: Introduction

1.1 Retroviruses

Retroviruses are a group of enveloped, single-stranded, positive-sense RNA viruses. These viruses 'reverse transcribe' their RNA genome into a complementary DNA (cDNA) using the enzyme Reverse Transcriptase (RT), an RNA-dependent-DNA polymerase. The cDNA thus generated is then integrated into the host genome for transcription and translation (Blomberg et al., 2012). Howard Temin and David Baltimore first reported the presence of RT in Rous Sarcoma Virus (RSV) and Murine Leukemia Virus (MLV), respectively, in 1970 (Baltimore, 1970; Mizutani and Temin, 1970). Several groups subsequently reported the presence of RT in many different retroviruses. More than 50 retroviruses have been classified today into seven genera under the *Retroviridae* family. These viruses infect a diverse host range, spanning birds, fishes, reptiles, and mammals. Among retroviruses, only five can infect human beings: the Human T-Lymphotropic Virus (HTLV) Types-1, -2, -3 and the Human Immunodeficiency Virus Types-1 and -2 (Blomberg et al., 2012). The Human Immunodeficiency Virus Type-1 is the most studied retrovirus among these.

1.2 The Human Immunodeficiency Virus Type-1

The Human Immunodeficiency Virus Type-1 (HIV-1) is the causative agent of Acquired Immunodeficiency Syndrome (AIDS) (Deeks et al., 2015). The virus primarily infects the CD4$^+$ T-cells of the immune system. Since the discovery of HIV-1 in 1983, the AIDS pandemic has claimed the lives of 32.7 million people globally (UNAIDS HIV 2020 Fact Sheet). Extensive research over the past three decades has provided valuable insights into the molecular biology of the virus and has enabled the development of combinatorial anti-retroviral therapy (cART) for the efficient management of the disease today. Widespread awareness campaigns and the accessibility to ART have ensured the progressive decline of the burden of HIV-1 since 2004. Nevertheless, there are still 37.7 million people living with HIV-1 as of 2020 (UNAIDS HIV 2020 Fact Sheet). Increasing access to drugs and ensuring that patients are aware of their HIV-1 status are current target areas for the World Health Organization (WHO).

1.2.1 Molecular Biology and Life Cycle of HIV-1

The HIV-1 genome consists of two copies of a positive-sense RNA. The genome is approximately 9,500 bp long and encodes 15 proteins (Deeks et al., 2015; Frankel et al., 1998) that can be classified based on their functions, as follows:

(i) Structural proteins: Matrix, Capsid, Nucleocapsid, p6, and envelope (gp41, gp120)
(ii) Enzymes: Protease, Reverse Transcriptase, and Integrase
(iii) Regulatory proteins: Tat and Rev
(iv) Accessory proteins: Vpr, Vpu, Vif, and Nef

The arrangement of these genes in the genome is depicted (Figure-1). All viral proteins are expressed from a single promoter called the Long Terminal Repeat (LTR). The structural proteins, except for the envelope, and enzymes, are expressed as inactive polyproteins that are then subsequently cleaved by the viral protease to produce functional proteins. The viral envelope is also expressed as a polyprotein, but is cleaved by the host furin protease to generate two functional proteins, gp120, and gp41. The schematic structure of the HIV-1 virion is presented (Figure-2). The viral particles typically measure between 80 – 100 nm in diameter and contain a central viral core consisting of repeating units of the capsid protein (p24).

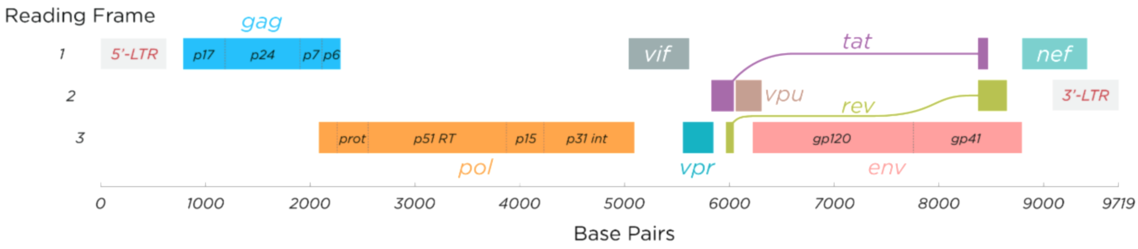


Figure-1: The HIV-1 Genome. A schematic presentation of the viral genome and proteins. Several viral proteins are initially expressed as polyproteins (color-coded) that are processed by viral or host proteases. The vertical dotted lines denote the protease cleavage sites. The individual proteins are labelled in the black text. The two exons of the *tat* and *rev* ORFs are connected by purple- and green-colored dotted lines, respectively. Image Attribution: HIV Genome by Thomas Splettstoesser is licensed under the Creative Commons by ShareAlike 3.0 Unported.

The viral core contains two copies of the viral genomic RNA and co-packages several host and viral proteins necessary for infection, such as the Nucleocapsid, RT, and Integrase. It is made up of the capsid protein (p24), and is encapsulated within a shell made up of the viral matrix (p17), which in turn is surrounded by the viral membrane containing the embedded envelope proteins.

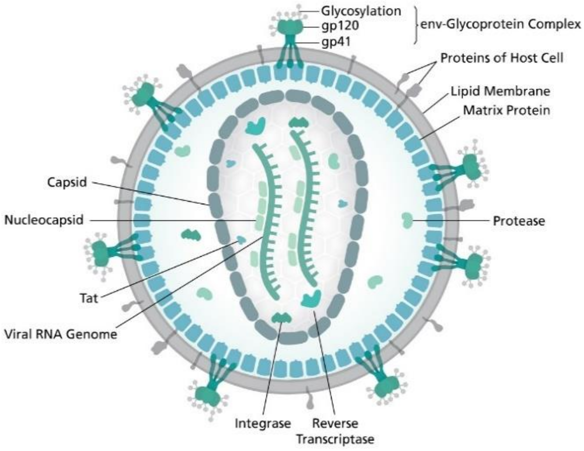


Figure-2: Structure of the HIV-1 Virion. The outermost lipid bilayer is shown in gray, in which envelope proteins gp41 and gp120 (in green) are embedded. The viral matrix (in blue) is immediately underneath the lipid bilayer, which houses the capsid (gray). The two copies of the viral RNA (in green) are depicted along with the rest of the viral proteins (RT, integrase, protease, nucleocapsid, in different shades of green) inside the capsid. Image Attribution: HI Virion Structure by Thomas Splettstoesser is licensed under the Creative Commons by ShareAlike 4.0 Unported.

A schematic diagram of the viral life cycle is depicted (Figure-3). Following the binding to the CD4 receptor on T-cells, the viral envelope undergoes a conformational change, enabling the fusion of the viral membrane with the host plasma membrane. The viral core is then inserted into the cytoplasm of the host cell, following which the viral core slowly begins to disintegrate. RT then reverse transcribes the viral RNA into a single copy of cDNA. The viral cDNA, as part of a multi-protein-cDNA complex called the pre-integration complex, translocates into the nucleus and integrates into the host genome, mediated by the activity of viral integrase. The virus can then remain transcriptionally silent (a phenomenon termed viral latency) or transcribe its genes to make more viral particles that bud out of the cell to initiate a fresh round of infection (Sundquist and Krausslich, 2012).

1.2.2 Classification of HIV-1 and Geographical Distribution of the viral genetic families

HIV consists of two closely related viruses, HIV-1 and HIV-2. They belong to the *Lentivirus* genus of the *Orthoretrovirinae* sub-family in *Retriviridae*. HIV-1 strains are derived from a Simian Immunodeficiency Virus (SIV$_{cpz}$) that infects Chimpanzees, whereas HIV-2 originated from an

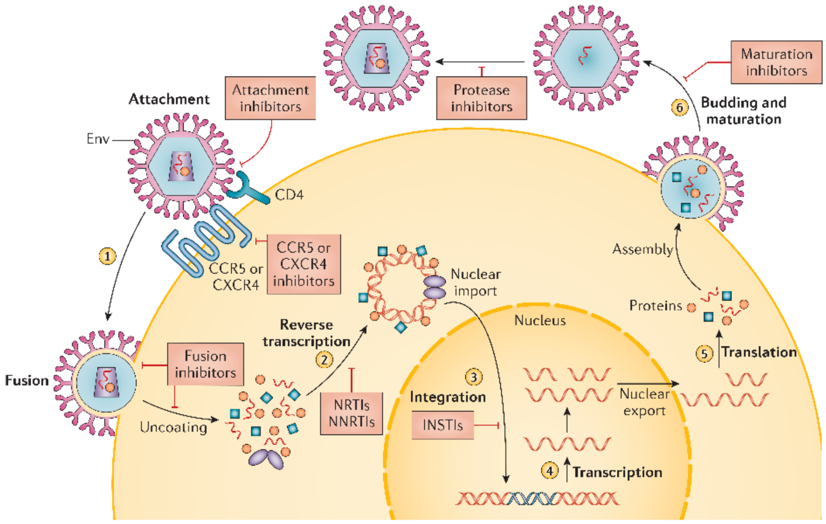


Figure – 3: Major Steps in the Life cycle of HIV-1. (1) The viral particle (shown in pink) enters the cell by using the CD4 and CCR5 receptors on the cell surface. (2) The capsid unpackages, and reverse transcription is completed to generate the pre-integration complex. (3) The pre-integration complex translocates to the nucleus to integrate the viral DNA into the host chromosome. (4) The viral genome is transcribed, and the unspliced, and multiply spliced RNAs are exported to the cytoplasm. (5) The viral proteins are translated, and they assemble to form new virions. (6) The viral particles bud out of the cell, and then undergo maturation to yield infectious virions. Different classes of Anti-Retroviral Drugs are shown in peach boxes adjacent to the step of the life cycle they target. Image reproduced from Deeks et. al (2015) with permission.

SIV that infects Sooty Mangabeys (SIV$_{smm}$) (P M Sharp et al., 2001; Paul M Sharp and Hahn, 2010). HIV-1 can be further classified into four groups – M, N, O, and P, each representing an independent zoonotic transmission event. Of the four, group M represents the most successful family, and is largely responsible for the global AIDS pandemic. The viral strains of the other groups of HIV-1 are mostly restricted to the central and southern African regions (Bbosa et al., 2019; Hemelaar et al., 2019).

Based on phylogenetic and genetic differences, Group M is further divided into 10 viral subtypes (A – D, F – H, J – L, Figure - 4). Several recombinant forms of these primary viral subtypes, labelled as Circulating Recombinant Forms (CRFs), are also found. As of March 2021, 103 different CRFs have been identified in the Los Alamos National Laboratory HIV sequence database. Apart from these, several other recombinant strains exist that have not established themselves in the population in large numbers. These are labelled as the Unique Recombinant Forms (URF).

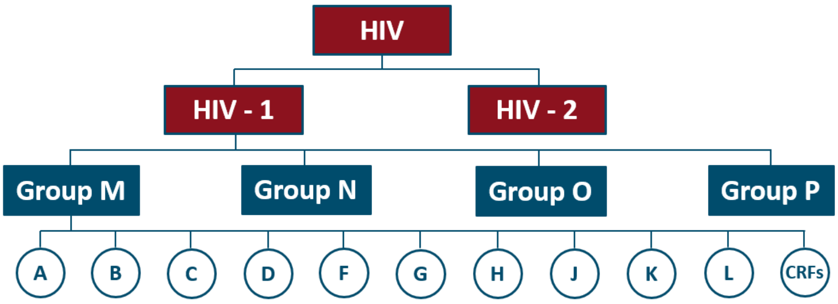


Figure-4: Classification of HIV. HIV is classified into two major types, HIV-1 and HIV-2 (red boxes). HIV-1 in turn is divided into four groups (depicted as blue boxes), and the various viral subtypes under group M are shown in blue circles. CRFs represent the Circulating Recombinant Forms of these primary viral subtypes.

The global distribution of these subtypes is non-uniform (Figure-5), with HIV-1B dominating the infections in the developed countries – North America, Europe, and Australia. HIV-1 A1 is mainly

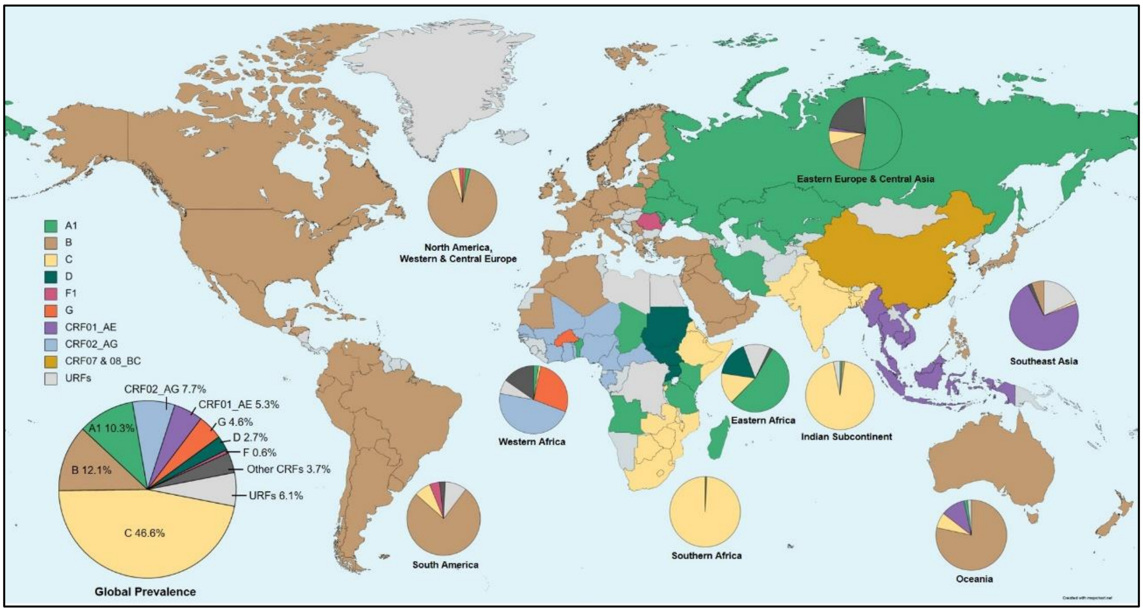


Figure-5: Global Distribution of HIV-1 Subtypes. Differential global distribution of HIV-1 subtypes. The pie chart depicts the relative global abundance of different HIV-1 subtypes. Note that HIV-1C (16,280,897 cases) causes nearly half (46.6%) of global HIV-1 infections (34,921,629). The other major subtypes in descending order of global infections are HIV-1B (12.1%; 4,235,299 cases), HIV-1A (10.3%; 3,587,003 cases), HIV-1 CRF02_AG (7.7%; 2,705,110 cases), HIV-1CRF01_AE (5.3%; 1,840,982 cases), HIV-1G (4.6%; 1,591,267 cases) and HIV-1D (2.7%; 926,255 cases). Subtypes F, H, J and K combined account for 0.9% (311,332 cases) of all infections

predominant in Russia and parts of eastern Europe, and HIV-1C is prevalent in the Indian subcontinent and southern and eastern Africa. These three subtypes together make up approximately 75% of global HIV-1 infections. The remaining seven subtypes collectively account for the rest of the 25 percent of infections, with each subtype being differentially endemic to specific geographical regions (Ariën et al., 2007; Hemelaar et al., 2019).

1.3 HIV-1C: Geographical Prevalence and Unique Molecular Features

HIV-1C is responsible for nearly half of the global infections and more than 90% of infections in India. Several groups have attempted to explain the causes underlying the uneven prevalence of viral subtypes of HIV-1. These explanations include demographic, socio-political, and molecular differences (Ariën et al., 2007; Gartner et al., 2020; Rodriguez et al., 2009; Salemi et al., 2005). Demographic factors point to a higher population density of regions such as the Indian subcontinent and the high HIV-1 seropositivity of places such as South Africa - both areas where HIV-1C is the predominant genetic subtype. Socio-political disturbances such as the large-scale migration of immigrants into South Africa during the turbulent 1970s and 80s could also have influenced the predominance of HIV-1C in this region. Behavioral differences, such as sexual promiscuity, may have contributed to these differences, especially considering that HIV-1 primarily spreads through heterosexual transmission (Deeks et al., 2015).

Although the socio-demographic factors may play some role in the prevalence of HIV-1C, these influences may not sufficiently explain the magnitude of the differential spread observed. Additionally, these factors also fail to explain the sudden spurt of HIV-1C infections in Brazil and other regions of southern America and parts of China. Further, the expansion rate of HIV-1C in Brazil is nearly double that of HIV-1B, alluding to a superior ability of HIV-1C to spread in a population. Ariën et al., (2007) and Tebit and Arts, (2013) attribute this to the lower virulence of HIV-1C among other subtypes in group M of HIV-1.

If the ‘low virulence’ quality attributed to HIV-1C is valid, subtype-specific molecular differences may underlie such variations. Several publications have underlined a possible association between subtype-specific molecular differences and a variable disease outcome. A few such differences have been well documented, including the genetic differences in the cytoplasmic tail of the envelope protein (Da Silva et al., 2016), the lack of the LYPX$_n$L motif in p6-Gag (Neogi et al., 2014; Patil and Bhattacharya, 2012), and the Cysteine to Serine variation at position 31 of the Tat protein that modulates its chemokine function, (Ranga et al., 2004), to name a few.

However, the lack of research on non-HIV-1B subtypes has ensured that the elucidation of crucial molecular features influencing differential biological properties of other subtypes of HIV-1 remains relatively low. The sequence-motif duplications in HIV-1 subtypes and their association with subtype-specific phenotypes is one such area and constitutes the primary objective of the present work.

HIV-1C appears to possess a unique ability to duplicate sequence motifs of biological significance, leading to a replication advantage. While other subtypes can also duplicate sequence motifs, their ability to this end appears limited compared to HIV-1C. Research from our laboratory identified two regions in the viral genome – the LTR and p6-Gag, where the frequency of sequence duplications is unique or significantly higher in HIV-1C (see sections 1.3.1 and 2.0), with implications for replication fitness advantage.

1.3.1 Sequence Duplications in HIV-1C

The LTR of HIV-1 serves as the promoter for viral transcription and harbors several transcription factor binding sites (TFBS). Many subtype-specific differences exist in the order, arrangement, and genetic makeup of these TFBS. The most prominent feature of the HIV-1C LTR is the presence of additional copies of the NF-κB binding site in the viral enhancer. The canonical LTR of HIV-1C contains three NF-κB motifs compared to only two in most other HIV-1 subtypes (Bachu, Mukthey, et al., 2012). Further, work from our laboratory demonstrated the emergence of promoter-variant viral strains containing four NF-κB binding sites (See figure-23). The additional copies of the NF-κB motif are also genetically distinct (Bachu, Yalla, et al., 2012). Additionally, the presence of four NF-κB binding sites in the LTR conferred a higher magnitude of transcriptional activity on the viral strain as compared to the conventional 3-κB viral strains, thus, enhancing viral replication fitness. The 4-κB variant viral strains of HIV-1C have also been reported in other countries (Obasa et al., 2019).

Recently, the emergence of yet more promoter variant viruses containing a multitude of TFBS duplications in and around the viral enhancer has been reported (Bhange et al., 2021). One common property of most of these variants is the presence of a second RBE-III binding site immediately upstream of the NF-κB binding sites. The biological characterization of the viral strains containing a duplicated RBE-III motif is underway.

A second region where HIV-1C frequently duplicates sequences is in the p6 protein located at the C-terminus of Gag (See figures-22,-24). p6-Gag plays a crucial role in viral budding by recruiting the host factor TSG-101. TSG-101 binds a four amino acid PT/SAP motif in p6-Gag, thereby enabling viral release using the Endosomal Sorting Complex Required for Transport (ESCRT) pathway (Demirov et al., 2002; Joshi et al., 2011; Sundquist and Krausslich, 2012). The duplication of the PTAP motif has been well documented in the context of HIV-1B (Ibe et al., 2003; Martins et al., 2011, 2015; Peters et al., 2001; Tamiya et al., 2004). However, PTAP duplication in HIV-1C differs from that of HIV-1B in two respects - duplication at a higher frequency and duplication of longer sequences, both features impacting viral replication fitness (Sharma et al., 2017, 2018).

The increased frequency of sequence motif duplication in HIV-1C is suggestive of unique molecular properties in the reverse transcriptase of this genetic subtype.

1.4 HIV-1 Reverse Transcriptase

HIV-1 Reverse Transcriptase is a 117 kDa protein expressed as a part of the Gag-Pol polyprotein. During the expression of the Gag protein, a ribosomal frameshift ensures the expression of the Pol polyprotein at approximately 5% of that of Gag transcripts. The viral enzymes protease, reverse transcriptase, and integrase are released from the Gag-Pol polyprotein by the activity of the viral protease. HIV-1 RT is a heterodimer of two subunits of 66 kDa and 51 kDa. p66 is the active subunit of 560 amino acids, and p51 performs the role of a structural component of the enzyme that contains the first 440 amino acids of p66. HIV-RT is capable of three functions – RNA-dependent DNA polymerization, DNA-dependent DNA polymerization, and has RNase H activity (Deeks et al., 2015; Hu and Hughes, 2012).

Since 1992, several groups have solved the crystal structure of HIV-1B RT (Kohlstaed et al., 1992). Today, the structure of RT is available at resolutions as low as 1.8 Å (Das et al., 2008). As with most polymerases, the structure resembles the human right hand. Therefore, the individual sub-domains of the p66 subunit have been named fingers, palm, thumb, connection, and RNase H. The p51 subunit lacks the RNase H domain and has a similar arrangement to the other domains as in p66. The fingers, palm, and thumb subdomains of the p66 subunit form the cleft that grips the template-primer complex during polymerization (Sarafianos et al., 2009). The structure and arrangement of the various domains are presented (Figure-6).

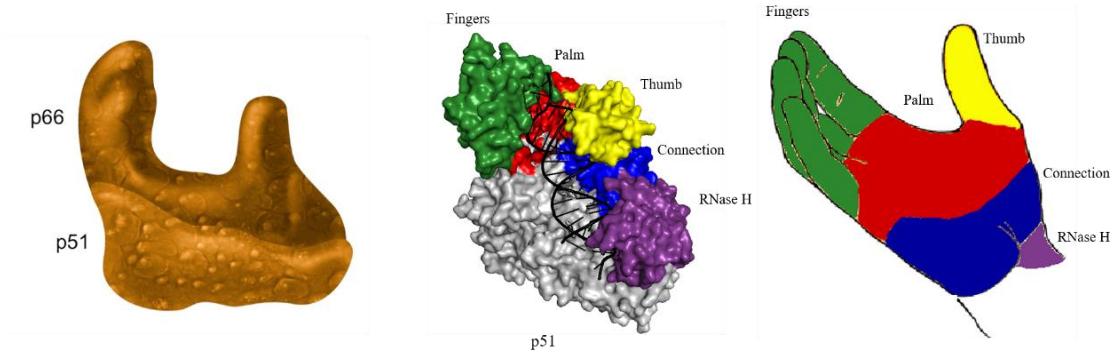


Figure-6: Structure of HIV-1 Reverse Transcriptase. The left panel depicts the relative arrangement of the two subunits of RT. The centre and the right panels present the various domains of RT and their resemblance to the human right hand. The domains are coloured as follows: fingers in green, palm in red, thumb in yellow, connection in blue and RNase H in violet. The central panel shows DNA in black and the p51 structural subunit in gray.

The RT of HIV-1, like other RNA-dependent polymerases, lacks a proofreading mechanism due to the absence of 3' – 5' exonuclease activity. This property underlies the extremely low fidelity of HIV-1 RT (Hu and Hughes, 2012). HIV-RT has an extremely high error rate. This is partly responsible for the excessively high sequence diversity observed among HIV-1 isolates. According to one estimate, the sequence diversity of HIV-1 in a single patient may be comparable to that of the global sequence diversity of the Influenza virus over a year (Korber et al., 2001).

A second molecular property of HIV-1 RT, the ability to switch between the two strands of the viral RNA during polymerization, also contributes to the high magnitude sequence diversity of the virus. Genetic recombination permits the generation of chimeric viral transcripts at a higher frequency. While some strand transfer events are intrinsic to the reverse transcription process, others may appear randomly, spanning the length of the viral genome. It is estimated that RT switches strands 2-3 times per round of reverse transcription (Jetzt et al., 2000; Klarmann et al., 2002)

1.5 HIV-1 Reverse Transcription

HIV-1 Reverse Transcription is initiated at the Primer Binding Site (PBS) present upstream of the U5 region in the viral RNA, where the $\text{tRNA}^{\text{Lys3}}$ (Figure - 7, shown in green) binds and acts as a primer to commence the minus-strand DNA synthesis. The minus-strand DNA (in red) is synthesized in the 5'– 3' direction until the 5' terminus of the viral RNA, after which the first strand transfer event occurs. This transfer shifts the newly synthesized cDNA to its complementary region at the 3' end of the genome, from where the minus strand synthesis proceeds to the PBS. As the viral RT proceeds along the genome, the RNase H activity degrades the template RNA concomitantly (shown as dashed lines) except for a short A-G rich region near the 3' LTR known as the polypurine tract (PPT). The PPT region remains bound to the cDNA and serves as the primer to initiate plus-strand synthesis. The plus-strand (shown in blue) is synthesized from this primer until RT polymerizes DNA corresponding to the PBS from the $\text{tRNA}^{\text{Lys3}}$ still attached to the minus-strand. This enables the second strand transfer event to occur, where the newly incorporated bases complement the PBS anneal to the PBS at the 5' end of the genome. RT then completes polymerization in both directions to yield the complete proviral DNA (Hu and Hughes, 2012; Ilina et al., 2012).

1.6 Models of Retroviral Recombination

HIV-1 is considered a 'pseudo-diploid' virus, since, despite having two copies of the genomic RNA, only a single copy of the viral DNA is made. The co-packaging of two RNA molecules in a viral particle permits viral recombination by RT switching the strands three- or four-times during cDNA synthesis. The transient nature of retroviral recombination has made it near impossible to identify the precise mechanism driving this process. Several experimental observations using electron micrography (Junghans et al., 1982), DNA sequencing etc, offered essential clues leading to the proposition of models to explain viral recombination during different stages of reverse transcription. The RNase H activity of HIV-1 RT serves as a backbone for most of the models describing recombination events that occur during minus-strand DNA synthesis. Of the three such models that are not mutually exclusive but complementary to one another, the dynamic copy choice model gained wide popularity as this model offers assertive leads describing viral recombination (Hwang et al., 2001; Onafuwa-Nuga & Telesnitsky, 2009; Rawson et al., 2018).

The recombination models are described below in brief:

(i) Forced-Copy-Choice or Pause-Driven Recombination: This model proposes that the intrinsic qualities of the viral RNA, such as nicks or strong secondary structures, cause the RT to pause and switch the strands. In other words, the RT is 'forced' to switch strands to regions of similar homology on the second strand. This model is primarily based on the observation that many RNA genomes isolated from retroviruses were nicked at multiple locations (Junghans et al., 1982). Other experiments performed using *in vitro* synthesized RNA revealed that stalling of RT at RNA secondary structures increases the frequency of recombination, thereby lending credibility to this model (Lanciault & Champoux, 2006).

(ii) Copy-Choice or Pause-Independent Recombination: The copy-choice model relies heavily on the RNase H activity of the RT. The model proposes the annealing of the acceptor RNA strand to the complementary DNA strand downstream of the RT. As RT proceeds along the donor RNA strand and the RNase H activity degrades the donor RNA, the cDNA is made available to hybridize with the acceptor RNA. Consequently, the acceptor RNA invades the polymerization bubble, causing RT to switch to the acceptor RNA strand and continue cDNA synthesis. The copy-choice model has a technical advantage over others that it permits viral recombination in the absence of nicks or RNA secondary structures; therefore, it can be applied to explain a larger number of recombination events in the HIV-1 genome.

(iii) Dynamic-Copy-Choice Recombination: The Dynamic-Copy-Choice model, a modification of the copy-choice model, depicts recombination as an outcome of an imbalance in the equilibrium between the rates of polymerization and RNase H activity. Any event that decreases the polymerization rate, such as an RNA secondary structure, a nick or purine-rich tracts or hybridization of the acceptor RNA, can lead to recombination as the equilibrium is disturbed. The dynamic copy choice model incorporates observations leading to the establishment of both the forced-copy choice and copy-choice models; therefore, it represents the most widely accepted model to explain viral recombination today.

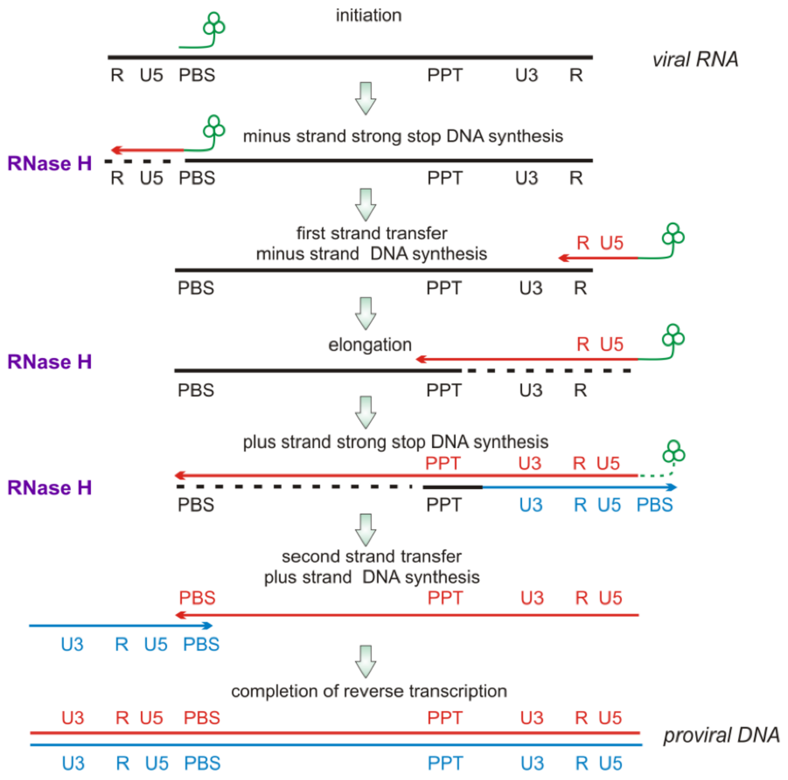


Figure-7: Steps of HIV-1 Reverse Transcription. Reverse transcription begins at the PBS near the 5' end of the genome where the tRNA^{Lys3} (shown in green) binds and serves as the primer. The viral RNA is depicted in black and the minus and plus strand DNA are shown as red and blue lines, respectively. Dotted lines indicate the degraded template RNA hydrolyzed by the RNase H activity of the RT. Image reproduced from Ilina et.al (2012).

1.6.1 Non-Homologous Recombination: Duplications, Deletions, and Insertions-in-Deletions

While most recombination events occur at homologous regions of the donor and acceptor RNA templates, a small frequency (0.1 – 1%) is estimated to be non-homologous, mediated via micro-homology domains (Zhang & Temin, 1993). Thus, in rare instances, the location on the donor RNA from where the RT dissociates and the location on the acceptor template from where it resumes polymerization are different. This discordance leads to changes in the length of the genome, and these changes can be of three different kinds (Onafuwa-Nuga & Telesnitsky, 2009):

(i) Duplication: When RT resumes polymerization on the acceptor strand at a location that it has already polymerized on the donor, RT copies the sequence again, leading to duplication.

(ii) Deletion: A deletion occurs when RT switches strands to the acceptor at a location further downstream than the one it has just vacated on the donor strand. This leads to stretches of nucleotides that the virus misses during polymerization, thus causing deletions.

(iii) Insertions-in-Deletions: Insertions-in-deletions result from the RT switching strands multiple times within a short region. These are a combination of both deletions and insertions described above and are not observed commonly in wild-type isolates. However, the frequency of their occurrence in experimental systems is high (Onafuwa-Nuga & Telesnitsky, 2009)

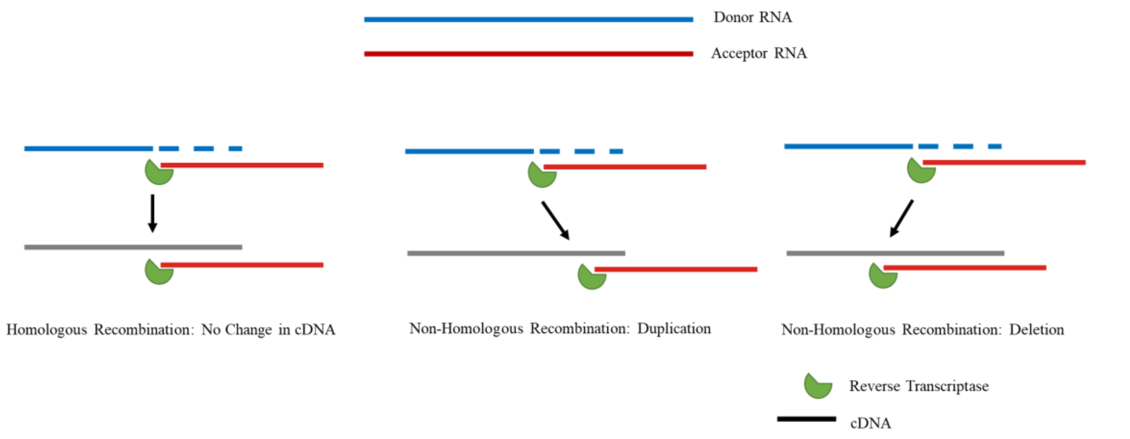

Donor RNA
Acceptor RNA
Homologous Recombination: No Change in cDNA
Non-Homologous Recombination: Duplication
Non-Homologous Recombination: Deletion
Reverse Transcriptase
cDNA

Figure – 8: Mechanisms of sequence duplication and deletion. The panel on the left depicts homologous recombination, the centre and the right panel show the mechanism of sequence duplications and deletions, respectively. RT is shown in green, and the cDNA is depicted in red. RNase H cleavage of the donor RNA is represented as dashed lines.

1.7 Sequence Duplication and Viral Evolution

Sequence duplications are often associated with an evolutionary cost. In both the sequence duplication hotspots that our laboratory identified (LTR and p6-Gag), sequence duplication is associated with a gain in replicative fitness (Bachu, Yalla, et al., 2012; Martins et al., 2015; Sharma et al., 2018). While the NF-κB duplication enhances transcription, the PTAP motif duplication appears to compensate for debilitating immune/drug escape mutations. The strong selection forces associated with these two regions make them stand out in sequence analyses from databases. Searching for sequences in extant databases can reveal that sequence motif duplication on the LTR and p6-Gag is qualitatively different in HIV-1C. These events in HIV-1C also appear at a significantly higher frequency, thus alluding to subtype-specific variations. However, drawing meaningful inferences beyond this preliminary analysis from sequence data alone becomes difficult. Several features remain unclear regarding subtype-specific differences in sequence motif duplication. Although RT plays the central role in reverse transcription, several additional *cis* and *trans* factors must influence sequence motif duplication of a specific nature. Additionally, we have reasons to believe that the two events of subtype-specific sequence motif duplication represent extremely rare occurrences (see section 2.2). Faithful recapitulations of the duplication in the laboratory would be exceedingly challenging in the absence of knowledge of the nature of the selection forces that permit the selection of the variant viral strains. Further, appreciating the biological significance of such duplication events at the population level is exceedingly challenging since such efforts need time.

Given the technical challenges, the investigations presented in the thesis rely primarily on regressive and limited experimental design while asking questions and drawing inferences. Based on the bioinformatic analysis of viral sequences downloaded from the sequence databanks, the magnitude of sequence duplication appears to be significantly higher in HIV-1C than in other subtypes. This observation warrants focusing on the RT, given the association between RT function and sequence duplications. If the RT function underlies differences in duplication frequency; such differences could have implications for several aspects of HIV-1 biology, including viral evolution, disease progression, virulence, drug resistance, and pathogenesis.

1.8 References

Ariën, K. K., Vanham, G., & Arts, E. J. (2007). Is HIV-1 evolving to a less virulent form in humans? Nature Reviews Microbiology, 5(2), 141–151. https://doi.org/10.1038/nrmicro1594

Bachu, M., Mukthey, A. B., Murali, R. V., Cheedarla, N., Mahadevan, A., Shankar, S. K., Satish, K. S., Kundu, T. K., & Ranga, U. (2012). Sequence Insertions in the HIV Type 1 Subtype C Viral Promoter Predominantly Generate an Additional NF-κB Binding Site. AIDS Research and Human Retroviruses, 28(10), 1362–1368. https://doi.org/10.1089/aid.2011.0388

Bachu, M., Yalla, S., Asokan, M., Verma, A., Neogi, U., Sharma, S., Murali, R. V., Mukthey, A. B., Bhatt, R., Chatterjee, S., Rajan, R. E., Cheedarla, N., Yadavalli, V. S., Mahadevan, A., Shankar, S. K., Rajagopalan, N., Shet, A., Saravanan, S., Balakrishnan, P., … Ranga, U. (2012). Multiple NF-κB Sites in HIV-1 Subtype C Long Terminal Repeat Confer Superior Magnitude of Transcription and Thereby the Enhanced Viral Predominance. Journal of Biological Chemistry, 287(53), 44714–44735. https://doi.org/10.1074/jbc.M112.397158

Baltimore, D. (1970). Viral RNA-dependent DNA polymerase: RNA-dependent DNA polymerase in virions of RNA tumour viruses. In Nature (Vol. 226, Issue 5252, pp. 1209–1211). https://doi.org/10.1038/2261209a0

Bbosa, N., Kaleebu, P., & Ssemwanga, D. (2019). HIV subtype diversity worldwide. Current Opinion in HIV and AIDS, 14(3), 153–160. https://doi.org/10.1097/COH.0000000000000534

Blomberg, J., Boeke, J. D., Coffin, J. M., Eickbush, T., Fan, H., Geering, A. D. W., Gerlich, W. H., Hahn, B., Hull, R., Kann, M., Loeb, D., Magnius, L., Mason, W. S., Mizokami, T., Neil, J., Norder, H., Quackenbush, S., Rethwilm, A., Sandmeyer, S. ., … Voytas, D. F. (2012). Ninth Report of the International Committee on Taxonomy of Viruses. In Virus Taxonomy (Vol. 9, pp. 477–495). https://doi.org/10.1201/9781351071642-11

Da Silva, E. S., Mulinge, M., Lemaire, M., Masquelier, C., Beraud, C., Rybicki, A., Servais, J. Y., Iserentant, G., Schmit, J. C., Seguin-Devaux, C., & Bercoff, D. P. (2016). The envelope cytoplasmic tail of HIV-1 Subtype C contributes to poor replication capacity through low viral infectivity and cell-To-cell transmission. PLoS ONE, 11(9), 1–29. https://doi.org/10.1371/journal.pone.0161596

Das, K., Bauman, J. D., Clark, A. D., Frenkel, Y. V., Lewi, P. J., Shatkin, A. J., Hughes, S. H., & Arnold, E. (2008). High-resolution structures of HIV-1 reverse transcriptase/TMC278 complexes: Strategic flexibility explains potency against resistance mutations. Proceedings of the National Academy of Sciences of the United States of America, 105(5), 1466–1471. https://doi.org/10.1073/pnas.0711209105

Deeks, S. G., Overbaugh, J., Phillips, A., & Buchbinder, S. (2015). HIV infection. Nature Reviews Disease Primers, October, 15035. https://doi.org/10.1038/nrdp.2015.35

Demirov, D. G., Orenstein, J. M., & Freed, E. O. (2002). The Late Domain of Human Immunodeficiency Virus Type 1 p6 Promotes Virus Release in a Cell Type-Dependent Manner. Journal of Virology, 76(1), 105–117. https://doi.org/10.1128/jvi.76.1.105-117.2002

Frankel, A. D., Francisco, S., & Young, J. A. T. (1998). HIV-1 : Fifteen Proteins and an RNA. 1–25.

Gartner, M. J., Roche, M., Churchill, M. J., Gorry, P. R., & Flynn, J. K. (2020). Understanding the mechanisms driving the spread of subtype C HIV-1. EBioMedicine, 53, 102682. https://doi.org/10.1016/j.ebiom.2020.102682

Hemelaar, J., Elangovan, R., Yun, J., Dickson-Tetteh, L., Fleminger, I., Kirtley, S., Williams, B., Gouws-Williams, E., Ghys, P. D., Abimiku, A. G., Agwale, S., Archibald, C., Avidor, B., Barbás, M. G., Barre-Sinoussi, F., Barugahare, B., Belabbes, E. H., Bertagnolio, S., Birx, D., … Zhang, R. (2019). Global and regional molecular epidemiology of HIV-1, 1990–2015: a systematic review, global survey, and trend analysis. The Lancet Infectious Diseases, 19(2), 143–155. https://doi.org/10.1016/S1473-3099(18)30647-9

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

Hwang, C. K., Svarovskaia, E. S., & Pathak, V. K. (2001). Dynamic copy choice: Steady state between murine leukemia virus polymerase and polymerase-dependent RNase H activity determines frequency of in vivo template switching. Proceedings of the National Academy of Sciences of the United States of America, 98(21), 12209–12214. https://doi.org/10.1073/pnas.221289898

Ibe, S., Shibata, N., Utsumi, M., & Kaneda, T. (2003). Selection of human immunodeficiency virus type 1 variants with an insertion mutation in the p6gag and p6pol genes under highly active antiretroviral therapy. Microbiology and Immunology, 47(1), 71–79. https://doi.org/10.1111/j.1348-0421.2003.tb02788.x

Ilina, T., Labarge, K., Sarafianos, S. G., Ishima, R., & Parniak, M. a. (2012). Inhibitors of HIV-1 Reverse Transcriptase-Associated Ribonuclease H Activity. Biology, 1(3), 521–541. https://doi.org/10.3390/biology1030521

Jetzt, A. E., Yu, H., Klarmann, G. J., Ron, Y., Preston, B. D., & Dougherty, J. P. (2000). High rate of recombination throughout the human immunodeficiency virus type 1 genome. Journal of Virology, 74(3), 1234–1240. https://doi.org/10.1128/jvi.74.3.1234-1240.2000

Joshi, A., Garg, H., Ablan, S., Freed, E. O., Nagashima, K., Manjunath, N., & Shankar, P. (2011). Targeting the HIV entry, assembly and release pathways for anti-HIV gene therapy. Virology, 415(2), 95–106. https://doi.org/10.1016/j.virol.2011.03.028

Junghans, R. P., Boone, L. R., & Skalka, A. M. (1982). Products of reverse transcription in avian retrovirus analyzed by electron microscopy. Journal of Virology, 43(2), 544–554. https://doi.org/10.1128/jvi.43.2.544-554.1982

Klarmann, G., Jetzt, A. E., Preston, B. D., Zhuang, J., Sun, G., Ron, Y., Dougherty, J. P., & Yu, H. (2002). Human Immunodeficiency Virus Type 1 Recombination: Rate, Fidelity, and Putative Hot Spots. Journal of Virology, 76(22), 11273–11282. https://doi.org/10.1128/jvi.76.22.11273-11282.2002

Kohlstaed, L. A., Wang, J., Friedman, J. M., Rice, P. A., & Steitz, T. A. (1992). Crystal structure at 3.5 Å resolution of h iv -1 reverse transcriptase complexed with an inhibitor. Structural Insights into Gene Expression and Protein Synthesis, 256(June), 254–261. https://doi.org/10.1142/9789811215865_0026

Korber, B., Gaschen, B., Yusim, K., Thakallapally, R., Kesmir, C., & Detours, V. (2001). Evolutionary and immunological implications of contemporary HIV-1 variation. British Medical Bulletin, 58, 19–42. https://doi.org/10.1093/bmb/58.1.19

Lanciault, C., & Champoux, J. J. (2006). Pausing during Reverse Transcription Increases the Rate of Retroviral Recombination. Society, 80(5), 2483–2494. https://doi.org/10.1128/JVI.80.5.2483

Martins, Angélica N., Arruda, M. B., Pires, A. F., Tanuri, A., & Brindeiro, R. M. (2011). Accumulation of P(T/S)AP Late Domain Duplications in HIV Type 1 Subtypes B, C, and F Derived from Individuals Failing ARV Therapy and ARV Drug-Naive Patients. AIDS Research and Human Retroviruses, 27(6), 687–692. https://doi.org/10.1089/aid.2010.0282

Martins, Angelica N., Waheed, A. A., Ablan, S. D., Huang, W., Newton, A., Petropoulos, C. J., Brindeiro, R. de M., & Freed, E. O. (2015). Elucidation of the Molecular Mechanism Driving Duplication of the HIV-1 PTAP Late Domain. Journal of Virology, 90(October), JVI.01640-15. https://doi.org/10.1128/JVI.01640-15

Mizutani, S., & Temin, H. M. (1970). An RNA-Dependent DNA Polymerase in Virions of Rous Sarcoma Virus. In Nature (Vol. 226, Issue 5252, pp. 1211–1213). https://doi.org/10.1101/sqb.1970.035.01.100

Neogi, U., Rao, S. D., Bontell, I., Verheyen, J., Rao, V. R., Gore, S. C., Sonif, N., Shet, A., Schülter, E., Ekstrand, M. L., Wondwossen, A., Kaiser, R., Madhusudhan, M. S., Prasad, V. R., & Sonnerborg, A. (2014). Novel tetra-peptide insertion in Gag-p6 ALIXbinding motif in HIV-1 subtype C associated with protease inhibitor failure in Indian patients. Aids, 28(15), 2319–2322. https://doi.org/10.1097/QAD.0000000000000419

Obasa, A. E., Ashokkumar, M., Neogi, U., & Jacobs, G. B. (2019). Mutations in Long Terminal Repeats κB Transcription Factor Binding Sites in Plasma Virus Among South African People Living with HIV-1. AIDS Research and Human Retroviruses, 35(6), 572–576. https://doi.org/10.1089/aid.2018.0293

Onafuwa-Nuga, A., & Telesnitsky, A. (2009). The Remarkable Frequency of Human Immunodeficiency Virus Type 1 Genetic Recombination. Microbiology and Molecular Biology Reviews, 73(3), 451–480. https://doi.org/10.1128/MMBR.00012-09

Patil, A., & Bhattacharya, J. (2012). Natural deletion of L35Y36 in p6 gag eliminate LYPXnL/ALIX auxiliary virus release pathway in HIV-1 subtype C. Virus Research, 170(1–2), 154–158. https://doi.org/10.1016/j.virusres.2012.08.020

Peters, S., Muñoz, M., Yerly, S., Lopez-galindez, C., Perrin, L., Larder, B., Cmarko, D., Fakan, S., Noz, M. M. U., & Perrin, L. U. C. (2001). Resistance to Nucleoside Analog Reverse Transcriptase Inhibitors Mediated by Human Immunodeficiency Virus Type 1 p6 Protein Resistance to Nucleoside Analog Reverse Transcriptase Inhibitors Mediated by Human Immunodeficiency Virus Type 1 p6 Protein. 75(20), 9644–9653. https://doi.org/10.1128/JVI.75.20.9644

Ranga, U., Shankarappa, R., Siddappa, N. B., Ramakrishna, L., Nagendran, R., Mahalingam, M., Mahadevan, A., Jayasuryan, N., Satishchandra, P., Shankar, S. K., & Prasad, V. R. (2004). Tat Protein of Human Immunodeficiency Virus Type 1 Subtype C Strains Is a Defective Chemokine. Journal of Virology, 78(5), 2586–2590. https://doi.org/10.1128/jvi.78.5.2586-2590.2004

Rawson, J. M. O., Nikolaitchik, O. A., Keele, B. F., Pathak, V. K., & Hu, W. S. (2018). Recombination is required for efficient HIV-1 replication and the maintenance of viral genome integrity. Nucleic Acids Research, 46(20), 10535–10545. https://doi.org/10.1093/nar/gky910

Rodriguez, M. A., Ding, M., Ratner, D., Chen, Y., Tripathy, S. P., Kulkarni, S. S., Chatterjee, R., Tarwater, P. M., & Gupta, P. (2009). High replication fitness and transmission efficiency of HIV-1 subtype C from India: Implications for subtype C predominance. Virology, 385(2), 416–424. https://doi.org/10.1016/j.virol.2008.12.025

Salemi, M., De Oliveira, T., Soares, M. A., Pybus, O., Dumans, A. T., Vandamme, A. M., Tanuri, A., Cassol, S., & Fitch, W. M. (2005). Different epidemic potentials of the HIV-1B and C subtypes. Journal of Molecular Evolution, 60(5), 598–605. https://doi.org/10.1007/s00239-004-0206-5

Sarafianos, S. G., Marchand, B., Das, K., Himmel, D. M., Parniak, M. a., Hughes, S. H., & Arnold, E. (2009). Structure and Function of HIV-1 Reverse Transcriptase: Molecular Mechanisms of Polymerization and Inhibition. Journal of Molecular Biology, 385(3), 693–713. https://doi.org/10.1016/j.jmb.2008.10.071

Sharma, S., Aralaguppe, S. G., Abrahams, M.-R., Williamson, C., Gray, C., Balakrishnan, P., Saravanan, S., Murugavel, K. G., Solomon, S., & Ranga, U. (2017). The PTAP sequence duplication in HIV-1 subtype C Gag p6 in drug-naive subjects of India and South Africa. BMC Infectious Diseases, 17(1), 95. https://doi.org/10.1186/s12879-017-2184-4

Sharma, S., Arunachalam, P. S., Menon, M., Ragupathy, V., Satya, R. V., Jebaraj, J., Aralaguppe, S. G., Rao, C., Pal, S., Saravanan, S., Murugavel, K. G., Balakrishnan, P., Solomon, S., Hewlett, I., & Ranga, U. (2018). PTAP motif duplication in the p6 Gag protein confers a replication advantage on HIV-1 subtype C. Journal of Biological Chemistry, 293(30), 11687–11708. https://doi.org/10.1074/jbc.M117.815829

Sharp, P M, Bailes, E., Chaudhuri, R. R., Rodenburg, C. M., Santiago, M. O., & Hahn, B. H. (2001). The origins of acquired immune deficiency syndrome viruses: where and when? Philosophical Transactions of the Royal Society B: Biological Sciences, 356(1410), 867–876. https://doi.org/10.1098/rstb.2001.0863

Sharp, Paul M, & Hahn, B. H. (2010). The evolution of HIV-1 and the origin of AIDS. Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences, 365(1552), 2487–2494. https://doi.org/10.1098/rstb.2010.0031

Sundquist, W. I., & Krausslich, H. G. (2012). HIV-1 assembly, budding, and maturation. Cold Spring Harbor Perspectives in Medicine, 2(7), 1–24. https://doi.org/10.1101/cshperspect.a006924

Tamiya, S., Mardy, S., Kavlick, M. F., Yoshimura, K., & Mistuya, H. (2004). Amino Acid Insertions near Gag Cleavage Sites Restore the Otherwise Compromised Replication of Human Immunodeficiency Virus Type 1 Variants Resistant to Protease Inhibitors. Journal of Virology, 78(21), 12030–12040. https://doi.org/10.1128/jvi.78.21.12030-12040.2004

Tebit, D. M., & Arts, E. J. (2013). From Simian to Human Immunodeficiency Viruses (SIV to HIV): Emergence from Nonhuman Primates and Transmission to Humans. In The Role of Animals in Emerging Viral Diseases. Elsevier. https://doi.org/10.1016/B978-0-12-405191-1.00009-0

Zhang, J., & Temin, H. (1993). Rate and mechanism of nonhomologous recombination during a single cycle ofretroviral replication. 727(1984).

Chapter – 2: Review of Literature

2.0 Sequence Duplications in HIV-1C

Sequence duplication is a well-documented occurrence in retroviruses (Bachu, Mukthey, et al., 2012; Neil et al., 1991; Zheng et al., 1997). In nearly all these reports, sequence duplication has been associated with a replicative gain. This could be because the duplication events observed in sequence databases are those which have been positively selected. When sequence duplications cause a deleterious effect, such as a frameshift mutation, negative selection may have eliminated those viral strains.

As described in the previous chapter, sequence duplications appear to confer an evolutionary advantage on HIV-1. Although several publications have reported sequence duplications in various regions spanning the HIV-1 genome (Carl et al., 2000; Dang & Hu, 2001; Ji et al., 2018), most of these reports represent sporadic and isolated occurrences. In contrast, two regions in the viral genome - the LTR and the p6-Gag protein, appear to be the 'hotspots' of sequence duplication. A previous analysis of full-length HIV-1B and -C genome sequences available in the extant databases found that approximately 17% of the deposited sequences contain sequence duplications in these regions (Bachu, Mukthey, et al., 2012; Sharma et al., 2017). The frequency of such duplications, however, is not uniform between the subtypes, with HIV-1C containing sequence duplications at higher numbers than HIV-1B. While the LTR duplications are observed at 14.1% and 19.1% in HIV-1B and HIV-1C, respectively, the p6-Gag PTAP duplication-frequency difference is more pronounced, at 6.4% and 29%, respectively. While our laboratory examined the biological significance of the 4-κB LTR duplications in HIV-1C (Bachu, Yalla, et al., 2012), the significance of PTAP duplications has been evaluated by several groups (Martins et al., 2011, 2015; Peters et al., 2001; Tamiya et al., 2004), including ours (Sharma et al., 2017, 2018).

In essence, it appears that HIV-1C is endowed with a natural propensity to generate sequence duplications at a higher frequency than other HIV-1 genetic subtypes. However, the molecular causes underlying the phenomenon have not been uncovered. The present chapter reviews the available evidence related to viral recombination and sequence duplication, highlighting the gaps in our understanding of this area and how the present work aims to fill this gap.

2.1 Subtype-Specific Differences

The significantly high frequency of sequence motif duplications observed in HIV-1C suggests that cis- and trans-acting elements may function differently in this subtype. While the basic mechanisms of sequence duplication by non-homologous recombination have been well established (Onafuwa-Nuga & Telesnitsky, 2009; Zhang & Temin, 1993), the underlying causes have not been characterized to the same extent. Thus, the function of factors accounting for subtype-specific differences in sequence motif duplication remains unexplored.

To this end, here, an attempt has been made to examine the data available in the literature relevant to sequence duplication, and where possible, attention is drawn to subtype-specific differences. However, one technical challenge is the limited availability of reports using non-HIV-1B genetic subtypes. Nevertheless, we took advantage of several publications that have examined crucial differences among viral subtypes.

2.1.1 HIV-1C RT – Properties

The inherent properties of the Reverse Transcriptase of HIV-1C may influence the propensity of the enzyme to duplicate sequences. Multiple studies have highlighted various subtype-specific mutations in the RTs of the different HIV-1 subtypes. However, few publications have experimentally analyzed the effects of these mutations on HIV-1C-RT function. The analysis of subtype-specific variations has primarily served function of epidemiological classification of viral subtypes. Of note, a couple of publications analyzed the biochemical properties of HIV-1C RT in comparison with those of HIV-1B. Despite differences in several amino acid residues, the biochemical properties, including specific activity, processivity, error rate, and RNase H activity, were comparable between the RTs of HIV-1B and HIV-1C (Xu et al., 2010). In contrast, Iordanskiy et al. (2010) demonstrated that despite similar biochemical properties, viruses containing chimera forms of B and C RT gene in a subtype B backbone displayed slower rates of reverse transcription and, consequently, reduced replication. Interestingly, the same does not hold when B-RT or its chimera forms were substituted for C-RT in an HIV-1C backbone. However, interpretation of these results is challenging as the viral vectors contained additional genetic variations in other regions, including protease, integrase, and *vif*. Thus, an elegant experimental design is crucial to delineate intrinsic but subtle differences in the functioning of RTs of diverse viral subtypes that may govern subtype-specific phenotype differences. Nevertheless, the conclusion of the authors that the two RTs, or their fragments do not function at full capacity in a heterologous backbone cannot be dismissed entirely. If so, it would suggest that while the properties of the two RTs maybe the same, the two enzymes use different molecular mechanisms to achieve that end. This could explain why even fragments of C-RT do not function optimally in an HIV-1B backbone.

2.1.2 Viral recombination in HIV-1

Recombination is a prerequisite for sequence duplication. Several research groups over the years estimated recombination frequency in HIV-1 using various methods. A broad level consensus has emerged that HIV-1 RT switches strands on an average of 3-5 times per viral reverse transcription (Klarmann et al., 2002; Lanciault & Champoux, 2006; Onafuwa-Nuga & Telesnitsky, 2009). The earlier experimental strategies relied on mutant viral strains acquiring marker genes to estimate recombination frequencies (Anderson et al., 1998; Zhang & Temin, 1993). Recombination frequency in a cell-culture system was expressed as the fraction of cells that displayed a desired phenotype among the infected cells. Functional restoration of a reporter gene by recombination between two defective domains serves as a more popular and contemporary experimental strategy to estimate recombination frequency in HIV-1. The two genomic RNAs co-packaged into the viral particle harbor debilitating point mutations in two different regions of the reporter gene. The gene function can be rescued only when a recombination event occurs between the two mutations, thus bringing the intact functional domains together to restore the reporter activity, which can then be quantitated. The function-restoration strategy has been used widely to explore the role of factors that may govern or modulate viral recombination. The works of Hwang et al. (2001) brought an early insight into the process of HIV-1 recombination. Based on the findings, the authors proposed the dynamic-copy-choice model to explain viral recombination and highlighted the existence of a fine balance between the rate of polymerization and RNase H cleavage. The dynamic-copy-choice recombination model provides a valuable framework to identify factors that can influence recombination. The model helped explain the previous observations that viral recombination rate is significantly higher at RNA break points or secondary structures (Fan et al., 2007; Lanciault & Champoux, 2006), as both these events could slow polymerization rate.

Of note, misincorporation of nucleotides promotes recombination. Several publications have attested to this fact since the first publication of Palaniappan et al., (1996). While Schlub et al (2014) ascertained this using high-throughput sequencing, Chin et al. (2007) confirmed this by using recombination assays. Although the degree of correlation between misincorporation and recombination is controversial, it is widely accepted that misincorporation would indeed promote recombination. This observation is consistent with the dynamic-copy-choice recombination model (Hwang et al., 2001), since misincorporation reduces the polymerization rate significantly. The direct relevance of these studies to the present work is described in greater detail in the discussion section.

The viral nucleocapsid protein (NC) is an important factor that can modulate reverse transcription, particularly recombination. The NC protein acts as a nucleic acid chaperone and breaks RNA secondary structures to facilitate viral reverse transcription. NC significantly augments DNA strand transfers *in vitro* (Chen et al., 2003; Mougel et al., 2009; Racine et al., 2016; Rodriguez-Rodriguez et al., 1995). Since specific regulatory regions of the viral RNA, such as the Transactivation Response element, TAR, the dimerization signal, and DIS, can fold into stable RNA secondary structures, the role of NC is crucial in breaking these structures and promoting recombination to drive viral reverse transcription to completion.

Rhodes et al. (2005) ascertained the absence of the influence of viral accessory proteins, including Vpr, Vpu, Vif, and Nef, on HIV-1 recombination. They also reported that the recombination rate does not vary significantly between primary CD4+ cells and various T cell lines. However, subsequent work by David Levy and coworkers showed a significant variation between T-cells and cells of other lineages, such as monocytes (Levy et al., 2004). Thus, the rate of viral recombination may vary significantly depending on several factors, including cell lineage.

Recombination rates are greatly enhanced in the presence of anti-retroviral drugs. The enhanced recombination frequency may be ascribed to the purifying selection that the antiretroviral therapy (ART) may confer on drug-resistant mutations, such as K65R ((K. A. Delviks-Frankenberry et al., 2010; Nikolenko et al., 2007). Several anti-retroviral drugs may target the polymerization domain of the RT reducing the polymerization rate, resulting in an imbalance to the equilibrium proposed by the dynamic-copy-choice model, to favor recombination. The relevance of recombination to drug resistance is explored in detail in the following section.

A predominant research strategy common to several publications cited above is the use of HIV-1B as an experimental model to study HIV-1 recombination. Only a single publication measured recombination in HIV-1C using a sub-genomic viral strain derived from the pMJ4 molecular clone (Chin et al., 2005). The authors report no significant difference in recombination rates between B and C subtypes. This assertion merits re-investigation using other molecular clones of HIV-1C due to the high genetic variability HIV-1 displays. In other words, a single molecular clone need not necessarily be representative of a genetic family. Further, to recapitulate the natural conditions faithfully, the evaluation must be repeated using a full-length viral clone, not a sub-genomic vector.

2.1.3 Drug Resistance

Perhaps the most profound impact of reported sequence duplications is on antiretroviral drug resistance of HIV-1. While there is no direct evidence of the PTAP duplication augmenting drug resistance, multiple publications reported an enhanced frequency of duplications among drug-resistant HIV-1 strains. This observation has led to a speculation that PTAP duplication may play a compensatory role in countering the loss of function due to drug-resistant mutations (Brindeiro et al., 2002; Martins et al., 2011; Peters et al., 2001; Tamiya et al., 2004). Martins et al. attributed this compensation to the p6-pol reading frame by demonstrating that the PTAP duplication enables superior proteolytic cleavage of the NC-Sp2-p6 junction site in patients harbouring drug-resistant mutations (Martins et al., 2015). While this observation certainly explains the high frequency of duplications in drug-exposed subjects, it fails to account for high-frequency PTAP duplication observed in drug naïve patients reported by our laboratory (Sharma et al., 2017), suggesting that a more complex mechanisms might underlie these duplication frequencies.

An earlier analysis by Martins et al. compared the frequency of PTAP motif duplication in three subtypes (B, C, and F) between drug naïve subjects and patients under treatment failure (Martins et al., 2011). In this analysis, HIV-1C demonstrated an unusually high frequency of PTAP motif duplication, with more than half the patients (54%) in the drug failure group harbouring the PTAP motif duplication compared to 9.3% and 17.6% in HIV-1B and F, respectively. Additionally, the percentage of viral strains containing duplications in the drug naïve group was also highest in HIV-

1C (23%), compared to B (3.7%) and F (0%). This observation ascertains further that HIV-1C is endowed with viral factors that enable high-frequency duplication of sequences.

Further, the differential amino acid substitutions and the evolutionary trajectories the viral subtypes adapt towards drug resistance are of relevance. While some drug resistance mutations are common to all subtypes, such as M184V, K103M, and E138K (Singh et al., 2014), HIV-1C harbors certain unique mutations specific to this genetic family. For instance, the K65R mutation, which confers resistance to nearly all the Nucleoside Reverse Transcriptase Inhibitors (NRTIs), except for Zidovudine, is observed in all subtypes. However, in HIV-1C, the frequency of occurrence of K65R is remarkably high (Garforth et al., 2010; Nikolenko et al., 2004; Singh et al., 2014). While two studies attributed this difference to the nature of the HIV-1C template (Coutsinos et al., 2009, 2011), the role of RT itself cannot be ruled out. A few other mutations uniquely associated with drug resistance in HIV-1C include G190A, Y181C, Y188L, and N348I. In contrast, mutations such as K103N/Q are significantly under-represented in HIV-1C, although they are readily represented in HIV-1B and HIV-1 CRF02_AG, further alluding to subtype-specific differences in the RT function (Singh et al., 2014). (K. A. Delviks-Frankenberry et al., 2013)Subtype-specific differences of RT function were illustrated elegantly when the drug-resistant mutations were introduced into HIV-1B or HIV-1C RTs; HIV-1C displayed a higher magnitude of reporter gene expression (K. A. Delviks-Frankenberry et al., 2013).

These observations collectively allude to the hypothesis that RT of different viral subtypes may be endowed with unique properties leading to unique or preferential phenotypic differences, such as the frequencies of sequence motif duplication. However, the fundamental nature of sequence duplications makes it extremely challenging to evaluate this question, as explored in the following section.

2.2 The Frequency of Sequence-motif Duplication

Sequence-motif duplications represent the outcome of a series of rare events occurring during reverse transcription of the HIV-1 RNA template. Sequence-motif duplication occurs when RT switches strands leading to inter-molecular recombination. As the cDNA dissociates from the donor RNA, misaligns with the acceptor RNA strand, and begins to polymerize, a sequence of RNA that has already been incorporated into the cDNA will be copied for the second time. It is an established fact that, on average, RT switches between the strands three to five times per round of viral replication. Assuming that the frequency of recombination is uniform across the viral genome, despite the existence of recombination hotspots (Klarmann et al., 2002), the probability of recombination could be estimated to be 0.03 – 0.05% per residue of a viral genome of 9,500 bases. Additionally, nonhomologous recombination, responsible for sequence motif duplication, occurs at a significantly reduced frequency, around 100 – 1000 times lower than that of homologous recombination (Zhang & Temin, 1993). Approximately half of these recombination events are expected to lead to duplications and the other half to deletions. Considering all these factors, the mean rate of sequence duplication occurring at a given location on the genome is approximately 0.000015% - 0.0025%.

Following an event of duplication, the selection represents an inevitable step in the process of viral evolution. Numerous other factors may further influence the replication fitness of a variant viral strain containing a duplication. More than 95% of proviruses are replication-deficient owing to large sequence deletions or debilitating mutations in essential genes (Imamichi et al., 2020; Pollack et al., 2017). Therefore, a variant viral strain containing a duplicated sequence must be among the remaining 5% of the intact viral strains that can establish productive infection to be selected. Additional factors, such as the site of integration, the immune response, etc, further lower the survival probability of a viral strain harbouring a duplication.

Despite the low frequencies and adverse events acting against them, sequences of variant viral strains containing motif duplication are not sparse in the sequence databases. This phenomenon may be ascribed to two factors. Firstly, the remarkable replication dynamics of HIV-1 ensure that approximately 10^{10} to 10^{12} new virions are produced each day in an infected person (Perelson et al., 1996). Of note, these numbers represent new virions produced by already infected cells. The number of new cells that are infected each day and, as a consequence, the number of reverse transcription events that occur daily, is still a subject of debate but has been postulated to be less than 10^9 (Coffin & Swanstrom, 2013). These large numbers make up, in part, for the extremely low frequencies of the occurrence of sequence duplications. Secondly, some of these variants, as has been described earlier, possess a replication advantage over the wild-type viral strains; hence, they are subject to strong positive selection. Work from our laboratory demonstrated that in a pair-wise competition of viral variants discordant for a PTAP duplication, the double-PTAP variants dominated the single-PTAP strains under several experimental conditions, including natural infection (Sharma et al., 2018). Thus, the strong positive selection pressure exerted by these variants ensures their propagation despite their low frequencies. This dominance may also be reflected in the transmission potential of these variants, a phenomenon that is presently under investigation in our laboratory.

The balance between the low frequency of occurrence of sequence duplications, accompanied by the strong positive selection once they appeared, ensures the replication fitness of these variant viral strains.

2.3 Rationale underlying the present study

The present work aims at filling an essential gap in the field of HIV-1 molecular biology. Several publications examining sequence duplications among subtypes attest to the natural propensity of HIV-1C to cause a higher frequency of duplications (Bachu, Mukthey, et al., 2012; Bachu, Yalla, et al., 2012; Boullosa et al., 2014; Martins et al., 2011; Sharma et al., 2017, 2018). However, the molecular causes underlying the higher duplication frequency in HIV-1C have not been uncovered. Since recombination is an essential pre-requisite for sequence duplication, there is a strong likelihood that C-RT recombines at a higher rate than that of other subtypes.

Several publications described unique amino acid polymorphisms in HIV-1C RT (Brenner et al., 2006; Iordanskiy et al., 2010; Nagata et al., 2017; Singh et al., 2014; Snoeck et al., 2006; Xu et al., 2010) however, none of these studies examined the molecular mechanisms causing sequence duplication or the impact of such events on viral replication fitness, except on drug resistance. Therefore, it becomes imperative to understand how these polymorphisms may play a role in affecting the properties of HIV-1C. Notably, of the more than 100 crystal structures of HIV-1 RT available in the Protein Data Bank, not one belongs to C-RT. This paucity makes it challenging to ascertain how amino acid variation in HIV-1C RT may modulate the properties of the viral enzyme.

These observations collectively warrant a comprehensive and detailed analysis of C-RT, with a special emphasis on its ability to duplicate sequences, considering their clinical significance. Despite the low number of publications examining HIV-1C, the molecular differences underlying the high-rate sequence duplication in this viral subtype are expected to be subtle. This assumption is based on the fact that several groups failed to see dramatic differences in biochemical properties of RT of diverse viral subtypes. Therefore, it is unlikely that the scientific community would have missed a dramatic difference in over 38 years of research on HIV-1.

Of note, the appearance of a rare event, such as a unique sequence-motif duplication, is expected to be selected immediately in natural infection, when the variation confers a significant replication advantage. To recapitulate the magnitude of viral replication is practically impossible in a cell culture experiment despite the propagation of the viral strain for several generations. Further, the various *cis* and *trans* factors that may regulate subtype-specific recombination, duplication, and selection processes are ill-defined, making it challenging to design an appropriate experiment to recapitulate the event. Unpublished work from our laboratory failed to detect the emergence of a 4 NF-κB variant viral strain even after four months of serial passaging of the Indie molecular clone of HIV-1C.

Thus, one must rely on indirect evidence and attempt to draw conclusions from probable causes rather than directly demonstrate the mechanisms underlying the high rate of sequence duplication in HIV-1C.

The present work examines the biological significance of a natural amino acid variation (T359) to several functions of HIV-1C RT. We show that the unique amino acid polymorphism may be responsible for the high frequency duplications observed in this subtype.

2.4 References

Anderson, J. A., Bowman, E. H., and Hu, W.-S. (1998). Retroviral Recombination Rates Do Not Increase Linearly with Marker Distance and Are Limited by the Size of the Recombining Subpopulation. J. Virol., 72(2), 1195–1202. http://jvi.asm.org/cgi/content/abstract/72/2/1195

Bachu, M., Mukthey, A. B., Murali, R. V., Cheedarla, N., Mahadevan, A., Shankar, S. K., Satish, K. S., Kundu, T. K., & Ranga, U. (2012). Sequence Insertions in the HIV Type 1 Subtype C Viral Promoter Predominantly Generate an Additional NF-κB Binding Site. AIDS Research and Human Retroviruses, 28(10), 1362–1368. https://doi.org/10.1089/aid.2011.0388

Bachu, M., Yalla, S., Asokan, M., Verma, A., Neogi, U., Sharma, S., Murali, R. V., Mukthey, A. B., Bhatt, R., Chatterjee, S., Rajan, R. E., Cheedarla, N., Yadavalli, V. S., Mahadevan, A., Shankar, S. K., Rajagopalan, N., Shet, A., Saravanan, S., Balakrishnan, P., … Ranga, U. (2012). Multiple NF-κB Sites in HIV-1 Subtype C Long Terminal Repeat Confer Superior Magnitude of Transcription and Thereby the Enhanced Viral Predominance. Journal of Biological Chemistry, 287(53), 44714–44735. https://doi.org/10.1074/jbc.M112.397158

Brindeiro, P. A., Brindeiro, R. M., Mortensen, C., Hertogs, K., De Vroey, V., Rubini, N. P. M., Sion, F. S., De Sá, C. A. M., Machado, D. M., Succi, R. C. M., & Tanuri, A. (2002). Testing genotypic and phenotypic resistance in human immunodeficiency virus type 1 isolates of clade B and other clades from children failing antiretroviral therapy. Journal of Clinical Microbiology, 40(12), 4512–4519. https://doi.org/10.1128/JCM.40.12.4512-4519.2002

Brenner, B. G., Oliveira, M., Doualla-Bell, F., Moisi, D. D., Ntemgwa, M., Frankel, F., Essex, M., & Wainberg, M. A. (2006). HIV-1 subtype C viruses rapidly develop K65R resistance to tenofovir in cell culture. Aids, 20(9), 9–13. https://doi.org/10.1097/01.aids.0000232228.88511.0b

Carl, S., Daniels, R., Iafrate, A. J., Easterbrook, P., Greenough, T. C., Skowronski, J., & Kirchhoff, F. (2000). Partial “repair” of defective NEF genes in a long-term nonprogressor with human immunodeficiency virus type 1 infection. Journal of Infectious Diseases, 181(1), 132–140. https://doi.org/10.1086/315187

Chen, Y., Balakrishnan, M., Roques, B. P., & Bambara, R. A. (2003). Steps of the acceptor invasion mechanism for HIV-1 minus strand strong stop transfer. Journal of Biological Chemistry, 278(40), 38368–38375. https://doi.org/10.1074/jbc.M305700200

Chin, M. P. S., Chen, J., Nikolaitchik, O. A., & Hu, W. S. (2007). Molecular determinants of HIV-1 intersubtype recombination potential. Virology, 363(2), 437–446. https://doi.org/10.1016/j.virol.2007.01.034

Chin, M. P. S., Rhodes, T., Chen, J., Fu, W., & Hu, W.-S. (2005). Identification of a Major Restriction in HIV-1 Inter-subtype recombination. Proc Natl Acad Sci U S A, 102(25), 9002–9007. https://doi.org/10.1073/pnas.0502522102

Coffin, J., & Swanstrom, R. (2013). HIV pathogenesis: Dynamics and genetics of viral populations and infected cells. Cold Spring Harbor Perspectives in Medicine, 3(1). https://doi.org/10.1101/cshperspect.a012526

Coutsinos, D., Invernizzi, C. F., Moisi, D., Oliveira, M., Martinez-Cajas, J. L., Brenner, B. G., & Wainberg, M. a. (2011). A template-dependent dislocation mechanism potentiates K65R reverse transcriptase mutation development in subtype C variants of HIV-1. PLoS ONE, 6(5). https://doi.org/10.1371/journal.pone.0020208

Coutsinos, D., Invernizzi, C. F., Xu, H., Moisi, D., Oliveira, M., Brenner, B. G., & Wainberg, M. a. (2009). Template usage is responsible for the preferential acquisition of the K65R reverse transcriptase mutation in subtype C variants of human immunodeficiency virus type 1. Journal of Virology, 83(4), 2029–2033. https://doi.org/10.1128/JVI.01349-08

Dang, Q., & Hu, W. (2001). Effects of homology length in the repeat region on minus-strand DNA transfer and retroviral replication. Journal of Virology, 75(2), 809–820. https://doi.org/10.1128/JVI.75.2.809-820.2001

Delviks-Frankenberry, K. A., Lengruber, R. B., Santos, A. F., Silveira, J. M., Soares, M. A., Kearney, M. F., Maldarelli, F., & Pathak, V. K. (2013). Connection subdomain mutations in HIV-1 subtype-C treatment-experienced patients enhance NRTI and NNRTI drug resistance. Virology, 435(2), 433–441. https://doi.org/10.1016/j.virol.2012.09.021

Delviks-Frankenberry, K. A., Nikolenko, G. N., & Pathak, V. K. (2010). The "connection" between HIV drug resistance and RNase H. Viruses, 2(7), 1476–1503. https://doi.org/10.3390/v2071476

Fan, J., Negroni, M., & Robertson, D. L. (2007). The distribution of HIV-1 recombination breakpoints. Infection, Genetics and Evolution, 7(6), 717–723. https://doi.org/10.1016/j.meegid.2007.07.012

Garforth, S. J., Domaoal, R. A., Lwatula, C., Landau, M. J., Meyer, A. J., Anderson, K. S., & Prasad, V. R. (2010). K65R and K65A substitutions in HIV-1 reverse transcriptase enhance polymerase fidelity by decreasing both dNTP misinsertion and mispaired primer extension efficiencies. Journal of Molecular Biology, 401(1), 33–44. https://doi.org/10.1016/j.jmb.2010.06.001

Hwang, C. K., Svarovskaia, E. S., & Pathak, V. K. (2001). Dynamic copy choice: Steady state between murine leukemia virus polymerase and polymerase-dependent RNase H activity determines frequency of in vivo template switching. Proceedings of the National Academy of Sciences of the United States of America, 98(21), 12209–12214. https://doi.org/10.1073/pnas.221289898

Imamichi, H., Smith, M., Adelsberger, J. W., Izumi, T., Scrimieri, F., Sherman, B. T., Rehm, C. A., Imamichi, T., Pau, A., Catalfamo, M., Fauci, A. S., & Clifford Lane, H. (2020). Defective HIV-1 proviruses produce viral proteins. Proceedings of the National Academy of Sciences of the United States of America, 117(7), 3704–3710. https://doi.org/10.1073/pnas.1917876117

Iordanskiy, S., Waltke, M., Feng, Y., & Wood, C. (2010). Subtype-associated differences in HIV-1 reverse transcription affect the viral replication. Retrovirology, 7(1), 85. https://doi.org/10.1186/1742-4690-7-85

Ji, Y., Han, X., Tian, W., Gao, Y., Jin, S., Zhang, L., & Shang, H. (2018). V4 region of the HIV-1 envelope gene mediates immune escape and may not promote the development of broadly neutralizing antibodies. Vaccine, 36(50), 7700–7707. https://doi.org/10.1016/j.vaccine.2018.10.084

Klarmann, G., Jetzt, A. E., Preston, B. D., Zhuang, J., Sun, G., Ron, Y., Dougherty, J. P., & Yu, H. (2002). Human Immunodeficiency Virus Type 1 Recombination: Rate, Fidelity, and Putative Hot Spots. Journal of Virology, 76(22), 11273–11282. https://doi.org/10.1128/jvi.76.22.11273-11282.2002

Lanciault, C., & Champoux, J. J. (2006). Pausing during Reverse Transcription Increases the Rate of Retroviral Recombination. Society, 80(5), 2483–2494. https://doi.org/10.1128/JVI.80.5.2483

Levy, D. N., Aldrovandi, G. M., Kutsch, O., & Shaw, G. M. (2004). Dynamics of HIV-1 recombination in its natural target cells. Proceedings of the National Academy of Sciences, 101(12), 4204–4209. https://doi.org/10.1073/pnas.0306764101

Martins, Angélica N., Arruda, M. B., Pires, A. F., Tanuri, A., & Brindeiro, R. M. (2011). Accumulation of P(T/S)AP Late Domain Duplications in HIV Type 1 Subtypes B, C, and F Derived from Individuals Failing ARV Therapy and ARV Drug-Naive Patients. AIDS Research and Human Retroviruses, 27(6), 687–692. https://doi.org/10.1089/aid.2010.0282

Martins, Angelica N., Waheed, A. A., Ablan, S. D., Huang, W., Newton, A., Petropoulos, C. J., Brindeiro, R. de M., & Freed, E. O. (2015). Elucidation of the Molecular Mechanism Driving Duplication of the HIV-1 PTAP Late Domain. Journal of Virology, 90(October), JVI.01640-15. https://doi.org/10.1128/JVI.01640-15

Mougel, M., Houzet, L., & Darlix, J.-L. (2009). When is it time for reverse transcription to start and go? Retrovirology, 6, 24. https://doi.org/10.1186/1742-4690-6-24

Nagata, S., Imai, J., Makino, G., Tomita, M., & Kanai, A. (2017). Evolutionary analysis of HIV-1 pol proteins reveals representative residues for viral subtype differentiation. Frontiers in Microbiology, 8(NOV), 1–10. https://doi.org/10.3389/fmicb.2017.02151

Neil, J. C., Fulton, R., Rigby, M., & Stewart, M. (1991). Feline leukaemia virus: Generation of pathogenic and oncogenic variants. Current Topics in Microbiology and Immunology, 171, 67–93. https://doi.org/10.1007/978-3-642-76524-7_4

Nikolenko, G. N., Svarovskaia, E. S., Delviks, K. A., & Pathak, V. K. (2004). Antiretroviral Drug Resistance Mutations in Human Immunodeficiency Virus Type 1 Reverse Transcriptase Increase Template-Switching Frequency. Journal of Virology, 78(16), 8761–8770. https://doi.org/10.1128/JVI.78.16.8761-8770.2004

Nikolenko, Galina. N., Delviks-Frankenberry, K. A., Palmer, S., Maldarelli, F., Fivash, M. J., Coffin, J., & Pathak, V. K. (2007). Mutations in the connection domain of HIV-1 reverse transcriptase increase 3’-azido-3’-deoxythymidine resistance. 104(1), 317–322.

Onafuwa-Nuga, A., & Telesnitsky, A. (2009). The Remarkable Frequency of Human Immunodeficiency Virus Type 1 Genetic Recombination. Microbiology and Molecular Biology Reviews, 73(3), 451–480. https://doi.org/10.1128/MMBR.00012-09

Palaniappan, C., Wisniewski, M., Wu, W., Fay, P. J., & Bambara, R. A. (1996). Misincorporation by HIV-1 reverse transcriptase promotes recombination via strand transfer synthesis. Journal of Biological Chemistry, 271(37), 22331–22338. https://doi.org/10.1074/jbc.271.37.22331

Perelson, A. S., Neumann, A. U., Markowitz, M., Leonard, J. M., & Ho, D. (1996). HIV-1 dynamics in vivo: Virion clearance rate, infected cell life-span and viral generation time. Science, 271(March), 1582–1586. www.sciencemag.org

Peters, S., Muñoz, M., Yerly, S., Lopez-galindez, C., Perrin, L., Larder, B., Cmarko, D., Fakan, S., Noz, M. M. U., & Perrin, L. U. C. (2001). Resistance to Nucleoside Analog Reverse Transcriptase Inhibitors Mediated by Human Immunodeficiency Virus Type 1 p6 Protein Resistance to Nucleoside Analog Reverse Transcriptase Inhibitors Mediated by Human Immunodeficiency Virus Type 1 p6 Protein. Journal of Virology, 75(20), 9644–9653. https://doi.org/10.1128/JVI.75.20.9644

Pollack, R. A., Jones, R. B., Pertea, M., Bruner, K. M., Martin, A. R., Thomas, A. S., Capoferri, A. A., Beg, S. A., Huang, S. H., Karandish, S., Hao, H., Halper-Stromberg, E., Yong, P. C., Kovacs, C., Benko, E., Siliciano, R. F., & Ho, Y. C. (2017). Defective HIV-1 Proviruses Are Expressed and Can Be Recognized by Cytotoxic T Lymphocytes, which Shape the Proviral Landscape. Cell Host and Microbe, 21(4), 494-506.e4. https://doi.org/10.1016/j.chom.2017.03.008

Racine, P.-J., Chamontin, C., de Rocquigny, H., Bernacchi, S., Paillart, J.-C., & Mougel, M. (2016). Requirements for nucleocapsid-mediated regulation of reverse transcription during the late steps of HIV-1 assembly. Scientific Reports, 6(1), 27536. https://doi.org/10.1038/srep27536

Rhodes, T. D., Nikolaitchik, O., Chen, J., Powell, D., & Hu, W.-S. (2005). Genetic Recombination of Human Immunodeficiency Virus Type 1 in One Round of Viral Replication: Effects of Genetic Distance, Target Cells, Accessory Genes, and Lack of High Negative Interference in Crossover Events. Journal of Virology, 79(3), 1666–1677. https://doi.org/10.1128/JVI.79.3.1666-1677.2005

Rodriguez-Rodriguez, L., Tsuchihashi, Z., Fuentes, G. M., Bambara, R. A., & Fay, P. J. (1995). Influence of Human Immunodeficiency Virus Nucleocapsid Protein on Synthesis and Strand Transfer by Reverse Transcriptase in Vitro. Journal of Biological Chemistry, 270(25), 15005–15011.

Schlub, T. E., Grimm, A. J., Smyth, R. P., Cromer, D., Chopra, A., Mallal, S., Venturi, V., Waugh, C., Mak, J., & Davenport, M. P. (2014). Fifteen to Twenty Percent of HIV Substitution Mutations Are Associated with Recombination. Journal of Virology, 88(7), 3837–3849. https://doi.org/10.1128/jvi.03136-13

Sharma, S., Aralaguppe, S. G., Abrahams, M.-R., Williamson, C., Gray, C., Balakrishnan, P., Saravanan, S., Murugavel, K. G., Solomon, S., & Ranga, U. (2017). The PTAP sequence duplication in HIV-1 subtype C Gag p6 in drug-naive subjects of India and South Africa. BMC Infectious Diseases, 17(1), 95. https://doi.org/10.1186/s12879-017-2184-4

Sharma, S., Arunachalam, P. S., Menon, M., Ragupathy, V., Satya, R. V., Jebaraj, J., Aralaguppe, S. G., Rao, C., Pal, S., Saravanan, S., Murugavel, K. G., Balakrishnan, P., Solomon, S., Hewlett, I., & Ranga, U. (2018). PTAP motif duplication in the p6 Gag protein confers a replication advantage on HIV-1 subtype C. Journal of Biological Chemistry, 293(30), 11687–11708. https://doi.org/10.1074/jbc.M117.815829

Singh, K., Flores, J. A., Kirby, K. A., Neogi, U., Sonnerborg, A., Hachiya, A., Das, K., Arnold, E., McArthur, C., Parniak, M., & Sarafianos, S. G. (2014). Drug resistance in non-B subtype HIV-1: Impact of HIV-1 reverse transcriptase inhibitors. Viruses, 6(9), 3535–3562. https://doi.org/10.3390/v6093535

Tamiya, S., Mardy, S., Kavlick, M. F., Yoshimura, K., & Mistuya, H. (2004). Amino Acid Insertions near Gag Cleavage Sites Restore the Otherwise Compromised Replication of Human

Immunodeficiency Virus Type 1 Variants Resistant to Protease Inhibitors. Journal of Virology, 78(21), 12030–12040. https://doi.org/10.1128/jvi.78.21.12030-12040.2004

Xu, H.-T., Quan, Y., Asahchop, E., Oliveira, M., Moisi, D., & Wainberg, M. A. (2010). Comparative biochemical analysis of recombinant reverse transcriptase enzymes of HIV-1 subtype B and subtype C. 1–11.

Zhang, J., & Temin, H. (1993). Rate and mechanism of nonhomologous recombination during a single cycle ofretroviral replication. 727(1984).

Zheng, Y. H., Sentsui, H., Nakaya, T., Kono, Y., & Ikuta, K. (1997). In vivo dynamics of equine infectious anemia viruses emerging during febrile episodes: insertions/duplications at the principal neutralizing domain. Journal of Virology, 71(7), 5031–5039. https://doi.org/10.1128/jvi.71.7.5031-5039.1997

Chapter – 3: Materials and Methods

3.0 Common materials used in the present work

In all experiments described in this section, we used the NL4-3 (GenBank accession number: AF324493.2) and Indie-C1 (AB023804.1) molecular clones as representatives of HIV-1B and -1C, respectively. The NL4-3 and NL4-3 ΔEnv EGFP molecular clones were obtained through the NIH-AIDS reagent program of the Division of AIDS, NIAID, NIH (Cat. Nos: ARP-2852 and ARP-11100, contributed by Dr. M Martin, and Dr. Haili Zhang, Dr. Yan Zhou and Dr. Robert Siliciano, respectively). The Indie-C1 molecular clone was a kind gift from Dr. Masashi Tatsumi of the International Medical Center of Japan, Tokyo. The Indie ΔEnv EGFP reporter vector was generated in our laboratory.Prof. Vinayaka Prasad, Albert Einstein College of Medicine, New York donated the RT expression vectors p6-HRT and p6-HRT-p51, that were originally generated by Stuart LeGrice and Fiona Gruninger-Leitch (1990).

All primers used in the thesis were synthesised at Sigma-Aldrich, Bangalore, India. All the enzymes used for cloning and PCR purposes were purchased from New England Biolabs, USA. The remaining sources of specific components used in some assays are indicated in parenthesis as necessary.

3.1 Ethical Clearance

The Institutional Biosafety Committee, JNCASR, approved the work performed in this thesis (Office order no – JNC/IBSC/2020/RUK-02 dt. 05/06/2020). A copy of the approval certificate is attached in the appendix section of this thesis.

3.2 Bioinformatics analysis

HIV-1 sequences were downloaded from the Los Alamos National Laboratory Sequence Database and aligned using Clustal Omega multiple sequence alignment software using the default settings provided. The results were analysed using BioEdit v. 7.2.5. The RT structure was visualized using the PyMol software v. 2.4.1. A high-resolution crystal structure of HIV-1B RT was downloaded from the Protein Data Bank (PDB ID – 5J2M). The HIV-1C RT signature amino acid residues were mapped on this structure using the mutate option of the program, followed by an energy minimization analysis. The bond length of the T359 residue with the PO_4 group of the DNA was determined by using the resources provided in the software.

3.3 Construction of RT expression vectors, EGFP reporter viral variants, and full-length viral variant strains.

All plasmid vectors were cloned in *E. coli* XL-10 Gold® cells. The cells were maintained at 30°C overnight. A detailed description of the methodology employed for the construction of each group of vectors is provided below,

- **Full-Length variant viral strains:** Panels of four full-length infectious viral variants were generated representing each of the NL4-3 and Indie molecular clones. The panel members are discordant for a single amino acid residue at position 359 of RT – Glycine, Threonine, Serine, or Alanine. The variations were introduced into the RT by an overlap-PCR using Phusion^{TM} DNA polymerase and the primer sets described in Table-1. The first PCR utilized an Outer Forward Primer (OFP) and an Inner Reverse Primer (IRP). The corresponding Inner Forward Primer (IFP) and Outer Reverse Primer (ORP) were used for the second PCR. The inner primers contained the mutations necessary to introduce the indicated change in the RT (see Table-1). The two PCR products were gel-purified and used as templates for the overlap-PCR using the outer primer pair. Restriction enzyme-mediated fragment substitution was employed to engineer a specific amino acid at position 359 using the AgeI and EcoRI sites and two PflMI sites for NL4-3 and Indie, respectively. Recombinant clones were identified by restriction digestion and confirmed by Sanger sequencing. The eight variant molecular clones thus generated were designated with a suffix 359 followed by the amino acid mutated at that position – for example, the Glycine variant of Indie was labelled Indie – 359G.

- **RT-expression Vectors:** vectors p6HRT and p6HRT-p51 express the p66 and p51 subunits, respectively, of the HIV-1 BH10 reverse transcriptase under inducible control of the *lac* promoter. These plasmids also harbour an N-terminal 6 X Histidine tag to enable purification of the expressed proteins. The p66 subunit was replaced with the corresponding NL4-3 or Indie RT variant using the BamHI and SalI enzyme sites on the p6HRT vector backbone, while the p51 subunits were cloned using BamHI and HindIII sites on the p6HRT-p51 plasmid.

Molecular Clone	AA at 359	Primer Name	Primer Sequence (5' – 3')
pIndie	-	N 4024 I – OFP	TTCTAAAAGAA**CCAGTACATGG**AGTATATTATGACCCAT
	-	N 4025 I – ORP	TATATCTTCTCAATCT**CCATTCTATGG**AGACTCCAT
	Ala	N 3097 IA – IRP	ATTAGTGTGGGCAGCCCTCCTTTTTG
		N 3098 IA – IFP	CAAAAAGGAGGGCTGCCCACACTAAT
	Ser	N 3099 IS – IRP	ATTAGTGTGGGCAGACCTCCTTTTTG
		N 4000 IS – IFP	CAAAAAGGAGGTCTGCCCACACTAAT
	Gly	N 4003 IG – IRP	ATTAGTGTGGGCCCTCCTCCTTTTTG
		N 4004 IG – IFP	CAAAAAGGAGGGGTGCCCACACTAAT
pNL4-3	-	N 4009 N-OFP	GGAGATTCTAAAAGA**ACCGGT**ACATGGAG
	-	N 4020 N – ORP	ACAGCAGTTGTTGCA**GAATTC**TTATTATGG
	Ala	N 4010 NA – IRP	ATTAGTGTGGGCAGCCTTCATTCTTGCAT
		N 4011 NA – IFP	ATGCAAGAATGAAGGCTGCCCACACTAAT
	Ser	N 4014 NS – IRP	ATTAGTGTGGGCACTCTTCATTCTTGCAT
		N 4015 NS – IFP	ATGCAAGAATGAAGAGTGCCCACACTAAT
	Thr	N 4016 NT – IRP	ATTAGTGTGGGCAGTCTTCATTCTTGCAT
		N 4017 NT – IFP	ATGCAAGAATGAAGACTGCCCACACTAAT

Table-1: The primers used for generating the full-length RT variant viral strains. The restriction enzyme sites of PflMI (for Indie), AgeI and EcoRI (for NL4-3) are highlighted in bold. The primer label represents a unique identification number, suffixed with I (for Indie) or N (for NL4-3), and the single letter code of the amino acid substituted. The sequences of the reverse primers represent the reverse-complement. OFP – Outer Forward Primer, ORP – Outer Reverse Primer, IFP – Inner forward Primer, IRP – Inner Reverse Primer.

Sixteen RT-expression vectors (2 RT subunits x 4 variants x 2 subtypes) were generated using Phusion^{TM} DNA polymerase to amplify the p66 and p51 subunits from the eight full-length viral variants described above. A unique restriction enzyme site was engineered into the reverse primer after the stop codon as a surrogate mark to represent a specific amino acid residue at position 359 (see Table-2). All the Indie RT vectors contained an additional SacI site before the unique restriction enzyme site to distinguish them from the vectors of the NL4-3 panel. The primer details used for cloning the RT expression constructs are presented (Table-2). Recombinant clones were identified by restriction digestion and confirmed by Sanger sequencing.

The 16 RT-variant expression vectors were designated as follows. The p6HRT and p6HRT-p51 derivatives were labelled with N or I (for NL4-3 or Indie, respectively), followed by a single letter representing the unique amino acid residue at position 359. For example, the p66 clone of the NL4-3 RT Serine variant was labelled as p6HRT-NS.

Molecular Clone	AA at 359	Primer Name	Primer Sequence (5' – 3')
pIndie	-	N 2603 I FP	CATCAC**GGATCC**CAGCTTCCAATCAGTCCCATTGAAACTGTACCAGTAAAATTAAAGC
	WT	N 2604 I66 WT RP	TAGGTA**GTCGAC**GAGCTCCTATAGCACTTTCCTGAT TCCACTACTTACTAATTTATCTACT
		N 2605 I51 WT RP	TAGGTA**AAGCTT**GAGCTCCTAGAAAGTTTCTACTCCTGCTATGGGATCTTTCTCCAGCTGG
	Gly (MluI)	N 2979 IG66 RP	TAGGTA**GTCGAC**GAGCTCACGCGTCTATAGCACTTTCCTGATTCCACTACTTACTAATTTATCTACT
		N 2978 IG51 RP	TAGGTA**AAGCTT**GAGCTCACGCGTCTAGAAAGTTTCTACTCCTGCTATGGGATCTTTCTCCAGCTGG
	Ala (BsiWI)	N 4005 IA66 RP	TAGGTA**GTCGAC**GAGCTCCGTACGCACTTTCCTGATTCCACTACTTACTAATTTATCTACT
		N 4006 IA51 RP	TAGGTA**AAGCTT**GAGCTCCGTACGCTAGAAAGTTTCTACTCCTGCTATGGGATCTTTCTCCAGCTGG
	Ser (EagI)	N 4007 IS66 RP	TAGGTA**GTCGAC**GAGCTCCGGCCGCTATAGCACTTTCCTGATTCCACTACTTACTAATTTATCTACT
		N 4008 IS51 RP	TAGGTA**AAGCTT**GAGCTCCGGCCGCTAGAAAGTTTCTACTCCTGCTATGGGATCTTTCTCCAGCTGG
pNL4-3	-	N 3021 N FP	ATATAT**GGATCC**CCCATTAGTCCTATTGAGACTGTACC AGTAAAA
	WT	N 3024 N66 WT RP	ATATAT**GTCGAC**TACTTTCCTGATTCCAGCACTGACC
		N 3025 N51 WT RP	ATATAT**AAGCTT**CTAGAAAGTTTCTGCTCCTATTATGGG TTCTTTCTC
	Thr (SpeI)	N 2985 NT66 RP	TAGGTA**GTCGAC**ACTAGTTCATAGTACTTTCCTGATTCC AGCACTGACCAATTTATCTACT
		N 2984 NT51 RP	TAGGTA**AAGCTT**ACTAGTCTAGAAAGTTTCTGCTCCTAT TATGGGTTCTTTCTCTAACTGGTA
	Ala (BsiWI)	N 4018 NA66 RP	TAGGTA**GTCGAC**CGTACGTCATAGTACTTTCCTGATTCC AGCACTGACCAATTTATCTACT
		N 4019 NA51 RP	TAGGTA**AAGCTT**CGTACGCTAGAAAGTTTCTGCTCCTAT TATGGGTTCTTTCTCTAACTGGTA
	Ser (EagI)	N 4021 NS66 RP	TAGGTA**GTCGAC**CGGCCGTCATAGTACTTTCCTGATTCC AGCACTGACCAATTTATCTACT
		N 4022 NS 51 RP	TAGGTA**AAGCTT**CGGCCGCTAGAAAGTTTCTGCTCCTA TTATGGGTTCTTTCTCTAACTGGTA

Table–2: The primers used for generating the RT-expression vectors. The restriction enzyme sites of BamHI (for the forward primer), and SalI and HindIII (for the p66 and p51 reverse primers, respectively) are highlighted in bold. The diagnostic RE sites (underlined) serve as surrogate marks for the unique amino acid residue present at position 359–. The Indie RT panel contains an additional SacI enzyme site (also underlined) immediately upstream of the RE mark. The primer label includes a unique number for identity, suffixed with I (for Indie) or N (for NL4-3), followed by the RT subunit they encode (66 or 51). The sequences of the reverse primers represent the reverse-complement. FP – Forward Primer, RP – Reverse Primer.

- **EGFP-reporter variant viral panels:** Two reporter vectors - NL4-3 ΔEnv EGFP and Indie ΔEnv EGFP were modified sequentially at the RT and the GFP regions to generate the final reporter constructs. Point mutations were introduced into NL4-3 RT by domain swapping the AgeI and EcoRI restriction fragment of the reporter vector NL4-3 ΔEnv EGFP with the corresponding RT mutant generated for the full-length viral variants described above. A similar strategy was used to construct the Indie ΔEnv EGFP reporter panel by swapping PflMI restriction fragments with the full-length RT viral variants.

  Each of the eight intermediate RT variant viral reporters thus generated was used as the parental vector to construct two EGFP mutant ORFs. These two mutants contained a frameshift mutation at amino acid position 4 or 204 of the EGFP. The mutations in the Indie panel at pos. 4 and 204 were introduced using a direct and an overlap PCR, respectively. In the NL4-3 panel, both the mutations were introduced using an overlap-PCR. The PCR fragments were cloned directionally using the EcoRI and NheI sites or the SphI and StuI sites into NL4-3 ΔEnv EGFP and Indie ΔEnv EGFP, respectively (Figure-9). The primer sequences used for cloning the EGFP variants are presented (Table-3). Recombinant clones were identified by restriction digestion and confirmed by Sanger sequencing.

  The plasmids were then labelled to indicate the amino acid variation at position 359, followed by a hyphen denoting the position of the mutation in GFP (0 for WT, 4, or 204). For instance, the Alanine variant containing a mutation at position 204 of Indie ΔEnv EGFP was designated pIndie ΔEnv EGFP A204.

3.4 Expression and Purification of Recombinant RTs

The p6HRT and p6HRT-p51 RT expression constructs were transformed separately into *E. coli* M15 competent cells (Stuart & Friona, 1990). The protein expression was induced by adding IPTG to a final concentration of 1 mM when the cells reached an OD_{600} of 0.4. After a 5-hour induction with shaking at 37°C, the cells were harvested and resuspended in 50 mM Sodium Phosphate buffer (pH 8.0) containing 150 mM NaCl and 1 mM PMSF. The cell suspension was maintained at 4°C from this point forth. The cells were lysed by sonication (VCX - 130, Sonics and Materials Inc, USA) using a probe of 1.3 cm diameter (Vibracell 630-0569) and 30 cycles of 10 s on-off pulses and at an amplitude of 52 W. After sonication, the cell lysate was pelleted at 12,000 g for 10 min using a Sorvall Evolution RC centrifuge.

The cell-free supernatants of the two subunits were pooled and mixed with the Ni-NTA resin (cat. no – 786-940, G-Biosciences, USA) at a 0.5 ml Resin/200 ml cell lysate. The protein was allowed to bind to the resin overnight by incubation at 4°C with gentle agitation. The lysate-resin mix was then poured into a chromatographic column, and the unbound fraction was allowed to flow through. The resin was washed with 50 ml wash buffer containing 50 mM Sodium-Phosphate (pH 8.0), 20 mM Imidazole, 150 mM NaCl and 1 mM PMSF. The resin was incubated with 10 ml of an elution buffer containing 50 mM -Phosphate (pH 8.0), 250 mM Imidazole, and 150 mM NaCl, for 10 min, before the elute was collected. Four sequential eluted fractions were collected, pooled, and dialysed against 2 litres of dialysis buffer containing 50 mM Tris-Cl (pH 7.0), 25 mM NaCl, 1 mM EDTA, and 20% glycerol. The dialysed protein was purified further using Ion-Exchange Chromatography. One hundred µl of DEAE-Sepharose resin was added to the purified protein extract, and the suspension was incubated at 4°C with gentle agitation for 4 hours. The protein-resin mixture was then applied to a chromatography column, and the flow-through was collected. The enzyme was expected to be present in the non-binding fraction while most contaminants remain bound to the resin. The protein was dialysed against four litres of the storage buffer containing 50 mM Tris-Cl (pH 8.0), 150 mM NaCl, 1 mM DTT, and 50% glycerol (Stahlhut, Mark; Olsen, 1996), and concentrated using an Amicon® protein concentrator (cat. no ACS 505024, Merck-Millipore, USA) to a final volume of 500 µl. The protein concentration of the preparation was estimated by densitometric analysis against a BSA standard and stored in aliquots at -20°C until further use.

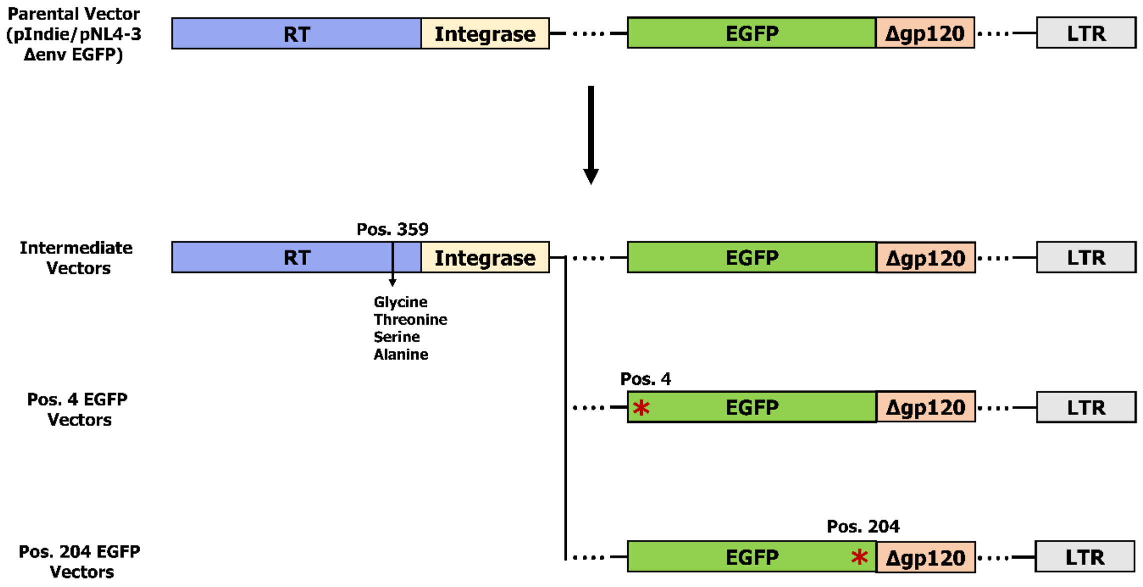


Figure-9: Schematic representation of the construction of panels of variant EGFP reporter viral vectors. Using parental vectors Indie or NL4-3, a panel of four variant strains was generated containing one of four amino acid residues at pos. 359 of RT (shown in blue). The two panels were further engineered to produce a pair of EGFP mutants of each RT variant, harbouring a frameshift mutation (represented by red asterisks) at amino acid position 4 or 204 in EGFP reading frame, as depicted. Only a productive recombination event between the two debilitating mutations can produce EGFP expression.

Molecular Clone	Primer Name	Primer Sequence (5' – 3')
Indie	N 4340 I – OFP	TACTCTGT**GCATGC**GCCACCATGGTGAGC
	N 4341 I4 – FP	ACTCTGT**GCATGC**GCCACCATGGTGAGCAAGGGACGAGGAGCTGTT
	N 4354 I - ORP	ATGCTAGT**AGGCCT**CTACTTGTACAGCTCGTCCATGCCGAGAGTG
Common	N 4353 Pos. 204 - IRP	TTGCTCAGGGCGGACATGGGTGCTCAGGTAG
	N 4352 Pos. 204 - IFP	CTACCTGAGCACCCATGTCCGCCCTGAGCAA
NL4-3	N 4355 N - OFP	GGAAGCCATAATAA**GAATTC**TGCAACAACTGCTGTTT
	N 4357 Pos. 4 - IRP	GGTGAACAGCTCCTCGTCCCTTGCTCACCATG
	N 4356 Pos. 4 - ORP	CATGGTGAGCAAGGGACGAGGAGCTGTTCACC
	N 4358 N - ORP	ATGCTTGT**GCTAGC**CTACTTGTACAGCTCGTCCATGCCGAGAG

Table-3: The primers used for generating the near-full-length reporter viral panels. The restriction enzyme sites of SphI and StuI (for the pIndie ΔEnv EGFP panel), EcoRI and NheI (for the pNL4-3 ΔEnv EGFP panel) are highlighted in bold. The primer label includes a unique number, suffixed with I (for Indie) or N (for NL4-3), followed by a number indicating the amino acid position of the frameshift mutation. The sequences of the reverse primers represent the reverse-complement. FP – Forward Primer, OFP – Outer Forward Primer, ORP – Outer Reverse Primer, IFP – Inner forward Primer, IRP – Inner Reverse Primer.

3.5 Determination of RT Activity

Four µg of polyuridylic acid (cat. no – P9528, Sigma-Aldrich, USA) was annealed with 1 µg (≈ 80 pmol.) of polydA$_{20}$ primer by maintaining the mixture at 95°C before rapid chilling on ice. The polyU/dA template primer pair thus generated was used in a 20 µl reaction mixture consisting of 25 mM Tris-Cl (pH 8.0), 3 mM $MgCl_2$, 100 mM KCl, and 250 µM dATP (cat. no – FC10HJ, GeNei Labs, India,) supplemented with 1 µCi of [α - ^{32}P] labelled dATP (cat. no – LCP103, Board of Radiation & Isotope Technologies, India). The reaction was initiated by adding varying amounts (0.1, 0.25, 0.5, 1 µl) of the RT enzyme and allowed to incubate at 37°C for 10 min. The reaction was stopped by adding 2 µl of 0.5 M EDTA.

Five µl of the reaction mixture was spotted on a nylon nucleic acid transfer membrane (Hybond-N+, Amersham Lifesciences, USA) of approximately 1.5 cm x 1.5 cm and allowed to air dry. The strips were combined and washed twice with 30 ml of wash buffer containing 150 mM NaCl, 20 mM Na-Citrate, and 1% SDS. The strips were then washed twice with 100% ethanol and air-dried. The dried strips were placed in 1.5 ml microcentrifuge tubes containing 0.5 ml of scintillation fluid (Ultima Gold^{TM} XR, Perkin Elmer®, USA). The quantity of [α - ^{32}P] labelled dATP incorporated was estimated by liquid scintillation spectrometry using a MicroBeta^2 liquid scintillation counter (Perkin Elmer, USA). The specific activity was calculated by estimating the number of enzyme units based on the amount of dATP incorporated and normalising it with the unit mass (1 µg) of the enzyme (Stuart, Le Grice, Craig E Cameron, 1995). One unit of enzyme activity was defined as an amount of enzyme that incorporates 1 pmol of the radioactive precursor into the final product in 10 min at 37°C.

The rate of cDNA synthesis at a fixed ATP concentration was determined using the same experimental conditions described above but keeping the volume of RT constant at 1 µl and measuring the amount of [α - ^{32}P] labelled dATP incorporated at 5-minute intervals.

3.6 Cell Culture

The T-cell lines, CEM-CCR5 and Jurkat, were maintained in RPMI 1640 medium (cat. no – R4130, Sigma-Aldrich, USA), and the epithelial cell lines, HEK 293T and TZM-bl were cultured in DMEM (cat. no – D1152, Sigma-Aldrich, USA). The culture media were supplemented with 10% fetal bovine serum (FBS, cat. no - 04-121-1A, Life Technologies, India), 2 mM Glutamine (cat. no – G6392, Sigma-Aldrich, USA), and 100 µg/ml each of penicillin G (cat. no – 13752, Sigma-Aldrich, USA) and Streptomycin (cat. no S9137, Sigma-Aldrich, USA). Peripheral Blood Mononuclear Cells (PBMCs) were isolated from 20 ml of fresh blood from healthy donors. The CD8+ cells were depleted using the RosetteSep™ Human CD8 Depletion Cocktail (cat. no – 15663, Stemcell Technologies, Canada). PBMCs were maintained for 72 hours in RPMI 1640 supplemented with 20 U/ml of Interleukin - 2 (cat. no – H7041, Sigma-Aldrich, USA), 5 µg/ml of Phytohaemagglutinin – P (PHA-P, cat. no – L1668, Sigma-Aldrich, USA) and FBS, Glutamine and the antibiotics at concentrations described above. Cells were maintained without PHA-P supplementation for viral infection. All cells were maintained at 37°C and 5% CO_2.

3.7 Virus production and titration of viral stocks

Infectious viral stocks of the eight molecular clones were produced by transient transfection of the HEK 293T cells. 3 x 10^6 cells were seeded in a 100 mm dish, and 20 µg of the plasmid DNA and 30 ng of pCMV-TdTomato (an internal control for transfection efficiency) were co-transfected using the calcium-phosphate protocol (Kingston et al., 2003).

The pseudotyped viral variants used for the recombination assay were produced in 6-well dishes. 0.3 x 10^6 HEK 293T cells were transfected with a mixture consisting of 3 µg of the reporter vector, 1 µg of the pCMV VSV-G envelope expression plasmid, and 10 ng of pCMV-TdTomato. Virus particles copackaging the two RNA strands, each containing one of the two debilitating mutations in the GFP ORF, were produced by mixing equal amounts (1.5 µg) of the vectors. The viral stocks of each member of the panel were produced as described above.

The culture medium was replaced with fresh medium 6 hours following transfection. Cell-culture supernatant containing the virus was harvested after 48 hours, passed through a 0.22 µM syringe filter, and stored in aliquots at -80°C until further use.

The amount of p24 protein was quantitated using p24 ELISA (4th generation p24 ELISA kit, J. Mitra and Co. Pvt Ltd., India). Viral infection was confirmed by measuring luciferase secretion into the medium following the infection of TZM-bl cells. Briefly, 10^4 cells were seeded per well of a 96-well plate in a total volume of 100 µl. A 4-fold serial dilution of the viral stock was used to infect the cells in the presence of 10 µg/ml of DEAE-dextran. The culture medium was replaced with complete DMEM 8 hours after infection. The plates were incubated at 37°C and at a CO_2 concentration of 5% for 48 hours. The medium was aspirated, and 100 µl of 1X passive lysis buffer (Catalogue. no – E1941, Promega Corporation, USA) was added to each well to lyse the cells. The wells were incubated with the lysis buffer for 10 minutes. Ten µl of the cell lysate was mixed with 10 µl of firefly luciferase substrate (Catalogue. no – E1500, Promega Corporation, USA). The luminescence was measured using a Varioskan Lux multimode reader (Thermo Fisher Scientific, USA).

3.8 The Polymerase Stall Assay

A 155 bp region from both HIV-1B and -1C LTRs was selected for amplification encompassing the RBE-III, NF-κB, and Sp1 binding sites of both the subtypes (HXB2 coordinates: 290 – 444). This region was amplified using forward primers containing the T7 promoter sequence (See Table-4). The PCR product was purified by the phenol-chloroform extraction method and used as a template for *in vitro* transcription. The assay was performed using the HiScribe™ T7 In Vitro Transcription Kit (cat. no – E2030, New England Biolabs, USA). The *in vitro* transcribed RNA was purified using QIAamp viral RNA mini kit (Catalogue. no – 52904, Qiagen, Germany) and stored in aliquots at -80°C until use.

Two µg of the reverse primers N2873 and N2875 were end-labelled using 1 Unit of T4 Polynucleotide Kinase (cat. no – M0201S, New England Biolabs, USA) and 2 µCi of [γ – ^{32}P] ATP (cat. no – LCP101, Board of Radiation & Isotope Technologies, India) by incubating the vial at 37°C for 1 hr. The enzyme was heat-inactivated by incubating the vial at 65°C for 20 min, and the labelled primers were purified by gel filtration through a Sephadex® G-50 column (cat. no – G50150, Sigma-Aldrich, USA). The labelling efficiency was estimated by Cerenkov counting using a MicroBeta^2 liquid scintillation counter (Perkin Elmer, USA).

A probe amount equivalent to 30,000 CPM was hybridised to 2 µg of appropriate template RNA by incubating the mixture at 80°C followed by rapid chilling on ice. The reaction was initiated by adding 1 Unit of HIV-1B or -1C RT in the presence of 25 mM Tris-Cl (pH 8.0), 3 mM $MgCl_2$, 100 mM KCl and 250 µM dATP (Catalogue. no – FC10HJ, GeNei Labs, India) in a reaction volume of 20 µl. The tubes were incubated in a PCR machine for 20 min at 37°C, and the reaction was stopped by adding 2 µl of 0.5 M EDTA.

Five µl of the reaction mixture was resolved on a 12% Urea-denaturing polyacrylamide-sequencing gel, and the bands were visualised using autoradiography.

3.9 The Recombination Assay

Ten ng p24 equivalent of the NL4-3 and Indie ΔEnv EGFP variant viral strains described in section 2.5 were used independently to infect 0.3 x 10^6 CEM-CCR5 or Jurkat cells. The culture medium was supplemented with 10 µg/ml of DEAE-dextran in a total volume of 2 ml in a 6-well dish. The medium was replaced with fresh RPMI 6 hrs post-infection. Twenty-four hours before the flow-

Molecular Clone	Primer Name	Primer Sequence (5' – 3')
pIndie	N 2872 I - FP	**TAATACGACTCACTATAGG**CATGGCCCGCGAGCTACATCCG
	N 2873 I - RP	GAAAAGCAGCGGCTTATATGCCGCA
pNL4-3	N 2874 N - FP	**TAATACGACTCACTATAGG**CATGGCCCGAGAGCTGCATCCG
	N 2875 N - RP	AAAAAGCAGCTGCTTATATGTAGCA

Table-4: The primers used for generating the templates for *in vitro* transcription. The T7 promoter is highlighted in bold. The primer label includes a unique number, suffixed with I (for Indie) or N (for NL4-3), followed by FP – Forward Primer or RP - Reverse Primer. The sequences of the reverse primers represent the reverse-complement.

cytometry analysis, the cells were activated with a cocktail of activators containing TNF-α (10 ng/ml, cat. no - 130-094-014, Miltenyi Biotec, USA), PMA (5 ng/ml, cat. no – P1585, Sigma-Aldrich, USA) and HMBA (5 mM, cat. no – H4663, Sigma-Aldrich, USA). The cells were centrifuged at 500 g, washed twice with Phosphate-Buffered Saline (PBS) solution, and resuspended in 250 µl of PBS containing 2% FBS before their analysis on a BD FACS ARIA III (Becton, Dickinson and Company, USA) flow cytometer.

3.10 The Strand-Transfer Assay

Two regions in the Indie genome (HXB2 coordinates: Regions 256 to 462 and 473 to 603) were amplified using forward primers containing the T7 promoter (see Table-5) to facilitate their use in *in vitro* transcription. A sequence of 20 bp at the 3' terminal of the second region was homologous to the first 20 bp at the 5' end of the first region to permit strand transfer during reverse transcription. The RNAs were generated and purified as described in section 2.8. The two RNA fragments were mixed at an equimolar ratio, and cDNA synthesis was initiated using 1 Unit of RT and an antisense primer N4864 that anneals to RNA-1. The reaction conditions were as described previously in section 2.8. Two µl of this reaction product was diluted 1,000-fold and used as the template for two separate Real-Time PCRs. The primer pair N4863 and N4864 amplified the RNA-1 template. A different PCR used the sense primer (N4865) binding the RNA-2 template and the anti-sense primer (N4864) annealing to the RNA-2 template. The PCR will be possible only when the RT successfully switches the strands during reverse transcription using the 20 bp overlap between the two RNA templates (see Figure-10).

3.11 The Mismatch Extension Assay

A 105 bp region of the Indie genome (HXB2 coordinates - 1809 to 1914) was synthesised as an oligonucleotide (Sigma-Aldrich, India) and used as the template in the assay. A target region of four sequential bases on the template – ATGC was selected for initiation of polymerization. Sixteen different reverse primers, divided into four groups, were designed such that each group of primers initiated polymerization at one of the four different sequential bases on the template. The four primers within each group were identical except for the terminal base at the 3' end of the primer. The 16 template-primer pairs thus generated represent, at the 3' end of the primer, the 16 different combinations of base pairs that are possible with the four bases. Of these, four pairings represent correct (A-T, T-A, G-C, C-G) matches, while the rest of the twelve are mismatches. The primer details are presented in Table-6. Four µg of the template was annealed to 1 µg of the primer as described in section 2.8. The primer extension was performed using the eight variant RT preparations described in section 2.8.

Primer Name	Primer Sequence (5' – 3')
N 4859 RNA 1 T7 FP	A**TAATACGACTCACTATAGA**CTCTGGTAACTAGAGATCAGAAGTATTAAAGT GGAAGTTTGACAGTCAGCT
N 4860 RNA 1 T7 RP	AGAGACCCAGTACAAGCGAAAAGCAGC
N 4861 RNA 2 T7 FP	A**TAATACGACTCACTATAG**ATCTGAGCCTGGGAGCTCTCTG
N 4862 RNA 2 T7 RP	TCTGAGGGATCTCTAGTTACCAGAGTCACACAATAGAC
N 4863 RNA 1 FP	AGAAGTATTAAAGTGGAAGTTTGACAGTC
N 4865 RNA 2 FP	GATCTGAGCCTGGGAGCTCTCTG

Table – 5: The primers used for generating the templates for the strand transfer assay. The T7 promoter is highlighted in bold. The 20 bp homology region is underlined in the FP of RNA 1. The primer label includes a unique number for internal reference, suffixed with thetemplate identity – RNA 1 or 2, followed by FP – Forward Primer or RP - Reverse Primer.

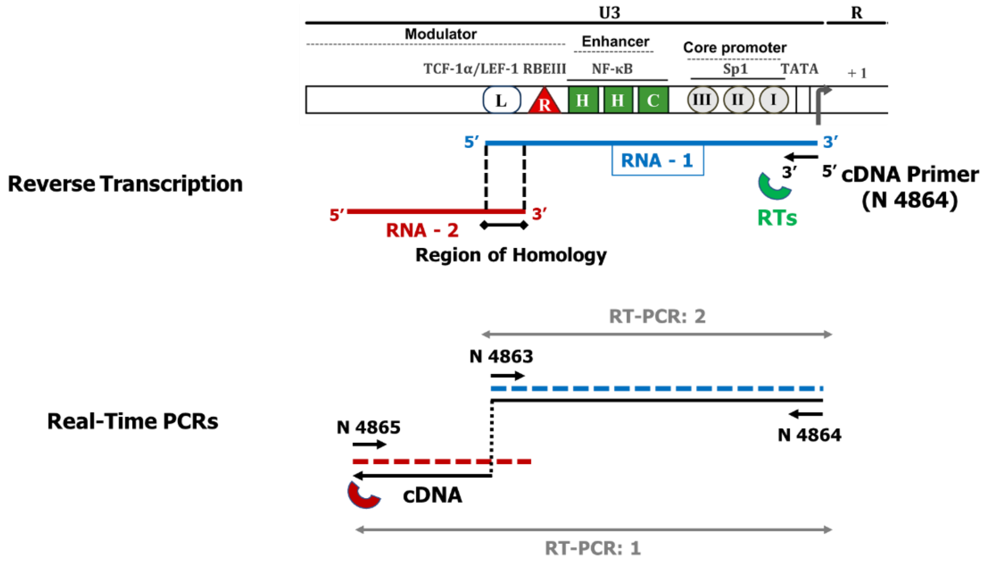


Figure – 10: Schematic of the Strand Transfer assay. The schematic depicts the two RNA molecules (green and blue) containing a 20 bp overlap (highlighted with the vertical dashed lines and spanning the LTR as shown). After the reverse transcription is initiated by the reverse primer N4864, the full-length PCR of primer pair N4865 and N4864 would be possible only when the RT successfully switches to RNA template-2 when it reaches the end of RNA template-1. The PCR of primer pair N4863 and N4864 is used as a normalization control in the assay. The three primers, the target locations of binding and the orientation of the primers are depicted as black arrows. the RT-PCR amplicons are shown in grey and the cDNA is shown in black

Primer Group	Primer Name	Primer Sequence (5' – 3')
-	N 4921 - Template	CAGTATGGTACTGTTTGCTTGGCTCATTGCCTCAGCCAACACTC TTGCTTTGTGGCCAGGTCCTCCCACTCCCTGAC**ATGC**TGTCATC ATCTCTTCTAATGAAGC
Group I	N 4922 C – T	GCTTCATTAGAAGAGATGATGACAT
	N 4923 C – G	GCTTCATTAGAAGAGATGATGACAG
	N 4924 C – A	GCTTCATTAGAAGAGATGATGACAA
	N 4925 C – C	GCTTCATTAGAAGAGATGATGACAC
Group II	N 4926 G – T	CTTCATTAGAAGAGATGATGACAGT

	N 4927 G – G	CTTCATTAGAAGAGATGATGACAGG
	N 4928 G – A	CTTCATTAGAAGAGATGATGACAGA
	N 4929 G – C	CTTCATTAGAAGAGATGATGACAGC
Group III	N 4930 T – T	TTCATTAGAAGAGATGATGACAGCT
	N 4931 T – G	TTCATTAGAAGAGATGATGACAGCG
	N 4932 T – A	TTCATTAGAAGAGATGATGACAGCA
	N 4933 T – C	TTCATTAGAAGAGATGATGACAGCC
Group IV	N 4934 A – T	TCATTAGAAGAGATGATGACAGCAT
	N 4935 A – G	TCATTAGAAGAGATGATGACAGCAG
	N 4936 A – A	TCATTAGAAGAGATGATGACAGCAA
	N 4937 A – C	TCATTAGAAGAGATGATGACAGCAC

Table – 6: The primers used for the mismatch extension assay. The four sequential bases ATGC on the template where the four groups of primers initiate polymerisation are highlighted in bold. The reverse primer Groups I, II, III, IV initiate polymerisation at C, G, T and A, respectively, on the template. The primer label includes a unique number for reference, followed by the template-primer base pair that the primer forms.

3.12 References

Kingston, R. E., Chen, C. A., & Okayama, H. (2003). Foreword. In *Current Protocols in Molecular Biology*. https://doi.org/10.1016/b978-0-12-394380-4.50005-0

Stahlhut, Mark; Olsen, D. (1996). Expression and Purification of Retroviral HIV-1 Reverse Transcriptase. Methods in Enzymology, 275, 122–133.

Stuart, Le Grice, Craig E Cameron, S. J. B. (1995). Purification and Characterization of Human Immunodeficiency Virus Type 1 Reverse Transcriptase. Methods in Enzymology, 262(1987), 130–144.

Stuart, L., & Friona, G.-L. (1990). Rapid purification of homodimer and heterodimer HIV-1 reverse transcriptase by metal chelate affinity chromatography. European Journal of Biochemistry, 314(187), 307–314. https://doi.org/10.1111/j.1432-1033.1990.tb15306.x

Chapter – 4: Results

4.1 The Subtype identity of the template, not RT, determines polymerase stalling.

Several groups demonstrated an increase in recombination frequency at specific locations on the viral RNA, where RT stalls during polymerisation (Chen et al., 2003; Lanciault & Champoux, 2006; Negroni & Buc, 1999; Viguera et al., 2001). This observation led us to speculate that the difference in the duplication frequencies spanning the genomes of HIV-1B and -C could be due to differential stalling locations since recombination is an essential pre-condition for sequence duplication. The transcription factor binding site variations in the enhancer region between the two subtypes, such as the presence of an additional NF-κB site in HIV-1C, further support this hypothesis since these changes could allow the formation of different RNA secondary structures. These secondary structures could further lead to stall sites distinct between the two templates, leading to increased recombination. Such an increased recombination frequency at specific genome locations may cause the duplication of sequence motifs, which may be subjected to Darwinian selection if the gain-of-function advantage is significant. Of note, although the subtype-associated differences in the RT function may also underlie or influence such differences, the available evidence rules out such an outcome. Additionally, the possibility of discordant stalling of the two RTs when the RNA template remains the same also need to be examined.

We performed a primer extension assay to examine whether the high-frequency duplication of select sequence motifs observed in HIV-1C, compared to HIV-1B, could be explained by the differential stalling of RTs on the template RNA. In the assay, we used recombinant HIV-1B and -1C RTs and two RNA templates transcribed in vitro as templates representative of the two HIV-1 subtypes. The template region centered around the duplication hotspots of the LTR region (HXB2 coordinates: 290-443), encompassing the RBE-III, NF-κB, and Sp1 binding sites of both the subtypes (NL4-3 and Indie-C1). The forward primers used in the PCR contained a T7 RNA polymerase promoter (Figure-11A and B). The two RNA templates were transcribed in vitro and purified using commercial kits. Primer extension was performed using four combinations of the two in vitro synthesized RNA templates combined with either of the two purified RTs. Two subtype-specific reverse primers, each radio-labelled at the 5’ end, were employed in the assay. The end-products were resolved on a 10% urea denaturing polyacrylamide gel and visualized by autoradiography (Figure-11D).

Between the free primers at the bottom of the gel and the fully extended products at the top, several partially polymerized extension products were visible in all four lanes (Figure-11D). While some of the bands were faint, others were quite prominent - alluding to strong stalling of polymerization at the corresponding sites in the template. Importantly, the profiles of the bands were nearly identical when the template RNA was the same, regardless of the RT difference (Figure-11C, compare lanes 1 vs. 2 or 3 vs. 4). Lanes 1 and 2 represent the RNA template of HIV-1B, whereas lanes 3 and 4 are that of HIV-1C. Therefore, the stalled products were similar for a given template irrespective of the RT used.

Importantly, multiple differences were evident when the stalling patterns of the two templates were compared. (Figure-11C, lanes 1 or 2 vs 3 or 4). For example, lanes 1 and 4 represented the templates of HIV-1B and -1C, respectively, polymerized by B-RT. The profiles of the extended products in these two lanes were different despite being catalysed by the same RT. Likewise, the profiles of the extended products of lanes 2 and 3 differed in a few locations even though C-RT catalysed both. In summary, the locations of stalling sites on each template, were dictated by the nature of the RNA template, and not the subtype identity of the RT.

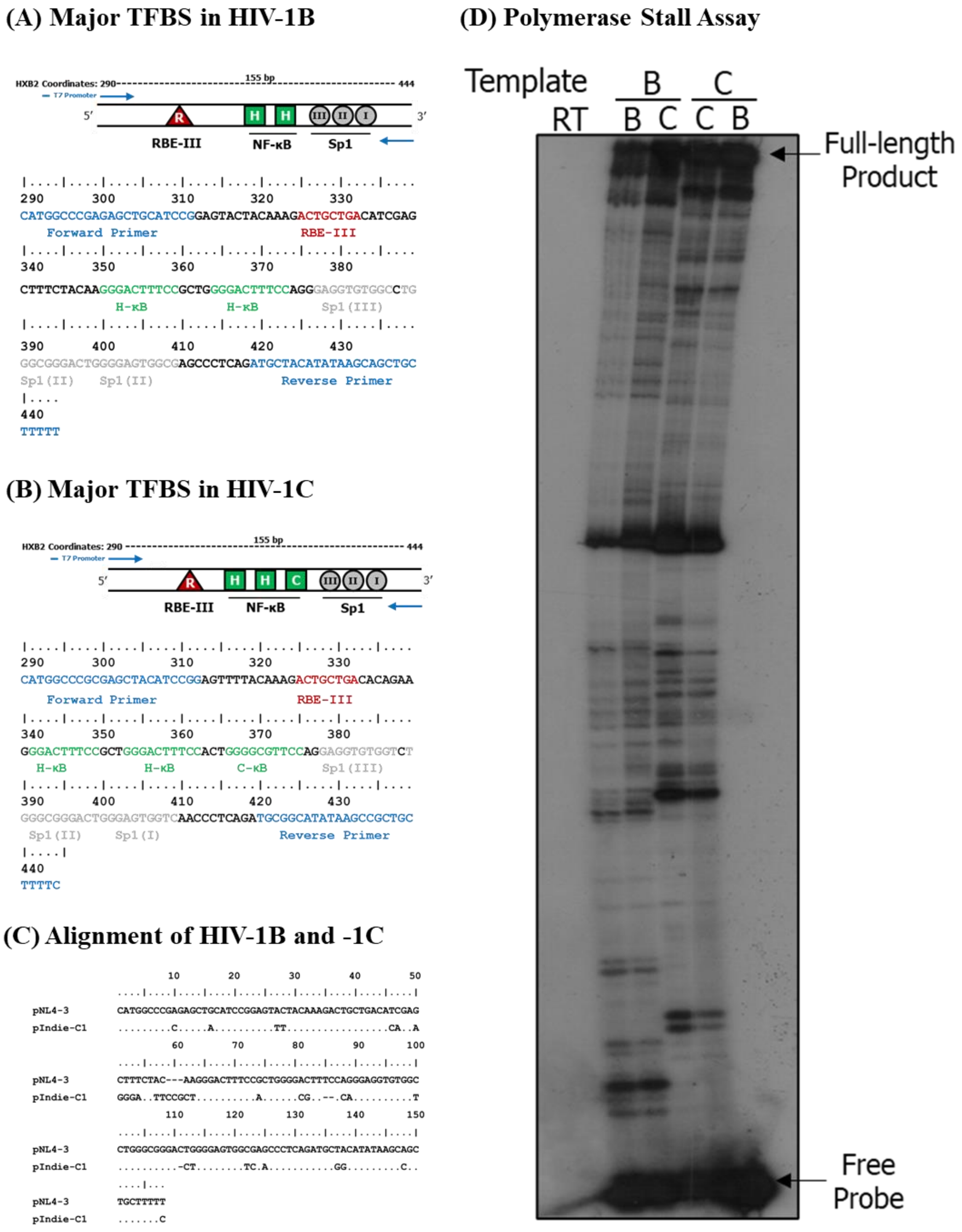

(A) Major TFBS in HIV-1B
HXB2 Coordinates: 290
155 bp
444
T7 Promoter
5′
R
H
H
III
II
I
3′
RBE-III
NF-κB
Sp1
290
300
310
320
330
CATGGCCCGAGAGCTGCATCCGGAGTACTACAAAGACTGCTGACATCGAG
Forward Primer
RBE-III
340
350
360
370
380
CTTTCTACAAGGGACTTTCCGCTGGGGACTTTCCAGGGAGGTGTGGCCTG
H-κB
H-κB
Sp1(III)
390
400
410
420
430
GGCGGGACTGGGGAGTGGCGAGCCCTCAGATGCTACATATAAGCAGCTGC
Sp1(II)
Sp1(II)
Reverse Primer
440
TTTTT
(B) Major TFBS in HIV-1C
HXB2 Coordinates: 290
155 bp
444
T7 Promoter
5′
R
H
H
C
III
II
I
3′
RBE-III
NF-κB
Sp1
290
300
310
320
330
CATGGCCCGCGAGCTACATCCGGAGTTTTACAAAGACTGCTGACACAGAA
Forward Primer
RBE-III
340
350
360
370
380
GGGACTTTCCGCTGGGACTTTCCACTGGGGCGTTCCAGGAGGTGTGGTCT
H-κB
H-κB
C-κB
Sp1(III)
390
400
410
420
430
GGGCGGGACTGGGAGTGGTCAACCCTCAGATGCGGCATATAAGCCGCTGC
Sp1(II)
Sp1(I)
Reverse Primer
440
TTTTC
(C) Alignment of HIV-1B and -1C
10
20
30
40
50
pNL4-3
CATGGCCCGAGAGCTGCATCCGGAGTACTACAAAGACTGCTGACATCGAG
pIndie-C1
.........C.....A..........TT.................CA..A
60
70
80
90
100
pNL4-3
CTTTCTAC---AAGGGACTTTCCGCTGGGGACTTTCCAGGGAGGTGTGGC
pIndie-C1
GGGA..TTCCGCT..........A......CG..--.CA..........T
110
120
130
140
150
pNL4-3
CTGGGCGGGACTGGGGAGTGGCGAGCCCTCAGATGCTACATATAAGCAGC
pIndie-C1
..........-CT........TC.A...........GG.........C..
pNL4-3
TGCTTTTT
pIndie-C1
.......C
(D) Polymerase Stall Assay
Template
RT
B
C
B
C
C
B
Full-length Product
Free Probe

Figure-11: The subtype identity of the template, not RT, determines polymerase stalling. The transcription factor binding sites (TFBS) in the region targeted by the assay are shown for HIV-1 B (panel A) and -1C (panel B). The primers used for *in vitro* transcription are highlighted in blue. The TFBS RBE-III, NF-κB and Sp1 are depicted using red, green, and grey colours, respectively. Panel C shows the multiple sequence alignment of the NL4-3 and Indie templates used in the polymerase stall assay. Dots and dashes represent sequence identity and gaps, respectively. Panel D depicts the autoradiogram of the polymerase stall assay. Two µg of the corresponding *in vitro* transcribed RNA template was extended using 30,000 CPM equivalent of ^{32}P labelled primer and 2 Units of HIV-1B or C RT (red). The products were resolved on a 12% Urea-denaturing Polyacrylamide gel and visualized using autoradiography. The free probe and full-length products are indicated using black arrows.

However, the scope of the polymerase stall assay is restricted by two major technical limitations. Firstly, the assay was performed in the absence of several host and viral factors known to influence reverse transcription, including the viral nucleocapsid protein. Secondly, the assay could span only a limited area of the viral genome, given the limited resolution capacity of the acrylamide gel.

Additionally, the minor qualitative differences observed in the stall patterns between the two templates in the assay cannot explain the significant quantitative differences seen in the sequence duplication frequencies between the subtypes. Given that the template identity did not have a substantial impact, we decided to examine whether significant leads could be obtained from possible differences in the RT sequences.

4.2 The G359T substitution in C-RT may form an additional hydrogen bond with the nascent DNA

The higher magnitude of duplications observed in HIV-1C genome sequences available in the database points towards subtype-specific molecular properties of the RT. We, therefore, performed a comprehensive examination of p66 RT sequences representing different genetic subtypes of HIV-1 downloaded from the HIV-1 LANL database.

We first downloaded full-length HIV-1 genome sequences of each subtype separately. Only one sequence per patient was considered from drug naïve individuals for the analysis to avoid bias. The sequences under HIV-1C were divided into two groups depending on the discordant NF-κB motif copy number (3 vs. 4 copies), and the RT sequences were compared between the two groups. This analysis did not identify any amino acid residues of RT to be unique for this phenotype. A similar analysis performed based on the copy number difference of the PTAP motif (1 vs. 2 copies) also failed to identify any significant differences in the RT sequence between the groups. These results appeared to suggest that inter-subtype, not intra-subtype, sequence differences be examined for leads underlying high-frequency duplications in HIV-1C. Therefore, we next compared p66 RT sequences downloaded from the database among different subtypes of HIV-1.

We compared 100 RT sequences of each of the major subtypes of HIV-1 – A1, B, C, D, and AE using the Viral Epidemiology Signature Pattern Analysis (VESPA) tool available on the HIV-LANL website (https://www.hiv.lanl.gov/content/sequence/VESPA/vespa.html). The sequences of the other subtypes were not included in the analysis as not many sequences representing those viral families were available. The comparative analysis identified several amino acid residues of the RT unique for each of the five subtypes (Table-7).

The comparative analysis identified amino acid residues to be unique at 29 positions in HIV-1C RT. The number of the unique positions in HIV-1C RT was narrowed by applying two additional measures. First, the threshold frequency values of query versus background values were increased further. Additionally, only those positions where amino acid residues were unique for HIV-1C were considered. Following the added stringency measures, we found seven positions in the RT that could be considered signature residues for HIV-1C: A36, E39, T48, A173, A200, T359, and S530. While the A36, E39, and T48 residues lie in the fingers domain; A173 and A200 are located in the palm; T359 in the connection domain, and the R530 residue in the RNase H domain (Table-8).

A detailed investigation was warranted to evaluate the functional significance of each of these seven amino acid residues to HIV-1C RT-associated phenotypes, including sequence motif duplication differences. Notably, the presence of a Threonine residue at position 359 is in HIV-1C of interest as explained below. A bioinformatic analysis identified the presence of four different amino acid residues –Alanine, Glycine, Serine, and Threonine, at position 359 of RT among the five HIV-1 subtypes compared here (Figure-12A). The relative proportion of each of the four amino acid residues varies significantly within each subtype. For example, at this position, Glycine is most commonly found in two subtypes – B and D, while being absent among the other three- A1, C, and AE. In contrast, Serine is majorly present in HIV-1 A1 and AE, representing 91% (653/710) and 99% (2,072/2,092) of the sequences, respectively, while representing a minor component among the other subtypes. Of note, in HIV-1C, a Threonine residue is the predominant amino acid residue at this location, representing 85% (3,502/4,092) of the sequences. In contrast, Serine, Alanine, and Glycine are present in only 6.8% (280/4,092), 4.4% (182/4,092), and 0.14% (6/4,092) of the sequences, respectively. Interestingly, the Threonine residue is nearly absent from the other subtypes except for HIV-1 A1, where it is present in a minority of 2.3% (17/710) of sequences. The replacement of Glycine (preferred in HIV-1 B and D) with a Threonine in HIV-1C represents

RT Domain	AA Position	HIV-1 Subtype				
		A	B	C	D	AE
Fingers	6	E	E	E	E	D
	11	K	K	K	K	T
	35	T	V/I	T	T	T
	36	E	E	A	E	E
	39	T	T	E/D	T	K
	43	K	K	K	K	E
	48	S	S	T	S	S
	49	K	K	K	R	K
	60	I	V	I/V	I	V
	122	E	K	E	E	E
	123	S/D	D	G/D	D	S
	135	T	I	I	I	I
Palm	173	S	K	A	K	I
	174	K	Q	Q/K	Q	K
	177	E	D	E	E	E
	178	I	I	I	I	M
	179	I	V	V	V	V
	200	T	T	A	I/T	T
	207	A	Q	E	E	A
	211	S	R/K	K	K	S
Thumb	238	K	K	K	K	R
	245	Q	V	Q	K	E
	250	E	D	D	E	D
	272	A	P	P	P	A
	277	K	K	R/K	R	K
	282	L	L	L	C	L
	286	A	T	A	A	A
	291	D	E	D	E	D
	292	I/V	V	I	V	I
	294	T	P	P	P	P
	312	D	E	E	E	T
Connection	326	I	I	I	I	V
	329	I	I	I	I	V
	334	Q	Q	H/D/L	Q	Q
	335	D	G	D	D	D
	345	P	P	P	Q	P
	346	F	F	F	Y	F
	356	R	R	K	K	R
	357	K	M	M	M/L/R	K
	359	S	G	T	G	S
	366	K	K	K	K	R
	369	A/T	T	T	T	T
	371	V	A	A	A	V
	375	V	I	I	I	I
	376	M/T	T	T	T	T
	377	M	T	M/T	Q	T
	379	S	S	S	C	S
	390	K	R	R	R	R
	400	T	A	T	T	T
	403	M	T	T	T	M
	404	D	E	D	E	E
RNase H	432	D	E	E	E	D
	435	A/V	V	A	V/I	V/I
	447	N	N	N	N	S
	452	L	L	I	L	L
	466	V	V	I	V	V
	471	E	D	E	D	E
	480	H	Q	Q	Q	H
	483	H	Y	Q	N	H
	491	S	S	S	L	S
	512	R	K	K	K	R
	517	L	L	L	L	V
	519	N	N	N	S/N	N
	524	K	Q	Q	Q	E
	527	G	K	K	K	K
	529	D	E	E	E	E
	530	K	K	R	K	K
	554	S	S	S	N	S
	559	V	V	V	I	V

Table-7: VESPA of HIV-1 RT sequences. One hundred full-length RT sequences representative each major viral subtype were downloaded from the LANL database and analysed using the VESPA tool. The table shows the amino acids that are present at maximum frequency at the corresponding locations in each subtype. The amino acids indicated in red represent residues that are unique or maximally represented in that subtype.

Domain	Alignment position	AA in HIV-1C	Frequency in HIV-1C	Frequency in Other Subtypes	AA in Other Subtypes	Frequency in HIV-1C	Frequency in Other Subtypes
Fingers	36	A	0.76	0.01	E	0.23	0.97
	39	E	0.62	0.02	T	0.01	0.58
	48	T	0.93	0.02	S	0.05	0.98
	60	I	0.52	0.50	V	0.48	0.50
	123	G	0.42	0.04	D	0.29	0.43
Palm	173	A	0.68	0.04	K	0.03	0.48
	200	A	0.96	0.21	T	0.01	0.53
	207	E	0.68	0.24	A	0.07	0.36
	211	K	0.62	0.34	S	0	0.40
Thumb	245	Q	0.82	0.11	E	0.01	0.35
	272	P	0.8	0.45	A	0.15	0.46
	277	R	0.58	0.28	K	0.4	0.70
	291	D	0.94	0.43	E	0.06	0.56
	292	I	0.93	0.38	V	0.07	0.61
Connection	334	H	0.32	0.02	Q	0.29	0.84
	356	K	0.98	0.31	R	0.01	0.69
	357	M	0.76	0.28	K	0.01	0.43
	359	T	0.87	0.03	S	0.05	0.48
	371	A	0.96	0.48	V	0.04	0.51
	377	M	0.46	0.16	T	0.03	0.52
	403	T	0.95	0.46	M	0	0.46
	404	D	0.96	0.31	E	0.04	0.69
RNase H	435	A	0.63	0.14	V	0.2	0.55
	452	I	0.51	0.06	L	0.22	0.88
	466	I	0.63	0.03	V	0.37	0.95
	480	Q	0.97	0.49	H	0.01	0.49
	483	Q	0.75	0.03	H	0.04	0.44
	530	R	0.92	0.08	K	0.08	0.92

Table-8: The frequencies of the Signature Amino acid residues of HIV-1C. HIV-1C RT sequences were subjected to signature amino acid residue analysis using the VESPA tool, with the HIV-1 subtypes A1, B, D, and AE as background. Amino acids that were unique to HIV-1C, and those that passed a threshold frequency of 0.65 and 0.25 for the query and background, of both groups, respectively are highlighted in green.

a non-conservative amino acid substitution. These reasons prompted us to investigate the significance of the Threonine residue in the RT structure.

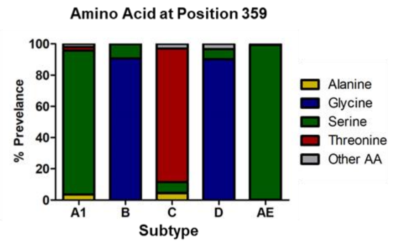
However, the lack of an RT crystal structure of HIV-1C in the Protein Data Bank posed a challenge. As a possible solution, we downloaded the high-resolution HIV-1B RT structure (PDB ID: 5J2M) deposited by Salie et al. (2016), replaced the Glycine residue at position 359 with Threonine, and performed an energy minimisation analysis of the residue to visualise the immediate micro-environment of this variation. The analysis suggested the possibility of a hydrogen bond formation at this position. The atomic distance measurement between the Hydrogen atom of the -OH group of Threonine and the Oxygen of the phosphate backbone of the nascent cDNA was predicted to be 2.4 Å, indicating a strong possibility of a hydrogen bond formation between the two atoms (Figure-12B).

Attempts to solve the crystal structure of C-RT by X-ray crystallography are in progress. Multiple attempts to crystallize the native recombinant protein have failed, possibly due to the floppy nature of the Fingers domain. We are currently attempting to optimize the crystallization conditions in the presence of an RT inhibitor to stabilize the complex, and aid in the crystal formation.

4.3 HIV-1C RT containing T359 demonstrates optimal catalytic activity

The bioinformatic analysis identified the preferential presence of a Threonine residue at position 359 of HIV-1C RT. The presence of a Threonine residue at this location contains the potential to form an additional hydrogen bond with the nascent DNA. Additionally, a substitution of an amino acid at a critical location such as the connection domain may have considerable influence on several other properties of the RT, including the strand switch activity, recombination frequency and consequently, sequence motif duplications. Each of diverse functions of the enzyme must be evaluated to understand the effect of the presence of Threonine on the overall activity of the RT.

(A) The Amino Acid Profile at Position 359



AA at Position 359	Subtype				
	A1	B	C	D	AE
Alanine	3.52	0.08	4.45	0.00	0.10
Glycine	0.14	90.60	0.15	90.08	0.14
Serine	91.97	9.09	6.84	6.52	99.04
Threonine	2.39	0.01	85.58	0.00	0.14
Other AA	1.97	0.22	2.98	3.40	0.57
N =	710	7252	4092	353	2092

(B) Hydrogen Bond Formation by Threonine at 359th Position

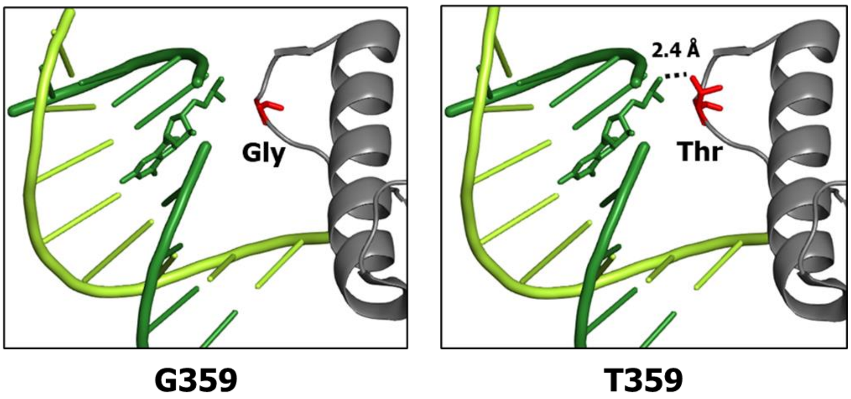


Figure-12. The G359T substitution may form an additional hydrogen bond. **(A)** HIV-1 p66 RT sequences of five different subtypes were downloaded from the HIV-LANL database and aligned using the BioEdit software. The chart presents the amino acid frequency profile at position 359 of RT among the subtypes using a colour code representing each amino acid. The number of sequences used in the analysis representing each subtype is shown. **(B)** HIV-1B RT containing a natural Glycine residue at position 359 was mapped and the immediate micro-environment of the residue was visualized (Left panel, PDB ID: 5J2M) or after substituting a Threonine at this position (Right panel). The template and the cDNA are depicted in light and dark green colours, respectively. The protein backbone is shown in the grey color, and Glycine and Threonine are highlighted in red and shown as sticks. The base on the cDNA strand that forms a hydrogen bond with the T359 residue is shown as sticks in both structures. The distance between the phosphate group and the OH-group on the C-RT was determined using the options provided in the software and is shown using a dotted line.

To this end, we constructed several panels of RT bearing four different amino acid residues at position 359 as recombinant proteins (Figure-13) or infectious molecular clones (Figure-14). As discussed above, the four amino acid residues – Alanine, Glycine, Serine, and Threonine, are naturally present in RT of different HIV-1 subtypes, although at variable frequencies among the subtypes (Figure-12A). Within each panel, except at position 359, the four RTs are genetically identical at all other locations. To generate each variant recombinant RT protein, the two subunits, p51 and p66, were individually engineered, bacterially expressed, and purified using two different and sequential chromatography procedures; the purified subunits were assembled in vitro, and the functional activity of the assembled RT was evaluated. The individual subunits of the RTs were expressed in *E. coli* M15 cells, purified by Ni-NTA chromatography followed by a second round of purification using ion-exchange chromatography, assembled in vitro as described in Materials and Methods, and used in the assays described below.

The polymerase activity of the RT protein was estimated as a measure of ^{32}P-labelled dATP incorporation in a reaction using a poly-U template and varying quantities of each enzyme. The time-course of dATP incorporation confirmed that all eight recombinant RT proteins were functional. Further, the polymerase activity of each enzyme increased progressively with time, although the polymerase kinetics of each enzyme varied significantly (Figure-13A). Among the four enzymes of the HIV-1B panel, the RTs containing Glycine and Serine, the two amino acids most common at position 359 in this subtype (Figure-12A), showed significantly superior polymerase activity than the other two variants containing Alanine or Threonine (Figure-13A, Top-left panel). At 30 min, for example, the wild-type Glycine-containing and the Serine variant incorporated 27.47 ± 1.11 and 26.26 ± 2.85 pmol of dATP, respectively. Threonine and the Alanine variants, on the other hand, displayed much lower values at 10.81 ± 1.16 and 11.81 ± 0.65 pmol, respectively. The contrasting activity differences among variants were more striking in HIV-1C. The Glycine and Serine variants demonstrated extremely high values of 179.28 ± 7.66 and 54.18 ± 5.19 pmol, respectively, than the wild-type Threonine containing RT that showed a much lower value of 17.1 ± 1.79 and the Alanine variant that demonstrated a value of 21.09 ± 1.80 pmol (Figure-13A, Bottom-left panel). In other words, the substitution of Threonine with Glycine in HIV-1C RT at position 359 resulted in a super active RT.

Several groups previously demonstrated that the specific activities of wild-type RTs of HIV-1B and -C are comparable (Iordanskiy et al., 2010; Xu et al., 2010). Consistent with these reports, we also found that the specific activities of the wild-type RT proteins were comparable in our assay. HIV-1B and -1C RTs showed 64.98 ± 5.33 U/µg and 76.15 ± 12.24 U/µg activities, respectively, the difference being insignificant. In B-RT, the substitution of wild-type Glycine with Serine significantly enhanced the specific activity of the protein from 64.98 ± 5.33 U/µg to 95.24 ± 8.23 U/µg. In contrast, replacement of Glycine with Alanine or Threonine reduced specific activity to 26.83 ± 8.71 U/µg and 34.65 ± 9.64 U/µg, respectively, a difference which was also statistically significant. However, the variations observed in the specific activities of B-RT variant proteins were only moderate, unlike in the C-RT proteins. The replacement of Threonine, the amino acid naturally present at position 359 in C-RT, with Glycine, the amino acid naturally present in B-RT, resulted in a profound augmentation of the specific activity to 510.66 ± 43.39 U/µg, a 6.7 folds enhancement. Likewise, the substitution of Serine at this position also enhanced the specific activity of the C-RT by 2.23-folds to 170.25 ± 27.23 U/µg. In summary, reciprocal exchange of the amino acid residues present in the wild-type RT proteins at position 359 of the two subtypes profoundly impacts HIV-1C, but not HIV-1B. Whether the dramatically enhanced specific activity of Glycine-containing C-RT could negatively impact the replication fitness of the viral strain needed to be evaluated.

Based on these data, we speculate that the presence of a Threonine residue at position 359 in C-RT is crucial to maintaining optimal RT activity by slowing down the otherwise profoundly faster enzyme kinetics. The formation of a potential hydrogen bond between Threonine and the nascent cDNA could play a pivotal role in attenuating the RT polymerase activity. Substitution of Threonine with Glycine precludes the formation of the additional hydrogen bond, thus, substantially increasing RT activity. Of note, a conservative replacement of Threonine with Serine at this position in C-RT increases RT activity only marginally, thus, further corroborating the hypothesis. The presence of Serine could also possibly cause the formation of a hydrogen bond, but of a lower strength given the increased distance of 3.2 Å between the RT and the cDNA.

The sequences of B- and C-RT proteins used here differ by 7.68% at the amino acid level. Further, our analysis identified several amino acid residues differentially conserved between the two subtypes. A variation to this extent could significantly impact the relative specific activities of the two proteins. Thus, the presence of Threonine at position 359 of the C-RT counterbalances the otherwise superior activity of the enzyme. Of note, reverse transcription is an exceptionally slow process. Depending on the cell type, the RT may take 8-33 hours to fully polymerise the 9 kb genome (Murray et al., 2011). Therefore, a dramatically enhanced RT activity, in all probability, could be detrimental to the overall kinetics of the reverse transcription process. Thus, it seems reasonable to conclude that the T359 residue offsets the detrimentally faster enzyme kinetics and reduces the RT activity of HIV-1C to optimal levels.

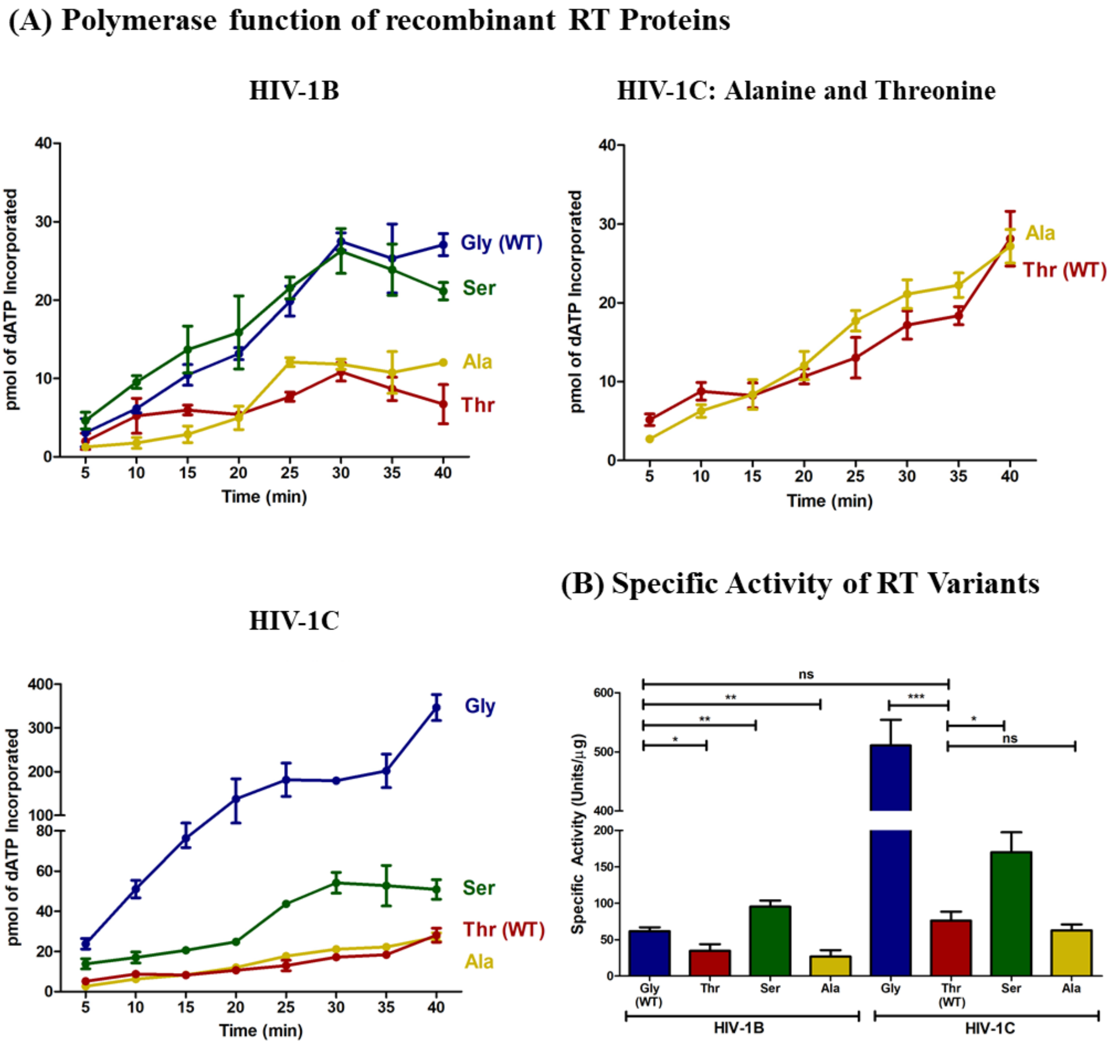


Figure-13: Polymerase function of the recombinant RT proteins. (A) cDNA synthesis rate of RT variants. Bacterially expressed recombinant subunits of each of four RT variants of HIV-1B (Left panel) and HIV-1C (Right panel) were assembled in vitro and used in the assay. The assay was performed using each of the RT proteins shown, ^{32}P-labelled dATP, a homopolymeric template, and a primer. The amount of dATP polymerized was estimated by liquid scintillation spectrometry by monitoring the enzyme activity at five-minute intervals up to 40 minutes. Data are presented as Mean ± SD, followed by two-way analysis of variance and Bonferroni post. hoc tests. **(B) Specific activities of the RT variant proteins**. One Unit of RT activity was defined as the amount of enzyme that incorporates 1 pmol precursor into the product in 10 minutes at 37°C. The amount of radioactive dATP incorporated in 10 min was used to calculate the specific activity of all RT variants. The data are representative of three independent experiments, and are plotted as Mean ± SD, followed by one-way analysis of variance and Tukey-Kramer post. hoc tests. Each RT variant is represented using a different color as depicted. . ***p < 0.001, **p < 0.01, *p < 0.05, ns = non-significant

4.4 Wild-Type viral strains outperform other RT-variants in replication.

In the biochemical assay described above, the wild-type recombinant RT proteins (containing Glycine and Threonine residues at position 359, in HIV-1B and HIV-1C, respectively) demonstrated a comparable specific activity, although the variant forms of the RT showed enhanced or reduced activities. Notably, the T359G variant form of HIV-1C RT demonstrated the highest magnitude of specific activity among all the variants assayed. These observations warranted the evaluation of the variant forms of RT in the backdrop of full-length and replication-competent molecular clones. To this end, we constructed two panels of four RT variant viral strains in each panel, analogous to the panels of RT recombinant proteins described above. We used the molecular clones NL4-3 and Indie-C1 as representatives of HIV-1B and HIV-1C, respectively (Figure-14).

Viral stock of each variant strain was produced in HEK293T cells using standard procedures. The infection titer of each viral stock was measured by infecting TZM-bl reporter cells with a stock equivalent of 10 ng of p24 and measuring the Tat-inducible luciferase secretion into the medium after 48 h. The rate of infectivity of the TZM-bl cells was comparable among all the variants tested, suggesting that all the stocks contained equivalent titers (Figure-14A). Subsequently, the replication profiles of the two panels of viral strains were determined in the PBMC of healthy donors. PBMCs were isolated from healthy donors by density-gradient centrifugation, CD8-depleted using a commercial kit, and activated with PHA-P (5 µg/ml) and IL-2 (20 U/ml) for 72 h, as described in the Methods section. The activated PBMC were infected with 10 ng p24 equivalent of the individual viral stocks in the presence of 10 µg/ml of DEAE-dextran for 8 hours. The cells were washed three times to remove residual p24. The infected cells were incubated in a CO_2 incubator. The levels of secreted p24 in the culture medium were estimated for 28 days at an interval of 4 days, using a commercial kit.

All the viral strains proliferated in PBMC, with the replication profile peaking on day 8 or 12, regardless of the difference in the molecular backbone (Figure-14B). The viral strains of NL4-3 panel typically produced more p24 than those of Indie-C1, as expected. Importantly, within each panel, the wild-type strain of the corresponding subtype produced more p24 than the other strains of the panel. For example, in the NL4-3 panel, the Glycine strain, representing the wild-type RT profile of HIV-1B produced 153.90 ± 39.63 ng/ml of p24 at day 8, compared to the Threonine, Serine, and Alanine variant strains that secreted 122.88 ± 18.15, 69.09 ± 28.03, and 38.04 ± 39.93 ng/ml of p24, respectively. Similarly, in the Indie-C1 panel, the wild-type strain containing Threonine out-performed the other variants, with p24 levels peaking on day 8 at 55.95 ± 12.09 ng/ml, and the Glycine, Serine, and Alanine variants displaying significantly lower values at 34.13 ± 13.64, 16.25 ± 2.07, and 14.83 ± 6.82 ng/ml, respectively.

This experiment revealed that the wild-type residue for each subtype (Glycine for B, Threonine for C) outperformed the other variants in p24 production. There was a marginal difference between the wild-type variant and Glycine/Threonine or Serine for both subtypes, with a marked difference in the p24 production when Alanine was substituted (Figure-14B). The superior performance of the wild-type variants of both the subtypes in the replication assay is not surprising given the near-ubiquitous presence of Glycine and Threonine at position 359 of HIV-1B and -1C RTs, respectively. Since both these enzymes display comparable specific activities, it seems prudent to conclude that RT activity must be maintained at an optimal level for viral replication to proceed effectively. Any deviation from the optimal level of performance is likely to have a detrimental effect on the virus. The marginal differences observed in the replication assay are likely to be magnified in natural infection, where immune pressure plays a vital role in shaping the host landscape.

Further, given the association between specific mutations in the connection domain and drug resistance, particularly against Nucleoside Reverse Transcriptase Inhibitors (Von Wyl et al., 2010), we examined a possible effect of amino acid variation at position 359 on viral proliferation in the presence of an anti-viral molecule. To this end, we monitored the replication of variant viral strain of the two panels in the presence of Tenofovir at a sub-lethal concentration. The results showed that the magnitude of viral proliferation was considerably lower in the presence of the anti-viral agent; however, the replication profiles of each viral strain remained consistent with or without Tenofovir (Figure-14C). The results did not allude to a possible association between the RT variant forms and Tenofovir resistance. Since the problem of drug resistance is not directly relevant to the present work, we did not pursue this line of experimentation further.

4.5 Comparable levels of recombination among HIV-1B and -1C RT variant viral strains

Higher levels of recombination frequency could explain the greater magnitude of sequence motif duplications observed in HIV-1C, as recombination is a precondition for duplication. Since the T359 residue appears important for regulating HIV-1C RT, we hypothesised that the additional hydrogen bond formed due to Threonine could stabilize the enzyme-cDNA complex, thus, enhancing recombination frequency. However, although limited severely, the available evidence precludes the possibility of higher magnitude recombination in HIV-1C. As mentioned previously, the recombination rate of HIV-1C RT was reported using a single molecular clone, MJ4, in a sub-genomic viral backbone (Chin et al., 2005).

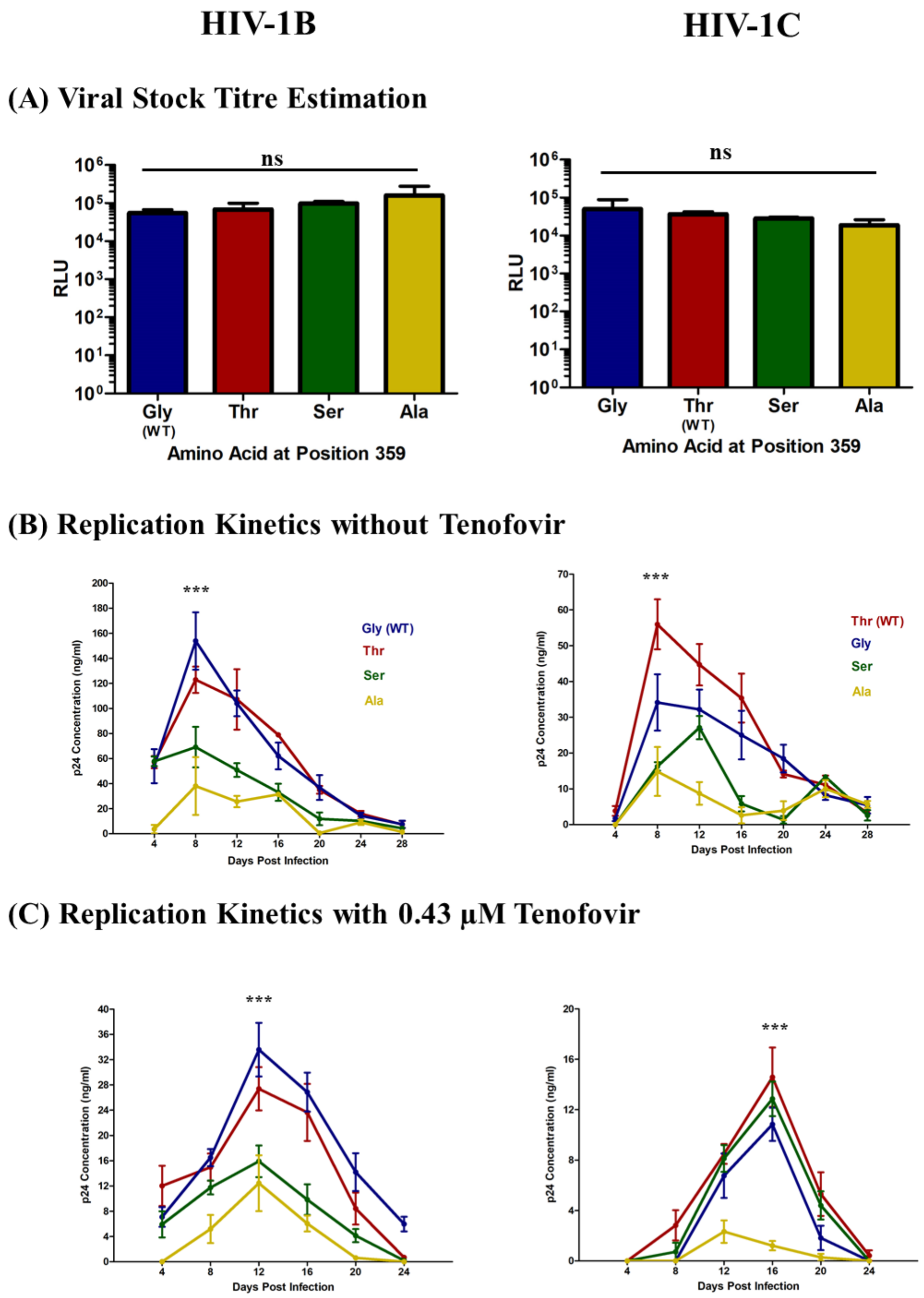


Figure-14: Replication profiles of infectious RT-variant viral strains. Two panels of replication-competent molecular strains, each containing Alanine, Glycine, Serine, or Threonine at position 359 of the RT were constructed in NL4-3 or Indie molecular clones, representing HIV-1B or HIV-1C, respectively. These strains are not defective in any of the viral genes; hence they can establish productive viral infection in target cells. **(A) Viral stock titer determination.** TZM-bl cells, 10^4 cells/well of a 96-cluster plate, were infected with 10 ng p24 equivalent of each of the eight viral stocks independently. The cells secrete luciferase into the medium when the virus-encoded Tat induces the LTR, serving as a surrogate for viral proliferation. The data are representative of three experiments and are presented as mean ± SD, and analysed using One-Way Analysis of Variance, followed by Tukey-Kramer post. hoc test. A different colour represents each amino acid residue at position 359. The color-code is consistent among all the panels and figures in the thesis. **Replication kinetics of RT-variant viral strains in the absence (B), and presence (C) of 0.043 µM Tenofovir in PBMC.** 3 x 10^6 PBMCs were infected with 10 ng p24 equivalent of each of the viral stocks separately. The viral proliferation was estimated by measuring the p24 production at 4-day intervals using a commercial kit. . ***p < 0.001, ns = non-significant

To understand the influence of the T359 residue and the three other amino acid residues naturally present at this location on viral recombination, we performed the fluorescent reporter (EGFP) recombination assay described previously (Levy et al., 2004; Rhodes et al., 2005). The assay measures the regeneration of a functional EGFP protein, by recombination, from two defective precursor fluorescent proteins harboring debilitating but complementary mutations in different regions of the reading frame. To this end, we generated two EGFP mutant forms containing a frame-shift mutation at amino acid residue 4 or 204; thus, neither of these proteins is fluorescent. Using the defective EGFP precursors, we constructed a panel of four RT variant viral strains containing Alanine, Glycine, Serine, or Threonine at position 359, using envelope-deficient, pseudotypable, and replication-competent molecular clones NL4-3 and Indie; a total of 24 viral strains were generated (Figure-15A). The eight viral strains containing the intact EGFP, but none of the 16 strains containing either of the frame-shift mutations, expressed fluorescence (Figure-15B).

Having confirmed that the viral strains harboring EGFP encoding debilitating mutations could not express a functional fluorescent protein, we performed the fluorescence complementation assay. Four different viral stocks – intact EGFP only, FSM4 only, FSM204 only, and FSM4 plus FSM204 - were produced in HEK293 cells for each of the four RT variations (Figure-15A). CEM-CCR5 T-cells were infected with each individual viral stock, and the percentage of fluorescent cells was determined by flow cytometry.

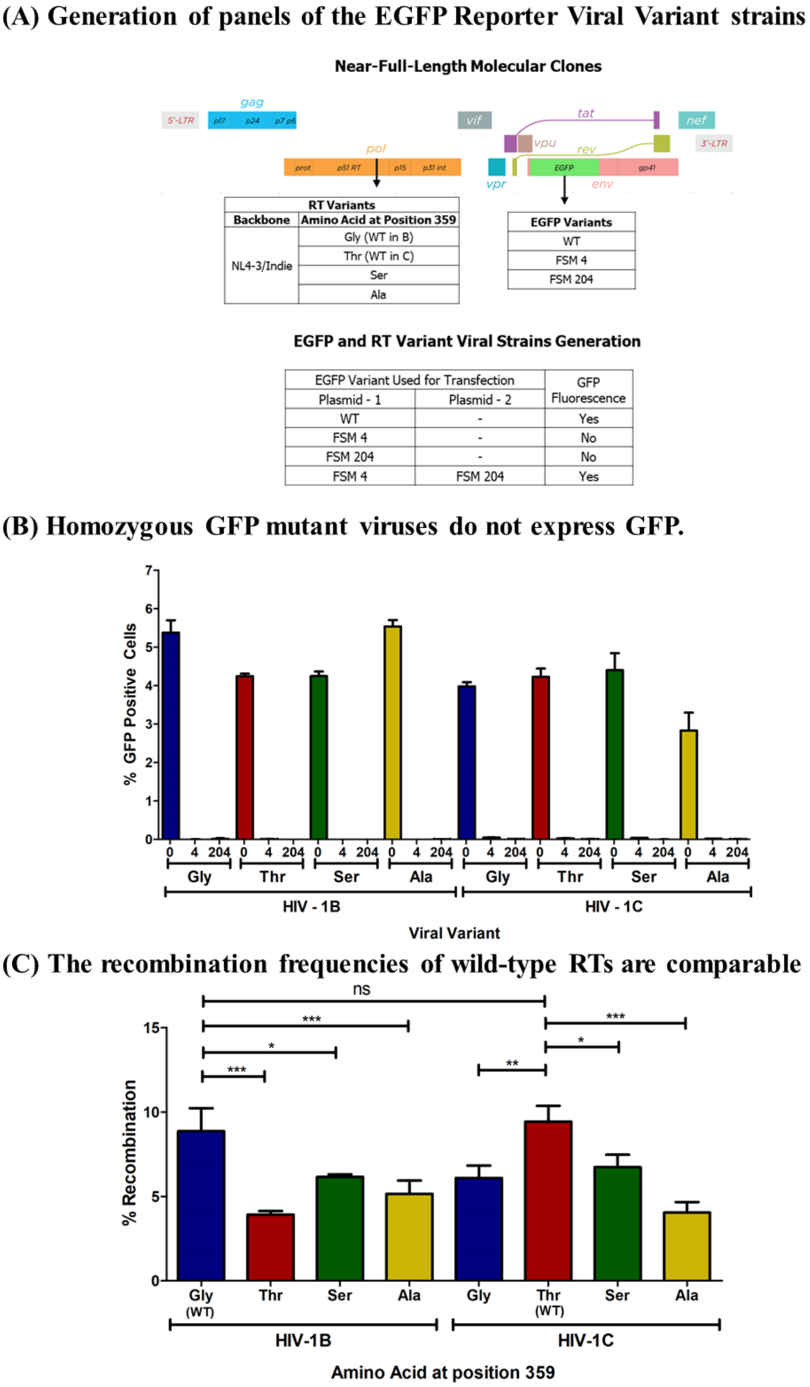
When a mix of FSM4 and FSM204 plasmid vectors was introduced into HEK293 cells, approximately 25% of the virus particles are expected to co-package the two viral RNA molecules harboring the frame-shift mutations in EGFP. In such viral particles, efficient recombination between the two template molecules, between the two debilitating mutations of EGFP, should produce a functional EGFP protein. Further, the magnitude of fluorescence regeneration is expected to be proportional to recombination, which in turn depends on the function of the RT variant.

In the assay, we found that the two viral strains harboring the respective canonical amino acid at 359 (Glycine and Threonine in HIV-1B and -1C, respectively) produced the highest percentage of fluorescent cells compared to the other RT variant strains. Importantly, the recombination rate of wild-type HIV-1B and -1C RTs was comparable at 8.86 ± 1.36% and 9.42 ± 0.94%, respectively (Figure-15C), consistent with the previous report (Chin et al., 2005). The recombination rates dropped significantly when a non-canonical amino acid was substituted for a canonical residue, in both subtypes. For example, the presence of Threonine in B-RT and Glycine in C-RT showed 3.91 ± 0.21% and 6.09 ± 0.73% GFP-positive cells, respectively. Serine in either subtype showed reduced percentages of 6.14 ± 0.15% and 6.72 ± 0.73% for HIV-1B and -1C, respectively. Our results are consistent with the dynamic-copy-choice recombination model; an increase in RT activity was associated with a decrease in the recombination rates for most variants studied, except the Alanine variant.

4.6 HIV-1C RT displays a superior strand-transfer potential.

Given the lack of a significant difference in recombination frequency between HIV-1B and HIV-1C, we asked whether the strand-transfer abilities of the two RTs would be different. To this end, we performed a strand-transfer assay using two RNA molecules generated by *in vitro* transcription that overlapped over 20 bases (Figure-16A). The full-length cDNA synthesis of 336 bp would be possible only following the successful strand transfer of the RT between the two template RNA molecules. The copy numbers of cDNA molecules produced by the RT variant enzymes of the two panels were determined using a PCR that amplified the 206 bp cDNA, which did not require strand transfer between the templates. The copy numbers of this PCR were used to normalize the second strand transfer assay results of the corresponding RTs.

The various RT variants of the two panels demonstrated equivalent reverse transcription ability while reverse transcribing the common RNA-2 template (Figure-16B). In contrast, the assay revealed that the HIV-1C variant RTs as a class showed significantly superior strand-transfer ability than HIV-1B RT variant enzymes. Importantly, within the panel of HIV-1C, the wild-type RT containing Threonine showed the highest level of strand transfer activity (Figure-16 B and C). While the wild-type HIV-1C RT demonstrated a strand-switch frequency of 0.096 ± 0.01, the Glycine, Serine, and Alanine variant RTs demonstrated slightly reduced frequencies of 0.057 ± 0.00, 0.07 ± 0.01, and 0.08 ± 0.01, respectively. HIV-1B RTs showed significantly lower activities than that of wild-type HIV-1C RT, with the Glycine (wild-type), Threonine, Serine, and Alanine variants demonstrating frequencies of 0.03 ± 0.00, 0.03 ± 0.00, 0.02 ± 0.01, and 0.06 ± 0.00,



RT Variants	
Backbone	Amino Acid at Position 359
NL4-3/Indie	Gly (WT in B)
	Thr (WT in C)
	Ser
	Ala

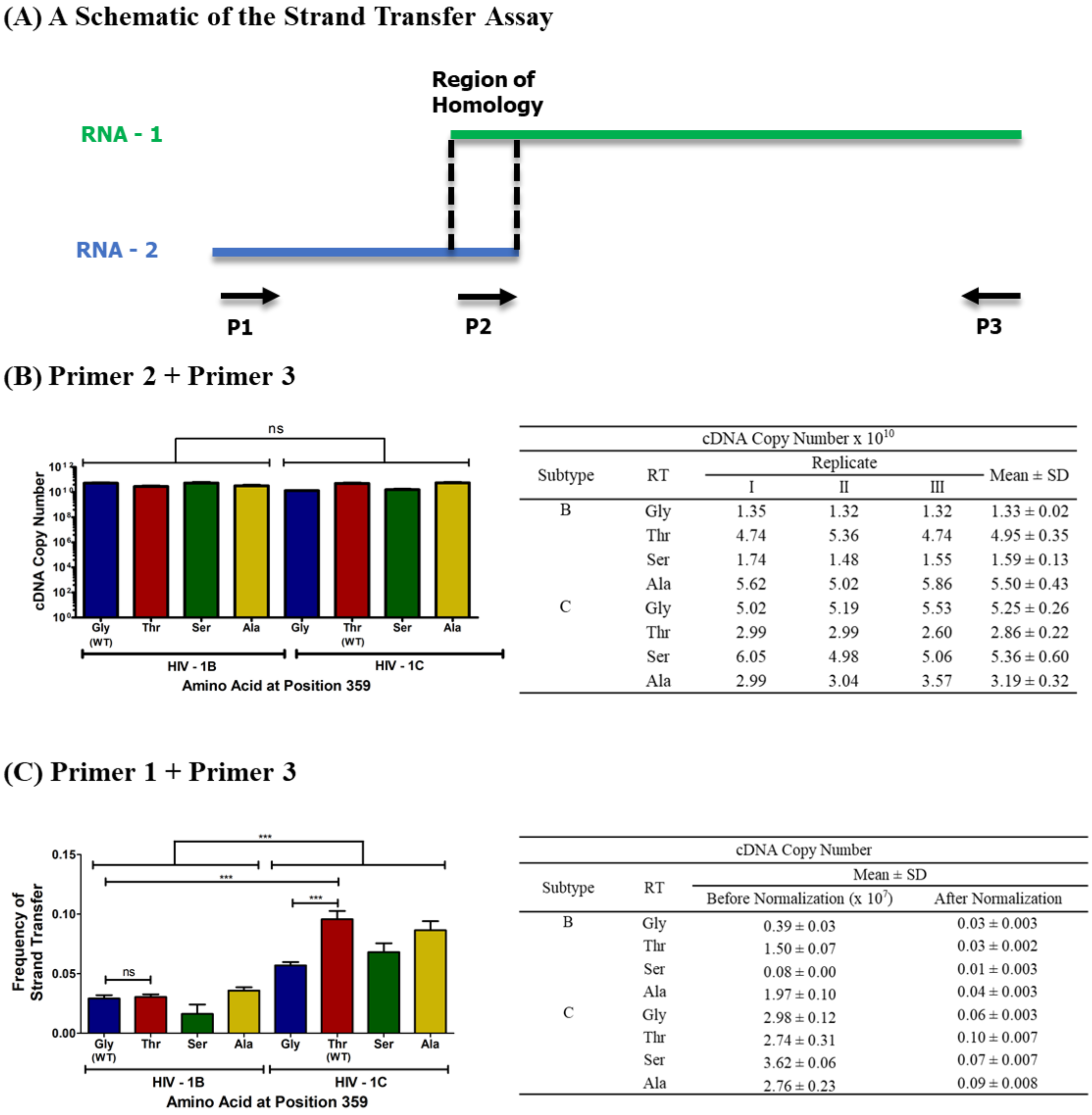
EGFP Variants
WT
FSM 4
FSM 204

EGFP Variant Used for Transfection		GFP Fluorescence
Plasmid - 1	Plasmid - 2	
WT	-	Yes
FSM 4	-	No
FSM 204	-	No
FSM 4	FSM 204	Yes

Figure-15: The fluorescence complementation assay for viral recombination. (A) Generation of panels of the EGFP Reporter Viral Variant strains. The top panel presents a schematic diagram of the viral genetic organization. Panels of four RT variants containing Alanine, Glycine, Serine, or Threonine at position 359 of RT were constructed using NL4-3 or Indie full-length molecular clones. The envelope of the viral strains has been substituted with the EGFP ORF; therefore, the strains can infect for a single round when pseudotyped. Further, the EGFP reading frame contains frame-shift mutations (FSM) at amino acid positions 4 or 204. Thus, we constructed 24 replication-competent viral strains for the assay. The table presents the four experimental conditions and the expected outcomes of the assay. **(B) Homozygous GFP mutant viruses do not express GFP.** Viral stocks of the 24 individual strains were produced in HEK293 cells and used to infect CEM-CCR5 T-cells. After 24 h, the cells were activated with a cocktail of cellular activators comprising TNF-α (10 ng/ml), HMBA (10 ng/ml), and PMA (5 mM). GFP fluorescence was measured after a 24 hour activation period. **(C) The recombination frequencies of wild-type RTs.** The viral stocks were produced in HEK293 cells using the mix of FSM4 (FSM, Frame-shift mutation) and FSM204, representing all the four RT variant viral strains of both subtypes. CEM-CCR5 cells were infected and fluorescence was monitored as described in (B) above. The recombination percentage is represented as the ratio of GFP-positive cells to all infected cells. The statistical analysis was performed using One-way Analysis of Variance followed by Tukey-Kramer post.-hoc test. The data are representative of three independent experiments. Each viral variant is represented using a different color. ***p < 0.001, **p < 0.01, *p < 0.05, ns = non-significant

respectively. As expected, the overall magnitude of reverse transcription of the 206 bp RNA template, which did not require strand transfer, was several orders higher than the amplification of the full-length product that depended on strand transfer.

In summary, HIV-1C RT in general and the wild-type Threonine-containing RT specifically are more efficient in strand-transfer. The additional hydrogen bond formed by Threonine possibly contributes to the superior strand-transfer ability of HIV-1C RT. Consistent with the model, substituting Threonine in C-RT with Glycine at position 359 leads to a two-fold reduction in the strand transfer activity. Of note, the positioning of a Serine residue at this location did not have such a severe negative impact on the strand transfer activity, probably due to the potential of Serine to make a weak hydrogen bond, although the behavior of the Alanine-containing RT is not entirely consistent with the proposed model. In essence, the strand-transfer assay results clearly demonstrate a superior ability of HIV-1C RT to switch to the acceptor strand, an essential prerequisite for sequence duplication.



(A) A Schematic of the Strand Transfer Assay

(B) Primer 2 + Primer 3

Subtype	RT	cDNA Copy Number x 10^{10}			
		Replicate			Mean ± SD
		I	II	III	
B	Gly	1.35	1.32	1.32	1.33 ± 0.02
	Thr	4.74	5.36	4.74	4.95 ± 0.35
	Ser	1.74	1.48	1.55	1.59 ± 0.13
	Ala	5.62	5.02	5.86	5.50 ± 0.43
C	Gly	5.02	5.19	5.53	5.25 ± 0.26
	Thr	2.99	2.99	2.60	2.86 ± 0.22
	Ser	6.05	4.98	5.06	5.36 ± 0.60
	Ala	2.99	3.04	3.57	3.19 ± 0.32

(C) Primer 1 + Primer 3

Subtype	RT	cDNA Copy Number	
		Mean ± SD	
		Before Normalization (x 10^{7})	After Normalization
B	Gly	0.39 ± 0.03	0.03 ± 0.003
	Thr	1.50 ± 0.07	0.03 ± 0.002
	Ser	0.08 ± 0.00	0.01 ± 0.003
	Ala	1.97 ± 0.10	0.04 ± 0.003
C	Gly	2.98 ± 0.12	0.06 ± 0.003
	Thr	2.74 ± 0.31	0.10 ± 0.007
	Ser	3.62 ± 0.06	0.07 ± 0.007
	Ala	2.76 ± 0.23	0.09 ± 0.008

Figure-16: HIV-1C RT demonstrates a superior strand-transfer ability. (A) A schematic of the strand-transfer assay. The two *in vitro* transcribed RNA templates, containing an overlap of 20 bp, are shown as green and blue lines, respectively. The arrows represent primers with the orientation indicated. Each of the recombinant RT variant enzymes (color-coded) was used in the assay separately. The Ct values of the real-time PCR were detected using SYBR Green. **(B) The input cDNA copy number.** The product of primers P2 and P3 is common to all the reactions. The cDNA copy numbers determined by a regression analysis were used for the normalisation of the results of the second PCR (P1 + P3). **(C) The strand-transfer frequencies of the RT variant proteins.** Successful amplification of the primer pair P1 and P3 is conditional to efficient strand transfer by the RT variant from template RNA-1 to RNA-2 using the region of overlap. The data represent the normalized strand-transfer frequencies relative to the total cDNA copies. The data are representative of three independent experiments and are presented as mean ± SD. The values used to make the figures are presented in a tabulated form. The data were analysed using a One-Way Analysis of Variance, followed by Tukey-Kramer post. hoc tests. ***p < 0.001, ns = non-significant.

4.7 HIV-1C RT can extend 3'-OH mismatches more efficiently.

The primary objective of the present work is to explore the molecular mechanisms underlying the higher frequency of specific sequence motif duplications observed in HIV-1C. The precondition for sequence motif duplication is the ability of the RT to switch between two templates and, importantly, hybridize to a sequence on the acceptor RNA molecule already been copied. However, the RT may often face two technical challenges while attempting to duplicate a sequence motif. First, the hybridization to a copied sequence is unlikely to involve absolute annealing conditions between the cDNA and the acceptor RNA template. Therefore, hybridization must be achieved via annealing to a shorter sequence, often a minimum of three bases – a process called hybridization by the Micro-Homology Domain (MHD). However, a hybridized structure so formed is expected to be highly unstable and transient, thus, making such events of recombination extremely rare. The additional technical challenge for successful sequence motif duplication is the possible base-pair mismatch at the growing end of the nascent cDNA, if not all times (Figure-17). Thus, the ability of the RT to resume polymerization, disregarding the base-pair mismatch at the growing end, is expected to be crucial.

While analyzing the sequences of PTAP motif duplication spanning several genetic subtypes of HIV-1, we observed subtype-specific variations in the nature of the duplication. The comparison subsequently was restricted to HIV-1B and -1C sequences, given the limited number of available sequences for other subtypes. The preferred length of the PTAP motif duplicated is different between the two subtypes, as reported previously (Sharma et al., 2018) - three or six amino acids for HIV-1B and six or seven amino acids for HIV-1C, although the total length could be up to 16 amino acids in both subtypes (Figure-18). In HIV-1C, additional spikes, although smaller, were evident at nine, twelve, and fourteen amino acids. The frequencies of the remaining amino acid duplications do not vary significantly within a subtype. Thus, in addition to inter-subtype differences in the length of motif duplication, intra-subtype variations also exist, which probably cannot be explained by the differences in the strand-transfer ability.

Importantly, the sequence analysis led to another crucial lead associated with the length of the PTAP motif duplication in the two HIV-1 subtypes. The spikes in the motif length in both subtypes are associated with a perfect match at the 3' end of the cDNA with the MHD of the acceptor RNA (Figure-18A). Since duplications result from MHD-mediated recombination, it is not necessary that the terminal nucleotide always base-pairs correctly. In other words, when the bases of the MHD are a perfect match, preferably all three bases, the frequency of the sequence duplication is higher, as depicted schematically (Figure-18A). This observation was consistent for nearly all the spikes of sequence motif duplication (six, seven, twelve, and fourteen amino acid duplications) of HIV-1C and (three and six amino acid duplications of) HIV-1B. For example, when the nascent cDNA switches to the acceptor RNA template by hybridizing via the MHD of 3 bp, this may create perfect annealing (Figure-17B) or a mismatch at the growing end of the cDNA (Figure-17C). The perfect annealing will initiate efficient polymerization leading to the sequence motif duplication. (Please see Discussion, Section 5.5 for a comprehensive classification of the sequence motif duplications into four categories, illustrated with examples from the natural context (p. 60).

Of note, when cDNA encounters a mismatch at the growing end after annealing to the acceptor RNA via the MHD, the differential ability of the RT to extend such mismatches could confer a great replication advantage to a specific HIV-1 subtype. To this end, we analyzed all the sequences of both the HIV-1 subtypes where RT must have extended a mismatched base and compared the duplication frequencies (Figure-17C). It is evident that HIV-1C RT extended all the mismatches at a significantly higher frequency than HIV-1B. For example, while HIV-1C duplicated a nine amino acid PTAP motif at a frequency of 1.45%, this frequency was only 0.19% for HIV-1B. The results of the analysis are consistent with the model that HIV-1C RT can extend base-pair mismatches more efficiently than that of HIV-1B.

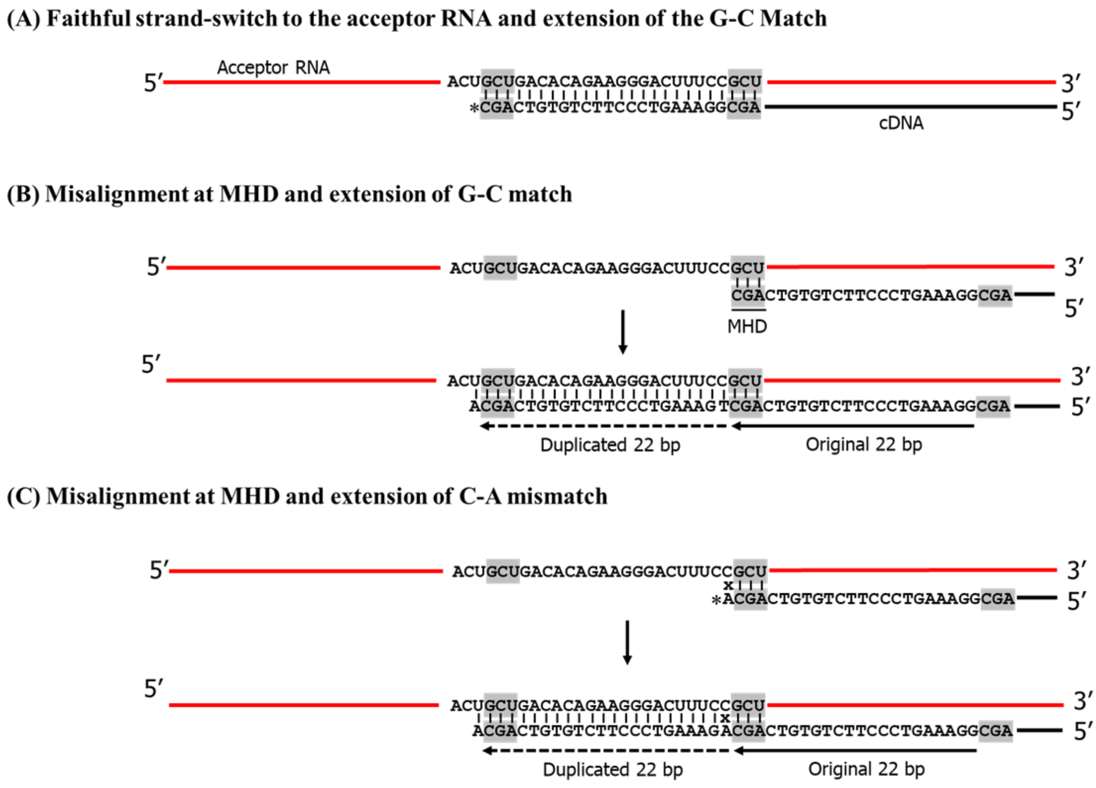
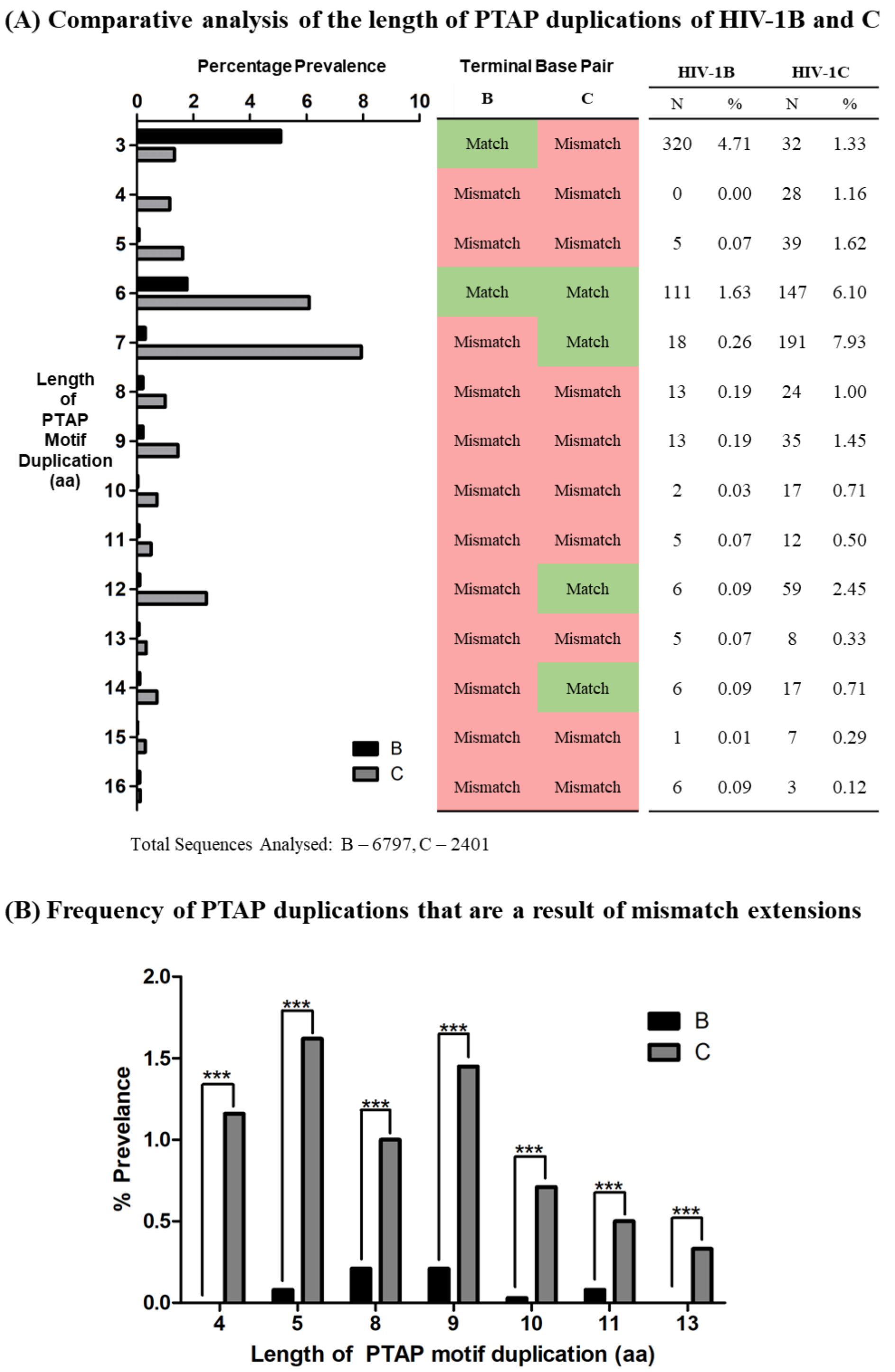

(A) Faithful strand-switch to the acceptor RNA and extension of the G-C Match
5′
Acceptor RNA
ACUGCUGACACAGAAGGGACUUUCCGCU
3′
*CGACTGTGTCTTCCCTGAAAGGCGA
5′
cDNA
(B) Misalignment at MHD and extension of G-C match
5′
ACUGCUGACACAGAAGGGACUUUCCGCU
3′
CGACTGTGTCTTCCCTGAAAGGCGA
5′
MHD
5′
ACUGCUGACACAGAAGGGACUUUCCGCU
3′
ACGACTGTGTCTTCCCTGAAAGTCGACTGTGTCTTCCCTGAAAGGCGA
5′
Duplicated 22 bp
Original 22 bp
(C) Misalignment at MHD and extension of C-A mismatch
5′
ACUGCUGACACAGAAGGGACUUUCCGCU
3′
*ACGACTGTGTCTTCCCTGAAAGGCGA
5′
5′
ACUGCUGACACAGAAGGGACUUUCCGCU
3′
ACGACTGTGTCTTCCCTGAAAGACGACTGTGTCTTCCCTGAAAGGCGA
5′
Duplicated 22 bp
Original 22 bp

Figure 17: A schematic representation of mismatch nucleotide extension. (A) Canonical viral recombination mediated by a faithful strand-switch to the acceptor RNA and the extension of the G-C match. The cDNA (black) aligns faithfully with the acceptor RNA (red), resulting in homologous recombination. The Micro-Homology Domain (MHD) and its associated triplet are shaded in grey. The asterisk represents the RT stall site. The matched G-C pair will mediate efficient recommencement of polymerization – an event not leading to sequence duplication. The donor RNA is not shown. **(B) Viral recombination via the cDNA annealing to the MHD and the extension of the G-C match**. The nascent cDNA mis-aligns at the MHD of the acceptor RNA, and the RT will extend the perfect G-C base pair match to resume polymerisation. **(C) Viral recombination via the cDNA annealing to the MHD and extension of the C-A mismatch**. Note that the RT stall site is 'UGCU', which ('GCU') is different from that of panel-B above. The hybridization of the cDNA to the acceptor RNA via the MHD results in mismatched base pairing (red arrow head) at the growing end. The ability of the RT to disregard the mismatch and extend the same efficiently could confer replication advantage. The solid and dashed arrows represent the 22 bp original and duplicated sequence motifs, respectively. The example illustrated here is adopted from Figure-23C in Discussion (See p. 64).

To this end, we experimentally evaluated mismatch extension by recombinant RT variants of HIV-1B and -C using a primer extension assay. All possible combinations of a single base-pair mismatch at the 3' end of a DNA primer hybridized to an appropriate template (Figure-19A). Four sets of four oligonucleotides in each set were designed to anneal to a synthetic, single-stranded DNA template of 105 bases (HXB2 coordinates - 1809 to 1914). The four oligonucleotides of each set anneal to the same sequence on the template; however, following the annealing, three oligonucleotides will have three different mismatches at the 3'- end. Thus, only four oligonucleotides represent a perfect match among the four sets. In contrast, the other 12 primers represent the 12 possible nucleotide mismatches, all annealing to the proximal target sequence of the same template DNA (Figure-19A). We did a primer-extension assay using the 16 primers and the four RT variant enzymes representing each of HIV-1B and HIV-1C subtypes (Figure-19B). The amount of ^{32}P-labelled dATP incorporation by each RT was quantitated in the assay by liquid scintillation spectrometry.

The results were plotted as the percentage of extension of the mismatched base-paired primer, using the corresponding correct base-paired primer as a reference (Figure-19B). The mismatch-extension assay demonstrated that wild-type HIV-1C RT displayed a superior ability to extend mismatches than HIV-1B RT. In nine out of the twelve mismatches tested (Template-Primer pairs: A-A, A-G, A-T, C-A, C-C, G-A, G-G, T-C, T-A), HIV-1C RT was superior in this function, of which four comparisons (C-A, C-C, T-C, T-T) were statistically significant (Figure-19B). Substituting Glycine (WT of HIV-1B) for Threonine (WT of HIV-1C) in HIV-1C RT reduced this ability in six out of the twelve mismatches (A-A, A-G, C-A, C-T, G-T, T-T). A reciprocal substitution of HIV-1B RT, Glycine-to-Threonine replacement, enhanced the ability of the enzyme to extend through the mismatches in ten of the twelve cases (A-A, A-C, A-T, C-A, C-C, C-T, G-A, G-G, G-T, T-T. Interestingly, the Glycine substitution appeared to enhance the mismatch extension ability of the



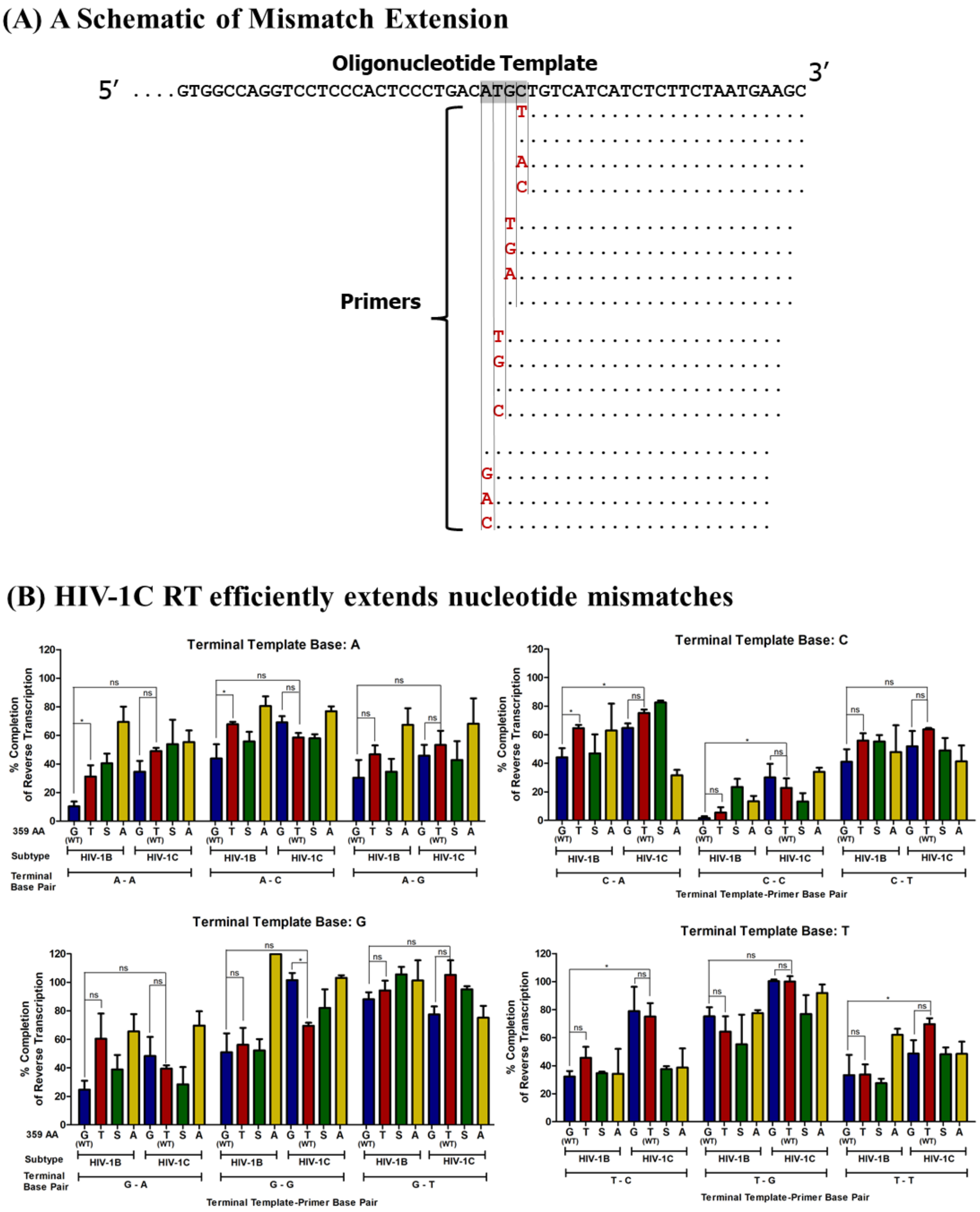
(A) Comparative analysis of the length of PTAP duplications of HIV-1B and C

Length of PTAP Motif Duplication (aa)	Terminal Base Pair B	Terminal Base Pair C	HIV-1B N	HIV-1B %	HIV-1C N	HIV-1C %
3	Match	Mismatch	320	4.71	32	1.33
4	Mismatch	Mismatch	0	0.00	28	1.16
5	Mismatch	Mismatch	5	0.07	39	1.62
6	Match	Match	111	1.63	147	6.10
7	Mismatch	Match	18	0.26	191	7.93
8	Mismatch	Mismatch	13	0.19	24	1.00
9	Mismatch	Mismatch	13	0.19	35	1.45
10	Mismatch	Mismatch	2	0.03	17	0.71
11	Mismatch	Mismatch	5	0.07	12	0.50
12	Mismatch	Match	6	0.09	59	2.45
13	Mismatch	Mismatch	5	0.07	8	0.33
14	Mismatch	Match	6	0.09	17	0.71
15	Mismatch	Mismatch	1	0.01	7	0.29
16	Mismatch	Mismatch	6	0.09	3	0.12

Total Sequences Analysed: B – 6797, C – 2401

(B) Frequency of PTAP duplications that are a result of mismatch extensions

Figure-18: Spikes of sequence motif duplication are associated with perfect base pairing: (A) Comparative analysis of the length of PTAP duplications of HIV-1B and -1C. HIV-1B and -1C *gag* sequences were downloaded from the HIV-1 LANL database and classified according to the length of the PTAP duplications in each subtype. The data are plotted as the percentage of sequences containing a duplication length of specific number of amino acids as shown, belonging to HIV-1B (black) and HIV-1C (grey). The table in the middle panel depicts the presence or absence of a perfect base-pair match between the nascent cDNA and the acceptor RNA molecule (color-coded). The table in the right panel depicts the numbers and the percentage prevalence of each duplication. **(B) A comparison of the frequencies of PTAP duplication resulting from mismatch extensions.** The duplication frequencies resulting from mismatch extensions of HIV-1B (Black bars) and HIV-1C (Grey bars) are represented for each of the motif duplications comprising four. five, eight, nine, ten, eleven, and thirteen amino-acid residues. The data were analysed using Fisher's exact test. ***p < 0.001.


(A) A Schematic of Mismatch Extension

(B) HIV-1C RT efficiently extends nucleotide mismatches

Figure-19: HIV-1C RT can extend 3'-OH mismatched bases more efficiently. (A) A schematic presentation of the template and primer complementarity. A synthesis single-stranded DNA of 105 bp region of HIV-1 (HXB2 coordinates - 1809 to 1914) was used as the template in the primer extension assay. Four sets of primers of four in each set anneal to the template as depicted. Three primers in each set contain a base-pair mismatch with the template at the 3'-end, as shown. All 12 possible mismatches are represented in the primer design. Dots represent a complimentary base. **(B)** HIV-1C RT efficiently extends nucleotide mismatches. Four µg of an oligonucleotide template was annealed to one µg of 16 different primers individually, each condition representing one primer-template base-pair combination. The primers were extended using 1 Unit of each RT in the presence of ^{32}P-labelled dATP, and the amount of labelled dATP incorporated was measured using liquid scintillation spectrometry. The terminal base on the template is indicated above each of the four panels shown above. The corresponding base pair on the primer that the RT extends is indicated below the X-axis using the single letter code for base pairs. The amino acid at position 359 is shown using the single letter amino acid code and are colour coded The data are representative of three independent experiments and are depicted as mean ± SD, and the performance of the wild-type RTs and the corresponding Glycine/Threonine variants was analysed using One-Way Analysis of Variance followed by Tukey-Kramer post. hoc tests. *p < 0.05, ns = non-significant.

RT in three cases (A-C, G-G, C-C). Serine substitution in HIV-1B and -1C RTs revealed trends similar to the Threonine variant and the wild-type C-RT, respectively. The Alanine variant of both RTs performed the best with all the primer-template combinations tested. The superior performance of the Alanine variant, consistent with its performance in the strand-transfer assay

(Figure-16), warrants further investigation once the crystal structure of HIV-1C RT has been solved.

The data collectively appear to allude to superior propensity of HIV-1C RT to extend most nucleotide mismatches compared to HIV-1 B RT, probably contributing to the C-RT ability in duplicating sequences efficiently (see Section 5.2 in Discussion). We are presently conducting an analogous experiment using an RNA template to evaluate the ability of different RTs to extend mismatches in a reverse transcription assay.

4.8 References

Bachu, M., Yalla, S., Asokan, M., Verma, A., Neogi, U., Sharma, S., Murali, R. V., Mukthey, A. B., Bhatt, R., Chatterjee, S., Rajan, R. E., Cheedarla, N., Yadavalli, V. S., Mahadevan, A., Shankar, S. K., Rajagopalan, N., Shet, A., Saravanan, S., Balakrishnan, P., … Ranga, U. (2012). Multiple NF-κB Sites in HIV-1 Subtype C Long Terminal Repeat Confer Superior Magnitude of Transcription and Thereby the Enhanced Viral Predominance. Journal of Biological Chemistry, 287(53), 44714–44735. https://doi.org/10.1074/jbc.M112.397158

Chen, Y., Balakrishnan, M., Roques, B. P., & Bambara, R. A. (2003). Steps of the acceptor invasion mechanism for HIV-1 minus strand strong stop transfer. Journal of Biological Chemistry, 278(40), 38368–38375. https://doi.org/10.1074/jbc.M305700200

Chin, M. P. S., Rhodes, T., Chen, J., Fu, W., & Hu, W.-S. (2005). Identification of a Major Restriction in HIV-1 Inter-subtype recombination. Proc Natl Acad Sci U S A, 102(25), 9002–9007. https://doi.org/10.1073/pnas.0502522102

Iordanskiy, S., Waltke, M., Feng, Y., & Wood, C. (2010). Subtype-associated differences in HIV-1 reverse transcription affect the viral replication. Retrovirology, 7(1), 85. https://doi.org/10.1186/1742-4690-7-85

Lanciault, C., & Champoux, J. J. (2006). Pausing during Reverse Transcription Increases the Rate of Retroviral Recombination. Society, 80(5), 2483–2494. https://doi.org/10.1128/JVI.80.5.2483

Levy, D. N., Aldrovandi, G. M., Kutsch, O., & Shaw, G. M. (2004). Dynamics of HIV-1 recombination in its natural target cells. Proceedings of the National Academy of Sciences, 101(12), 4204–4209. https://doi.org/10.1073/pnas.0306764101

Murray, J. M., Kelleher, A. D., & Cooper, D. A. (2011). Timing of the Components of the HIV Life Cycle in Productively Infected CD4+ T Cells in a Population of HIV-Infected Individuals. Journal of Virology, 85(20), 10798–10805. https://doi.org/10.1128/jvi.05095-11

Negroni, M., & Buc, H. (1999). Recombination during reverse transcription: an evaluation of the role of the nucleocapsid protein. Journal of Molecular Biology, 286(1), 15–31. https://doi.org/10.1006/jmbi.1998.2460

Rhodes, T. D., Nikolaitchik, O., Chen, J., Powell, D., & Hu, W.-S. (2005). Genetic Recombination of Human Immunodeficiency Virus Type 1 in One Round of Viral Replication: Effects of Genetic Distance, Target Cells, Accessory Genes, and Lack of High Negative Interference in Crossover Events. Journal of Virology, 79(3), 1666–1677. https://doi.org/10.1128/JVI.79.3.1666-1677.2005

Sharma, S., Arunachalam, P. S., Menon, M., Ragupathy, V., Satya, R. V., Jebaraj, J., Aralaguppe, S. G., Rao, C., Pal, S., Saravanan, S., Murugavel, K. G., Balakrishnan, P., Solomon, S., Hewlett, I., & Ranga, U. (2018). PTAP motif duplication in the p6 Gag protein confers a replication advantage on HIV-1 subtype C. Journal of Biological Chemistry, 293(30), 11687–11708. https://doi.org/10.1074/jbc.M117.815829

Verma, A., Rajgopalan, P., Lotke, R., Veerapaneni, S., Bachu, M., & Ranga, U. (2013). Unique molecular features of the HIV-1 subtype C enhancer and core promoter and their influence on the viral gene expression. Retrovirology, 10(Suppl 1), P94. https://doi.org/10.1186/1742-4690-10-S1-P94

Viguera, E., Canceill, D., & Ehrlich, S. D. (2001). Replication slippage involves DNA polymerase pausing and dissociation. EMBO Journal, 20(10), 2587–2595. https://doi.org/10.1093/emboj/20.10.2587

Von Wyl, V., Ehteshami, M., Demeter, L. M., Bürgisser, P., Nijhuis, M., Symons, J., Yerly, S., Böni, J., Klimkait, T., Schuurman, R., Ledergerber, B., Götte, M., & Günthard, H. F. (2010). HIV-1 reverse transcriptase connection domain mutations: Dynamics of emergence and implications for success of combination antiretroviral therapy. Clinical Infectious Diseases, 51(5), 620–628. https://doi.org/10.1086/655764

Xu, H.-T., Quan, Y., Asahchop, E., Oliveira, M., Moisi, D., & Wainberg, M. A. (2010). Comparative biochemical analysis of recombinant reverse transcriptase enzymes of HIV-1 subtype B and subtype C. 1–11.

Chapter – 5: Discussion

The basis of the present work lies in two hotspots of sequence-motif duplication in the genome of HIV-1C – p6-Gag and the LTR. Numerous previous publications have reported duplications or deletions of sequences and motifs in other regions spanning the viral genome of HIV-1 subtypes (Carl et al., 2000; Dang & Hu, 2001; Ji et al., 2018). However, sequence motif duplications in the two hotspots mentioned above in HIV-1C differ in several respects. The sequence duplication frequency is significantly higher (PTAP motif duplication, Sharma et al., 2017), longer and complete (PTAP motif duplication, Sharma et al., 2018), or unique (NF-κB motif duplication, Bachu, Mukthey, et al., 2012)). Importantly, sequence-motif duplication in both the hotspots appears to confer a profound replication-fitness advantage on the variant viral strains, thus, alluding to directional viral evolution and a positive selection of these genetic variations (Bachu, Yalla, et al., 2012; Sharma et al., 2018). The phenomenon of subtype-specific or -unique association of the sequence motif duplication in HIV-1C warranted the characterization of the phenomenon at the molecular level. The present thesis attempted to examine whether subtype-specific RT functions could underlie such differences.

5.1 The T359 signature amino acid residue may stabilize the RT functions of HIV-1C

` The bioinformatic analysis of RT sequences of different HIV-1 subtypes identified several signature amino acid residues unique to HIV-1C (Table-7, Chapter 4, p. 37). Of these subtype-specific variations, we evaluated the non-conservative amino-acid substitution from Glycine to Threonine at position 359 in the connection domain of RT. The presence of a Threonine residue at position 359 alluded to the formation of an additional hydrogen bond between the nascent DNA and RT, given the proximity between the two. To this end, a major technical limitation to confirm the presence of the hydrogen bond is the non-availability of a crystal structure of HIV-1C RT. While efforts are currently underway in our laboratory to confirm the formation of a hydrogen bond between the nascent DNA and RT due to the presence of T359 residue in the enzyme, two experimental observations support the possible formation of the hydrogen bond. First, a substantial increase in RT activity was observed (Figure-13, Chapter 4, p. 41) when T359 in HIV-1C RT was substituted with a Glycine residue that lacks a hydrogen bond-forming -OH group. Further, substitution with a Serine, which is predicted to form a weak hydrogen bond, resulted in a marginal increase in the activity of the RT. Second, the sequence identity between the RTs of HIV-1B and 1C ranges between 88 – 93%, therefore, the ability of the PyMol software to provide an accurate prediction of the formation of a hydrogen bond when Threonine is positioned at location 359 of the HIV-1B RT structure is significantly high.

Importantly, the specific activities of wild-type RTs of both HIV-1B and -1C are comparable (Figure-13, Chapter 4, p. 41). The specific activity of C-RT is many folds higher when Glycine, as compared to Threonine, is present at position 359 suggesting that the consequential formation of an additional hydrogen bond due to Threonine may ensure optimal functioning of the enzyme by guarding against hyperactivity. Since HIV-1 reverse transcription is an exceptionally slow process requiring multiple events to happen simultaneously, a hyper-active enzyme could be detrimental to viral replication. This differential enzyme activity influencing viral replication kinetics is consistent with our observation of the Threonine RT variant outperforming the Glycine variant in p24 production (Figure-14, Chapter 4, p. 43).

Of note, while the attention of the present work has been focused on the T359 residue and its potential to form an additional hydrogen bond with the nascent DNA molecule, the significance of the other signature amino acid residues identified in HIV-1C RT should not be undermined. Importantly, C-

RT containing a Glycine residue at position 359 demonstrates a six-fold higher polymerase activity than B-RT. This augmented enzyme activity may be ascribed to the presence of the other signature amino acid residues of C-RT. Thus, T359 may play a compensatory role in maintaining optimal polymerization of C-RT. We have identified five additional signature amino-acid residues, other than T359, in C-RT that may also influence the subtype-specific functions of the enzyme. Of these signature residues, E39 and T48 are of significance as they are located proximal to the dNTP binding pocket in the fingers domain. Given that the fingers domain makes several contacts with the viral RNA (Patel & Loeb, 2001; Warrilow et al., 2009), the presence of variations in this domain may modulate strand transfer and the mismatch-extension qualities of C-RT.

5.2 Enhanced strand-transfer and mismatch extension qualities of HIV-1C RT

Recombination is an essential process central to the propagation of nearly all life forms, particularly RNA viruses. This is because RNA viruses rely heavily on recombination to generate new viral variants or purge their genomes of deleterious mutations. This ability, coupled with the promiscuous nature of viral RNA polymerases, enables these viruses to produce quasi-species that serve to escape immune surveillance. However, recombination rates vary among viruses, with retroviruses displaying the highest magnitude of recombination. The recombination rates of retroviruses are maintained many folds higher than those of other RNA viruses, given the pseudo-diploid nature of retroviruses, the co-packaging of two RNA strands, and the intrinsic programming of strand-transfers during replication. Within Retrovirideae and among lentiviruses, the recombination rates of HIV-1 are at least 10-folds higher than other lentiviruses, including Murine Leukemia Virus or Spleen Necrosis Virus (Onafuwa et al., 2003).

Although the high recombination frequency of HIV-1 was initially ascribed to the biochemical properties of the reverse transcriptase, subsequent data showed that the differences in RNA packaging are more important. However, it is not clear if similar differences in recombination rates exist within different genetic subtypes of HIV-1. In HIV-1C, the frequency of sequence duplications is significantly higher than that in other HIV-1 subtypes (Bachu, Mukthey, et al., 2012; Bachu, Yalla, et al., 2012; Bhange et al., 2021; Martins et al., 2011; Sharma et al., 2017, 2018). Since recombination is a prerequisite for sequence duplication, it is reasonable to presume that HIV-1C recombines at a rate higher than other subtypes. In contrast, previous work from other groups (Chin et al., 2005; Galli et al., 2010) and data from the present work ascertain that recombination rates among different HIV-1 subtypes are not significantly different. Thus, variable recombination rates cannot explain the observed difference in the frequency of sequence motif duplication in HIV-1C. To this end, the data presented here offer crucial leads. While the results of the strand-transfer assay provide direct evidence (Figure-16, Chapter 4, p. 46), those of the mismatch extension assay offer indirect proof (Figure-19, Chapter 4, p. 50). Both these qualities of RT are probably reinforced by the prospects of an additional hydrogen bond with the nascent DNA molecule. Thus, all these properties collectively allude to the molecular basis of a higher frequency of sequence motif duplications observed in HIV-1C.

To the best of our knowledge, previous publications did not directly compare the ability of RTs of different subtypes to switch between strands. The previous studies predominantly used the cell culture-based EGFP complementation assay to compare recombination frequencies. This assay lacks the finer resolution needed to quantitate the subtle differences between RTs to switch between the RNA strands. Secondly, the EGFP assay is limited by its ability to estimate RT switch within a narrow window frame of only 600 bp. It is assumed that the RT switch is uniform across the viral genome and that the data of the 600 bp region can be extrapolated to the entire genome, thus possibly disregarding the recombination hotspots and RNA secondary structure differences. Our work, too, suffers from the same limitation that the frequency of RT switching has not been examined using a few other reporter genes.

However, sequence duplication is the product of micro-homologous recombination, a process estimated to occur at a frequency between 10 and 100-fold lower than homologous recombination (Zhang & Temin, 1993). The strand transfer assay, although not a perfect indicator, can provide a broader estimate of the ability of different RTs to affect micro-homologous recombination. This is because the assay measures the frequency of cDNA synthesis on the acceptor strand after a forced transfer and, therefore, mimics the conditions of a sequence duplication more closely.

Importantly, micro-homologous recombination could often lead to two technical problems. First, the intermediate secondary structure formed by the cDNA is expected to be extremely unstable and transient, given the absence of sufficient complementarity between the growing end of the nascent DNA and the micro-homology domain on the template. Additionally, in certain cases, the absence of complementarity between the residue located at the 3' growing end of the nascent DNA and the residue present at this location on the acceptor RNA could lead to a condition where three of four times a mismatch exists between them (Figures-23, -25, see below models MR-MME and NR-MME). This unique condition obviates that RT must be endowed with an ability to disregard mismatches and resume polymerization efficiently. Thus, an additional copy of a specific sequence motif can be created only when the two successive and rare events have been successfully accomplished.

Our data are suggestive that the presence of a Threonine residue at position 359 of the connection domain of C-RT and the formation of an additional hydrogen bond with the growing end of the nascent DNA can stabilize the unstable looped structure between the nascent DNA hybridized to the MHD on the recipient RNA template. Additionally, our data also confirm the superior efficiency of C-RT in extending a mismatched residue at the growing end of the nascent DNA (Figure-19, Chapter 4, p. 50) It has been established that a mismatched base can promote recombination (Chin et al., 2007; Palaniappan et al., 1996; Schlub et al., 2014). The ability of C-RT to extend mismatched bases could overcome this drawback, thereby enabling the enzyme to continue polymerization on the acceptor RNA without a second switch resulting in a sequence duplication.

In summary, by characterizing the two unique qualities of C-RT, our results elucidate the molecular causes underlying the higher frequency of sequence motif duplications observed in HIV-1C, reported by several groups. Additional work is warranted to delineate the functional significance of the other signature amino acid residues identified through this work to the unique functions of HIV-1C RT.

5.3 Sequence Duplications in HIV-1C: The role of Darwinian Selection.

A direct consequence of the high recombination frequency of HIV-1 RT is the generation of diverse viral recombinant forms of HIV-1. The recombinants of the primary viral subtypes have been subjected to extensive analyses given the relevance to drug resistance, global sequence diversity, and viral evolution (Taylor et al., 2008). Two major classes of recombinants may be identified – those that have established themselves in the population, classified as the Circulating Recombinant Forms (CRFs), and those isolated from sporadic cases, labelled as the Unique Recombinant Forms (URFs). Both CRFs and URFs are the products of homologous recombination, a process estimated to occur three to five times on average per round of viral replication (Onafuwa-Nuga & Telesnitsky, 2009). Of note, CRFs and URFs result from the superinfection of the same cell by viral variant strains belonging to two different subtypes. On the other hand, sequence duplications differ from these two recombinant forms in two aspects. Firstly, they do not require superinfection for their generation. Second, they are a product of micro-homologous recombination, not homologous. The relevance of the variant viral strains to drug resistance and disease management awaits stringent evaluation (discussed in Section 5.6 below).

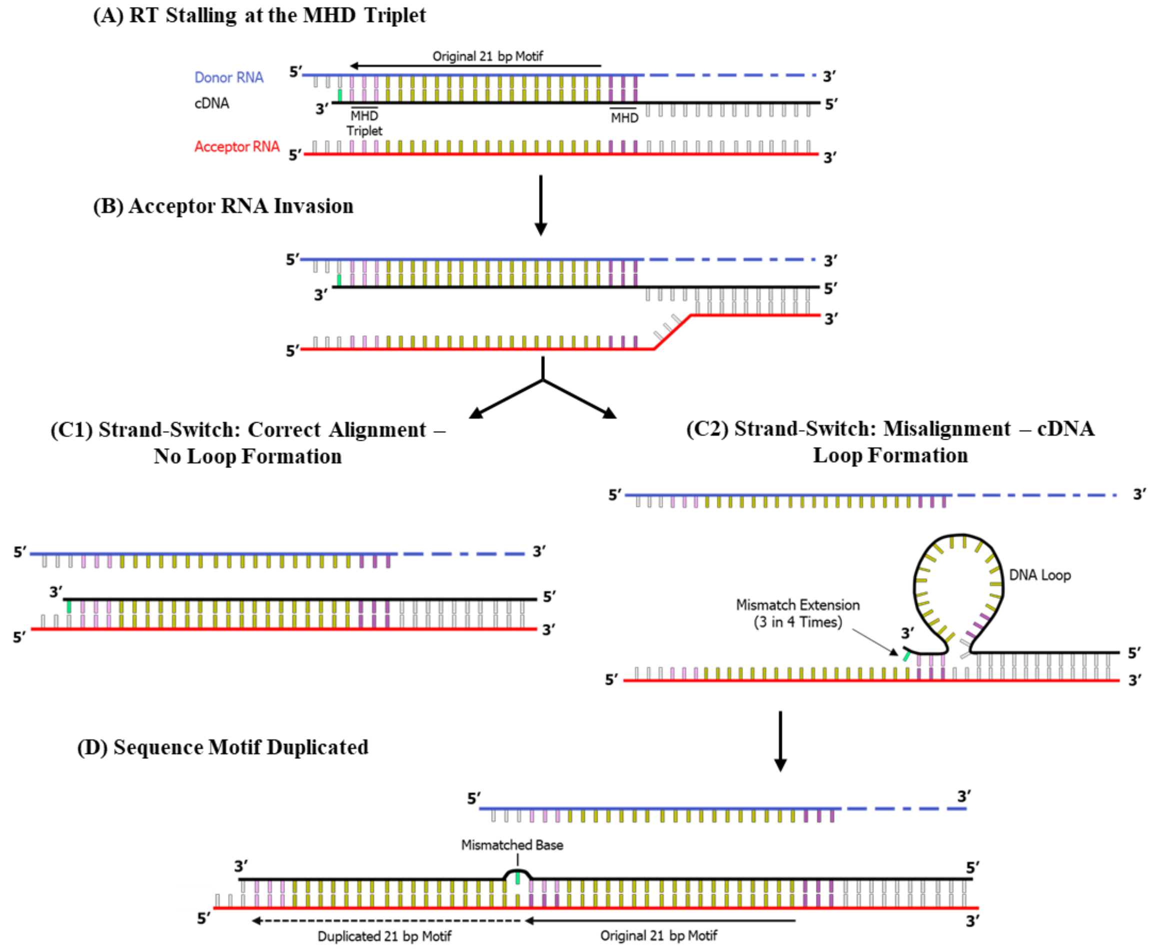
Of note, sequence duplications in HIV-1C may not be limited to the LTR and the p6 protein but may span the length of the viral genome. Since a positive correlation is expected between the frequencies of recombination and sequence duplication, recombination hotspots may also coincide with duplication hotspots. However, the limited analysis we performed, using the viral sequences available at the extant sequence databases, failed to find any additional sequence duplication hotspots in the viral genome apart from the two we reported in p6-Gag and the LTR. Importantly, previous work from our laboratory offered extensive experimental evidence that sequence duplication at the two hotspots in HIV-1C confers a significant replication advantage on the variant viral strains (Bachu, Yalla, et al., 2012; Sharma et al., 2018). Thus, our observations are suggestive that, unlike the genetic variations appearing at a uniform level in the other locations of the viral genome, the variations appearing in the two hotspots, LTR and p6-Gag, are subjected to immediate and strong positive selection. Such variations in the other regions of the viral genome may or may not be subjected to selection above a baseline level. In most cases, duplications or deletions occurring elsewhere in the genome would likely be lethal due to the strong possibility of frameshifts or deletions in important motifs. Thus, despite the natural propensity of the RT to create sequence duplications in various other regions of the viral genome, probably uniformly, these events may not have the same selection advantage as these two hotspots enjoy in HIV-1C. As a result, these two molecular events are selected at a significantly higher frequency in the population once they are generated. Therefore, while a higher propensity to create duplications will also be associated with a concomitant increase in the creation of defective viral variants, the shallow frequency of non-homologous recombination ensures that these defective viral strains form a tiny fraction of the reverse-transcribed viral genomes.

5.4 A model to explain how HIV-1C RT can produce sequence motif duplications at a higher frequency.

The dynamic-copy-choice recombination model remains the most widely accepted paradigm to explain homologous recombination in retroviruses (Hwang et al., 2001; Onafuwa-Nuga & Telesnitsky, 2009; Rawson et al., 2018). According to the model, a disturbance to the fine balance between the polymerase and the RNase H activities of the RT is crucial for a template switch. Mutations modulating either of the enzyme activities or the presence of RNA secondary structures can alter the rate of recombination. A corollary to the dynamic-copy-choice paradigm is the dock-and-lock model of acceptor strand invasion catalyzing nascent DNA hybridization to the acceptor RNA molecule, subsequently promoting strand switch (Matteo Negroni and Henri Buc, 2000; Roda et al., 2002, 2003). The dock-and-lock model proposes that the acceptor RNA hybridizes at a slow rate with the nascent cDNA that trails behind RT (Figure-20B). Ultimately, the invasion of the polymerization bubble will be complete with the acceptor RNA strand displacing the donor RNA molecule thereby forcing the enzyme to switch to the acceptor RNA-cDNA hybrid and continue polymerization (Negroni and Buc, 2001; Balakrishnan et al., 2003; Delviks-Frankenberry et al., 2011).

However, these two models can explain only homologous recombination, where the complementarity between the nascent DNA and the template RNA is expected to be significant or absolute. In contrast, microhomology-mediated recombination is expected to function in the absence of perfect complementarity, despite the highly unstable nature of the DNA-RNA hybrid. An additional challenge to this end is the possibility of a mismatch base extension. Thus, although sequence motif duplication in HIV-1 is rooted in the dynamic-copy-choice paradigm and the dock-and-lock model, a different and comprehensive schematic is required to explain how the challenges of sequence motif duplications are circumvented.

Here we propose a schematic model (Figures-20 and -21) to explain the mechanism of sequence motif duplications observed at two different locations in the HIV-1C genome – p6-Gag and the LTR. The proposed schematic has its foundation on the basic tenets of retroviral recombination blended with the experimental leads obtained through the present work. Further, our model portrays a broader framework of sequence motif duplication and lends itself to explain several subthemes of sequence



(A) RT Stalling at the MHD Triplet

(B) Acceptor RNA Invasion

(C1) Strand-Switch: Correct Alignment – No Loop Formation

(C2) Strand-Switch: Misalignment – cDNA Loop Formation

(D) Sequence Motif Duplicated

Figure-20. A generalized model portraying the various stages of the duplication of a sequence motif. The model portrays the duplication of 21 bp motif. The blue and red horizontal lines represent the two template RNA molecules, donor and acceptor, respectively, and the black line the nascent DNA being polymerized by RT. The dashed lines represent template RNA hydrolyzed due to the RNase H activity. Vertical lines between two strands represent hydrogen bonds formed due to complementarity. Three vertical lines in dark and light purple colours represent the micro-homology domain (MHD), and the MHD triplet, respectively, that consist of at least three base pairs. The 21 bp sequence flanked by the MHD and the MHD triplet is highlighted with a solid arrow which also shows the direction of polymerization. The terminal base that RT will extend is highlighted in green (**A)** The RT stalls after polymerizing the 21 bases, the three residues comprising the MHD triplet at the 3'-end of the cDNA are highlighted (light purple). Note that the three residues of MHD and MHD triplet are identical and flank the 21 bases to be duplicated. RT may resume polymerization on the same RNA template, in which case there will be no recombination (not shown). There are two other possible and mutually exclusive outcomes of the RT stalling as shown below. Of note, the RNA-DNA hybrid of approximately 20 residues is encompassed by the RT complex and protected. (**B)** The acceptor RNA (red) begins to invade the exposed 5' end of the cDNA (black) and makes an RNA-DNA hybrid outside the RT complex. (**C1)** Strand switch to the correct position on the acceptor RNA (Red). Following RT stalling, the RT complex disassembles and the donor RNA-DNA complex un-hybridizes. The invading acceptor RNA continues the process of invasion to form a new hybrid with the nascent DNA. A new RT molecule recognizes the fresh RNA-DNA hybrid and forms a new RT complex. The entire process represents the conventional viral recombination not leading to sequence motif duplication. (**C2)** Strand-switch followed by alignment of DNA via the MHD. The nascent DNA separated from the donor RNA and hybridizes with the acceptor RNA via the micro-homology domain of three residues (light and dark purple). The intervening sequence of 21 bases between the two hybridized regions forms a single-stranded loop. Note that the residue at the 3'end (shown in green) is a mismatch with the base on the template, as a result, the nascent DNA is not hydrogen-bonded with the template at the growing end. In a case like this, the RT must demonstrate a special ability to resume polymerization disregarding the base pair mismatch. The highly unstable RNA-DNA structure is recognized by a new RT molecule and a replication complex is formed. We propose that the T359 residue of HIV-1C RT located in the connection domain can stabilize the highly unstable RNA-DNA structure by forming an additional hydrogen bond with the nascent DNA. Thus, the unique qualities of C-RT to form an additional hydrogen bond with the nascent DNA and the ability to extend a mismatch may underlie the higher frequency of sequence motif duplication in HIV-1C. (**D)** A stretch of 21 base pairs has been duplicated, as highlighted using a broken arrow. Note that the base pair mismatch is fixed as a new variation. The model largely recapitulates the creation of the fourth copy of NF-κB binding motif of 22 base pairs in HIV-1C LTR (See Figure-23, p. 64).

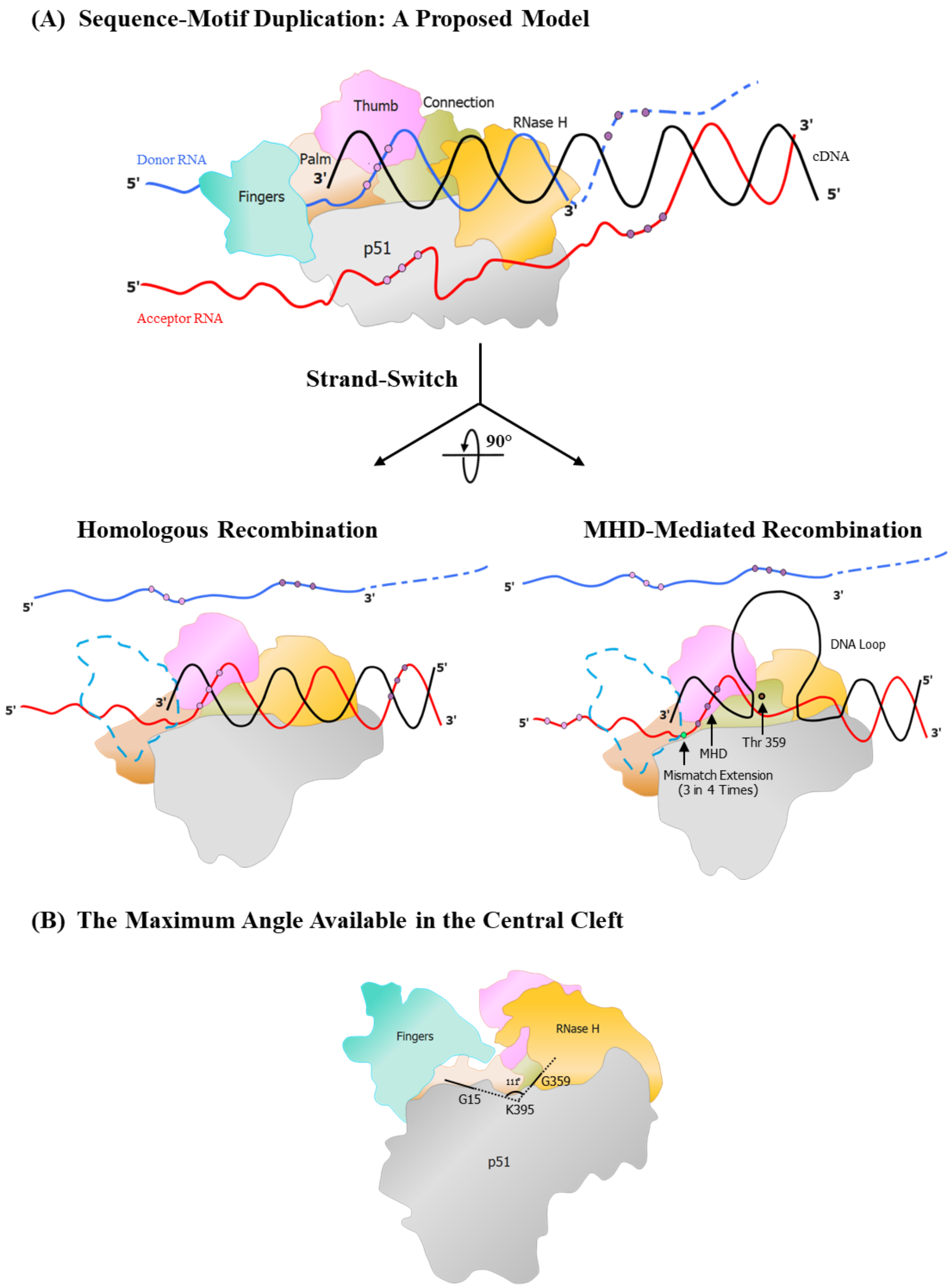
motif duplication observed in HIV-1C. The common paradigm of the model is presented in two overlapping schematics. The first schematic (Figure-20) depicts the nucleic acid hybridization between the nascent DNA and template RNA molecules. The second schematic (Figure-21) portrays the same theme in the context of the RT structure. Thus, the two complimentary schematics collectively offer a comprehensive view of the paradigm of microhomology-based sequence motif duplication in HIV-1C, and by extension, in the other genetic subtypes of HIV-1.

The model portrayed in Figures-20 and -21 presents a generalized theme of the duplication of a 21 base sequence motif flanked between three identical bases. The RT is expected to stall after copying the three bases located at the growing end of the nascent DNA. We label these three residues as the ‘microhomology domain (MHD) triplet’ to distinguish it from the identical triplet located at the 3’ end of the 21 bases to be duplicated. The latter is labelled as ‘MHD’. The number of residues between the MHD and MHD triplet could vary ranging from nine to forty-eight, depending on the longest duplication observed in the sequence database. The example of 21 bp duplication is reminiscent of the creation of the F-κB motif by sequence duplication in HIV-1C (Bachu, Yalla, et al., 2012).

At the time of RT stalling, the nascent DNA is hybridized to the donor RNA template and approximately 20 base pairs of the RNA-DNA hybrid are encompassed by the RT complex (Figure-20A). The downstream sequence of the template RNA, which has been reverse transcribed already, is hydrolyzed by the RNase H activity of the RT. The single-stranded DNA of this region is hybridized with the complementary region of the second RNA template, the acceptor RNA molecule, in accordance with the dock-and-lock model (Matteo Negroni and Henri Buc, 2000; Roda et al., 2002, 2003). Thus, at this time, the nascent DNA is hybridized to both the template RNA molecules in different regions (Figure-20B). The RT may resume reveres transcription using the same RNA template, in which case there will be no recombination. Alternatively, the RT complex may disassemble, and the RNA-DNA hybrid may disassociate, permitting the invasion of the donor RNA molecule to continue until a new RNA-DNA complex is formed. The new RNA-DNA hybrid is recognized by a new RT molecule to form a new RT complex (Figure-20C1). This event would lead to viral recombination, but not sequence duplication. Alternatively, after dissociating from the donor RNA molecule, the nascent DNA may bind the recipient RNA template via the three bases of the MHD (Figure-20C2), thus strand-switching to a location on the template RNA molecule that has already been copied. The 21 bases of the nascent DNA intervening between the two regions hybridized to the recipient RNA will form a loop as these residues do not find a complementary sequence on the template. It is often possible that an additional base is present on the nascent DNA at the growing end after the MHD triplet. When the nascent DNA strand switches to the bonafide location on the template RNA by homologous recombination, all the four bases at the 3’ end of the nascent DNA, as well as the other residues, will be a perfect match with the complementary bases on the template RNA. In the case of MHD-mediated recombination (MR), the ultimate base at the 3’ end of the nascent DNA is expected to find a perfect match only one in four times (Figure-20C2). Our results show that HIV-1C RT can extend base pair mismatches more efficiently (Figure-19, Chapter 4, p. 50). Given that only three base pairs are engaged in forming the RNA-DNA hybrid, such a hybrid structure is expected to be highly unstable.

However, when a new RT molecule binds to this looped RNA-DNA structure, the additional hydrogen bond formed between the nascent DNA and T359 of C-RT may stabilize the hybrid and permit the resumption of polymerization. The resumed reverse transcription will copy the sequence motif for the second time, thus creating an additional sequence motif and fixing the variation due to mismatch extension (Figure-20D). Thus, the two unique properties of C-RT through the present work, the ability to form a new hydrogen bond and mismatch extension, may collectively enable the enzyme to create sequence motif duplications at a significantly higher frequency. For additional clarity, the MR-mediated sequence duplication is presented in the context of the RT structure (Figure-21).

Our model predicts that the DNA loop, ranging up to 48 base pairs in length, extends outward, away from the central DNA binding groove of the RT, as is observed from viral sequences available from the extant databases. From the crystal structure of HIV-1B RT deposited to the Protein Data Bank (PDB ID: 5J2M), we determined the width of the DNA binding groove to be 16.3 Å at its minimum



(A) Sequence-Motif Duplication: A Proposed Model

Strand-Switch

Homologous Recombination

MHD-Mediated Recombination

(B) The Maximum Angle Available in the Central Cleft

Figure-22: A generalized model depicting sequence motif duplication. (A) Sequence duplication: A proposed model. The figure is an extension of Figure-21, and the colour codes of the nucleic acids are consistent between the two figures. The RT domains are depicted as follows – fingers (green), thumb (pink), RNase H (beige), palm (brown), connection (green), and the p51 subunit (grey). The three base pairs that form the MHD and the MHD triplet are shown as purple dots on the donor and acceptor RNA molecules. **(A)** RT stalled at the MHD and the acceptor RNA begins to invade from the 3' end of the nascent DNA, as described in Figure-21A and B. The two outcomes C1 and C2 of Figure 21 are shown using a bifurcated arrow. The RT complex is rotated by 90° for the clear presentation of the replication complex from the top view. In this view, only the outline of the fingers domain is depicted to aid the visualization of the polymerase active site. The central left panel depicts strand-switching by homologous recombination, where the cDNA and the acceptor RNA molecules align perfectly. The central right panel presents the alignment of the cDNA to the acceptor RNA template via the MHD, leading to the unaligned sequence forming a loop. The approximate location of the T359 residue is indicated using a brown circle. The mismatched base is shown using a grey dot on the DNA molecule. **(B) The maximum space available in the central RT cleft.** The RT is presented as it is seen by the RNA while entering the cleft. The direction of polymerization in (B) is perpendicular to the view presented. The angle available within the central cleft for the cDNA-RNA complex to rotate as a nucleotide is added during polymerization is measured spanning the G15, K 395, and G359 residues, as shown. Dashed lines represent the movement possible within the cleft before further polymerization is obstructed by the domains of the RT. See text for more details.

between the R78 and L289 residues of the p66 chain and 46.1 Å at its maximum between the G15 and D471 residues of the p51 and p66 chains, respectively. Therefore, the available space appears to be sufficient for the loop to extend outwards and away from the RT (Figure-21A, bottom right panel).

With continued reverse transcription, the enzyme translocates along with the template, in a way that the incorporation of each base causes the DNA to undergo an approximately 36° turn relative to the enzyme along the axis of polymerization. Consequently, with each incorporated nucleotide, the loop must also undergo a 36° turn along the axis. Since the orientation of the DNA within the groove permits an unhindered turn of only up to 111° (as measured across the G359, K395, and G15 residues), our model permits the RT to add a maximum of three nucleotides before the loop encounters the p51 subunit that debilitates reverse transcription further (Figure-21B). At this point, the RT complex must dissociate, releasing the DNA-RNA complex, and polymerization must be re-initiated by a different RT molecule. The process of three nucleotide addition and enzyme dissociation must repeat for multiple rounds until the loop has exited the RT complex, and the RT can resume polymerisation normally. Our model predicts that the incorporation of the initial 3-6 base pairs is more challenging for the enzyme due to the proximity of the loop to the active site and as the loop migrates further away from the active site the rate of polymerization becomes normal. Our predictions are consistent with the observation that the error rate is the highest in bases within three to six positions from the junction between the original and duplicated sequences.

Our data collectively suggest that the hydrogen bond formed due to the T359 residue of the C-RT appears to stabilize the RT-cDNA loop-acceptor RNA complex enabling the enzyme to duplicate sequence motifs more efficiently. In other genetic subtypes of HIV-1, the absence of this stabilizing hydrogen bond probably causes the RT to dissociate from the template when it encounters the loop leading to a significantly low frequency of sequence motif duplication.

5.5 Types of Sequence-Motif Duplication.

Sequence motif duplication in HIV-1C presents a highly complex and variable scenario that, despite the common framework that underlies the phenomenon, leads to diverse outcomes in various regions of the genome. For example, in the LTR, the sequence of the fourth copy of the NF-κB binding site, technically labelled as the F-κB motif, is highly conserved (Bachu, Yalla, et al., 2012). The creation of the F-κB motif consists of a faithful duplication of a 22 bp sequence (5’-GCTGACACAGAA*GGGACTTTCT*-3’) that is highly conserved and invariable among the variant viral strains, despite geographical differences and time of isolation. The seven base Ap-1 binding site (5’-CACAGAA-3’) is underlined and the 10 bp NF-κB binding site is italicized (5’-*GGGACTTTCT*-3’). Importantly, a ‘C-to-T’ substitution is invariably present at position 10 of the F-κB motif compared to the canonical H-κB motif (5’-GGGACTTTCC-3’).

In contrast, the duplication of the RBEIII motif in the modulatory region is highly variable both within and among different subtypes with respect to the length of the inserted sequence, which may span from 15 to 35 residues (Bhange et al., 2021). Although the core RBEIII site (5’-ACTGCTGA-3’) is conserved in the duplicated motif, the flanking sequences inserted are highly variable that a consensus motif may not be identified. Based on these differences, the molecular mechanisms involved in the duplication of the NF-κB and RBEIII motifs are probably different despite some possible overlap. The situation is further complicated by the sequence motif duplication of other Transcription Factor Binding Sites (TFBS) and their combinations leading to the identification of at least nine promoter-variant viral strains recently in India (Bhange et al., 2021). These sequence motif duplications consist of copy number variation of specific TFBS, genetic variation of the TFBS, and positional variation of the TFBS, rendering it difficult to explain all these duplication events using a single model. A similar situation may be identified in p6-Gag, where the duplication of the core PTAP motif is typically conserved but not the length and amino acids of the flanking sequences (Sharma et al., 2017, 2018).

Therefore, to explain the various types of sequence duplication observed in HIV-1C, a broader classification is required to depict the central and common theme of motif duplication, while

No	Region of HIV-1 Genome	Core Motif	Viral Strain	Type of Duplication	Reference
1	LTR	NF-κB	FHHC	MR-MME	Figure-23, (p. 64)
2		RBEIII	LRHR-HC	NR-MME	-
3			LRhR-HC		-
4			LRhR-HHC		Figure-25, (p.68)
5			LRXR-HHC		-
6			LRXR-HC		-
7	p6-Gag	PTAP	3 AA	MR-MME	-
8			5 AA	NR-MME	-
9			6 AA	NR-ME	Figure-24, (p. 66)
10			7 AA	MR-ME	Figure-22, (p. 63)
11			8 AA	NR-MME	-
12			9 AA	NR-MME	-
13			10 AA	NR-MME	-
14			11 AA	NR-MME	-
15			12 AA	MR-ME	-
16			13 AA	NR-ME	-
17			14 AA	MR-ME	-

Table-9: Classification of the sequence-duplication of various kinds observed in HIV-1C. The viral variants containing duplications of the NF-κB (Bachu et.al, 2012), RBEIII, (Bhange et. al 2021), and PTAP motifs (Sharma et. al 2018) are classified into four different categories as described. The terminology for categorizing the LTR variants is as follows – F, H, and C represent the genetically distinct NF-κB binding sites in the HIV-1C LTR, whereas h represents an NF-κB-like motif, R and L represent the RBEIII and the LEF-1/TCF-α motifs, respectively. MR-ME: MHD-mediated Recombination and Matched-base Extension, MR-MME: MHD-mediated Recombination and Mis-Matched base Extension, NR-ME: Non-MHD mediated Recombination and Matched-base Extension, NR-MME: Non-MHD mediated Recombination and Mis-Matched base Extension.

simultaneously permitting flexibility sufficient to accommodate crucial differences within the common theme. We found that the presence or absence of two different molecular features - the MHD triplet and a mismatch base pairing at the 3’ end of the nascent DNA, permit the organization of the various kinds of sequence-motif duplication into four categories, as presented below.

A sequence motif of at least three base pairs typically flanks the sequence motif marked for duplication. RT stalls after copying the MHD triplet located at the 3’end of the copied sequence and misaligns with the MHD triplet located at the 5’end of the already copied sequence on the recipient RNA. Following sequence motif duplication, the MHD triplet could be identified at three locations flanking both the copies of the sequence motif duplicated. Thus, the presence of MHD at either end of the sequence to be duplicated and misalignment of cDNA to the MHD at the 5’ end play a pivotal role in underlying sequence motif duplication. While the presence of the MHD triplet is evident in the case of several types of sequence motif duplication, we also identified the second kind of sequence motif duplications where a distinct MHD sequence was not present, for example – most of the double RBEIII motif duplications in the LTR (Bhange et al., 2021). Thus, based on the presence or absence of an MHD sequence, the sequence motif duplications could be classified into two categories. Of note, it is difficult to explain the mechanism(s) underlying sequence duplication in the absence of an evident MHD motif, as described below.

The second molecular feature helpful in classifying sequence motif duplications is the presence of a mismatched base at the 3' end of the cDNA. Duplication frequency appears to be significantly higher when the terminal base pair of the cDNA aligns correctly with the complementary residue on the recipient RNA template, as expected. Thus, based on these two molecular features, we classified the various sequence motif duplications of HIV-1C into four categories (Table-9), as depicted below.

(i) MR-ME: MHD-based recombination and matched-base extension
(ii) MR-MME: MHD-based recombination and mismatched-base extension
(iii) NR-ME: Non-MHD-based recombination and matched-base extension
(iv) NR-MME: Non-MHD-based recombination and mismatched-base extension

Below, we will describe the four recombination categories in detail with appropriate illustrations. However, in the absence of an MHD, the molecular mechanisms underlying the NR-ME and NR-MME categories are difficult to explain.

5.5.1 The MR-ME model – The seven amino acid PTAP motif duplication in HIV-1C.

In the example presented (Figure-22), a seven-amino acid sequence of the PTAP motif in p6-Gag is duplicated in several primary viral strains (Sharma et al., 2018). A sequence of 21 bp, (5 – GAG CCA ACA GCC CCA CCA GAA – 3') encoding a stretch of seven amino acids (EPTAPPA, the core PTAP motif, and the corresponding codons are underlined) is duplicated (Figure-22A). Examination of the sequence identifies the triplet 'GAG' to be present at both the ends of the sequence duplicated. It may be inferred that the RT stalls after copying the G-residue of the GAG and the nascent DNA switches strands. The 3'– CTC– 5' at the terminal end of the cDNA (complementary to 5'-GAG-3') misaligns with the MHD 5'-GAG-3' located on the acceptor RNA template immediately downstream of the 21 bases already copied (Figure-22B). Of note, there is no base-pair mismatch at the growing end of the cDNA in this case. As the RT resumes polymerization, the seven amino acid stretch is copied once again, leading to the creation of an additional copy of the PTAP motif of seven amino acids. The same model applies faithfully to the12 amino acid PTAP motif duplication in HIV-1C, which can be mediated by the presence of the 5'-AGG-3' triplet that codes for Arginine (not presented).

5.5.2 The MR-MME model – the generation of the F-κB motif in HIV-1C enhancer.

The F-κB motif is genetically distinct from the canonical H-κB motif, and its generation requires the creation of two different molecular features (Bachu et. al, 2012). The F-κB motif creation consists of the faithful duplication of 22 bp, invariable among various variant viral strains. Further, the duplication contains a 'C-to-T' variation at position 10 of the κB-motif. The 22 bp of this sequence (5'-GCTGACACAGAA*GGGACTTTCT*-3') comprises a co-duplication of binding sites for the members of two different TF families – AP-1 (seven bases underlined) and NF-κB (ten bases italicized), further; five additional bases of partial RBEIII binding core motif are located at the 5' end of the motif duplicated. Thus, although, for simplicity, the sequence duplications are labelled using the identity of the major TF family, the duplicated sequences typically represent conglomerations of multiple transcription factor families, suggesting combined and synergistic

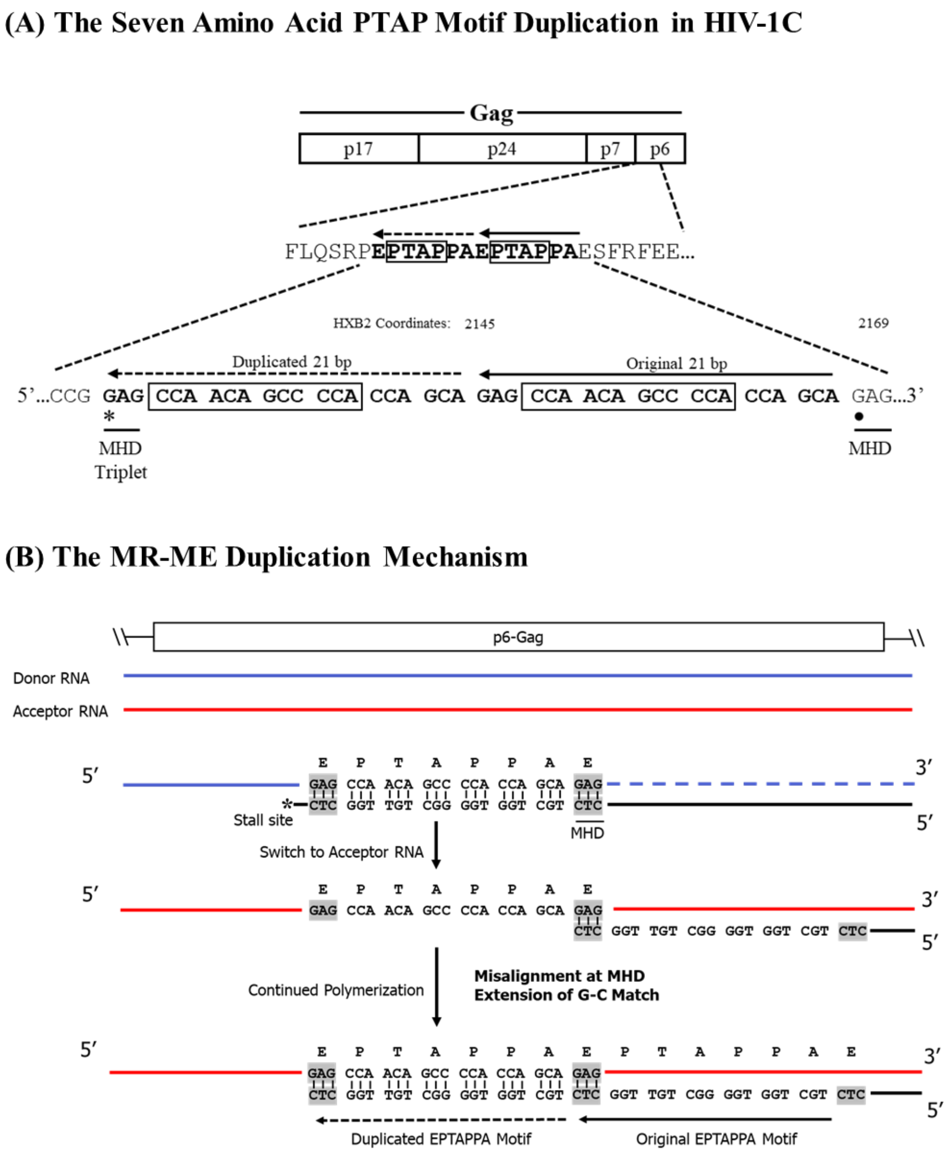

(A) The Seven Amino Acid PTAP Motif Duplication in HIV-1C
Gag
p17
p24
p7
p6
FLQSRPEPTAPPAEPTAPPAESFRFEE...
HXB2 Coordinates: 2145
2169
Duplicated 21 bp
Original 21 bp
5'...CCG GAG CCA ACA GCC CCA CCA GCA GAG CCA ACA GCC CCA CCA GCA GAG...3'
*
MHD
Triplet
MHD
(B) The MR-ME Duplication Mechanism
p6-Gag
Donor RNA
Acceptor RNA
5'
3'
E P T A P P A E
GAG CCA ACA GCC CCA CCA GCA GAG
CTC GGT TGT CGG GGT GGT CGT CTC
Stall site
MHD
Switch to Acceptor RNA
CTC GGT TGT CGG GGT GGT CGT CTC
Continued Polymerization
Misalignment at MHD
Extension of G-C Match
E P T A P P A E P T A P P A E
Duplicated EPTAPPA Motif
Original EPTAPPA Motif

Figure-22: The MR-ME model of sequence duplication. (A) The seven amino-acid duplication at the PTAP motif observed in HIV-1C p6-Gag. The seven amino acid residues of the PTAP motif, original and duplicated, are indicated using solid and broken arrows, respectively (Sharma et.al, 2018). The corresponding nucleotide sequence is presented where the core PTAP motif of four amino acids is highlighted using an open square box. The MHD and the GAG triplets, are underlined, and the RT stall and re-annealing sites are represented with an asterisk and a dot, respectively. (**B) A schematic to explain the duplication of the seven amino acids of the PTAP motif.** The donor and acceptor RNAs are shown using blue and red lines, respectively, and the cDNA with a black line. The MHD triplets and the complementary bases are shaded grey. The RT stall site is represented with an asterisk. Note that the growing end of the nascent DNA matches perfectly with the MHD of the acceptor RNA.

biological functions among the transcription factor families. Thus, sequence motif duplication in at least HIV-1C represents 'TFBS cluster duplication'.

The creation of the F-κB cluster duplication in C-LTR (Figure-23A) requires three processes to happen successively - RT stalling at a specific location after reverse transcribing the H-κB motif, the AP-1 motif, and a partial RBEIII core motif in that order; strand-switching to a specific MHD on the donor RNA template; and mismatch extension, if necessary, as depicted schematically (Figure-23). Depending on the RT stall site, *UGCU or *GCU, two mutually exclusive models of the five possible pathways can create the F-κB cluster duplication (Figure-23B).

The *UGCU-stall model: The model predicts that RT stalls at the 5' 'U' residue of the UGCU motif after reverse transcribing the H-κB motif, the AP-1 motif, and a partial RBEIII core motif (Figure-

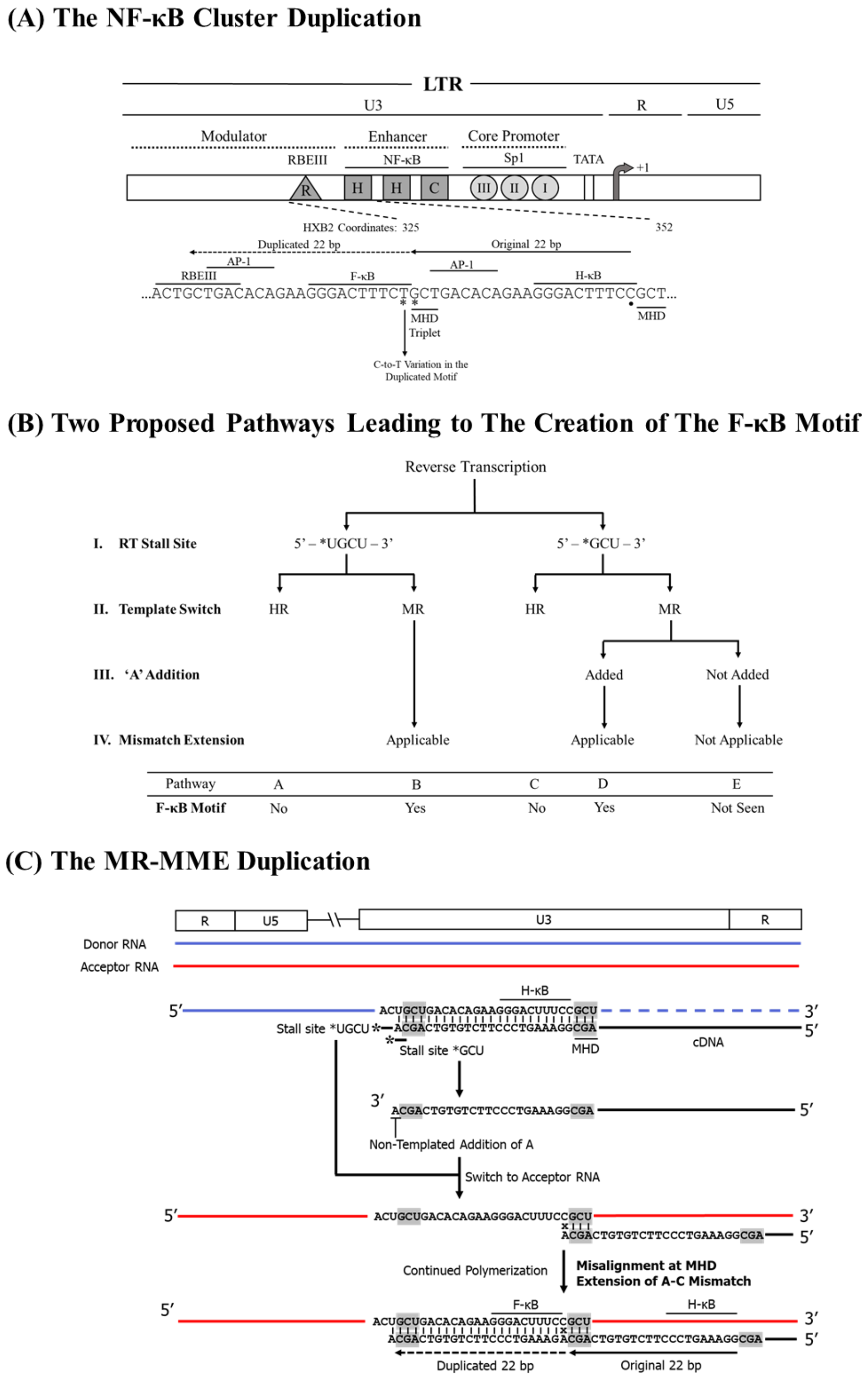


Figure-23: A schematic model explaining the NF-κB cluster duplication using the MR-MME model. (A) The NF-κB cluster duplication. Schematic representation of the HIV-1C LTR and the F-κB cluster duplication of 22 residues. The major transcription factor binding sites of AP-1, RBEIII, and NF-κB are labelled. The original and duplicated motifs are highlighted using solid and dashed arrows, respectively. The MHD and MHD triplet motifs flanking the duplicated motif are labelled. The two RT stall sites TGCT and GCT are depicted using asterisks and the re-annealing site using a dot. Note the C-to-T variation at position 10 of the F-κB motif**.** **(B) The two proposed pathways leading to the creation of the F-κB motif.** Five different pathways may be identified based on the four parameters illustrated, of which only two pathways can lead to the creation of the F-κB motif. Although the fifty pathway is theoretically feasible, no such LTR variant forms have been identified in the sequence databases. HR: homologous recombination, MR: MHD-mediated recombination. **(C) The MR-MME duplication.** Depending on the precise location of the RT stall site, *UGCA or *GCA, and the non-templated addition of an 'A' residue in the latter case, two different models can explain the creation of the F-κB motif. In the first model, when the RT stalls at *UGCA and strand-switches to the MHD on the acceptor RNA template, a base pair mis-match between the terminal 'A' at the 3' end of the cDNA and a 'C' located on the template is evident. Successful polymerization requires a base-pair mis-match extension. Alternatively, the RT may stall at *GCA and an 'A' residue is added to the 3'end of the cDNA by the non-templated polymerization activity of the RT. When such cDNA aligns with the recipient RNA template via the MHD, a base pair mismatch is identified again. Thus, either of these two models can cause the 22 base F-κB cluster duplication along with the C-to-T variation of the NF-κB motif.

23C). The 'nascent DNA-donor RNA' hybrid is released from the RT complex. Subsequently, the nascent DNA separates from the donor RNA molecule and hybridizes with the acceptor RNA template. If the strand-switch involves the alignment of the nascent DNA to homologous sequence on the acceptor RNA molecule, only homologous recombination ensues, without sequence duplication (Figure-23B, Path-A). A new RT molecule recognizes the 'nascent DNA-acceptor RNA' complex and resumes polymerization. Alternatively, the free nascent DNA molecule hybridizes with the acceptor RNA template using the MHD 'GCU' (Figure-23B, Path-B). The intervening sequence of the nascent DNA lying between the hybridized regions to the MHD triplet on the growing end and the sequences of acceptor RNA upstream will form a loop of unhybridized sequence. This unique structure of the 'nascent DNA-acceptor RNA' hybrid is recognized by a new RT molecule to form a 'nascent DNA-acceptor RNA-RT' complex. The hybridization of the nascent DNA to the acceptor RNA template through the MHD triplet of 'GCU' will create a mismatch between the 'A' residue on the DNA at the growing end and the 'C' residue present at this location on the acceptor RNA (Figure-23C). The superior mismatch-extension quality of HIV-1C RT ignores the mismatch and resumes polymerization efficiently, thus, causing the 'C-to-T' transition of the F-κB motif. In summary, the *UGCU-model proposed here not only explains the creation of the genetically distinct F-κB motif in C-LTR, but also how the exceptional molecular characteristics of HIV-1C RT cause the sequence motif duplication unique to HIV-1C. An alternative path is possible for the creation of the F-κB motif if the RT stalls one base short of the one proposed above, as depicted below.

The *GCU-stall model: The model predicts that RT stalls at the 5' 'G' residue of the GCU motif after reverse transcribing the H-κB motif, the AP-1 motif, and a partial RBEIII core motif. The paused RT tends to add a nucleotide at the growing end of the DNA in the RNA:DNA hybrid in a template-independent fashion, with the highest preference for an 'A' residue (W.Wu, B. M. Blumberg, P. J. Fay, 1995). The 'nascent DNA-donor RNA' hybrid is released from the 'RT-nascent DNA-donor RNA' complex. Nascent DNA dissociates from the donor RNA molecule and hybridizes with the acceptor RNA template. If the nascent DNA hybridizes to the homologous sequence on the acceptor RNA molecule, only homologous recombination ensues, without sequence duplication (Figure-23B, Path-C). A new RT molecule recognizes the 'nascent DNA-acceptor RNA' complex and resumes polymerization. Resumed polymerization should not generate a variation as the non-specifically added 'A' can pair with the natural 'U' present on the acceptor template RNA at this position. Alternatively, the free nascent DNA molecule may hybridize with the acceptor RNA template using the MHD 'GCU' (Figure-23B, Path-D). The intervening sequence of the nascent DNA lying between the hybridized regions to the MHD triplet on the growing end and the sequences of acceptor RNA upstream will form a loop of unhybridized sequence. This unique structure of the 'nascent DNA-acceptor RNA' complex is expected to be highly unstable. A new RT molecule identifies the highly unstable 'nascent DNA-acceptor RNA' complex. Importantly, the additional hydrogen bond formed between the nascent DNA and the RT identified through this work stabilizes the highly unstable structure. The subsequent steps of resumed RT polymerization are similar to the *UGCU model described above.

The creation of the F-κB motif in HIV-1C, thus, not only depends on where the RT stalls in the modulator region but also on the ability of the RT to add an 'A' residue to the growing end of the nascent DNA molecule non-specifically using the extendase activity, further reinforced by the stabilization the highly unstable DNA-RNA complex by the new-found hydrogen bond and the ability of the C-RT to extend mismatches efficiently. In summary, although the precise site where the RT stalls is not known, both the *UGCU- and *GCU- models can explain the creation of the genetically distinct F-κB motif in C-LTR using the MHD-mediated Recombination followed by the Mis-Match Extension mechanism.

5.5.3 The NR-ME model - the six amino acid 'EPTAPP' duplication in p6-Gag of HIV-1C.

As mentioned above, the events of sequence duplication occurring in the absence of an apparent micro-homology domain, flanking the sequences being duplicated, is difficult to explain. One prominent example is the six amino acid duplication of the PTAP motif of p6-Gag in HIV-1C. Importantly, the 'EPTAPP' motif duplication in HIV-1C is the second most common duplication seen in this region. Nearly 28% of HIV-1C Gag sequences comprising a duplication within the PTAP domain contain the six amino acid motif duplication. In contrast, more than 32% of such HIV-1C p6-Gag sequences contain a seven amino acid duplication (EPTAPPA) described above under the MR-ME model, which depends on the presence of an MHD motif (Figure-22). The six amino acid PTAP motif duplication is mediated by the RT stalling on the donor RNA at the 5' G of the 'GAG' codon, which codes for Aspartic Acid in the duplicated 'EPTAPP' motif. The RT then switches to the acceptor RNA and misaligns with the 'GCA' codon, adjacent to the 'CCA' codon of

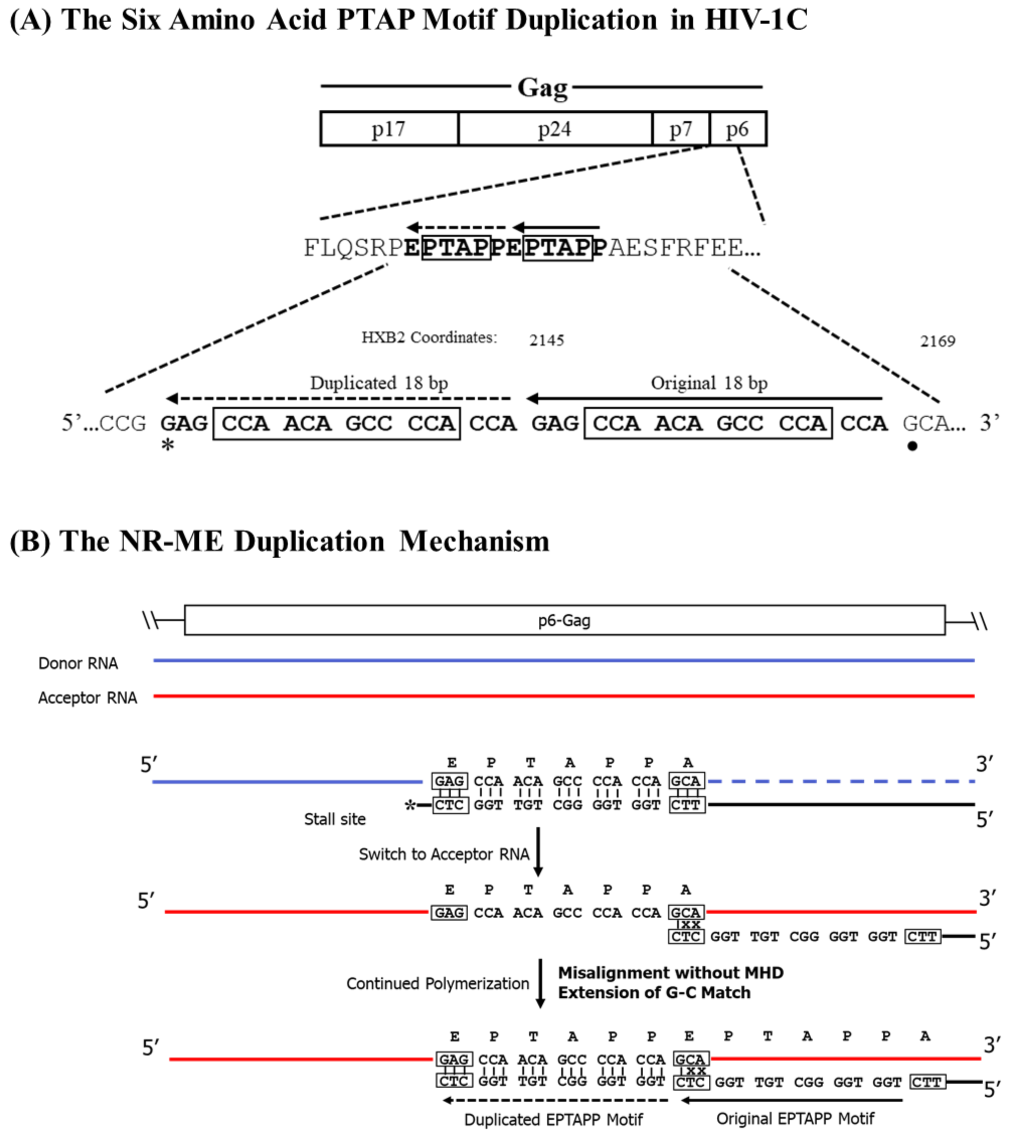

(A) The Six Amino Acid PTAP Motif Duplication in HIV-1C
Gag
p17
p24
p7
p6
FLQSRPEPTAPPEPTAPPAESFRFEE...
HXB2 Coordinates:
2145
2169
Duplicated 18 bp
Original 18 bp
5'...CCG GAG CCA ACA GCC CCA CCA GAG CCA ACA GCC CCA CCA GCA... 3'
*
●
(B) The NR-ME Duplication Mechanism
p6-Gag
Donor RNA
Acceptor RNA
5'
3'
E P T A P P A
GAG CCA ACA GCC CCA CCA GCA
CTC GGT TGT CGG GGT GGT CTT
Stall site
Switch to Acceptor RNA
Continued Polymerization
Misalignment without MHD
Extension of G-C Match
E P T A P P E P T A P P A
Duplicated EPTAPP Motif
Original EPTAPP Motif

Figure-24: The NR-ME model of sequence duplication. (A) The six amino acid duplication at the PTAP motif observed in HIV-1C p6-Gag. The six amino acid residues of the PTAP motif, original and duplicated, are indicated using solid and broken arrows, respectively. The corresponding nucleotide sequence is presented where the core PTAP motif of four amino acids is highlighted using an open square box. The RT stall location and re-annealing sites are represented with an asterisk and a dot, respectively**. (B) A schematic to explain the duplication of the six amino acids of the PTAP motif.** The donor and acceptor RNAs are shown using blue and red lines, respectively, and the cDNA with a black line. The absence of an MHD is shown by highlighting the corresponding bases using boxes. The RT stall site is represented with an asterisk. Note the extension of a single matched base at the growing end of the nascent DNA.

the terminal Proline in the duplicated motif. The misaligned sequence lacks an MHD but the terminal ‘C’ of the cDNA base pairs with the ‘G’ of the ‘GCA’ codon, leading to a Non-MHD-based Recombination Matched-base Extension duplication (Figure-24).

A second example of the NR-ME model is the partial duplication of the PTAP motif in HIV-1B. The biological significance of creating only a partial copy of the core PTAP motif is not understood. However, this event is the most predominant one seen in HIV-1B (Sharma et al., 2018). More than half of HIV-1B Gag sequences containing a duplication within the PTAP domain consist of the duplication of only three amino acid residues of ‘APP’ encoded by 5’- GCC CCA CCA -3’. Based on the nucleotide sequence, it may be possible to discern the RT stall site and the location of the strand-switch on the receptor RNA. RT is expected to stall at the 5’ G of the ‘GCC’ codon in the above sequence of the donor RNA and strand switch to the receptor RNA to the codon ‘GAA’ immediately downstream of the copied sequence, and resume polymerization from the ‘G’ residue of the ‘GAA’ codon, thus, creating an additional ‘APP’ amino acid triplet in the p6-Gag protein.

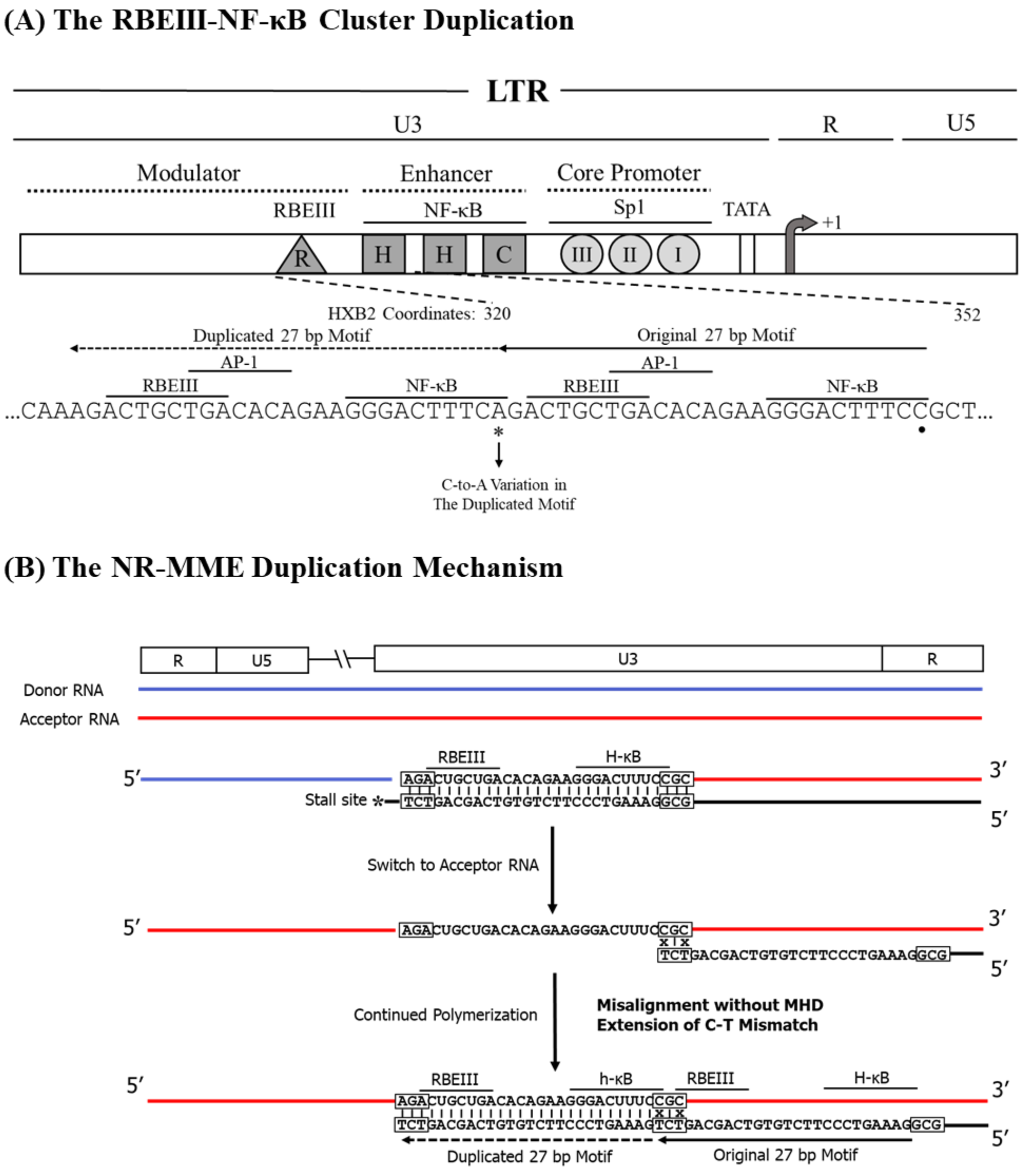
5.5.3 The NR-MME model - the RBEIII motif duplication in HIV-1C LTR.

As in the case of NF-κB motif duplication, the duplication of the RBEIII site in HIV-1C also represents a cluster duplication by creating binding sites for more than one TF family. Although there are several variant forms of the RBEIII cluster duplication, comprising copying sequences ranging from 15-35 bp, the creation of LRhR-HHC variant LTR is presented here as the prototype to represent this category of sequence motif duplications (Bhange et al., 2021). The LRhR-HHC variant consists of a duplication of a 27 bp sequence (5’- AG**ACTGCTGA**CACAGAA*GGGACTTTCA* – 3’) encompassing three TF families – RBEIII (bolded) AP-1 (underlined), and NF-κB (italicized). Importantly, there is a ‘C to A’ variation at position 10 of the NF-κB binding site (Figure-25A).

Inspection of the variant LTR sequence confirms the absence of an MHD proximal to the sequence duplicated. The RT is expected to stall at the ‘A’ residue (indicated by an asterisk, Figure-25A) after copying the RBEIII core motif on the donor RNA template. The RT appears to resume reverse transcription after switching to the acceptor RNA from the ‘C’ residue of the ‘CGC’ triplet located in the NF-κB binding site downstream by ignoring the ‘C-to-T’ mismatch at the growing end of the cDNA (Figure-25B). Continued reverse transcription leads to the copying of the 27 bp sequence once again and fixing the ‘C-to-A’ variation at position 10 of the newly created NF-κB motif.

5.6 Sequence Duplications: Global implications for disease management.

Our findings have significant implications for global HIV-1 disease management. It is well understood that different HIV-1 subtypes acquire resistance to specific antiretroviral drugs at different rates (Garforth et al., 2010; Singh et al., 2014). The initiation of ART, immediately following diagnosis, as is mandated by the ‘Test and Treat’ policy of the World Health Organization may lead to the emergence of variant viral strains. The generation of the PTAP motif duplication in HIV-1 subtypes, albeit at a faster rate in HIV-1C, could be a form of manifestation of drug resistance or its compensation. Several publications reported a significant association between ART initiation and PTAP duplication (Martins et al., 2011, 2015; Peters et al., 2001). While the molecular mechanisms driving the selection of the PTAP motif duplication following ART initiation remain enigmatic, preliminary data from our laboratory suggest the possibility of the motif duplication playing a compensatory role in restoring replication fitness after a drug-resistance mutation has been selected. Further, we previously demonstrated the dominance of double-PTAP motif variant strains over the wild-type single-PTAP motif variant strains in a mixed infection (Sharma et al., 2018). Given that ‘transmission-bottleneck’ imposes that only a small number of viral strains initiate a new infection, the probability of transmission of the double-PTAP variant is expected to be higher than that of the wild-type strain in proportion to their presence in a mixed infection. Bioinformatic



(A) The RBEIII-NF-κB Cluster Duplication

(B) The NR-MME Duplication Mechanism

R
U5
U3
R
Donor RNA
Acceptor RNA
5′
3′
5′
RBEIII
H-κB
AGACUGCUGACACAGAAGGGACUUUCCGC
Stall site *
TCTGACGACTGTGTCTTCCCTGAAAGGCG
Switch to Acceptor RNA
5′
3′
5′
AGACUGCUGACACAGAAGGGACUUUCCGC
TCTGACGACTGTGTCTTCCCTGAAAGGCG
Continued Polymerization
Misalignment without MHD
Extension of C-T Mismatch
5′
3′
5′
RBEIII
h-κB
RBEIII
H-κB
AGACUGCUGACACAGAAGGGACUUUCCGC
TCTGACGACTGTGTCTTCCCTGAAAGTCTGACGACTGTGTCTTCCCTGAAAGGCG
Duplicated 27 bp Motif
Original 27 bp Motif

Figure-25: The NR-MME model of sequence duplication. (A) The RBEII-NFκB Cluster duplication. Schematic representation of the HIV-1C LTR and the RBEIII-NF-κB cluster duplication of 27 residues. The major transcription factor binding sites of AP-1, RBEIII, and NF-κB are labelled. The original and duplicated motifs are highlighted using solid and dashed arrows, respectively. The proposed RT stall location and re-annealing sites are shown using an asterisk and a dot, respectively. Note the C-to-A variation at position 10 of the duplicated NF-κB motif. **(B) A schematic to explain the duplication of 27 bp using the NR-MME mechanism.** The donor and acceptor RNAs are shown using blue and red lines, respectively, and the cDNA with a black line. The absence of an MHD is shown by highlighting the corresponding bases using boxes. The RT stall site is represented with an asterisk. Note the extension of a mis-matched base at the growing end of the nascent DNA.

analysis of the HIV-1C sequences available in the extant databases showing that the prevalence of the double-PTAP variants has increased from 17.6% in 1996 to 31% 2015 in 2015 is consistent with the premise (Sharma et al., 2017).

The generation of diverse LTR variant viral strains could be of greater clinical significance since a single promoter regulates the expression of all the viral proteins and governs latency. The addition of an extra copy of the NF-κB site leads to a significantly higher magnitude of transcription and, consequently, to a higher plasma viral load in HIV-1C infection (Bachu, Yalla, et al., 2012). The duplication of the RBE-III motif, on the other hand, appears to resist latency reactivation of the variant viral strains. Unpublished work from our laboratory demonstrates that the known latency-reversing agents fail to activate the dual RBE-III, but not canonical, viral strains (Bhange D et al., manuscript in preparation). Whether the dual

RBE-III variant strains are likely to establish more stable viral reservoirs warrants urgent clinical evaluation.

Taking all these observations collectively, the ability of HIV-1C to duplicate sequences at a higher frequency is of significant concern to the global HIV-1 scenario considering that HIV-1C is responsible for nearly half of the infections in the world.

5.7 References

Bachu, M., Mukthey, A. B., Murali, R. V., Cheedarla, N., Mahadevan, A., Shankar, S. K., Satish, K. S., Kundu, T. K., & Ranga, U. (2012). Sequence Insertions in the HIV Type 1 Subtype C Viral Promoter Predominantly Generate an Additional NF-κB Binding Site. AIDS Research and Human Retroviruses, 28(10), 1362–1368. https://doi.org/10.1089/aid.2011.0388

Bachu, M., Yalla, S., Asokan, M., Verma, A., Neogi, U., Sharma, S., Murali, R. V., Mukthey, A. B., Bhatt, R., Chatterjee, S., Rajan, R. E., Cheedarla, N., Yadavalli, V. S., Mahadevan, A., Shankar, S. K., Rajagopalan, N., Shet, A., Saravanan, S., Balakrishnan, P., … Ranga, U. (2012). Multiple NF-κB Sites in HIV-1 Subtype C Long Terminal Repeat Confer Superior Magnitude of Transcription and Thereby the Enhanced Viral Predominance. Journal of Biological Chemistry, 287(53), 44714–44735. https://doi.org/10.1074/jbc.M112.397158

Balakrishnan, M., Roques, B. P., Fay, P. J., & Bambara, R. A. (2003). Template dimerization promotes an acceptor invasion-induced transfer mechanism during human immunodeficiency virus type 1 minus-strand synthesis. J Virol, 77(8), 4710–4721. https://doi.org/10.1128/JVI.77.8.4710-4721.2003

Bhange, D., Prasad, N., Singh, S., Prajapati, H. K., Maurya, S. P., Gopalan, B. P., Nadig, S., Chaturbhuj, D., Jayaseelan, B., Dinesha, T. R., Ahamed, S. F., Singh, N., Brahmaiah, A., Mehta, K., Gohil, Y., Balakrishnan, P., Das, B. K., Dias, M., Gangakhedkar, R., … Ranga, U. (2021). The Evolution of Regulatory Elements in the Emerging Promoter-Variant Strains of HIV-1 Subtype C. Frontiers in Microbiology, 12(November). https://doi.org/10.3389/fmicb.2021.779472

Carl, S., Daniels, R., Iafrate, A. J., Easterbrook, P., Greenough, T. C., Skowronski, J., & Kirchhoff, F. (2000). Partial “repair” of defective NEF genes in a long-term nonprogressor with human immunodeficiency virus type 1 infection. Journal of Infectious Diseases, 181(1), 132–140. https://doi.org/10.1086/315187

Chin, M. P. S., Rhodes, T., Chen, J., Fu, W., & Hu, W.-S. (2005). Identification of a Major Restriction in HIV-1 Inter-subtype recombination. Proc Natl Acad Sci U S A, 102(25), 9002–9007. https://doi.org/10.1073/pnas.0502522102

Dang, Q., & Hu, W. (2001). Effects of homology length in the repeat region on minus-strand DNA transfer and retroviral replication. Journal of Virology, 75(2), 809–820. https://doi.org/10.1128/JVI.75.2.809-820.2001

Delviks-Frankenberry, K., Galli, A., Nikolaitchik, O., Mens, H., Pathak, V. K., & Hu, W. S. (2011). Mechanisms and factors that influence high frequency retroviral recombination. Viruses, 3(9), 1650–1680. https://doi.org/10.3390/v3091650

Galli, A., Kearney, M., Nikolaitchik, O. A., Yu, S., Chin, M. P. S., Maldarelli, F., Coffin, J. M., Pathak, V. K., & Hu, W.-S. (2010). Patterns of Human Immunodeficiency Virus Type 1 Recombination Ex Vivo Provide Evidence for Coadaptation of Distant Sites, Resulting in Purifying Selection for Intersubtype Recombinants during Replication. J. Virol., 84(15), 7651–7661. https://doi.org/10.1128/JVI.00276-10

Garforth, S. J., Domaoal, R. A., Lwatula, C., Landau, M. J., Meyer, A. J., Anderson, K. S., & Prasad, V. R. (2010). K65R and K65A substitutions in HIV-1 reverse transcriptase enhance polymerase fidelity by decreasing both dNTP misinsertion and mispaired primer extension efficiencies. Journal of Molecular Biology, 401(1), 33–44. https://doi.org/10.1016/j.jmb.2010.06.001

Hwang, C. K., Svarovskaia, E. S., & Pathak, V. K. (2001). Dynamic copy choice: Steady state between murine leukemia virus polymerase and polymerase-dependent RNase H activity determines frequency of in vivo template switching. Proceedings of the National Academy of Sciences of the United States of America, 98(21), 12209–12214. https://doi.org/10.1073/pnas.221289898

Ji, Y., Han, X., Tian, W., Gao, Y., Jin, S., Zhang, L., & Shang, H. (2018). V4 region of the HIV-1 envelope gene mediates immune escape and may not promote the development of broadly neutralizing antibodies. Vaccine, 36(50), 7700–7707. https://doi.org/10.1016/j.vaccine.2018.10.084

Martins, Angélica N., Arruda, M. B., Pires, A. F., Tanuri, A., & Brindeiro, R. M. (2011). Accumulation of P(T/S)AP Late Domain Duplications in HIV Type 1 Subtypes B, C, and F Derived from Individuals Failing ARV Therapy and ARV Drug-Naive Patients. AIDS Research and Human Retroviruses, 27(6), 687–692. https://doi.org/10.1089/aid.2010.0282

Martins, Angelica N., Waheed, A. A., Ablan, S. D., Huang, W., Newton, A., Petropoulos, C. J., Brindeiro, R. de M., & Freed, E. O. (2015). Elucidation of the Molecular Mechanism Driving Duplication of the HIV-1 PTAP Late Domain. Journal of Virology, 90(October), JVI.01640-15. https://doi.org/10.1128/JVI.01640-15

Negroni, M, & Buc, H. (2001). Retroviral recombination: what drives the switch? Nature Reviews. Molecular Cell Biology, 2(2), 151–155. https://doi.org/10.1038/35052098

Negroni, Matteo, & Buc, H. (2000). Copy-choice recombination by reverse transcriptases: Reshuffling of genetic markers mediated by RNA chaperones. Proceedings of the National Academy of Sciences of the United States of America, 97(12), 6385–6390. https://doi.org/10.1073/pnas.120520497

Onafuwa-Nuga, A., & Telesnitsky, A. (2009). The Remarkable Frequency of Human Immunodeficiency Virus Type 1 Genetic Recombination. Microbiology and Molecular Biology Reviews, 73(3), 451–480. https://doi.org/10.1128/MMBR.00012-09

Onafuwa, A., An, W., Robson, N. D., & Telesnitsky, A. (2003). Human Immunodeficiency Virus Type 1 Genetic Recombination Is More Frequent Than That of Moloney Murine Leukemia Virus despite Similar Template Switching Rates. Journal of Virology, 77(8), 4577–4587. https://doi.org/10.1128/JVI.77.8.4577

Patel, P. H., & Loeb, L. (2001). Getting a grip on how DNA polymerases function. Nature Structural Biology, 8(8), 656–659. https://doi.org/10.1038/90344

Peters, S., Muñoz, M., Yerly, S., Lopez-galindez, C., Perrin, L., Larder, B., Cmarko, D., Fakan, S., Noz, M. M. U., & Perrin, L. U. C. (2001). Resistance to Nucleoside Analog Reverse Transcriptase Inhibitors Mediated by Human Immunodeficiency Virus Type 1 p6 Protein Resistance to Nucleoside Analog Reverse Transcriptase Inhibitors Mediated by Human Immunodeficiency Virus Type 1 p6 Protein. Journal of Virology, 75(20), 9644–9653. https://doi.org/10.1128/JVI.75.20.9644

Rawson, J. M. O., Nikolaitchik, O. A., Keele, B. F., Pathak, V. K., & Hu, W. S. (2018). Recombination is required for efficient HIV-1 replication and the maintenance of viral genome integrity. Nucleic Acids Research, 46(20), 10535–10545. https://doi.org/10.1093/nar/gky910

Roda, R. H., Balakrishnan, M., Hanson, M. N., Wöhrl, B. M., Le Gricell, S. F. J., Roques, B. P., Gorelick, R. J., & Bambara, R. A. (2003). Role of the reverse transcriptase, nucleocapsid protein, and template structure in the two-step transfer mechanism in retroviral recombination. Journal of Biological Chemistry, 278(34), 31536–31546. https://doi.org/10.1074/jbc.M304608200

Roda, R. H., Balakrishnan, M., Kim, J. K., Roques, B. P., Fay, P. J., & Bambara, R. A. (2002). Strand transfer occurs in retroviruses by a pause-initiated two-step mechanism. Journal of Biological Chemistry, 277(49), 46900–46911. https://doi.org/10.1074/jbc.M208638200

Sharma, S., Aralaguppe, S. G., Abrahams, M.-R., Williamson, C., Gray, C., Balakrishnan, P., Saravanan, S., Murugavel, K. G., Solomon, S., & Ranga, U. (2017). The PTAP sequence duplication in HIV-1 subtype C Gag p6 in drug-naive subjects of India and South Africa. BMC Infectious Diseases, 17(1), 95. https://doi.org/10.1186/s12879-017-2184-4

Sharma, S., Arunachalam, P. S., Menon, M., Ragupathy, V., Satya, R. V., Jebaraj, J., Aralaguppe, S. G., Rao, C., Pal, S., Saravanan, S., Murugavel, K. G., Balakrishnan, P., Solomon, S., Hewlett, I., & Ranga, U. (2018). PTAP motif duplication in the p6 Gag protein confers a replication advantage on HIV-1 subtype C. Journal of Biological Chemistry, 293(30), 11687–11708. https://doi.org/10.1074/jbc.M117.815829

Singh, K., Flores, J. A., Kirby, K. A., Neogi, U., Sonnerborg, A., Hachiya, A., Das, K., Arnold, E., McArthur, C., Parniak, M., & Sarafianos, S. G. (2014). Drug resistance in non-B subtype HIV-1: Impact of HIV-1 reverse transcriptase inhibitors. Viruses, 6(9), 3535–3562. https://doi.org/10.3390/v6093535

Taylor, B. S., Sobieszczyk, M. E., McCutchan, F. E., & Hammer, S. M. (2008). The Challenge of HIV-1 Subtype Diversity. New England Journal of Medicine, 358(15), 1590–1602. https://doi.org/10.1056/nejmra0706737

W.Wu, B. M. Blumberg, P. J. Fay, R. A. B. (1995). Strand Transfer Mediated by Human Immunodeficiency Virus Reverse Transcriptase in Vitro is Promoted by Pausing and Results in Misincorporation. The Journal of Biological Chemistry, 270(1), 325–332. http://www.jbc.org/content/270/1/325.short

Warrilow, D., Tachedjian, G., & Harrich, D. (2009). Maturation of the HIV reverse transcription complex: putting the jigsaw together. Reviews in Medical Virology, 19(6), 324–337. https://doi.org/10.1002/rmv

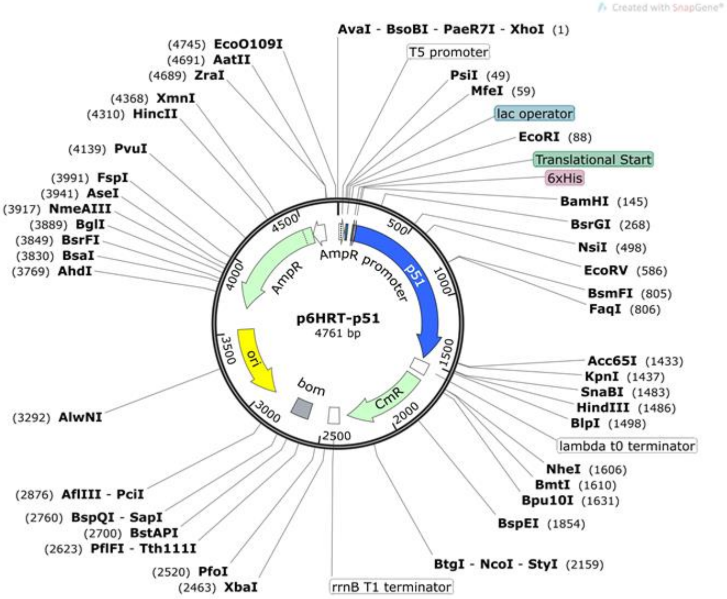
Zhang, J., & Temin, H. (1993). Rate and mechanism of nonhomologous recombination during a single cycle ofretroviral replication. 727(1984).

Appendix I: Plasmid Maps

1. RT p51 Expression Vectors

Plasmid Numbers: p945.1 – p945.8
Plasmid Name: p6HRT51 Variants
Features: RT Expression Vector, T5 Promoter, IPTG Inducible, N-terminal 6X. His. Tag

Keywords: NL4-3 and Indie RT p51 Expression Vectors, pos. 359 variants, Ampicillin resistance, T5 Promoter



Plasmid Number	Plasmid Name (p6HRT51-)	Unique RE
p945.1	N	-
p945.2	NT	SpeI
p945.3	NS	EagI
p945.4	NA	BsiWI
p945.6	I	SacI
p945.7	IG	MluI
p945.8	IS	EagI
p945.9	IA	BsiWI

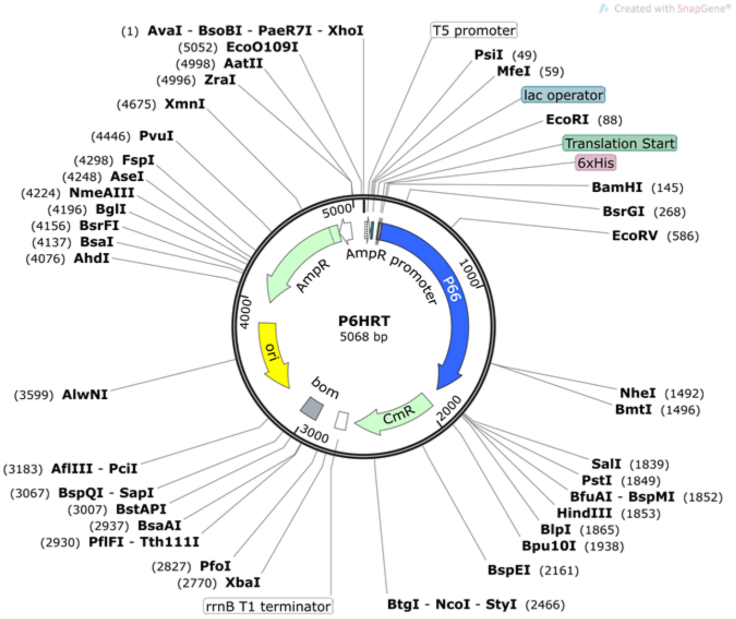
Key Points:

- This series of plasmids express the NL4-3 p51 RT subunit pos. 359 variants under IPTG inducible control of the T5 promoter.
- The p51 variants were generated by overlap PCR and cloned between the BamHI and HindIII sites on the plasmid backbone and confirmed by sequencing.
- Each variant has a unique restriction enzyme site engineered before the HindIII site to enable identification
- The expression vectors contain an N-terminal 6X Histidine Tag.
- A SacI site is engineered in all Indie RT expression vectors to differentiate them from NL4-3
- Due to the nature of the expression vectors, the amino acids MRGSHHHHHHGSQL are added before the first Amino Acid of RT (Proline).

2. RT p66 Expression Vectors

Plasmid Numbers: p946.1 – p946.4
Plasmid Name: p6HRT Variants
Features: RT Expression Vector, T5 Promoter, IPTG Inducible, N-terminal 6X. His. Tag

Keywords: NL4-3 and Indie RT p66 Expression Vectors, pos. 359 variants, Ampicillin resistance, T5 Promoter



Plasmid Number	Plasmid Name (p6HRT-)	Unique RE
p946.1	N	-
p946.2	NT	SpeI
p946.3	NS	EagI
p946.4	NA	BsiWI
p946.6	I	SacI
p946.7	IG	MluI
p946.8	IS	EagI
p946.9	IA	BsiWI

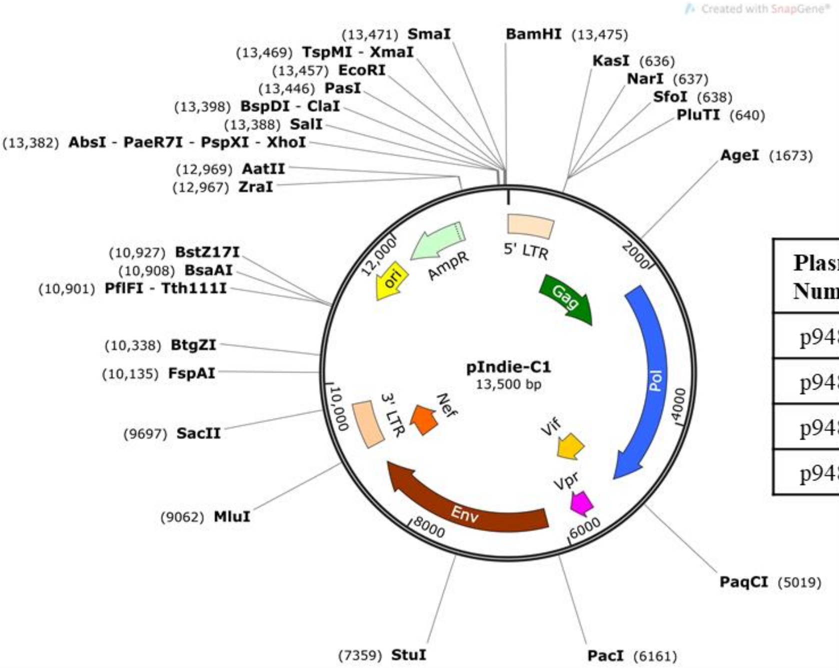
Key Points:

- This series of plasmids express the Indie p66 RT pos. 359 variants under IPTG inducible control of the T5 promoter.
- The p66 variants were generated by overlap PCR and cloned between the BamHI and SalI sites on the plasmid backbone and confirmed by sequencing.
- Each variant has a unique restriction enzyme site engineered before the SalI site to enable identification.
- A SacI site is engineered in all Indie RT expression vectors to differentiate them from NL4-3
- The expression vectors contain an N-terminal 6X Histidine Tag.
- Due to the nature of the expression vectors, the amino acids MRGSHHHHHHGSQL are added before the first Amino Acid of RT (Proline).

3. pIndie-C1 and Variants

Plasmid Numbers: p948.1 – p948.4
Plasmid Name: pIndie RT Variants
Features: Full length infectious viral variants, Mutations at Pos. 359 of RT.

Keywords: pIndie, subtype-C RT variants, full length infectious viruses, Ampicillin resistance.



Plasmid Number	Plasmid Name	Amino Acid at Pos. 359
p948.1	pIndie	Threonine
p948.2	pIndie G359	Glycine
p948.3	pIndie S359	Serine
p948.4	pIndie A359	Alanine

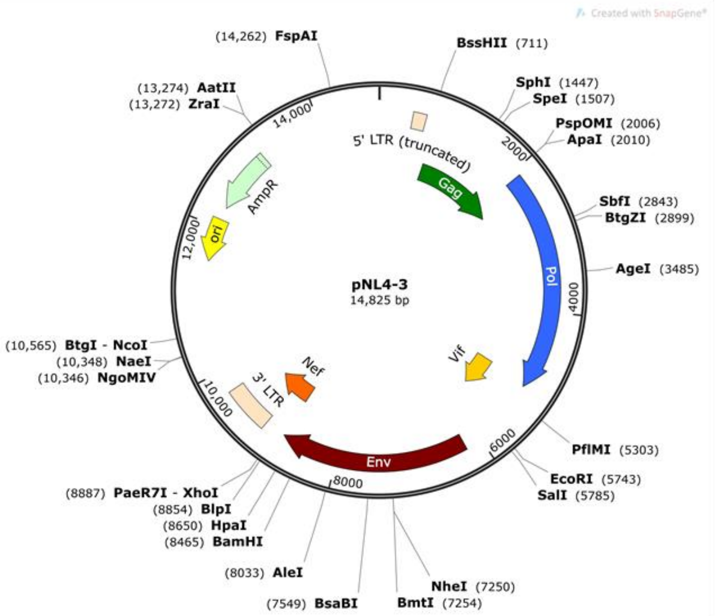
Key Points:

- This series of plasmids are derived from the pIndie full length molecular clone. The plasmids contain mutations at pos. 359 of RT.
- The mutations were introduced by using overlap PCR and cloned using the PflMI restriction sites
- The plasmids have been confirmed by sequencing.
- All viruses are infectious – infectivity was determined by TZM-bl assay and replication kinetics
- p948.1 is WT pIndie

4. pNL4-3 and Variants

Plasmid Numbers: p949.1 – p949.4
Plasmid Name: pNL4-3 RT Variants
Features: Full length infectious viral variants, pNL4-3, Mutations at Pos. 359 of RT.

Keywords: pNL4-3, subtype-B RT variants, full-length infectious viruses, Ampicillin resistance.



Plasmid Number	Plasmid Name	Amino Acid at Pos. 359
p949.1	pNL4-3	Glycine
p949.2	pNL4-3 G359	Threonine
p949.3	pNL4-3 S359	Serine
p949.4	pNL4-3 A359	Alanine

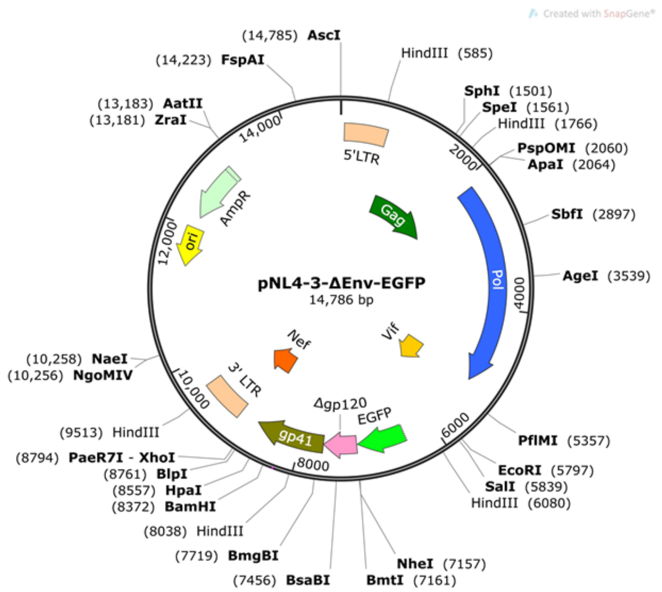
Key Points:

- This series of plasmids are derived from the pNL4-3 full length molecular clone. The plasmids contain mutations at pos. 359 of RT.
- The mutations were introduced by using overlap PCR and cloned using the AgeI and EcoRI restriction sites
- The plasmids have been confirmed by sequencing.
- All viruses are infectious – infectivity was determined by TZM-bl assay and replication kinetics
- p949.1 is WT NL4-3

5. pNL4-3-ΔEnv-EGFP Variants

Plasmid Numbers: p950.1 – p950.12
Plasmid Name: NL4-3-Δenv-EGFP
Features: Near-full-length reporter viral variants, pNL4-3, Mutations at Pos. 359 of RT.

Keywords: pNL4-3, subtype-B, Δenv, EGFP, frameshift mutants, near full-length viruses, Ampicillin resistance.



Plasmid Number	Plasmid Name (pNL4-3-ΔEnv EGFP-)
p950.1	G0
p950.2	G4
p950.3	G204
p950.4	T0
p950.5	T4
p950.6	T4
p950.7	S0
p950.8	S4
p950.9	S204
p950.10	A0
p950.11	A4
p950.12	A204

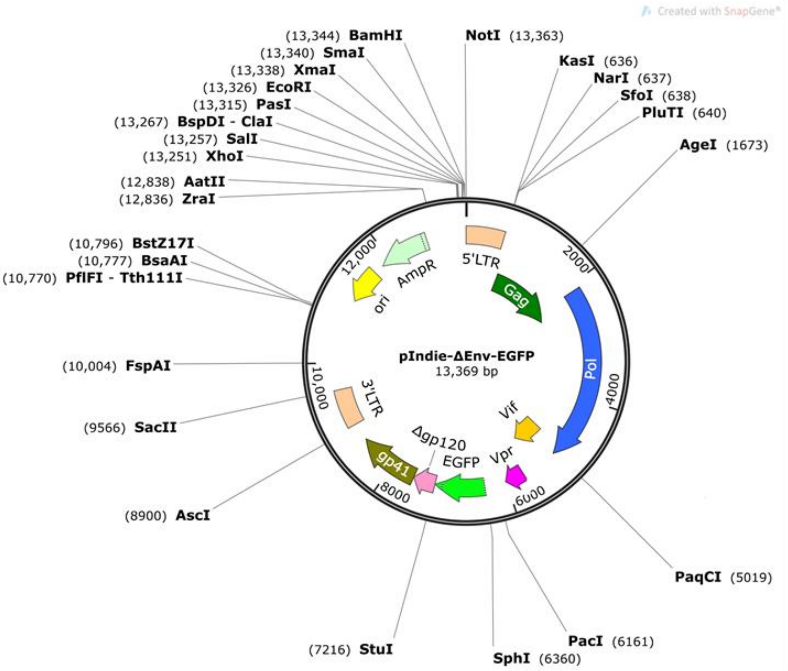
Key Points:

- This series of plasmids are derived from the pNL4-3-ΔEnv-EGFP molecular clone.
- The mutations were introduced in RT by domain swapping pNL4-3 RT variants using the AgeI and EcoRI restriction sites
- All plasmid sequences have been confirmed by sequencing.
- The EGFP mutations were introduced by overlap PCR and cloned using the SphI and StuI restriction enzyme sites.
- These molecular clones can be used for single round infection studies by pseudotyping with the VSV-G or HIV envelope.
- p950.1 is a modified version of pNL4-3-ΔEnv-EGFP developed by Dr. Haili Zhang, Dr. Yan Zhou and Dr. Robert Siliciano, such that the KDEL sequence has been deleted.

6. pIndie-ΔEnv-EGFP Variants

Plasmid Numbers: p951.1 – p951.12
Plasmid Name: pIndie-Δenv-EGFP
Features: Near-full-length reporter viral variants, pIndie, Mutations at Pos. 359 of RT.

Keywords: pIndie, subtype-C, ΔEnv, EGFP, frameshift mutants, near full-length viruses, Ampicillin resistance.



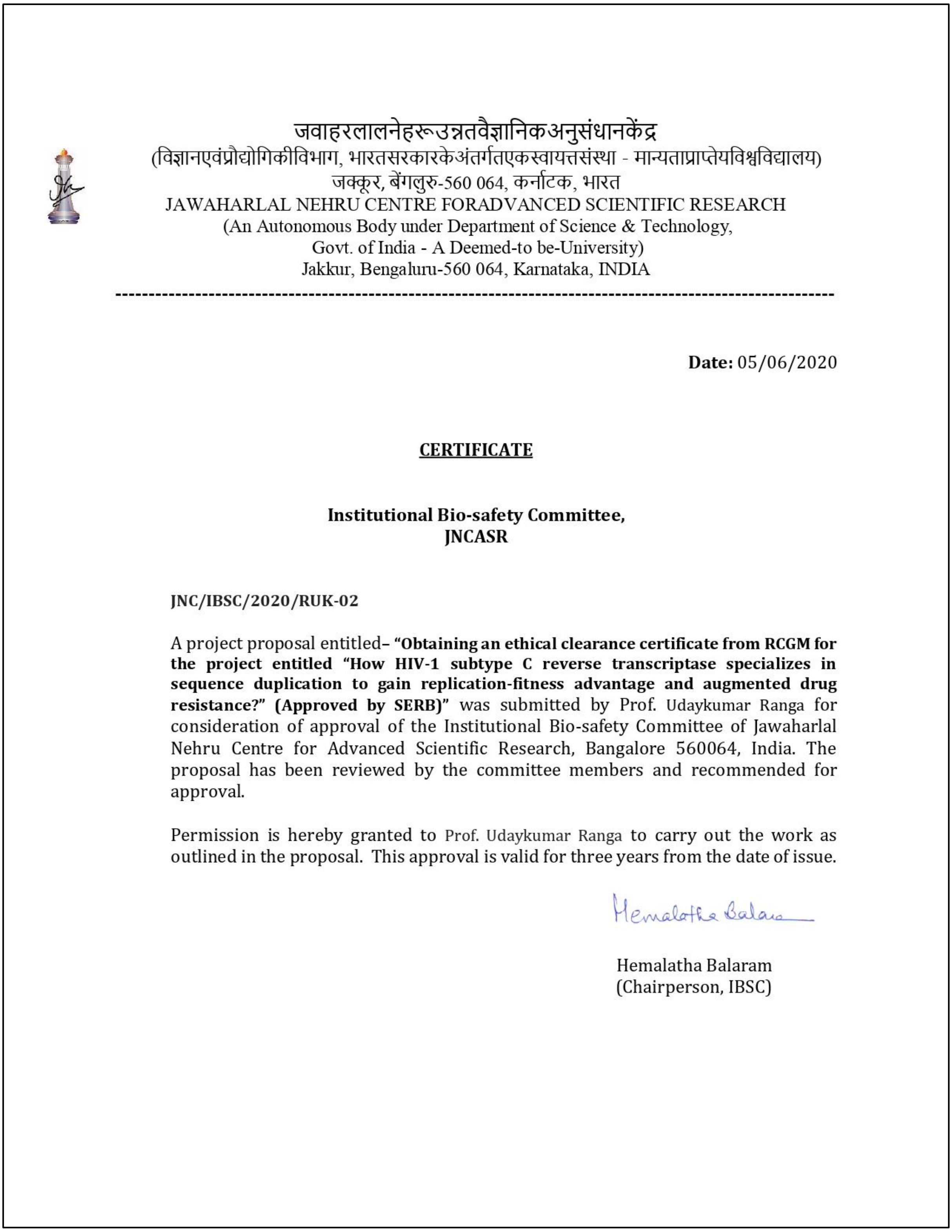
Plasmid Number	Plasmid Name (pIndie-ΔEnv EGFP-)
p951.1	G0
p951.2	G4
p951.3	G204
p951.4	T0
p951.5	T4
p951.6	T4
p951.7	S0
p951.8	S4
p951.9	S204
P951.10	A0
p951.11	A4
p951.12	A204

Key Points:

- This series of plasmids are derived from the pIndie-ΔEnv-EGFP molecular clone.
- The mutations were introduced in RT by domain swapping pIndie RT variants using the AgeI and PacI restriction sites
- The EGFP mutations were introduced by overlap PCR and cloned using the SphI and StuI restriction enzyme sites.
- All plasmid sequences have been confirmed by sequencing.
- These molecular clones can be used for single round infection studies by pseudotyping with the VSV-G or HIV envelope.

Appendix – II: Rights and Permissions

1. Institutional Bio-Safety Committee Approval



जवाहरलालनेहरूउन्नतवैज्ञानिकअनुसंधानकेंद्र
(विज्ञानएवंप्रौद्योगिकीविभाग, भारतसरकारकेअंतर्गतएकस्वायत्तसंस्था - मान्यताप्राप्तेयविश्वविद्यालय)
जक्कूर, बेंगलुरु-560 064, कर्नाटक, भारत
JAWAHARLAL NEHRU CENTRE FORADVANCED SCIENTIFIC RESEARCH
(An Autonomous Body under Department of Science & Technology,
Govt. of India - A Deemed-to be-University)
Jakkur, Bengaluru-560 064, Karnataka, INDIA

Date: 05/06/2020

CERTIFICATE

Institutional Bio-safety Committee,
JNCASR

JNC/IBSC/2020/RUK-02

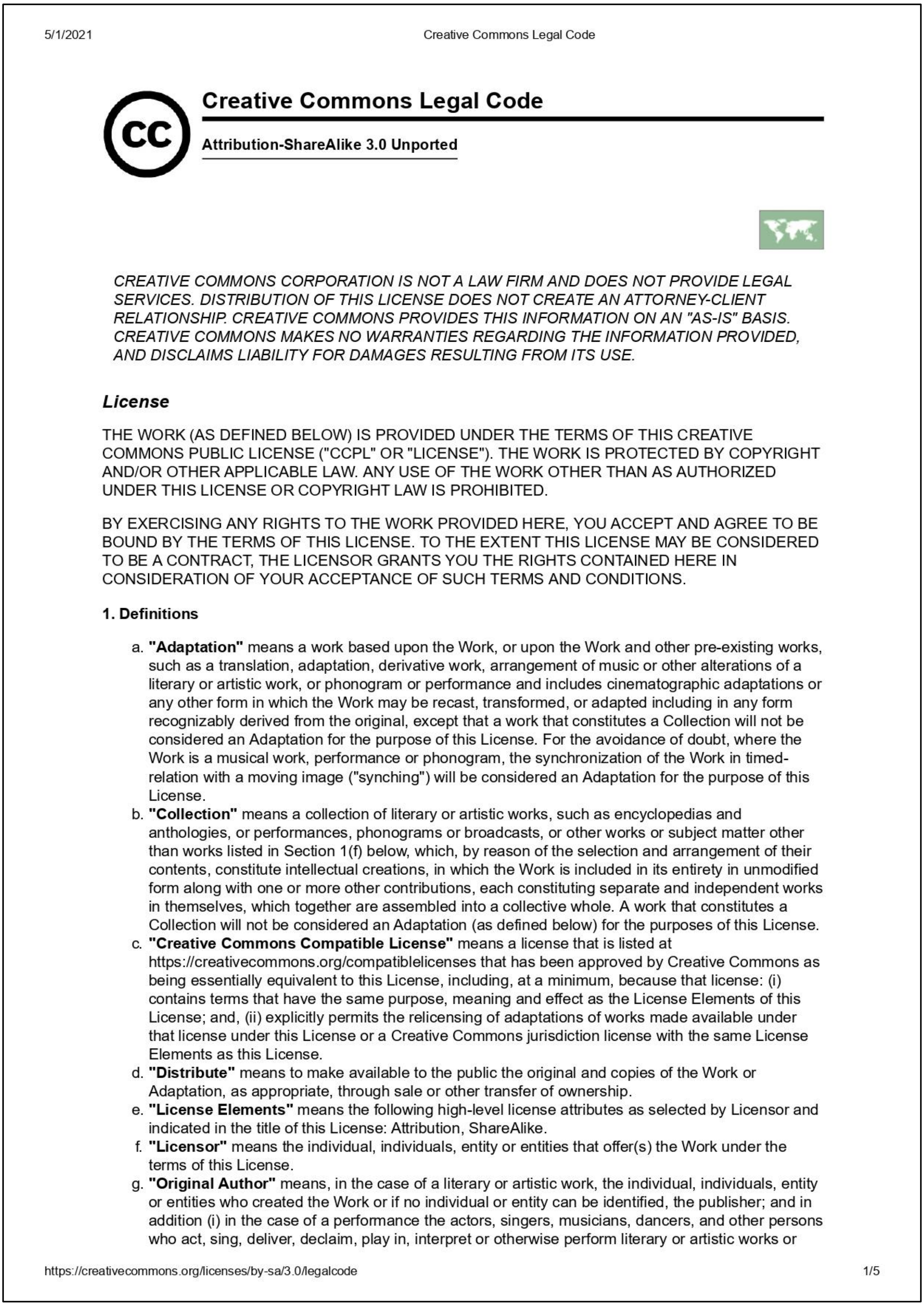
A project proposal entitled– **"Obtaining an ethical clearance certificate from RCGM for the project entitled "How HIV-1 subtype C reverse transcriptase specializes in sequence duplication to gain replication-fitness advantage and augmented drug resistance?" (Approved by SERB)"** was submitted by Prof. Udaykumar Ranga for consideration of approval of the Institutional Bio-safety Committee of Jawaharlal Nehru Centre for Advanced Scientific Research, Bangalore 560064, India. The proposal has been reviewed by the committee members and recommended for approval.

Permission is hereby granted to Prof. Udaykumar Ranga to carry out the work as outlined in the proposal. This approval is valid for three years from the date of issue.

Hemalatha Balaram

Hemalatha Balaram
(Chairperson, IBSC)

2. Creative Commons License



5/1/2021 Creative Commons Legal Code

Creative Commons Legal Code

Attribution-ShareAlike 3.0 Unported

CREATIVE COMMONS CORPORATION IS NOT A LAW FIRM AND DOES NOT PROVIDE LEGAL SERVICES. DISTRIBUTION OF THIS LICENSE DOES NOT CREATE AN ATTORNEY-CLIENT RELATIONSHIP. CREATIVE COMMONS PROVIDES THIS INFORMATION ON AN "AS-IS" BASIS. CREATIVE COMMONS MAKES NO WARRANTIES REGARDING THE INFORMATION PROVIDED, AND DISCLAIMS LIABILITY FOR DAMAGES RESULTING FROM ITS USE.

License

THE WORK (AS DEFINED BELOW) IS PROVIDED UNDER THE TERMS OF THIS CREATIVE COMMONS PUBLIC LICENSE ("CCPL" OR "LICENSE"). THE WORK IS PROTECTED BY COPYRIGHT AND/OR OTHER APPLICABLE LAW. ANY USE OF THE WORK OTHER THAN AS AUTHORIZED UNDER THIS LICENSE OR COPYRIGHT LAW IS PROHIBITED.

BY EXERCISING ANY RIGHTS TO THE WORK PROVIDED HERE, YOU ACCEPT AND AGREE TO BE BOUND BY THE TERMS OF THIS LICENSE. TO THE EXTENT THIS LICENSE MAY BE CONSIDERED TO BE A CONTRACT, THE LICENSOR GRANTS YOU THE RIGHTS CONTAINED HERE IN CONSIDERATION OF YOUR ACCEPTANCE OF SUCH TERMS AND CONDITIONS.

1. Definitions

a. **"Adaptation"** means a work based upon the Work, or upon the Work and other pre-existing works, such as a translation, adaptation, derivative work, arrangement of music or other alterations of a literary or artistic work, or phonogram or performance and includes cinematographic adaptations or any other form in which the Work may be recast, transformed, or adapted including in any form recognizably derived from the original, except that a work that constitutes a Collection will not be considered an Adaptation for the purpose of this License. For the avoidance of doubt, where the Work is a musical work, performance or phonogram, the synchronization of the Work in timed-relation with a moving image ("synching") will be considered an Adaptation for the purpose of this License.

b. **"Collection"** means a collection of literary or artistic works, such as encyclopedias and anthologies, or performances, phonograms or broadcasts, or other works or subject matter other than works listed in Section 1(f) below, which, by reason of the selection and arrangement of their contents, constitute intellectual creations, in which the Work is included in its entirety in unmodified form along with one or more other contributions, each constituting separate and independent works in themselves, which together are assembled into a collective whole. A work that constitutes a Collection will not be considered an Adaptation (as defined below) for the purposes of this License.

c. **"Creative Commons Compatible License"** means a license that is listed at https://creativecommons.org/compatiblelicenses that has been approved by Creative Commons as being essentially equivalent to this License, including, at a minimum, because that license: (i) contains terms that have the same purpose, meaning and effect as the License Elements of this License; and, (ii) explicitly permits the relicensing of adaptations of works made available under that license under this License or a Creative Commons jurisdiction license with the same License Elements as this License.

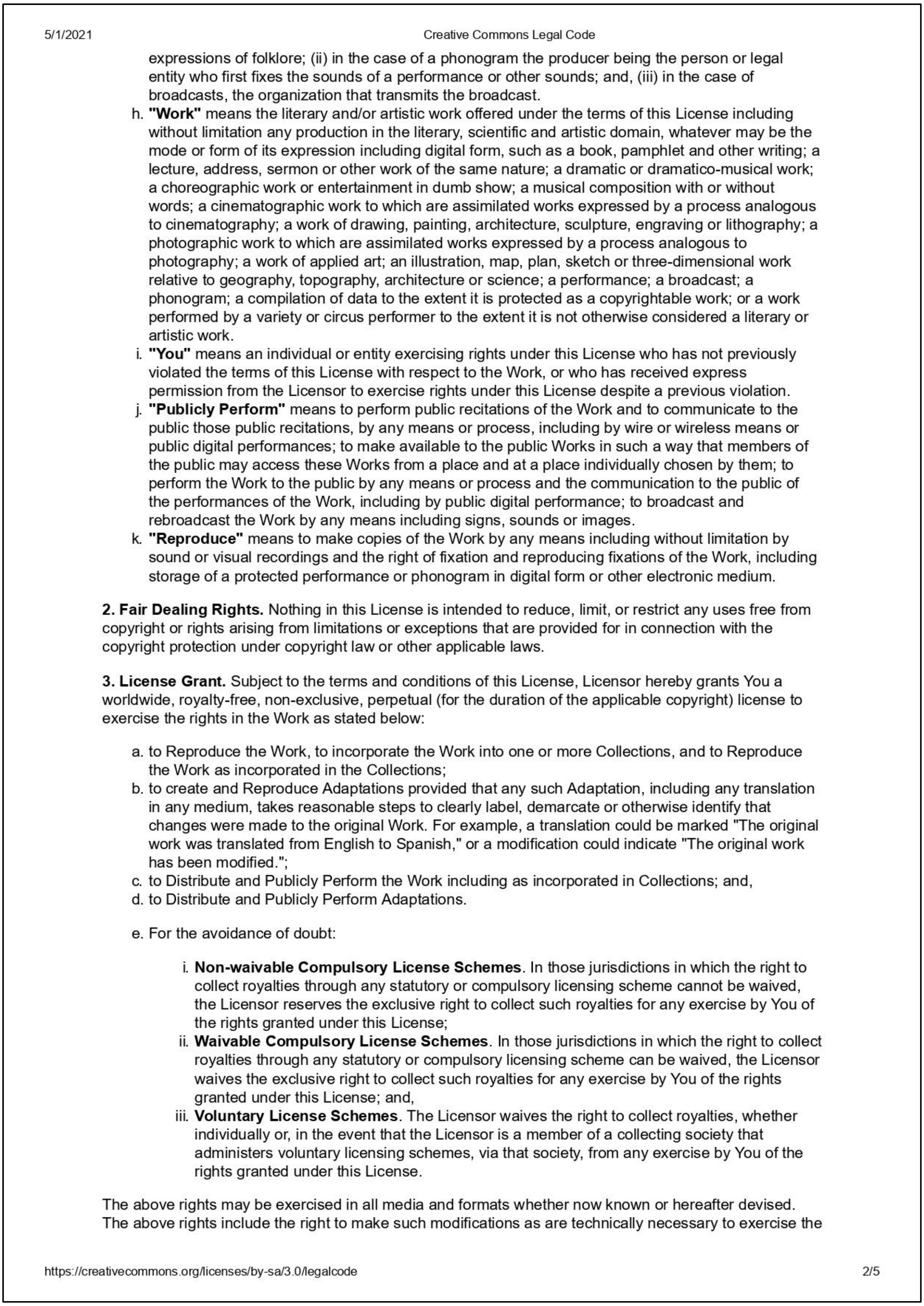
d. **"Distribute"** means to make available to the public the original and copies of the Work or Adaptation, as appropriate, through sale or other transfer of ownership.

e. **"License Elements"** means the following high-level license attributes as selected by Licensor and indicated in the title of this License: Attribution, ShareAlike.

f. **"Licensor"** means the individual, individuals, entity or entities that offer(s) the Work under the terms of this License.

g. **"Original Author"** means, in the case of a literary or artistic work, the individual, individuals, entity or entities who created the Work or if no individual or entity can be identified, the publisher; and in addition (i) in the case of a performance the actors, singers, musicians, dancers, and other persons who act, sing, deliver, declaim, play in, interpret or otherwise perform literary or artistic works or

https://creativecommons.org/licenses/by-sa/3.0/legalcode 1/5



5/1/2021 Creative Commons Legal Code

expressions of folklore; (ii) in the case of a phonogram the producer being the person or legal entity who first fixes the sounds of a performance or other sounds; and, (iii) in the case of broadcasts, the organization that transmits the broadcast.

h. **"Work"** means the literary and/or artistic work offered under the terms of this License including without limitation any production in the literary, scientific and artistic domain, whatever may be the mode or form of its expression including digital form, such as a book, pamphlet and other writing; a lecture, address, sermon or other work of the same nature; a dramatic or dramatico-musical work; a choreographic work or entertainment in dumb show; a musical composition with or without words; a cinematographic work to which are assimilated works expressed by a process analogous to cinematography; a work of drawing, painting, architecture, sculpture, engraving or lithography; a photographic work to which are assimilated works expressed by a process analogous to photography; a work of applied art; an illustration, map, plan, sketch or three-dimensional work relative to geography, topography, architecture or science; a performance; a broadcast; a phonogram; a compilation of data to the extent it is protected as a copyrightable work; or a work performed by a variety or circus performer to the extent it is not otherwise considered a literary or artistic work.

i. **"You"** means an individual or entity exercising rights under this License who has not previously violated the terms of this License with respect to the Work, or who has received express permission from the Licensor to exercise rights under this License despite a previous violation.

j. **"Publicly Perform"** means to perform public recitations of the Work and to communicate to the public those public recitations, by any means or process, including by wire or wireless means or public digital performances; to make available to the public Works in such a way that members of the public may access these Works from a place and at a place individually chosen by them; to perform the Work to the public by any means or process and the communication to the public of the performances of the Work, including by public digital performance; to broadcast and rebroadcast the Work by any means including signs, sounds or images.

k. **"Reproduce"** means to make copies of the Work by any means including without limitation by sound or visual recordings and the right of fixation and reproducing fixations of the Work, including storage of a protected performance or phonogram in digital form or other electronic medium.

2. Fair Dealing Rights. Nothing in this License is intended to reduce, limit, or restrict any uses free from copyright or rights arising from limitations or exceptions that are provided for in connection with the copyright protection under copyright law or other applicable laws.

3. License Grant. Subject to the terms and conditions of this License, Licensor hereby grants You a worldwide, royalty-free, non-exclusive, perpetual (for the duration of the applicable copyright) license to exercise the rights in the Work as stated below:

a. to Reproduce the Work, to incorporate the Work into one or more Collections, and to Reproduce the Work as incorporated in the Collections;

b. to create and Reproduce Adaptations provided that any such Adaptation, including any translation in any medium, takes reasonable steps to clearly label, demarcate or otherwise identify that changes were made to the original Work. For example, a translation could be marked "The original work was translated from English to Spanish," or a modification could indicate "The original work has been modified.";

c. to Distribute and Publicly Perform the Work including as incorporated in Collections; and,

d. to Distribute and Publicly Perform Adaptations.

e. For the avoidance of doubt:

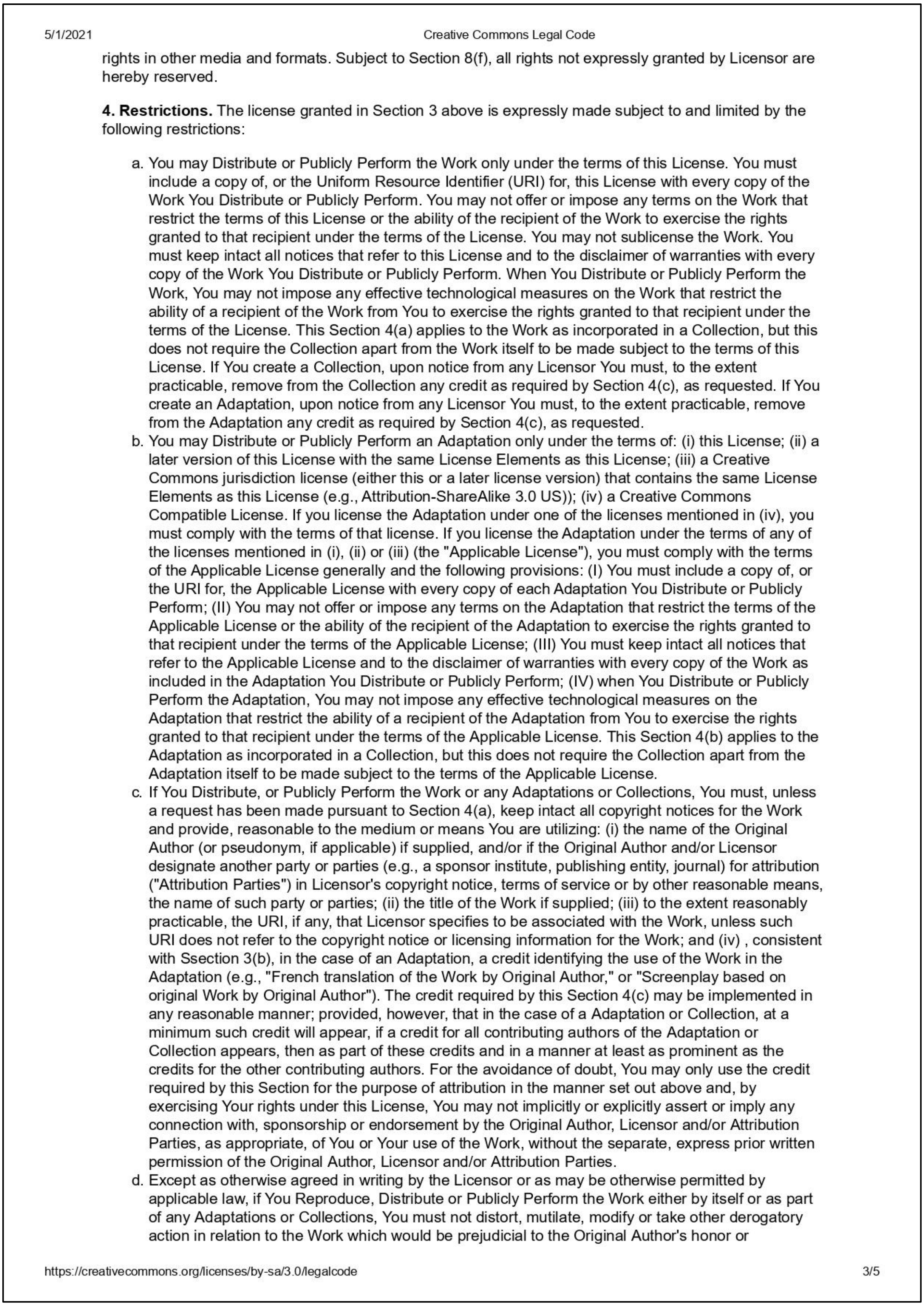
   i. **Non-waivable Compulsory License Schemes**. In those jurisdictions in which the right to collect royalties through any statutory or compulsory licensing scheme cannot be waived, the Licensor reserves the exclusive right to collect such royalties for any exercise by You of the rights granted under this License;

   ii. **Waivable Compulsory License Schemes**. In those jurisdictions in which the right to collect royalties through any statutory or compulsory licensing scheme can be waived, the Licensor waives the exclusive right to collect such royalties for any exercise by You of the rights granted under this License; and,

   iii. **Voluntary License Schemes**. The Licensor waives the right to collect royalties, whether individually or, in the event that the Licensor is a member of a collecting society that administers voluntary licensing schemes, via that society, from any exercise by You of the rights granted under this License.

The above rights may be exercised in all media and formats whether now known or hereafter devised. The above rights include the right to make such modifications as are technically necessary to exercise the

https://creativecommons.org/licenses/by-sa/3.0/legalcode 2/5



rights in other media and formats. Subject to Section 8(f), all rights not expressly granted by Licensor are hereby reserved.

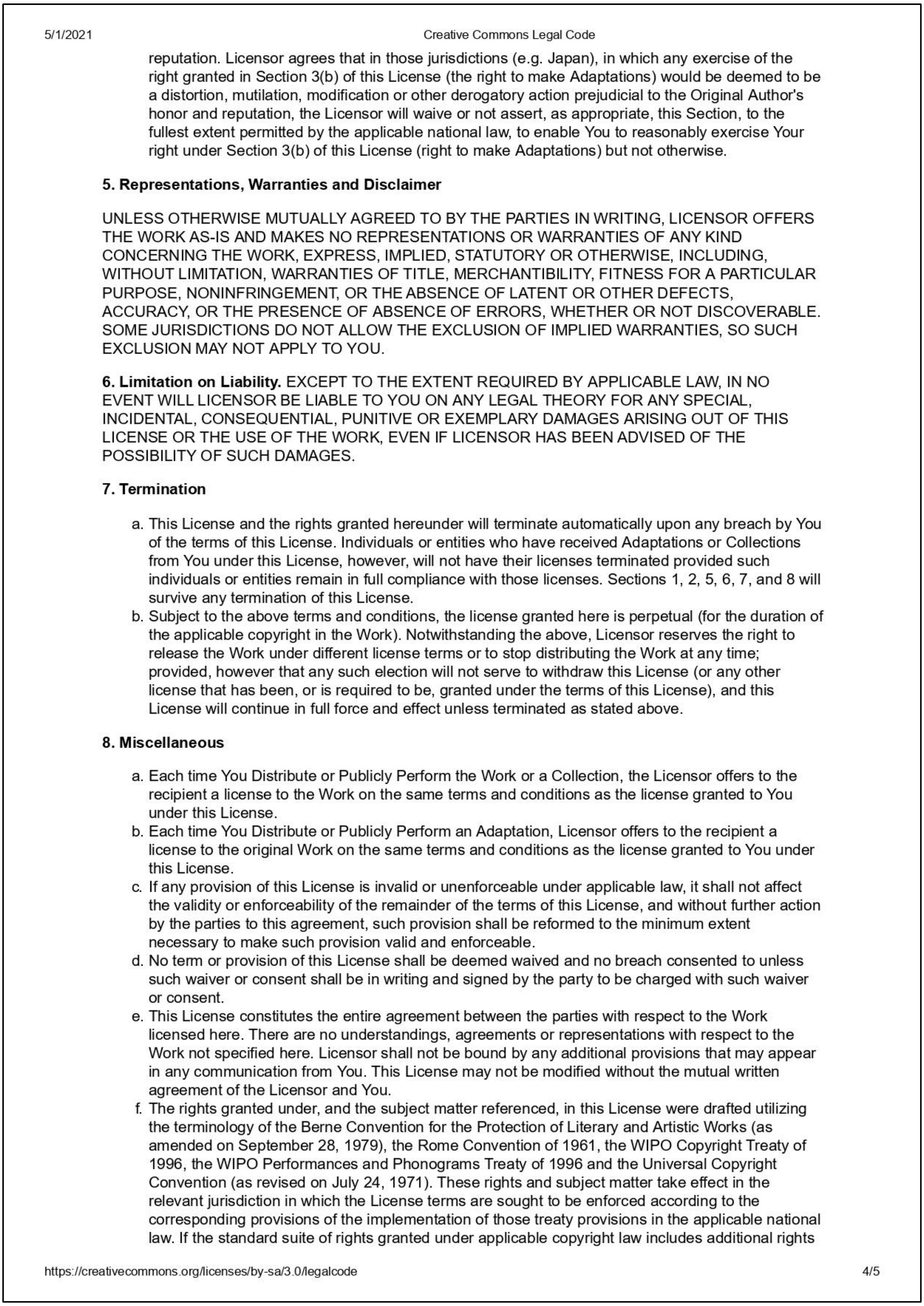
4. Restrictions. The license granted in Section 3 above is expressly made subject to and limited by the following restrictions:

a. You may Distribute or Publicly Perform the Work only under the terms of this License. You must include a copy of, or the Uniform Resource Identifier (URI) for, this License with every copy of the Work You Distribute or Publicly Perform. You may not offer or impose any terms on the Work that restrict the terms of this License or the ability of the recipient of the Work to exercise the rights granted to that recipient under the terms of the License. You may not sublicense the Work. You must keep intact all notices that refer to this License and to the disclaimer of warranties with every copy of the Work You Distribute or Publicly Perform. When You Distribute or Publicly Perform the Work, You may not impose any effective technological measures on the Work that restrict the ability of a recipient of the Work from You to exercise the rights granted to that recipient under the terms of the License. This Section 4(a) applies to the Work as incorporated in a Collection, but this does not require the Collection apart from the Work itself to be made subject to the terms of this License. If You create a Collection, upon notice from any Licensor You must, to the extent practicable, remove from the Collection any credit as required by Section 4(c), as requested. If You create an Adaptation, upon notice from any Licensor You must, to the extent practicable, remove from the Adaptation any credit as required by Section 4(c), as requested.

b. You may Distribute or Publicly Perform an Adaptation only under the terms of: (i) this License; (ii) a later version of this License with the same License Elements as this License; (iii) a Creative Commons jurisdiction license (either this or a later license version) that contains the same License Elements as this License (e.g., Attribution-ShareAlike 3.0 US)); (iv) a Creative Commons Compatible License. If you license the Adaptation under one of the licenses mentioned in (iv), you must comply with the terms of that license. If you license the Adaptation under the terms of any of the licenses mentioned in (i), (ii) or (iii) (the "Applicable License"), you must comply with the terms of the Applicable License generally and the following provisions: (I) You must include a copy of, or the URI for, the Applicable License with every copy of each Adaptation You Distribute or Publicly Perform; (II) You may not offer or impose any terms on the Adaptation that restrict the terms of the Applicable License or the ability of the recipient of the Adaptation to exercise the rights granted to that recipient under the terms of the Applicable License; (III) You must keep intact all notices that refer to the Applicable License and to the disclaimer of warranties with every copy of the Work as included in the Adaptation You Distribute or Publicly Perform; (IV) when You Distribute or Publicly Perform the Adaptation, You may not impose any effective technological measures on the Adaptation that restrict the ability of a recipient of the Adaptation from You to exercise the rights granted to that recipient under the terms of the Applicable License. This Section 4(b) applies to the Adaptation as incorporated in a Collection, but this does not require the Collection apart from the Adaptation itself to be made subject to the terms of the Applicable License.

c. If You Distribute, or Publicly Perform the Work or any Adaptations or Collections, You must, unless a request has been made pursuant to Section 4(a), keep intact all copyright notices for the Work and provide, reasonable to the medium or means You are utilizing: (i) the name of the Original Author (or pseudonym, if applicable) if supplied, and/or if the Original Author and/or Licensor designate another party or parties (e.g., a sponsor institute, publishing entity, journal) for attribution ("Attribution Parties") in Licensor's copyright notice, terms of service or by other reasonable means, the name of such party or parties; (ii) the title of the Work if supplied; (iii) to the extent reasonably practicable, the URI, if any, that Licensor specifies to be associated with the Work, unless such URI does not refer to the copyright notice or licensing information for the Work; and (iv) , consistent with Ssection 3(b), in the case of an Adaptation, a credit identifying the use of the Work in the Adaptation (e.g., "French translation of the Work by Original Author," or "Screenplay based on original Work by Original Author"). The credit required by this Section 4(c) may be implemented in any reasonable manner; provided, however, that in the case of a Adaptation or Collection, at a minimum such credit will appear, if a credit for all contributing authors of the Adaptation or Collection appears, then as part of these credits and in a manner at least as prominent as the credits for the other contributing authors. For the avoidance of doubt, You may only use the credit required by this Section for the purpose of attribution in the manner set out above and, by exercising Your rights under this License, You may not implicitly or explicitly assert or imply any connection with, sponsorship or endorsement by the Original Author, Licensor and/or Attribution Parties, as appropriate, of You or Your use of the Work, without the separate, express prior written permission of the Original Author, Licensor and/or Attribution Parties.

d. Except as otherwise agreed in writing by the Licensor or as may be otherwise permitted by applicable law, if You Reproduce, Distribute or Publicly Perform the Work either by itself or as part of any Adaptations or Collections, You must not distort, mutilate, modify or take other derogatory action in relation to the Work which would be prejudicial to the Original Author's honor or



5/1/2021 Creative Commons Legal Code

reputation. Licensor agrees that in those jurisdictions (e.g. Japan), in which any exercise of the right granted in Section 3(b) of this License (the right to make Adaptations) would be deemed to be a distortion, mutilation, modification or other derogatory action prejudicial to the Original Author's honor and reputation, the Licensor will waive or not assert, as appropriate, this Section, to the fullest extent permitted by the applicable national law, to enable You to reasonably exercise Your right under Section 3(b) of this License (right to make Adaptations) but not otherwise.

5. Representations, Warranties and Disclaimer

UNLESS OTHERWISE MUTUALLY AGREED TO BY THE PARTIES IN WRITING, LICENSOR OFFERS THE WORK AS-IS AND MAKES NO REPRESENTATIONS OR WARRANTIES OF ANY KIND CONCERNING THE WORK, EXPRESS, IMPLIED, STATUTORY OR OTHERWISE, INCLUDING, WITHOUT LIMITATION, WARRANTIES OF TITLE, MERCHANTIBILITY, FITNESS FOR A PARTICULAR PURPOSE, NONINFRINGEMENT, OR THE ABSENCE OF LATENT OR OTHER DEFECTS, ACCURACY, OR THE PRESENCE OF ABSENCE OF ERRORS, WHETHER OR NOT DISCOVERABLE. SOME JURISDICTIONS DO NOT ALLOW THE EXCLUSION OF IMPLIED WARRANTIES, SO SUCH EXCLUSION MAY NOT APPLY TO YOU.

6. Limitation on Liability. EXCEPT TO THE EXTENT REQUIRED BY APPLICABLE LAW, IN NO EVENT WILL LICENSOR BE LIABLE TO YOU ON ANY LEGAL THEORY FOR ANY SPECIAL, INCIDENTAL, CONSEQUENTIAL, PUNITIVE OR EXEMPLARY DAMAGES ARISING OUT OF THIS LICENSE OR THE USE OF THE WORK, EVEN IF LICENSOR HAS BEEN ADVISED OF THE POSSIBILITY OF SUCH DAMAGES.

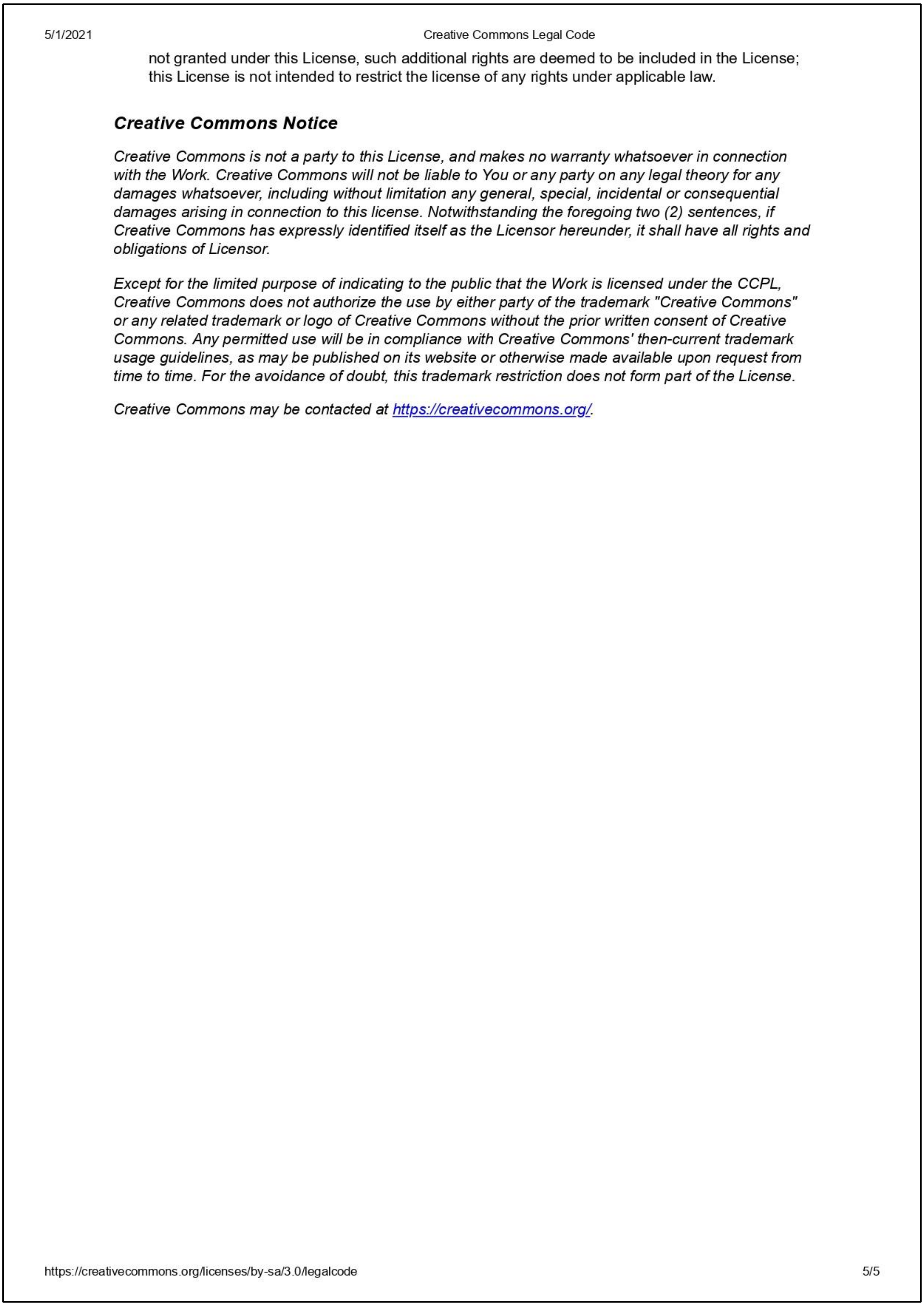
7. Termination

a. This License and the rights granted hereunder will terminate automatically upon any breach by You of the terms of this License. Individuals or entities who have received Adaptations or Collections from You under this License, however, will not have their licenses terminated provided such individuals or entities remain in full compliance with those licenses. Sections 1, 2, 5, 6, 7, and 8 will survive any termination of this License.
b. Subject to the above terms and conditions, the license granted here is perpetual (for the duration of the applicable copyright in the Work). Notwithstanding the above, Licensor reserves the right to release the Work under different license terms or to stop distributing the Work at any time; provided, however that any such election will not serve to withdraw this License (or any other license that has been, or is required to be, granted under the terms of this License), and this License will continue in full force and effect unless terminated as stated above.

8. Miscellaneous

a. Each time You Distribute or Publicly Perform the Work or a Collection, the Licensor offers to the recipient a license to the Work on the same terms and conditions as the license granted to You under this License.
b. Each time You Distribute or Publicly Perform an Adaptation, Licensor offers to the recipient a license to the original Work on the same terms and conditions as the license granted to You under this License.
c. If any provision of this License is invalid or unenforceable under applicable law, it shall not affect the validity or enforceability of the remainder of the terms of this License, and without further action by the parties to this agreement, such provision shall be reformed to the minimum extent necessary to make such provision valid and enforceable.
d. No term or provision of this License shall be deemed waived and no breach consented to unless such waiver or consent shall be in writing and signed by the party to be charged with such waiver or consent.
e. This License constitutes the entire agreement between the parties with respect to the Work licensed here. There are no understandings, agreements or representations with respect to the Work not specified here. Licensor shall not be bound by any additional provisions that may appear in any communication from You. This License may not be modified without the mutual written agreement of the Licensor and You.
f. The rights granted under, and the subject matter referenced, in this License were drafted utilizing the terminology of the Berne Convention for the Protection of Literary and Artistic Works (as amended on September 28, 1979), the Rome Convention of 1961, the WIPO Copyright Treaty of 1996, the WIPO Performances and Phonograms Treaty of 1996 and the Universal Copyright Convention (as revised on July 24, 1971). These rights and subject matter take effect in the relevant jurisdiction in which the License terms are sought to be enforced according to the corresponding provisions of the implementation of those treaty provisions in the applicable national law. If the standard suite of rights granted under applicable copyright law includes additional rights

https://creativecommons.org/licenses/by-sa/3.0/legalcode 4/5



not granted under this License, such additional rights are deemed to be included in the License; this License is not intended to restrict the license of any rights under applicable law.

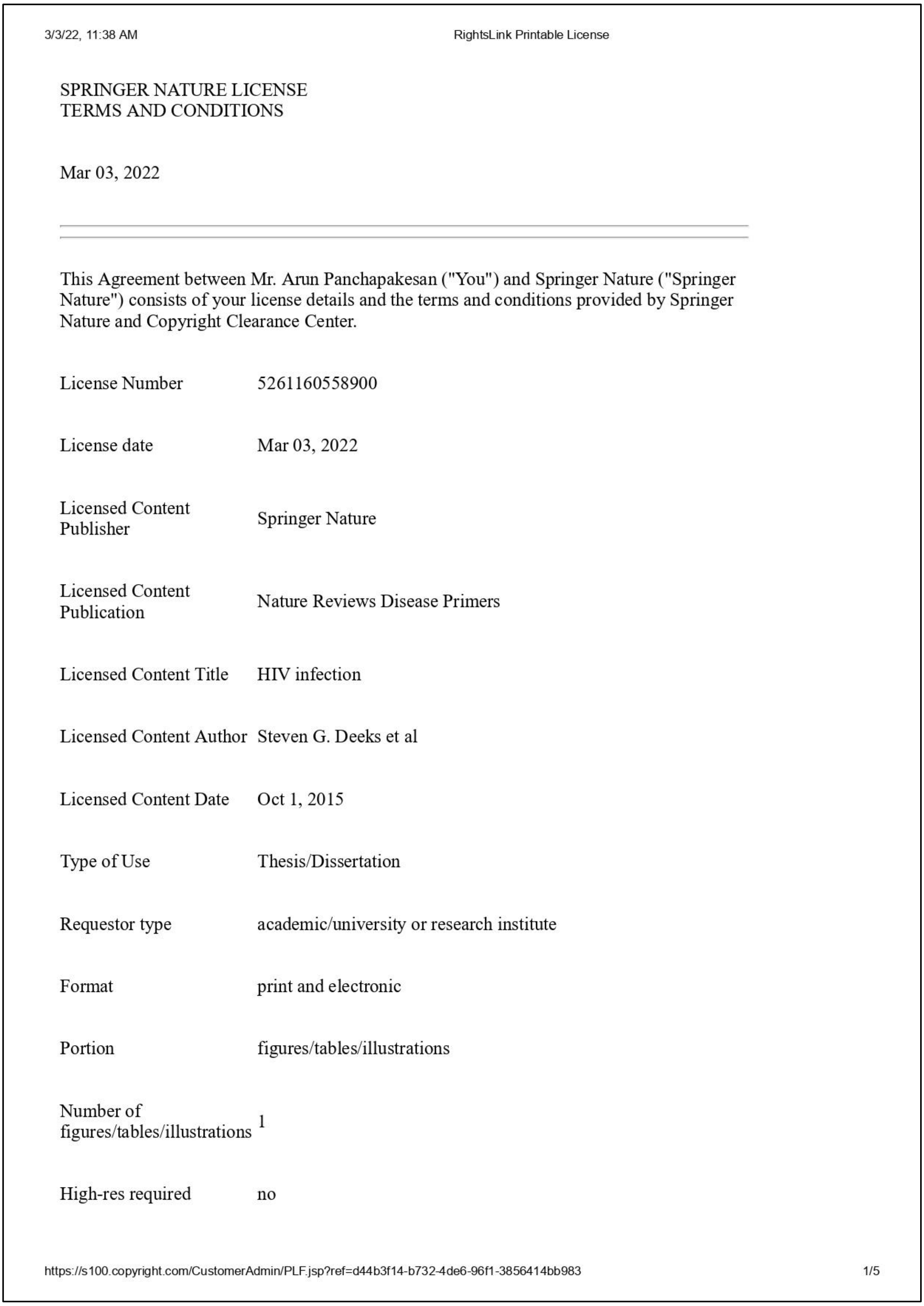
Creative Commons Notice

Creative Commons is not a party to this License, and makes no warranty whatsoever in connection with the Work. Creative Commons will not be liable to You or any party on any legal theory for any damages whatsoever, including without limitation any general, special, incidental or consequential damages arising in connection to this license. Notwithstanding the foregoing two (2) sentences, if Creative Commons has expressly identified itself as the Licensor hereunder, it shall have all rights and obligations of Licensor.

Except for the limited purpose of indicating to the public that the Work is licensed under the CCPL, Creative Commons does not authorize the use by either party of the trademark "Creative Commons" or any related trademark or logo of Creative Commons without the prior written consent of Creative Commons. Any permitted use will be in compliance with Creative Commons' then-current trademark usage guidelines, as may be published on its website or otherwise made available upon request from time to time. For the avoidance of doubt, this trademark restriction does not form part of the License.

Creative Commons may be contacted at https://creativecommons.org/.

3. Figure – 3: Major Steps in the Life cycle of HIV-1.



3/3/22, 11:38 AM RightsLink Printable License

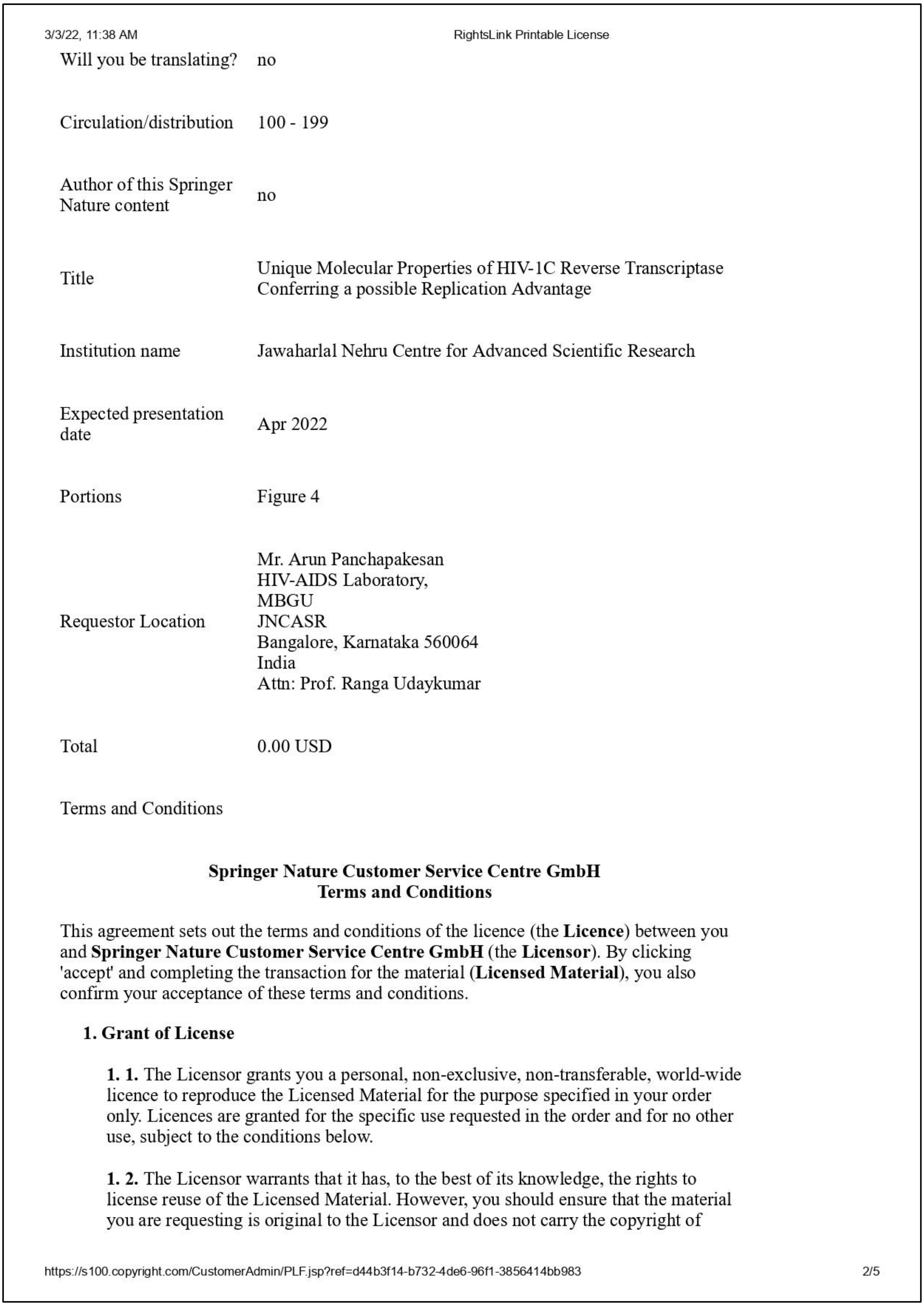
SPRINGER NATURE LICENSE
TERMS AND CONDITIONS

Mar 03, 2022

This Agreement between Mr. Arun Panchapakesan ("You") and Springer Nature ("Springer Nature") consists of your license details and the terms and conditions provided by Springer Nature and Copyright Clearance Center.

License Number	5261160558900
License date	Mar 03, 2022
Licensed Content Publisher	Springer Nature
Licensed Content Publication	Nature Reviews Disease Primers
Licensed Content Title	HIV infection
Licensed Content Author	Steven G. Deeks et al
Licensed Content Date	Oct 1, 2015
Type of Use	Thesis/Dissertation
Requestor type	academic/university or research institute
Format	print and electronic
Portion	figures/tables/illustrations
Number of figures/tables/illustrations	1
High-res required	no

https://s100.copyright.com/CustomerAdmin/PLF.jsp?ref=d44b3f14-b732-4de6-96f1-3856414bb983 1/5



Will you be translating?	no
Circulation/distribution	100 - 199
Author of this Springer Nature content	no
Title	Unique Molecular Properties of HIV-1C Reverse Transcriptase Conferring a possible Replication Advantage
Institution name	Jawaharlal Nehru Centre for Advanced Scientific Research
Expected presentation date	Apr 2022
Portions	Figure 4
Requestor Location	Mr. Arun Panchapakesan HIV-AIDS Laboratory, MBGU JNCASR Bangalore, Karnataka 560064 India Attn: Prof. Ranga Udaykumar
Total	0.00 USD
Terms and Conditions	

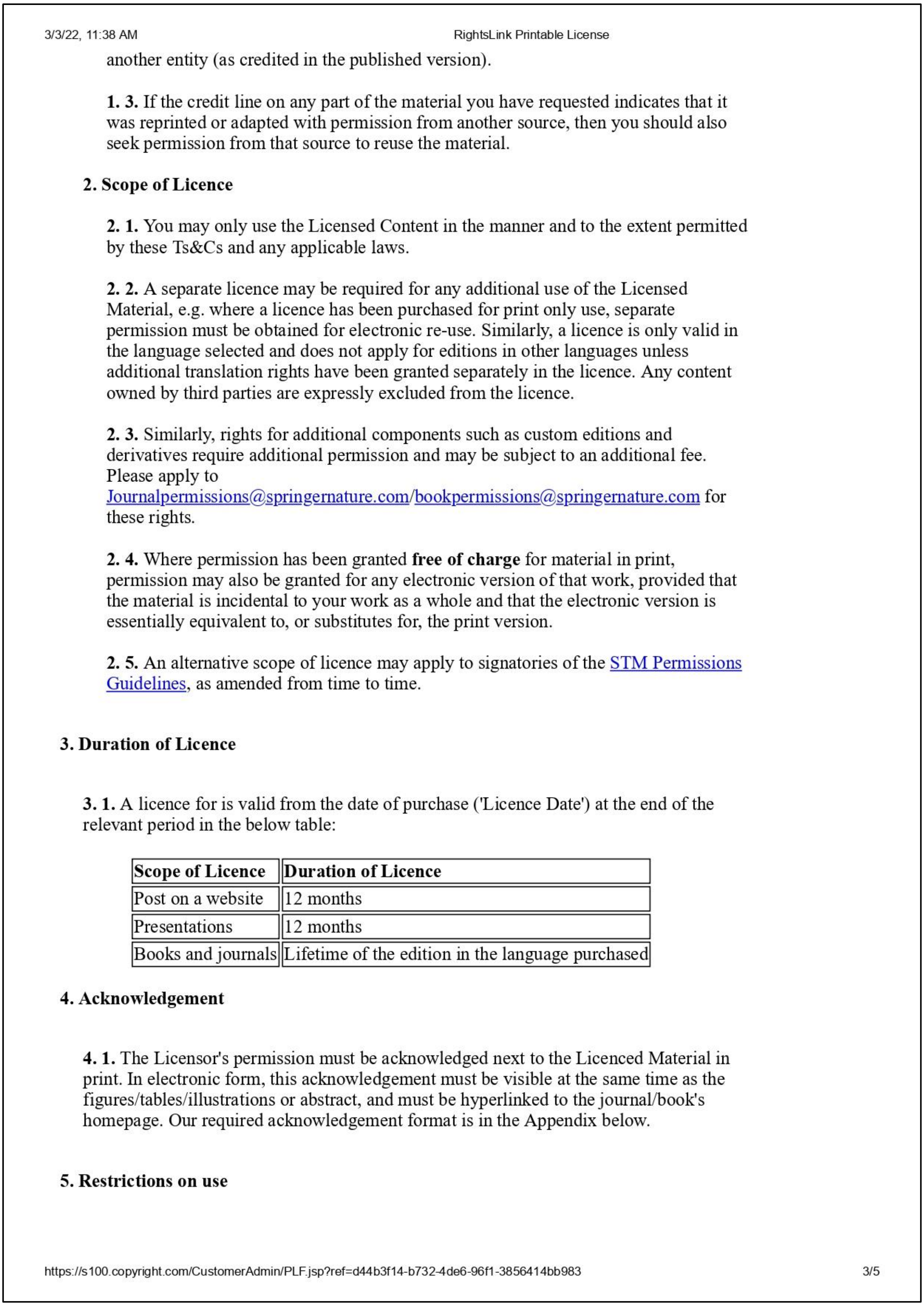
Springer Nature Customer Service Centre GmbH
Terms and Conditions

This agreement sets out the terms and conditions of the licence (the **Licence**) between you and **Springer Nature Customer Service Centre GmbH** (the **Licensor**). By clicking 'accept' and completing the transaction for the material (**Licensed Material**), you also confirm your acceptance of these terms and conditions.

1. **Grant of License**

1. 1. The Licensor grants you a personal, non-exclusive, non-transferable, world-wide licence to reproduce the Licensed Material for the purpose specified in your order only. Licences are granted for the specific use requested in the order and for no other use, subject to the conditions below.

1. 2. The Licensor warrants that it has, to the best of its knowledge, the rights to license reuse of the Licensed Material. However, you should ensure that the material you are requesting is original to the Licensor and does not carry the copyright of



another entity (as credited in the published version).

1. 3. If the credit line on any part of the material you have requested indicates that it was reprinted or adapted with permission from another source, then you should also seek permission from that source to reuse the material.

2. Scope of Licence

2. 1. You may only use the Licensed Content in the manner and to the extent permitted by these Ts&Cs and any applicable laws.

2. 2. A separate licence may be required for any additional use of the Licensed Material, e.g. where a licence has been purchased for print only use, separate permission must be obtained for electronic re-use. Similarly, a licence is only valid in the language selected and does not apply for editions in other languages unless additional translation rights have been granted separately in the licence. Any content owned by third parties are expressly excluded from the licence.

2. 3. Similarly, rights for additional components such as custom editions and derivatives require additional permission and may be subject to an additional fee. Please apply to Journalpermissions@springernature.com/bookpermissions@springernature.com for these rights.

2. 4. Where permission has been granted **free of charge** for material in print, permission may also be granted for any electronic version of that work, provided that the material is incidental to your work as a whole and that the electronic version is essentially equivalent to, or substitutes for, the print version.

2. 5. An alternative scope of licence may apply to signatories of the STM Permissions Guidelines, as amended from time to time.

3. Duration of Licence

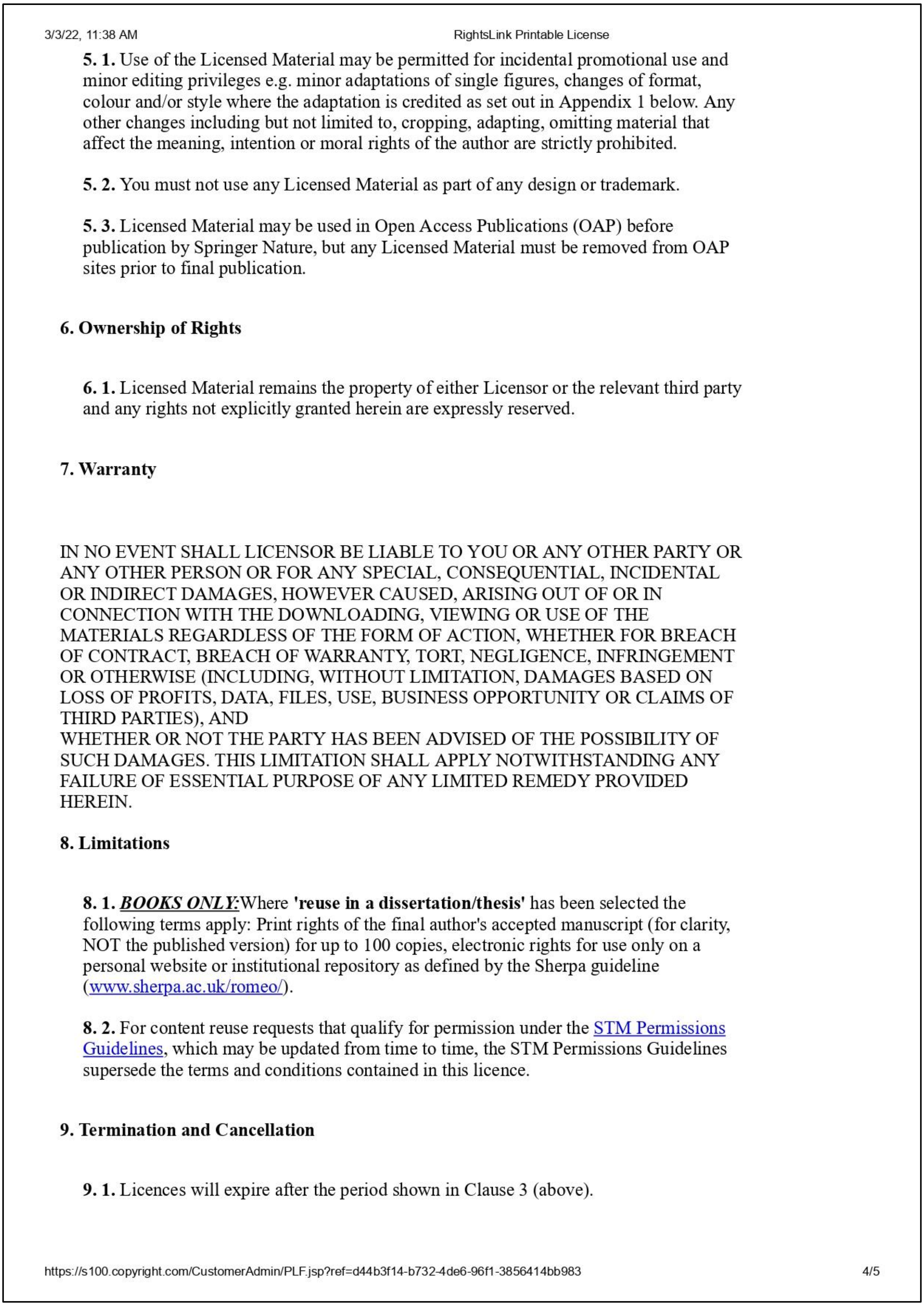
3. 1. A licence for is valid from the date of purchase ('Licence Date') at the end of the relevant period in the below table:

Scope of Licence	Duration of Licence
Post on a website	12 months
Presentations	12 months
Books and journals	Lifetime of the edition in the language purchased

4. Acknowledgement

4. 1. The Licensor's permission must be acknowledged next to the Licenced Material in print. In electronic form, this acknowledgement must be visible at the same time as the figures/tables/illustrations or abstract, and must be hyperlinked to the journal/book's homepage. Our required acknowledgement format is in the Appendix below.

5. Restrictions on use



5. 1. Use of the Licensed Material may be permitted for incidental promotional use and minor editing privileges e.g. minor adaptations of single figures, changes of format, colour and/or style where the adaptation is credited as set out in Appendix 1 below. Any other changes including but not limited to, cropping, adapting, omitting material that affect the meaning, intention or moral rights of the author are strictly prohibited.

5. 2. You must not use any Licensed Material as part of any design or trademark.

5. 3. Licensed Material may be used in Open Access Publications (OAP) before publication by Springer Nature, but any Licensed Material must be removed from OAP sites prior to final publication.

6. Ownership of Rights

6. 1. Licensed Material remains the property of either Licensor or the relevant third party and any rights not explicitly granted herein are expressly reserved.

7. Warranty

IN NO EVENT SHALL LICENSOR BE LIABLE TO YOU OR ANY OTHER PARTY OR ANY OTHER PERSON OR FOR ANY SPECIAL, CONSEQUENTIAL, INCIDENTAL OR INDIRECT DAMAGES, HOWEVER CAUSED, ARISING OUT OF OR IN CONNECTION WITH THE DOWNLOADING, VIEWING OR USE OF THE MATERIALS REGARDLESS OF THE FORM OF ACTION, WHETHER FOR BREACH OF CONTRACT, BREACH OF WARRANTY, TORT, NEGLIGENCE, INFRINGEMENT OR OTHERWISE (INCLUDING, WITHOUT LIMITATION, DAMAGES BASED ON LOSS OF PROFITS, DATA, FILES, USE, BUSINESS OPPORTUNITY OR CLAIMS OF THIRD PARTIES), AND
WHETHER OR NOT THE PARTY HAS BEEN ADVISED OF THE POSSIBILITY OF SUCH DAMAGES. THIS LIMITATION SHALL APPLY NOTWITHSTANDING ANY FAILURE OF ESSENTIAL PURPOSE OF ANY LIMITED REMEDY PROVIDED HEREIN.

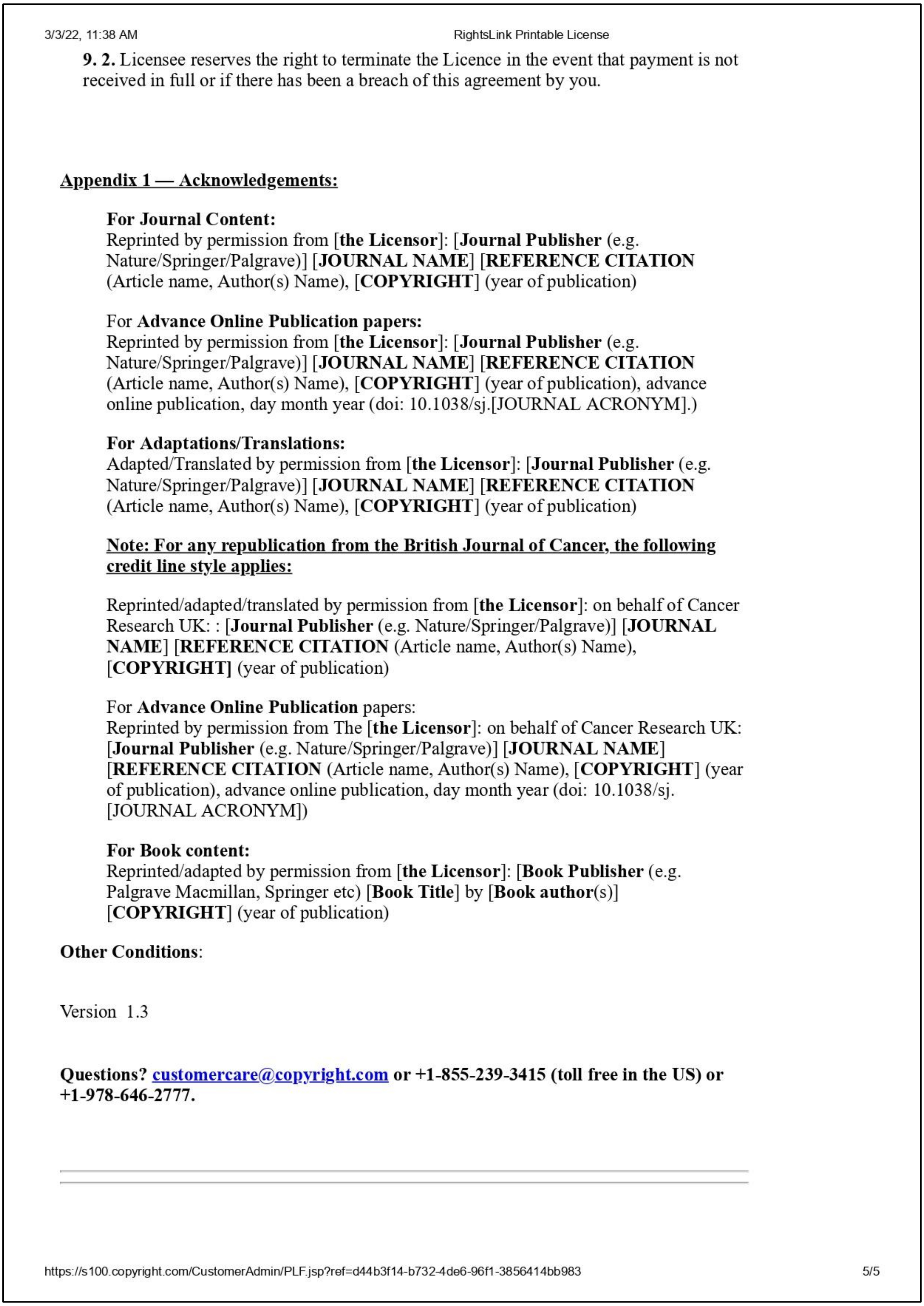
8. Limitations

8. 1. ***BOOKS ONLY:***Where **'reuse in a dissertation/thesis'** has been selected the following terms apply: Print rights of the final author's accepted manuscript (for clarity, NOT the published version) for up to 100 copies, electronic rights for use only on a personal website or institutional repository as defined by the Sherpa guideline (www.sherpa.ac.uk/romeo/).

8. 2. For content reuse requests that qualify for permission under the STM Permissions Guidelines, which may be updated from time to time, the STM Permissions Guidelines supersede the terms and conditions contained in this licence.

9. Termination and Cancellation

9. 1. Licences will expire after the period shown in Clause 3 (above).



9. 2. Licensee reserves the right to terminate the Licence in the event that payment is not received in full or if there has been a breach of this agreement by you.

Appendix 1 — Acknowledgements:

For Journal Content:
Reprinted by permission from [**the Licensor**]: [**Journal Publisher** (e.g. Nature/Springer/Palgrave)] [**JOURNAL NAME**] [**REFERENCE CITATION** (Article name, Author(s) Name), [**COPYRIGHT**] (year of publication)

For **Advance Online Publication papers:**
Reprinted by permission from [**the Licensor**]: [**Journal Publisher** (e.g. Nature/Springer/Palgrave)] [**JOURNAL NAME**] [**REFERENCE CITATION** (Article name, Author(s) Name), [**COPYRIGHT**] (year of publication), advance online publication, day month year (doi: 10.1038/sj.[JOURNAL ACRONYM].)

For Adaptations/Translations:
Adapted/Translated by permission from [**the Licensor**]: [**Journal Publisher** (e.g. Nature/Springer/Palgrave)] [**JOURNAL NAME**] [**REFERENCE CITATION** (Article name, Author(s) Name), [**COPYRIGHT**] (year of publication)

Note: For any republication from the British Journal of Cancer, the following credit line style applies:

Reprinted/adapted/translated by permission from [**the Licensor**]: on behalf of Cancer Research UK: : [**Journal Publisher** (e.g. Nature/Springer/Palgrave)] [**JOURNAL NAME**] [**REFERENCE CITATION** (Article name, Author(s) Name), [**COPYRIGHT]** (year of publication)

For **Advance Online Publication** papers:
Reprinted by permission from The [**the Licensor**]: on behalf of Cancer Research UK: [**Journal Publisher** (e.g. Nature/Springer/Palgrave)] [**JOURNAL NAME**] [**REFERENCE CITATION** (Article name, Author(s) Name), [**COPYRIGHT**] (year of publication), advance online publication, day month year (doi: 10.1038/sj. [JOURNAL ACRONYM])

For Book content:
Reprinted/adapted by permission from [**the Licensor**]: [**Book Publisher** (e.g. Palgrave Macmillan, Springer etc) [**Book Title**] by [**Book author**(s)] [**COPYRIGHT**] (year of publication)

Other Conditions:

Version 1.3

Questions? customercare@copyright.com or +1-855-239-3415 (toll free in the US) or +1-978-646-2777.