Control of HIV-1C Replication Both in Vivo and in Vitro

The Harvard community has made this article openly available. Please share how this access benefits you. Your story matters

<table>
<thead>
<tr>
<th>Citation</th>
<th>Xie, Wen. 2015. Control of HIV-1C Replication Both in Vivo and in Vitro. Doctoral dissertation, Harvard University, Graduate School of Arts &amp; Sciences.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Citable link</td>
<td><a href="http://nrs.harvard.edu/urn-3:HUL.InstRepos:17467379">http://nrs.harvard.edu/urn-3:HUL.InstRepos:17467379</a></td>
</tr>
<tr>
<td>Terms of Use</td>
<td>This article was downloaded from Harvard University’s DASH repository, and is made available under the terms and conditions applicable to Other Posted Material, as set forth at <a href="http://nrs.harvard.edu/urn-3:HUL.InstRepos:dash.current.terms-of-use#LAA">http://nrs.harvard.edu/urn-3:HUL.InstRepos:dash.current.terms-of-use#LAA</a></td>
</tr>
</tbody>
</table>
Control of HIV-1C Replication Both In Vivo and In Vitro

A dissertation presented

by

Wen Xie

to

The Committee on Higher Degrees in Biological Sciences in Public Health

in Partial Fulfillment of the Requirements

for the degree of

Doctor of Philosophy

in the subject of

Biological Sciences in Public Health

Harvard University
Cambridge, Massachusetts
Aug, 2014
Control of HIV-1C replication both in vivo and in vitro

Abstract

In this study, we established an HIV latency model using pluripotent stem cells *in vitro*. We found that a HIV accessory protein *nef* is responsible for full-length HIV silencing in human pluripotent stem cells. This effect is mediated by host factors NMT1, NMT2, Hck, PAK2, ACOT8. We further suggested that deprivation of host factors can only partially reactivate the virus.

We performed a genome-wide association study (GWAS) on Botswana treatment-naïve AIDS patients infected by HIV-1C. By applying a statistical method using functional principal component analysis to approximate the underlying trajectories of longitudinal CD4 and VL, we showed three SNPs mapped to genes of HCG22, ZBTB7C and CCNG1 are significantly associated with AIDS disease progression.

Our study provided insight in HIV latency mechanisms *in vitro* and also suggested new mechanisms of host controlling viral replication in African genetic backgrounds. Our result is important for both understanding the pathogenesis of HIV as well as treatment prioritization for AIDS intervention.
Preface

While this thesis will present the results of my academic work in graduate school, I first want to step back and reflect on those who have given me the most crucial support as I pursued my dream. I would like to thank Dr. Max Essex, my thesis advisor, for his support of all aspects of my PhD. His accomplishments have given me the strongest encouragement every day to become a scientist who, like him, significantly impacts both the basic science of HIV and global AIDS public health interventions. I also want to thank Dr. Tun-Hou Lee, my PhD supervisor. Since my first day at Harvard, he has mentored me in every aspect of my study. I want to acknowledge Dr. George Daley, my thesis co-supervisor. I will never regret joining his lab, as it is the place which truly opened my mind in science and showed me the joy of being a scientist. I want to thank Dr. Tianxi Cai, with whom I closely collaborated on my GWAS project. I feel privileged to have been their student!

I also want to thank many people who provided me advice or collaborated with me during my PhD program: Drs. Richard Marlink, Marianne Wessling-Resnick, Marc Lipsitch, Laurie Glimcher, Barry Bloom, Oliver Hoffmann and Rory Kirchner.

I also gratefully acknowledge my Dissertation Defense Committee, Drs. George Seage, Stephen J. O’Brien and Monty Montano, for offering me their expertise and guidance. And finally, I thank my family members and friends for their great support for me during my PhD program—especially my mother, Lijian Wang, my father, Jingyan Xie, and my fiancé, Yusheng Qi.
TABLE OF CONTENTS

Control of HIV-1C Replication Both In Vivo and In Vitro ........................................ iii
Preface ......................................................................................................................... iv

LIST OF FIGURES WITH CAPTIONS ........................................................................... ix
LIST OF TABLES WITH CAPTIONS ............................................................................... xi

Summary ..................................................................................................................... 1

Chapter I  Introduction ............................................................................................... 4

I.1 Overview of AIDS pandemic ................................................................................ 5
  a. Discovery of HIV and early studies on host-viral interaction ......................... 5
  b. Current AIDS pandemic ................................................................................... 5

I.2 Current research on HIV/AIDS intervention ...................................................... 7
  a. HIV transmission prevention ............................................................................. 8
  b. Before HAART .................................................................................................... 10
  c. Highly Active Antiretroviral therapy (HAART) ................................................ 12

I.3 Challenge of curing AIDS ...................................................................................... 13
  a. Proviral reservoir ............................................................................................... 13
  b. Lack of effective biomarker for host-viral interaction .................................... 14

I.4 Host restriction to HIV in vitro ............................................................................ 15
  a. Host restriction at pre-integration level ............................................................. 15
  b. Host restriction at post-integration level .......................................................... 15

I.5 Mechanism of HIV transcription restriction by viral genes interacting with host genes ................................................................................................................. 18

I.6 HIV host gene restriction in vivo--GWAS studies of European males infected with subtype-B ................................................................................................................. 18

Chapter II  A novel HIV model of human pluripotent stem cells identified Nef gene as silencing factor for provirus ................................................................. 20

II.1 Abstract ................................................................................................................. 21

II.2 Introduction ............................................................................................................ 21

II.3 Material and Methods ........................................................................................ 23
  a. Cell culture ......................................................................................................... 23
  b. Virus preparation: ............................................................................................ 24
c. SYBR-Green real-time RT–PCR .................................................. 25

d. Reprogramming assays .............................................................. 25

e. Immunostaining ........................................................................ 26

f. Integration site sequencing ........................................................ 26

g. Nef cloning ................................................................................ 27

h. Western blot .............................................................................. 28

i. ChIP-PCR: .................................................................................. 28

j. EB differentiation: ....................................................................... 28

k. Myeloid and Erythroid lineage differentiation ............................ 29

l. Fibroblast differentiation .............................................................. 29

II.4 Result ......................................................................................... 30

a. HIV-Δenv is silenced at transcriptional level in human pluripotent stem cells.. 30

b. Evaluating the silencing kinetics of envelope interrupted Full-length HIV(HIV-Δenv) in human pluripotent stem cells .......................................................... 32

c. Knocking out different HIV accessory genes affect the extent of viral silencing in human pluripotent stem cells ........................................................................ 34

d. Interrupting Nef at genomic level abolishes the transcriptional silencing and cytotoxicity of HIV .................................................................................... 36

e. Establishing single integrated HIV pluripotent stem cells via somatic reprogramming ............................................................................................... 36

f. ChIP-PCR revealed the epigenetic signature of silenced HIV proviral genome in pluripotent stem cells ................................................................. 38

g. Nef suppressing HIV expression is not through genome-wide integration sites selection .................................................................................... 40

h. Over-expressing Nef imposes silencing to HIV proviral gene in human pluripotent stem cells .................................................................................... 42

i. HIV expression and Nef-induced proviral silencing is linked to cell pluripotency ............................................................................................... 45

j. Identifying host factors that interact with Nef .................................... 48

II.5 Discussion ................................................................................. 50

a. Model of HIV latency in human pluripotent stem cells .................. 50

b. Function of Nef protein .................................................................. 52
c. HIV cytotoxicity in human pluripotent stem cells ........................................... 54

d. HIV long-term latency in human pluripotent stem cells .................................. 55

e. HIV latency in other cell lineages .................................................................. 56

Chapter III  A Genome-wide association study reveals multiple new genes for host control of HIV-1 Subtype-C AIDS disease progression in an African population ... 58

III.1 Abstract ........................................................................................................... 59

III.2 Introduction ................................................................................................... 60

III.3 Material and Methods .................................................................................. 62

a. Cohorts and patients eligibility ...................................................................... 62

b. Genotyping process and quality control ......................................................... 65

c. EIGENSTRAT method to perform population stratification ......................... 68

d. Functional principal component analysis steps .............................................. 68

e. Linear regression model to characterize CD4 and VL trajectories............... 69

f. Testing ............................................................................................................ 70

III.4 Results .......................................................................................................... 71

a. Population characteristics and genetic stratification ...................................... 71

b. Phenotype definition using Functional Principal Component Analysis (fPCA) to reconstitute of CD4 and VL trajectory .......................................................... 72

c. GWAS using BHP011 as primary cohort and BHP010 as replication cohort ... 74

d. Adjusting by age and sex ............................................................................ 78

e. Using the dominant model ........................................................................... 79

f. Linear model for phenotype approximation .................................................. 80

g. Phenotype of patients with and without minor alleles ................................ 83

III.5 Discussion ..................................................................................................... 86

a. HCG22 involved in HIV restriction in African population but not others....... 86

b. P21/WAF1 pathway may be highly involved in HIV host restriction in the African population ............................................................................................................. 87

c. Multiple genes below significance level may still be associated with host control of HIV/AIDS progression ................................................................. 88

d. PCDH11X on X chromosome may be associated with HIV restriction in vivo 90

e. New SNPs with small p-value after adjusting for age and sex ...................... 91
f. Using fPCA allows better simulation of small variations in longitudinal AIDS clinical data and identifies new SNPs ................................................. 93

Chapter IV Discussion ........................................................................................................... 96

IV.1 Using human pluripotent stem cell as an in vitro model of HIV latency ........ 97
IV.2 Implication of host restriction to HIV replication by transcription factors .... 101
IV.3 The collaboration of host genes and viral genes in controlling HIV replication and affecting CD4 and viral RNA ........................................................................ 106
IV.4 Using genetic markers as a public health indication for AIDS treatment prioritization .................................................................................. 109

Bibliography .......................................................................................................................... 113
LIST OF FIGURES WITH CAPTIONS

Figure 1. World Map of Prevalence of HIV Infection 2011 ......................................................... 7
Figure 2. Four major areas of HIV/AIDS research ................................................................. 8
Figure 3. HIV-1 life cycle. .................................................................................................. 10
Figure 4. A typical natural course of HIV infection. ............................................................. 12
Figure 5. Retroviral constructs used in the experiments ....................................................... 30
Figure 6. HIV-Δenv is transcriptionally silenced and cytotoxic to human pluripotent stem cells. .......................................................................................................................... 32
Figure 7. HIV-Δenvis transcriptionally silenced in human pluripotent stem cells while HIV-Δenv/Δef is not. .................................................................................................................. 34
Figure 8. Deleting HIV accessory proteins reactivates HIV by cell sorting ......................... 35
Figure 9. HIV-Δenv is silenced upon somatic reprogramming while HIV-Δenv/Δef is not ........................................................................................................................................... 38
Figure 10. Epigenetic signature difference between HIV-Δenv and HIV-Δenv/Δef proviral LTR in human pluripotent stem cells ................................................................................. 39
Figure 11. High-throughput integration sites sequencing strategy ........................................ 41
Figure 12. HIV-Δenv and HIV-Δenv/Δef are integrated into similar regions of human genome ........................................................................................................................................ 42
Figure 13. Supplementing nef into HIV-Δenv/Δef infection can silence the virus in human pluripotent stem cells ........................................................................................................... 44
Figure 14. HIV-Δenv latency is linked to cell pluripotency. ..................................................... 46
Figure 15. Induce differentiation of iPS colonies bearing HIV-Δenv latent proviral gene to blood lineage ................................................................................................................... 47
Figure 16. Inducing differentiation into CD34+ hematopoietic progenitor cells by EB differentiation .................................................................................................................................. 48
Figure 17. Strategic description of shRNA assay for Nef host partners .................................... 49
Figure 18. Knocking down Nef partners in pluripotent stem cells impairs Nef role of silencing HIV proviral genome in trans ....................................................................................... 50
Figure 19. Suggested model of HIV latency in human pluripotent stem cells ....................... 52
Figure 20. Population and sampling comparison between BHP010 and BHP011 ......... 63
Figure 21. Functional principle component analysis to estimate Eigenfunctions of CD4 and VL trajectories ................................................................................................................. 73
Figure 22. Extremity of intercept and slope values using linear model ......................... 74
Figure 23. Average of trajectory of significant SNPs that are eliminated due to opposite direction in primary and secondary cohorts generated by fPCA model. ....................... 76
Figure 24. GWAS results at 852889 SNPs with using combined VL and CD4 fPCA scores as phenotype under a dominant model ........................................................................ 76
Figure 25: The QQ plot and Manhattan plot after adjustment by age and sex if including sexual chromosome in association analysis (dominant model) for fPCA. ................. 79
Figure 26. QQ plots and Manhattan plots using non-dominant model before stratifying by age and sex ........................................................................................................... 80
Figure 27. QQ plots and Manhattan plots using different phenotype simulation models before adjusting for age and sex (excluding sexual chromosome) .............................................. 82
Figure 28. Average trajectories of VL and CD4 are highly corrected with the top significant SNP using dominant model .......................................................................................... 83
Figure 29. Coefficients measuring the trajectory of VL and CD4 are highly corrected with the top significant SNP using dominant model ........................................................................ 84
Figure 30. Coefficients measuring the trajectory of VL and CD4 are highly corrected with the top significant SNP using dominant model ........................................................................ 84
Figure 31. Phenotype presentation of patients with different alleles of top SNPs identified by linear mixed. ...................................................................................................................... 85
Figure 32. The QQ plot and Manhattan plot after adjusted by age and sex if including sex chromosome in association analysis (dominant model) for fPCA ................................................. 91
Figure 33. Strategy of using iPS cells as patient specific HIV model for both drug screen and basic research .................................................................................................................. 101
Figure 34. In vitro and In vivo combined finding of HIV viral factors collaborating with host factors to regulate HIV replication and host cell cycle. .......................................................... 105
Figure 35. Epidemiology modeling of HIV-1 infection ........................................................................................................... 107
Figure 36 Association between proviral load and plasma viral load in HIV-1C infection ............................................................................................................................................ 108
Figure 37. Extreme phenotypes of CD4 decline and genotype of the three significant SNPs ........................................................................................................................................ 112
LIST OF TABLES WITH CAPTIONS

Table 1 Information of cohorts ................................................................. 64
Table 2 Average baseline and median CD4 and VL (BHP010) ......................... 64
Table 3 Average baseline and median CD4 and VL (BHP011) .......................... 64
Table 4 Genotyping platforms used for BHP010 and BHP011 .......................... 65
Table 5 Sample and SNP filtering for this study ........................................... 67
Table 6 Significant SNPs that have opposite effects on BHP010 and BHP011 ...... 77
Table 7 SNPs significant* after replication using fPCA model (excluding X and Y chromosomes) .................................................................................. 77
Table 8 Genome-wide significant loci that are significant in both cohorts by fPCA and Linear mixed phenotype simulation .......................................................... 78
Table 9 SNPs significant in one cohort that fail to replicate as significant in second cohort ........................................................................................................ 78
Table 10 SNPs significant* after replication using Linear mixed model (excluding X and Y chromosomes) ........................................................................... 82
Table 11 The promising SNPs below genomic significant level .......................... 89
Table 12 New SNPs with small p-values after adjusting for age and sex (using fPCA, dominant model) ................................................................................. 93
Table 13 P-value of SNPs suggested before by other GWAS ............................ 95
Summary

After 33 years since the recognition of acquired immunodeficiency syndrome (AIDS) in 1981, there are still no effective therapies to cure AIDS. HAART (highly active antiretroviral therapy) interruption results in a recurrence of viremia requiring life-long adherence to this therapy. Interestingly, without antiviral drug treatment, AIDS patients also show variation in the disease progression to AIDS and clinical outcome after infection with Human immunodeficiency virus (HIV) (Fellay J, 2007) and a portion of patients are able to maintain low viremia for a long time without treatment. These phenomena suggest that both host factors and viral factors play important roles in controlling the viremia which leads to long-term HIV latency. Understanding how HIV establishes its latency and why people display different levels of HIV viremia control is important for both developing an effective treatment to eradicate the HIV long-term reservoir, as well as prioritizing patients predisposed to early development of AIDS for initiation of treatment.

Due to the lack of an efficient HIV latency cell model in vitro, we established an HIV latency model using pluripotent stem cells. We found that full-length HIV defective of envelope gene (HIV-Δenv) is silenced in human pluripotent stem cells but not other differentiated lineages (fibroblast, blood). By knocking out a HIV accessory protein Nef or supplementing Nef into Nef defective HIV infection (HIV-Δenv/Δnef), we demonstrated that Nef is responsible for silencing the virus at an epigenetically transcriptional level by recruiting histone modification players to perform methylation at H3K9 and
demethylation at H3K4 on the HIV genome. This effect is independent of viral integration site selection. Host factors (NMT1, NMT2, Hck, PAK2, ACOT8) that have been reported to interact with Nef are also highly involved in establishing the latency. We further suggested that Nef leads to long-term HIV silencing, and deprivation of host factors can only partially reactivate the virus. To further test the hypothesis that Nef and its associated host factors are responsible for HIV latency in vivo, we need to replicate the experiment in natural HIV reservoirs such as hematopoietic stem cells (HSCs).

To study HIV latency and host genetic restriction to viral replication in vivo, we performed a genome-wide association study (GWAS) on Botswana treatment-naïve AIDS patients. We applied a statistical method using functional principal component analysis to approximate the underlying trajectories of longitudinal CD4 and VL. By conducting an initial GWAS analysis on a Tswana HIV-1-Subtype-C cohort sized 317 and replicating it in an independent Tswana HIV-1C cohort sized 239, we showed three SNPs mapped to genes of HCG22, ZBTB7C and CCNG1 are significantly associated with AIDS disease progression. Adjusting for age and sex does not change our result. Our result suggested that HLA mediated HIV restriction plays an important role in viral control in African genetic backgrounds, and we revealed that the p53/p21 pathway may be an important element in viral replication restriction in vivo. Exon sequencing and in vitro knockdown assays may be needed to further analyze the causal relationship.

In summary, by establishing a new in vitro model of HIV latency using human pluripotent stem cells, we demonstrated that HIV-nef, together with both pluripotency
associated cellular factors and factors that interact with Nef, may contribute to establishing the HIV epigenetic silencing in vitro. By performing a GWAS on an HIV-1C African population, we further identified three host genes that may contribute to establishing long-term HIV host control in treatment naïve patients. Our in vitro study raised important host and viral targets that can be drug targets to reactivate the HIV latent reservoir for patients on HAART, while our GWAS study provided insight in mechanisms of HIV host restriction in vivo, which is important for both HIV vaccine design and treatment prioritization.
Chapter I  Introduction
I.1 Overview of AIDS pandemic

a. Discovery of HIV and early studies on host-viral interaction

The year of 2011 marked the thirtieth year of the AIDS pandemic. In 1981, the world first recognized the existence of AIDS when most patients with AIDS died within 2 years before the discovery that AIDS was caused by a new immunodeficiency virus known as HIV-1 (Rothenberg et al., 1987; Pomerantz and Horn, 2003). The first big breakthrough in AIDS happened in 1983 when scientists first identified viruses associated with AIDS (Barre-Sinoussi et al., 1983) and subsequently discovered AIDS was caused by a retrovirus HIV-1 (Gallo et al., 1984; Popovic et al., 1984). The methodology used to link animal and human retroviruses to immunosuppressive disease led to the development and implementation of a blood-screening antibody test indicating the presence or absence of HIV-1 (Kitchen et al., 1984; Lee et al., 1984a; Sarngadharan et al., 1984; Allan et al., 1985b; Silberner, 1985; Archibald et al., 1986). The understanding of the virus rapidly progressed as scientists decoded the functional units of the HIV genome (Lee et al., 1984b; Wain-Hobson et al., 1985; Wong-Staal et al., 1985; Haseltine and Patarca, 1986; Lee et al., 1986). The discovery of specific host immune response to a particular region on viral genome (gp120/160) was another big step towards understanding of host-viral interaction that later on greatly facilitated the development of anti-viral therapy and vaccines (Allan et al., 1985a; Barin et al., 1985).

b. Current AIDS pandemic
In 2012, UNAIDS estimated that 35.3 (32.2–38.8) million people were living with HIV. There were 2.3 (1.9–2.7) million new HIV infections globally, showing a 33% decline in the number of new infections from 3.4 (3.1–3.7) million in 2001 (UNAIDS, 2013b). At the same time the number of AIDS deaths is also declining in that there were 1.6 (1.4–1.9) million AIDS deaths in 2012 while there were 2.3 (2.1–2.6) million in 2005 (UNAIDS, 2013b).

However, diversity of the AIDS pandemic among countries continue to persist. Sub-Saharan African counties are still the most affected, followed by Eastern Europe and the Caribbean (Figure 1). The AIDS pandemic is especially high in southern Africa with an overall prevalence among adults of up to 31% in Swaziland, 25% in Botswana, and 17% in South Africa (UNAIDS, 2013b). Despite the fact that global investment in the AIDS response jumped from US$3.8 billion in 2002 to US$18.9 billion in 2012 (Zablotska, 2013), countries with less sustained investments and less effective systems to deliver antiretrovirals to women and infant feeding-based prevention programmes still experience an increase in AIDS infection (Kuehn, 2006), with 50% of eligible patients not receiving HAART (UNAIDS, 2012). It is worth noticing that there are two common features of those countries: first, the HIV-1 subtype C infection has become hyperendemic in those countries rather than subtype B; second, the average income of those countries are low. These two facts pointed to the necessity to prioritize treatment resources to patients more susceptible to AIDS development and transmission based on the knowledge of host genetic restriction to HIV, especially the African population infected by HIV subtype C.
I.2 Current research on HIV/AIDS intervention

Research of HIV/AIDS intervention has focused on four main steps: prevention of transmission and diagnostics, disease progression monitoring before HAART, effectiveness of HAART treatment, and opportunistic outcomes (Figure 2).
Figure 2 Four major areas of HIV/AIDS research

a. HIV transmission prevention

In the area of acquisition/transmission prevention, the research topics include developing an effective HIV vaccine and effective methods to prevent viral transmission including mother-to-child transmission and transmission through sexual activities. Recent studies showed that "treatment as prevention" as a new and promising direction (Cohen et al., 2011a; Padian et al., 2012; Shapiro et al., 2013).

Another area in HIV transmission prevention research is development of an HIV vaccine. Despite 30 years of effort, there is still no effective HIV-1 vaccine. The HIV-1 infection life cycle begins with host cell recognition followed by entry mediated by CD4 and CCR5 or CXCR4 on the host cell surface with the HIV-1 envelope on HIV virion(Figure
3). Studies suggested that during acute infection with HIV-1, serum viral loads peak approximately 3 weeks after transmission and then fall to a set point determined by the host immune system (Daar et al., 1991; Cohen et al., 2011b). Importantly, people have different viral load set point levels suggesting host genes may play a role in viral load control. Previous GWAS studies as well as other genetic research suggested that in rare individuals who carry specific HLA alleles such as HLA-B*5701, T cell responses can even reduce viremia to undetectable levels for prolonged periods of time (Migueles et al., 2000). This result raised the importance of T cell responses for partial control of viral replication in the early phases of HIV-1 infection. However, HLA-B*5701 protective allele can only explain less than 10% of the variation (Fellay et al., 2007), indicating other mechanisms involved in natural host control of HIV. Previously it has also been shown that patients with a homozygous defective CCR5 gene (CCRΔ32) are resistant to HIV infection (Huang et al., 1996; O'Brien et al., 1997; Quillent et al., 1998). Understanding the host genetic restriction to viral replication will be essential for developing a successful HIV vaccine.
Figure 3HIV-1 life cycle.
The virus is recognized by host CD4 and CCR5/CXCR4 on the surface of cells and after binding and fusing to the host cell, the virion core enters the cell. Reverse transcriptase performs reverse transcription that synthesizes the proviral double-stranded DNA, which will be transported to the nucleus and integrated into the host genome. The viral genome is then transcribed and the virus proteins are translated using host ribosomes. New virions are created by assembly and budding through the infected cell membrane. The figure is adapted from reference (Pomerantz and Horn, 2003)

b. Before HAART
The most heterogeneity in disease progression was displayed before patients started HAART. The disease course of AIDS is shown in Figure 4. However, there is considerable heterogeneity in virus control and progression rates that cannot be fully explained by environmental or viral factors (An and Winkler, 2010). These facts lead to the hypothesis that the heterogeneity of host genetics are responsible for explaining some variation in AIDS disease progression. The extreme phenotypes that deviate from the expected response to HIV exposure can be categorized into the following three kinds (An and Winkler, 2010):

(i) exposed uninfected (EU) individuals who are resistant to HIV acquisition even after multiple, high risk exposures (O'Brien and Nelson, 2004; Shacklett, 2006)

(ii) long-term non-progressors (LTNP) who maintain stable CD4 levels and low virus load for ten or more years (Kumar, 2013) or elite controllers (EC) (1% of HIV-infected patients), who are infected with replication-competent virus but control HIV replication to <50 copies/ml (Walker, 2007; Miura et al., 2008).

(iii) fast progressors who cannot control viremia and develop AIDS within three years of infection (O'Brien and Nelson, 2004)
Figure 4. A typical natural course of HIV infection.
Acute phase lasts for 6–12 weeks with flu-like symptoms, peak virus load and drop in CD4+ T cells. Chronic asymptomatic phase lasting on average 7–10 years; following the acute phase virus replication reaches a steady level known as the ‘set point’. This is followed by AIDS onset associated with increasing virus replication and declining CD4 cell counts to <200/mm³; time to AIDS onset varies between individuals from as little as 2 years to more than 15 years after seroconversion. Figure modified from references (Fauci et al., 1996; An and Winkler, 2010)

**c. Highly Active Antiretroviral therapy (HAART)**

After the patient CD4 drops to a certain threshold (typically 200-350/µl), they are eligible for HAART treatment(Dybul et al., 2002). Despite the development of multidrug resistance (Bangsberg et al., 2000; Hirsch et al., 2000), patients cannot be cured by HAART due to the viral latency. “Retroviral latency” is defined as integrated replication competent provirus with no active transcription(Pomerantz, 2002a). Multiple host genetic factors are involved in silencing the proviral genome reviewed in I.4b. However, upon suspension of HAART, latent proviral genes get reactivated and the mechanism of viral reactivation and the variation of the timing of viral reactivation is largely unknown. Studies of the mechanism of host genetic control of the HIV proviral genome using a
good *in vitro* cell model is needed to understand the HIV reactivation process and to understand possible mechanisms for the eradication of the proviral reservoir.

I.3 Challenge of curing AIDS
   a. Proviral reservoir

As discussed above, HAART can substantially reduce the viral load within weeks of initiation. However, upon stopping HAART, the viral remission can last for only a short period and virus will eventually reactivate through unknown mechanisms (Saez-Cirion et al., 2013). Even after early initiation of HAART in adults, only 5-15% could potentially lead to prolonged remission while the rest experience rapid viral load rebound after stopping HAART (Saez-Cirion et al., 2013). In the case of the famous "Mississippi baby", an HIV-infected baby who received HAART within 30 h of delivery, stayed on HAART for 18 months and seemed to be cured (Persaud et al., 2013; Shiau and Kuhn, 2014). However after 27 months off HAART, the viral load rebounded (NIAID, 2014). This fact that HIV can remain hidden for years and eventually rebound implicated that host control of HIV provirus is limited even at very early initiation of HAART and remission of viral replication.

Latent replication-competent HIV-1 provirus exists mainly in resting memory (CD45RO) CD4+ T lymphocytes (Pomerantz, 2002b). Yet it has been suggested that other cell types such as monocytes and resting naive (CD45RA) CD4+ T lymphocytes can also serve as proviral reservoirs in patients receiving virally suppressive HAART (further reviewed in II.2) (Sonza et al., 2001). Naive CD4+ T lymphocytes are rarely directly infected,
therefore the naive CD4+T lymphocytes harboring HIV proviral genes were reported to be generated by means of reversion from memory CD4+ T cells (Pierson et al., 2000).

It has been suggested that the proviral reservoir can also serving as a source for documentation of viral mutations. Viruses isolated from resting CD4+ T lymphocytes from the peripheral blood of patients with undetectable levels of viral RNA in plasma may harbor resistance mutations in the reverse-transcriptase(RT) and protease (PR) genes of HIV-1(Wong et al., 1997).

Outside the peripheral blood, lymphoid tissues can also bear latent HIV-1 proviral genes. It has been suggested that gut-associated lymphoid tissue also appears to be an important site for early HIV-1 replication in these mucosal lymphoid tissues, thus serving as an important reservoir of HIV-1 outside of the peripheral blood for HIV-1(Veazey et al., 1998). In addition to lymphoid tissues, tubular epithelial cells of the kidney has also been reported as another potential reservoir for HIV-1(Winston et al., 2001).

b. Lack of effective biomarker for host-viral interaction

Another challenge in curing HIV is the lack of a reliable biomarker for the total body burden of replication-competent HIV that persists in people on HAART. As virus may develop antidrug mutations, without an effective biomarker, it is hard to monitor the burden and reactivation of provirus from reservoirs in the patients after long term
antiretroviral treatment. The "Mississippi baby" (Persaud D, 2000; Persaud et al., 2013) had a viral rebound after 27 months of HAART suspension, and two patients infected with HIV in Boston, MA, USA who received a stem-cell transplant for lymphoma (Henrich et al., 2013) and had virus rebound 12 and 32 weeks after stopping HAART. In all these three cases, the currently available assays failed to detect evidence of HIV return. These tragedies indicated the importance of identifying a robust biomarker that will predict time to viral rebound once HAART is stopped.

I.4 Host restriction to HIV in vitro

a. Host restriction at pre-integration level

Many detailed molecular mechanisms of retrovirus or lentivirus latency have reported that both the host factors and viral factors are highly involved. Reported host pre-integration restriction factors include APOBEC3G that inhibits HIV-Δvif by digesting the reverse transcriptional machinery (Huthoff H, 2008), Trim5α that interacts with capsids to prohibit early entry viral-life cycle (Stremlau M, 2004; Huthoff H, 2008) and p21 which interacts with integrase (Zhang J, 2007). Mutations in the cellular proteins that facilitate viral entry such as CCR5-Δ32 (Loetscher P, 2000) and integrations such as HMG1 (Farnet CM, 1997), LEDGF (Llano M, 2006; Shun MC, 2007b), emerin (Jacque JM, 2006; Shun MC, 2007a), microtubules (McDonald D, 2002) and PIC can also lead to HIV latency in both field research and cell models.

b. Host restriction at post-integration level
For post-integration mechanisms, host microRNAs (miR-28, miR-125b, miR-150, miR-223 and miR-382) (Omoto S, 2004; Han Y, 2007a; Huang J, 2007) were shown to affect HIV mRNA stabilities. However, most mechanisms of post-integration latency operate at the transcriptional level (Mayte Coiras, 2009). Proviral integration site and orientation could both affect the expression of proviral genes (Mayte Coiras, 2009; Shan L, 2011), though HIV-1 tends to insert in host transcriptionally active sites (Schröder AR, 2002) and most inducible HIV-1 proviruses can be found in intronic regions of highly expressed host genes (Han Y, 2004; Lewinski MK, 2009). The orientation and position that the HIV viruses integrate into the host genome affects transcription. Insertion sites that are divergent, convergent or parallel relative to a proximate host promoter could result in transcriptional interference (Callen BP, 2004; Crampton N, 2006; Han Y, 2008; Lenasi T, 2008; Lewinski MK, 2009). These findings indicate that chromosomal environment at the site of integration and the accessibility of host and viral transcriptional factors have substantial influence on viral latency.

Epigenetic silencing is a major mechanism of retroviral transcriptional silencing and histone deacetylation and histone methylation are both highly involved in the HIV model of control in Hela cells (du Chéné I, 2007; Tyagi M, 2007), resting T cells (Archin NM, 2009), Jurkat cells (Tyagi M, 2007; Pearson R, 2008; Friedman J, 2011), the murine leukemia virus (MLV) model of mouse embryonic stem cells (mES) (Wolf D, 2009) and the L1 retrotransposon in human embryonic carcinoma cells (Garcia-Perez JL, 2010). Late SV40 factor (Ylisastigui L, 2004), Yin-yang1 (YY1) (Gordon S, 2006), C-promoter binding factor 1 (CBF-1) (Tyagi M, 2007) and chicken ovalbumin upstream promoter-
transcription factor-interacting protein 2 (BCL-11B) (Marban C, 2007) can bind to HIV LTR and recruit HDACs which will suppress transcription, while Tat can recruit HAT proteins such as CREB binding protein (CBP), CBP-associated factor (PCAF) and human general control of amino acid synthesis protein 5 (hGCN5) to hyper-aceytelate 5’ LTR and lead to viral expression (du Chéné I, 2007). Histone-lysine N-methyltransferase suppression of variegation 3-9 homologue 1 (SUV39H1) could recruit heterochromatin protein 1 homologue-γ (HP1γ) and was shown to correlate with H3K9me3 and suppression of viral transcription (Grewal SI, 2003). On the other hand, host proteins that participate in the assembling of transcriptional machinery could positively regulate viral transcription, such as nuclear factor-κB (NF-κB) (Nabel G, 1987), specificity protein 1 (SP1) (Jones KA, 1986), nuclear factor of activated T cells (NFAT) (Corthésy B, 1994) and positive transcription elongation factor b (PTEFb) (Garber ME, 1998; Zhou M, 2000). Recently, S. Goff’s group identified Trim28 to be recruited to the MLV proline tRNA primer binding site (P-PBS) by adaptor protein zinc finger protein 809 (ZFP809) with suppression of MLV expression in mES via further recruiting HP-1γ to perform H3K9me2 (Teich NM, 1977; Yamauchi M, 1995; Wolf D, 2007; Wolf D, 2008; Wolf D, 2009). Upon mES differentiation, MLV expression dramatically increased as Trim28 level is positively correlated with cell pluripotency (Wolf D, 2007). G. Daley and Y. Loh’s labs further showed Trim24 interacts with Trim28, which is crucial for viral silencing (data not published). This model raises the importance of Trim family proteins in silencing the retroviruses in embryonic stem cells via binding to PBS region.
I.5 Mechanism of HIV transcription restriction by viral genes interacting with host genes

Besides PBS and LTR as discussed above, some viral factors such as Nef (Venkatesan, 1988) and mutated Tat (Tyagi M, 2010) are also responsible for suppressing viral genome transcription. Nef was previously reported to be important for HIV pathogenesis in inducing apoptosis and negatively regulating transcription by binding to the LTR region in the SW480 cell line (Zauli G, 1999) while a nef mutant could maintain enhanced replication in T lymphocytes (Ahmad N, 1988). Yet these phenomena are controversial as other reports showed the opposite that deleting or inhibiting Nef reduces viral replication but does not reduce the cytotoxicity (Terwilliger E, 1986; Chowers MY, 1994; Spina CA, 1994; Pizzato M, 2008; Emert-Sedlak L, 2009; Ptak RG, 2010). Nef was also shown to facilitate the early step of viral replication by reverse transcription of the viral RNA genome (Aiken C, 1995; Chowers MY, 1995; Schwartz O, 1995). Field work showed elite controllers are infected with HIV-1 exhibiting defective nef alleles (Alexander L, 2000; Blankson JN, 2007; Zaunders J, 2011) and deletion of the SIV nef gene reduced viral replication in vivo and delayed onset of AIDS-like disease in macaques (Kestler HW 3rd, 1991), possibly through abolishing Nef's role of down-regulating cell surface MHC1, CD4, CD8, and CD28 (Kestler et al., 1991; Jordan A, 2001; Kim YK, 2006; Blankson JN, 2007).

I.6 HIV host gene restriction in vivo--GWAS studies of European males infected with subtype-B
Host cellular factors that contribute to post-integration latency are counted for 64-83% of the total variation in virus production among individuals (Ciuffi A, 2004), indicating the importance of genome wide association studies (GWAS) to identify the host factors that control viral replication and viremia. Previous GWAS for disease progression rate showed that most hits are on HLA-B*5701 and HLA-C (Migueles SA, 2000; Fellay J, 2007; Dalmasso C, 2008; van Manen D, 2009), raising a possible mechanism of certain HLA alleles’ immunity to HIV nef which selectively down-regulates the expression of HLV-A and –B but not –C on infected cells (Cohen GB, 1999). Other hits include ring finger protein 39 (RFN39) and zinc ribbon domain-containing 1 (ZNRD1) which encodes an RNA PolIII subunit (Fellay J, 2007). Despite the above results, other meta-analysis using permutation procedures identified genes such as Notch4, C6orf12 and Trim10 (Fellay J, 2009) which were also associated with viral set point.
Chapter II  A novel HIV model of human pluripotent stem cells identified Nef gene as silencing factor for provirus
II.1 Abstract

The failure of HAART on curing AIDS highlights the existence of HIV reservoir and the importance of learning host restriction to viral replication. The epigenetic restriction to HIV genome can lead to HIV latency and its mechanism remains to be studied. In this study, we used a unique model of human pluripotent stem cells to model HIV latency and reactivation, and its relationship with cell types. We found that envelope-defective full HIV genome can be silenced completely in human pluripotent stem cell. An HIV accessory protein—Nef, is able to silence the HIV in human pluripotent stem cells via interacting with host factors of NMT1, NMT2, ACOT8, Hck and PAK2. Over-expression Nef in human pluripotent cells infected with nef-defective HIV can silence the expressing proviral gene. The effect of Nef silencing HIV is subjective to cell pluripotency changes.

II.2 Introduction

Since 1995, highly active anti-retroviral therapy (HAART) has been widely adopted in the world. However, AIDS patients on HAART, despite maintaining low HIV viral load, cannot be cured due to HIV latent reservoirs (Gulick et al., 1997; Hammer et al., 1997; Davey et al., 1999; Imamichi et al., 2001; Palmer et al., 2008; Harrison et al., 2010). The HIV latent reservoirs are primarily in CD4+ resting T cells (CD69 and CD25 negative) (Chun TW, 1995; Chun et al., 1997; Bailey et al., 2006), yet increasing evidence indicates that latent HIV is also detected in other subsets of T cells including central memory, transitional memory, effector memory, and naïve cells (Chomont et al., 2009; Wightman et al., 2010). In addition to CD4+ T cells, there are accumulating evidence...
showing macrophages (Zhang et al., 1999b; Igarashi et al., 2001), dendritic cells (McIlroy et al., 1995; Donaghy et al., 2001), monocytes (Lambotte et al., 2000; Sonza et al., 2001; Zhu et al., 2002; Ellery et al., 2007; Jaworski et al., 2007; Hasegawa et al., 2009) and hematopoietic stem cells (Carter CC, 2010) could also serve as reservoirs for HIV replication both in vivo and in vitro.

The previous study on host restriction to HIV replications in vitro are fruitful yet limited in explaining how multiple epigenetic machinery is recruited onto the HIV genome that leads to tat-functional yet silenced HIV provirus. Other mechanisms of host silencing retroviral or retroviral elements points to the association of cell pluripotency with retrovirus silencing Murine leukemia virus (MLV) in mouse embryonic stem cells (mES) (Wolf D, 2009) and the L1 retrotransposon in human embryonic carcinoma cells (Garcia-Perez JL, 2010). In mouse ES cells, Trim28 can be recruited to the MLV Proline tRNA primer binding site (P-PBS) by adaptor protein Zinc finger protein 809 (ZFP809) and suppresses the MLV expression in mES via further recruiting HP-1γ to perform H3K9me2 (Teich NM, 1977; Yamauchi M, 1995; Wolf D, 2007; Wolf D, 2008; Wolf D, 2009). Yet in human pluripotent stem cells the mechanism of retroviral or retroviral elements latency is to be determined.

The ideal ex vivo HIV latency model should be the latent T cells isolated from HIV patients. However, due to horizontal transfer of activated HIV, temporal limits of ex vivo culture, low frequency of the latently infected cells and lack of cellular surface marker to distinguish latent cells, the mechanisms discussed above were discovered using
immortalized human cell lines infected with genetically engineered HIV. The most widely used model is Hela and Jurkat cell clones infected with a vector carrying a tat or tat H13L mutants linked to GFP (Jordan A, 2001). It performs strong HIV silencing which can be reactivated by treatment with TNF-α (Kim YK, 2006; Tyagi M, 2007). Other cell lines used include the ACH2 T-cell line (Folks TM, 1989), U1 promonocytic cell line (Folks TM, 1987), JΔK T-cell line (Antoni BA, 1994). However, transformed cells may alter the global cellular epigenetic status and the presence of viral mutations has raised questions about the relevance of using these lines to mimic latency in vivo. Therefore, developing better in vitro cell models for HIV latency is essential to probe the mechanisms of establishment of HIV reservoir.

In this study, we established an HIV latency model using human pluripotent stem cells and found that pseudotyped HIV that is defective of nef abolished the silencing of HIV in human pluripotent stem cells. The integrated nef-defective HIV can be silenced via introducing nef back into the cells. HIV nef leading to HIV latency via epigenetic modification, which is subjective to cell pluripotency status, is possibly through interacting with host factors. This study explores the molecular basis underlying the association of cell pluripotency and retroviral latency. Our result also showed a new function of nef in recruiting host epigenetic machinery to proviral gene. This suggests that antiviral approaches that could render nef gene defective should be explored as a means to reactivate latent HIV reservoir in HAART treated AIDS patients.

II.3 Material and Methods
   a. Cell culture
293T cells were cultured in DMEM with 10% inactivated fetal serum (IFS), 50 U/ml penicillin, 50 mg/ml streptomycin, and 1 mM L-glutamine. All human embryonic stem cells and Induced Pluripotent Stem Cells (iPS) were maintained in undifferentiated state on inactivated primary mouse embryonic fibroblasts (MEFs) (Millipore) in hES medium (80% DMEM/F12, 20% KO Serum Replacement, 10 ng/ml bFGF, 1 mM L-glutamine, 100 mM nonessential amino acids, 100 mM 2-mercaptoethanol, 50 U/ml penicillin, and 50 mg/ml streptomycin). The dH1f cells were cultured in alpha-MEM containing 10% IFS, 50 U/ml penicillin, 50 mg/ml streptomycin, and 1 mM L-glutamine.

b. Virus preparation:

pNL4.3-dE-EGFP was from NIH AIDS reagent bank (catalogue no. 11100). pNL-dE-mCherry, MMLV-EGFP, MMLV-mCherry, pHRI’HIV-tat-WT and pHRI’HIV-tat-H13L plasmids were gift from Dr. Stephen Goff (Columbia University). 293T cells were plated at a density of $1.4 \times 10^7$ cells per 15-cm dish. The next day 293T cells were transfected with 6.25 µg viral vector, 5.625 µg lenti reprogramming viruses pSIN4-EF2-O2S and pSIN4-CMV-K2M (Addgene catalogue no. 21162, 21164), FUWdGW-Tomato (Addgene catalogue no. 22771), pHRI’HIV-tat-WT and pHRI’HIV-tat-H13L. For pNL-4.3-dE-EGFP and for pNL-dE-mCherry, only viral vector and VSV-G plasmid were added. Supernatant was collected 48 h and 72 h post-transfection and filtered through 45-µm pore size filters. Viral supernatants were mixed with PEG3350 solution (Sigma P3640, dissolved in PBS, 10% final concentration) for concentration and left overnight at 4 °C. Supernatants were centrifuged at 2,500 r.p.m. for 25 min, and then the pellets were
re-suspended in right amount of PBS to achieve 50x concentrated viruses. All virus
titration was performed on 293T cells.

c. SYBR-Green real-time RT–PCR

Total genomic DNA was extracted using DNeasy Blood & Tissue Kit (Qiagen). Total
RNA was extracted using TRIzol and RNeasy Mini kit coupled with an RNase-free
DNase set (Qiagen) and reverse transcribed by Superscript III (Invitrogen). The resulting
complementary DNAs were used for PCR using SYBR-Green Master PCR mix (Applied
Biosystems) in triplicates. All quantifications were normalized to an endogenous β-actin
control. The relative quantification value for each target gene compared to the calibrator
for that target is expressed as $2^{-(\text{Ct} - \text{Cc})}$ (Ct and Cc are the mean threshold cycle
differences after normalizing to β-actin). The H9 HIV was amplified by primers annealing
to LTR region. For iPS colonies, the HIV was amplified by primer annealing to the gag
gene. Single copy control of HIV-Δenv/Δef is 8E-5 cell line, which is a subclone of HIV-
infected CD4+ CEM-derived human T-cell line with single copy of viral gene insertion
and confirmed by southern blot. Single copy control of HIV-SIN was obtained by single
cell clone of 293T infected with MOI=0.1 viruses and 80% of the colonies reached the
lowest level of viral integration by gDNA qPCR (data not shown), indicating that they
are single copy according to Poisson distribution.

d. Reprogramming assays

The reprogramming assay was performed as described (Onder TT, 2012) with
modification: dH1f cells were first infected with MLV viruses at high multiplicity for
labeling and HIV viruses at low multiplicity of infection (m.o.i=0.1) to ensure majority of infected cells did not receive more than one HIV vector. Twenty-five thousands infected dH1f cells were then plated per well in 12-well plates and infected overnight with lentiviral (100–200 µl supernatant before concentration)(Yu et al., 2007). For late infection group, the infected dH1f with MLV were re-plated at density of 25,000 cells per well in an 12 well plates and infected with low MOI of HIV five days post- infection with reprogramming viruses. Six days post- infection with OSKM lentiviruses, cells were trypsinized and re-plated 1:6 onto six-well plates with MEF. Medium was changed to hES medium daily until day 28 when colonies were picked. All colonies were split for Tra-1-60 staining before expanding.

e. **Immunostaining**

The iPS colonies cells were fixed with 4% paraformaldehyde and stained with biotin-anti-Tra-1-60 (eBioscience, catalogue no. 13-8863-82, 1:250) and streptavidin horseradish peroxidase (Biolegend, catalogue no. 405210, 1:500) diluted in PBS (3%), FCS (0.3%) and Triton X-100. Staining was developed with the Vector labs DAB kit (catalogue no. SK-4100).

f. **Integration site sequencing**

Genomic DNA (2.5ug) were fragmented using DraI in 100ul reaction at 37°C overnight, followed by pheno-chloroform extraction. Digested DNA was ligated to an adaptor in Universal Genome Walker kit (BD Biosciences Clontech, Palo Alto, CA). The ligation was done in a 8ul reaction at 16°C overnight. First and second round PCR were
performed using primers specific to HIV 3'LTR (Ciuffi and Barr, 2011) and AP1/AP2 provided by Universal Genome Walker kit (BD Biosciences Clontech, Palo Alto, CA). Both PCR were performed in 25μl reaction and 1μl of PCR1 products were used for second round PCR. The PCR amplicons were purified using QIAquick PCR Purification Kit (Qiagen) before used for Illumina MiSeq. The BLAT program (University of California, Santa Cruz, Human Genome Project working draft Feb 2009; http://www.genome.ucsc.edu/) (Karolchik et al., 2009) was used to map integration sites. We include only amplicons with the three criteria (MacNeil et al., 2006) into the analysis: (i) contained the terminal 3'end of the HIV-1 LTR; (ii) matched genomic DNA within five bps of the end of the viral 5' LTR; (iii) matched a single human genetic locus with at least 95% homology across the entire sequenced region. The genes that HIV is integrated into (exon, intron) were further analyzed for their expression level in human pluripotent stem cells.

g. Nef cloning

The pSIN-EF2-Lin28A-puro (Addgene, catalogue no. 16580) is digested with EcoRI and BamHI for 3 hours at 37°C and purified using QIAquick Gel Extraction Kit (Qiagen). HIV-1 nef gene and EGFP gene were amplified from pNL4-3-dE-EGFP plasmid with primers carrying EcoRI and BamHI sites. Amplicons were subcloned into pCR2.1 using TOPO® TA Cloning® Kit for Subcloning (Invitrogen, 450641) and digested with EcoRI and BamHI. After gel purification, inserts and vectors were ligated at room temperature for 1 hour using T4 DNA Ligase (1 U/μL) (Invitrogen, 15224017). The ligated plasmids
were then amplified using One Shot® Stbl3™ Chemically Competent E. coli (Invitrogen, C737303).

**h. Western blot**

The Western blot was performed as described (Zhu et al., 2011). Primary antibodies used were anti-Nef (Millipore, MAB899) and anti-alpha tubulin (Cell Signaling, #3873) antibodies. Secondary antibodies were sheep-anti-mouse or donkey-anti-rabbit HRP-conjugated secondary antibody (GE Healthcare).

**i. ChIP-PCR:**

ChIP analyses were performed as described (Cesana et al., 2011). Briefly, chromatin was extracted from cell culture according to manufacturer’s specifications (MAGnifyChIP - Invitrogen) with the following antibodies: Anti-trimethyl-Histone H3 (Lys27) antibody (Millipore, 07-449), Anti-acetyl-Histone H4 antibody (Millipore, 06-598), Anti-trimethyl-Histone H3 (Lys4) antibody (Millipore, 07-473), Anti-trimethyl-Histone H3 (Lys9) antibody (abcam, ab8898). A standard curve was generated for each primer pair testing 4-point dilutions of input sample. Fold enrichment was quantified using qRT-PCR(QuantiTect SYBR Green - QIAGEN) and calculated as a percentage of Inputchromatin (% Inp).

**j. EB differentiation:**

The EB differentiation is performed as described(Cerdan et al., 2007). Basically the iPS cells infected with HIV were collected as large aggregates and resuspended in EB
differentiation medium composed of 80% DMEM, 20% FCS (Stem Cell Technologies 06900), 50 Pg/ml ascorbic acid, 0.2Pg/ml holotransferrin on low attachment dishes. The next day, cytokines were added: hSCF (300 ng/ml), hFlt3L (300 ng/ml), IL-3 (10 ng/ml), IL-6 (10 ng/ml), G-CSF (50 ng/ml), BMP4 (50 ng/ml). Media containing cytokines was replaced every three days for 14-16 days. Then the EB were disassociated as described (Cerdan et al., 2007).

k. Myeloid and Erythroid lineage differentiation

The disassociated EB was sorted as described (Doulatov et al., 2013). Sorted cells were resuspended in IMDM with 2% IFS at a concentration of 10K/300ul. Five thousand cells were plated with 1.5ml MethoCult™ (StemCell Technologies H4434) and CFU assay was performed according to manufacturer's protocols. After 14 days the cells were harvested and flow cytometry was performed as described (Doulatov et al., 2013).

l. Fibroblast differentiation

Pluripotent H9 or iPS cells were passaged onto plates coated with gelatin (0.2%) by CollagenaseIV. Cells were cultured in medium containing 80% KO DMEM (Invitrogen 10829-018), 20% FCS, 0.67% 2-ME, 50ug/ml Ascorbic acid, 200ug/ml h-transferrin (Sigma T0665-1G). The cells were passaged for four passages with trypsin when reaching confluence. The differentiated cell lines were confirmed by both morphology and the loss of the pluripotent marker (Tra-1-60).
II.4 Result

a. HIV-Δenv is silenced at transcriptional level in human pluripotent stem cells

Murine leukemia virus (MLV) undergoes silencing in a time dependent manner in mouse embryonic stem cells (Schlesinger and Goff, 2013), and can also be well-silenced in human pluripotent stem cells (Kamata et al., 2010). Whether full-length HIV can also be silenced in embryonic stem cells is unknown. To study the extent the HIV is silenced in human pluripotent stem cells, we used an envelope-interrupted full-length HIV-1 vector (Zhang et al., 2004) (Figure 5A), pseudotyped with VSV-G envelope to infect human embryonic H9 cells (WiCell, Madison, WI).

Figure 5. Retroviral constructs used in the experiments.
Original plasmid name was indicated in parentheses. (A) A full-length HIV backbone with envelope interrupted with EGFP (Zhang et al., 2004). (B) A full-length HIV backbone with a frame-shift mutation on envelope and interrupted Nef gene with mCherry. (C) A self-inactivating lentiviral vector with hUbC promoter driven dTomato. (D) A Moloney murine leukemia virus LTR driven mCherry vector, depleted with gag/pol and with an inactivated envelope.

To better observe the MOI threshold of viral silencing, the infection was done with MOI starting from 0.005 to 2.5 (titrated by 293T, Figure 6A) and percent of HIV positive cells expressing HIV were measured by FACS on day 12 post infection. Cell counts were also
surveyed to monitor cytotoxicity. Due to the limitation of FACS to differentiate the cytotoxic effect versus viral silencing effect, which will both show reduced percent positive cells, the viral DNA were collected on day 12 post infection and the integration number were measured by qPCR (Figure 6A). As a positive control, we used an MLV reporter vector (Figure 5D) that is known to be silenced in human pluripotent stem cells. As a negative control we used an self-inactivated lentiviral vector (HIV-SIN) with a ubiquitinated hUbC promoter that is known not be silenced in human pluripotent stem cells (Figure 5C)(Ma et al., 2003; Rompani and Cepko, 2008; Norrman et al., 2010). We found that compared to controls, HIV-Δenv is completely silenced in H9 cells at all MOI day 12 post infection (Figure 6B) with a comparable level to MLV. While all viruses are cytotoxic at high MOIs, HIV-Δenv is cytotoxic at MOI below 0.1 both upon infection and over time(Figure 6D, Figure 6G).
Figure 6. HIV-Δenv is transcriptionally silenced and cytotoxic to human pluripotent stem cells.
(A) Schematic presentation of viral infection into H9 cells. (B) Percent of positive cells gated on Tra-1-60 positive cells day 12 post infection. Dark blue: HIV-Δenv, red: HIV-Δenv/nef, green: MMLV, purple: HIV-SIN. (C) The viral integration level day 12 post infection. Single copy control of HIV-SIN was obtained by single cell clone of 293T infected with MOI=0.1 viruses and 80% of the colonies reached the lowest level of viral integration by gDNA qPCR (data not shown), indicating they are single copy according to Poisson distribution. Single copy control of HIV-Δenv/nef is 8E-5 cell line, which is a subclone of HIV-infected CD4+ CEM-derived human T-cell line with single copy of viral gene insertion and confirmed by southern blot. Color representatives are the same as A. (D) Cell count of day 2 post infection, X-axis is on a log2.5. Both HIV-Δenv/nef and HIV-Δenv showed cytotoxicity at high MOI yet only HIV-Δenv showed significant silencing at low MOI. (E) Percent of positive cells gated on Tra-1-60 positive cells infected at MOI=0.1. Color representatives are the same as A. (F) The viral integration level at MOI=0.1 infection surveyed on different days post infection. Color same as A. (G) cell count of infection at MOI=0.1; Light blue: the Mock infection. Other color representatives same as A. D and G showed that though all viruses are killing cells at high MOI, HIV-Δenv is extremely toxic to cells even at low MOI. The x-axis is shown on a log2.5 scale to reveal low MOI. Viral integration level in H9 cells negative fraction sorted on Day 12 post infection by MOI=0.1 viruses.

b. Evaluating the silencing kinetics of envelope interrupted Full-length

HIV(HIV-Δenv) in human pluripotent stem cells
To better dissect the silencing phenotype, the percent positive cells and proviral integration level were also surveyed on different days (Day 2, 5, 8 and 12) post infection for the infection of MOI=0.1 (Figure 6A). For all vectors, the proviral integration level decreases dramatically from an average high copy number on day 2 to copy numbers as low as one on day 5 (Figure 6F, Figure 7B), indicating cytotoxicity of all vectors at high MOI and the fact that H9 is more infectable than 293T. The percent of cells expressing HIV-Δenv decreased over time (Figure 6E) and correlated with the decrease of integration level over time (Figure 6F), providing little evidence of overtime silencing of HIV-Δenv vector (Figure 6E, Figure 6F, Figure 7B). Therefore the decrease in positive cells of HIV-Δenv derived vectors is likely due to cytotoxic effect.

To further confirm that the observed post-integration viral silencing is at transcriptional level rather than post transcription level, the mRNA of cells infected with MOI=0.1 were collected at different days post infection (Figure 6A) and viral transcript level was measured by quantitative reverse transcription PCR (qRT-PCR). The qRT-PCR results measuring the HIV-Δenv viral mRNA expression level at different day post infection and different MOI all correlate well with the percent positive cells (Figure 7B), indicating all silencing is exerted at transcription level.
Figure 7. HIV-Δenv is transcriptionally silenced in human pluripotent stem cells while HIV-Δenv/nef is not.

(A) Insertion and expression of viral genes extracted from H9 cells infected with different virus of increasing MOI, day 12 sampling. Orange: genomic DNA shown in copy number; Purple: cDNA shown relative to Actin. This figure showed that compared to HIV-SIN (which is not silenced) and MMLV (which is silenced), HIV-Δenv/nef shows no sign of silencing while HIV-Δenv shows some degree of silencing in H9. (B) Insertion and expression of MLV and HIV based vector proviral genes from H9 infected with MOI=0.1 viruses and sampled on different days post infection. Orange: genomic DNA shown in copy number; Purple: cDNA shown relative to Actin. This figure presents that cDNA level of MLV correlates with the positive percent of cells while gDNA doesn’t, meaning that over time the reduction in positive percent of cells is largely due to silencing of the virus. Compared to MLV, the level of HIV-Δenv cDNA correlates with gDNA and the percent of positive cells, meaning that there is no overtime silencing effect and the reduction in positive percent cells is all due to the cytotoxicity.

c. Knocking out different HIV accessory genes affect the extent of viral silencing in human pluripotent stem cells.

To study the role of HIV accessory proteins in silencing HIV in human pluripotent stem cells, we infected the H9 cells separately at MOI=0.1 with the four vectors shown in Figure 5 and two vectors shown in Figure 8A (Pearson R, 2008). We chose these vectors because the vector in Figure 5B and the two vectors in Figure 8A have multiple accessory protein deleted including nef, vif, vpr. After culturing for 12 days, we collected the cells and sort for Tra-1-60+ and viral fluorescence negative fraction. Genomic DNA is collected for the negative fraction. We found that infected with similar amount of viruses,
there are less proviral genes in the negative fraction, indicating that deletion of vif, vpr and nef results in partially active HIV expression (Figure 8). Significantly, the HIV vector which has an interrupted nef gene via mCherry and inactivated env otherwise same with HIV-Δenv (Figure 5B) is completely depleted of viral integration in the negative fraction, meaning that there is no residue HIV-Δenv/nef in the negative fraction of MOI=0.1 infection (Figure 8).

**Figure 8. Deleting HIV accessory proteins reactivates HIV by cell sorting.**
A. The constructs with HIV accessory protein deletions. Upper: HIV construct without vif, nef, vpr but has a functional tat. Lower: Tat is inactivate with a H13L mutation otherwise same as upper.  B. Schematic presentation of cell sorting strategy. C. Genome DNA qPCR of negative fraction sorted on Day12 post viral infection. Copy number was calculated based on qPCR result of single copy 8E-5 cells. As expected, mutating that would cause viral silencing, while deleting multiple HIV accessory genes result in partially proviral gene reactivation. Significantly, HIV-Δenv/nef vector has no integrated virus in the negative fraction.
d. Interrupting Nef at genomic level abolishes the transcriptional silencing and cytotoxicity of HIV

To better document expression pattern of HIV-Δenv/Δ nef in human pluripotent stem cells, we performed the experiment shown in Figure 6A with HIV-Δenv/Δ nef. Strikingly, compared to HIV-Δenv, HIV-Δenv/Δ nef has similar proviral integration level (Figure 6C), yet is highly expressed in H9 cells even at MOI lower than 0.1 and over time (Figure 6B, Figure 6E). The percent of cells expressing HIV-Δenv/Δ nef is at a comparable level to HIV-SIN (Figure 6B, Figure 6E), indicating the lack of transcriptional restriction in H9 cells to HIV-Δenv/Δ nef. While the viral expression is compromised at high MOI by comparing HIV cDNA and gDNA (Figure 7A), RT-qPCR result showed a high correlation between the mRNA and gDNA level of HIV-Δenv/Δ nef over days at MOI=0.1 (Figure 7B). Our data suggest that deletion of nef gene from HIV genome can completely reverse transcriptional silencing of provirus in human pluripotent stem cells. Additionally, deleting nef abolished the cytotoxicity of HIV-Δenv at low MOI and over time. Interestingly, we also observed that both HIV-Δenv and HIV-Δenv/Δ nef showed a saturation of integration level compared to HIV-SIN and MLV, indicating strong pre-integration restriction at high MOI to full-length HIV genome.

e. Establishing single integrated HIV pluripotent stem cells via somatic reprogramming
MLV is rapidly silenced upon somatic reprogramming (Kamata et al., 2010). To test whether reprogramming fibroblast infected with HIV-Δenv can also recapitulate the phenotype of silencing HIV genome, and to establish a colonic pluripotent stem cell model with single copy of HIV integrated, we performed reprogramming experiment. We reprogrammed HIV infected dH1f cells and generated iPS cells comparable to H9 cells in regard to cell pluripotency (Park IH, 2008). We infected human differentiated fibroblast cells dH1f (Park IH, 2008) with low MOI of MLV and the HIV vectors (Figure 5) carrying different fluorescence color, and sequentially infect the cells with high MOI of Oct4, Sox2, c-Myc and Klf4 delivered by HIV-SIN vectors (Yu et al., 2009) (Figure 9A). Before reprogramming, the dH1f cells showed positivity for both HIV and MLV. Four types (MLV positive or negative, HIV positive or negative) of colonies emerged upon day 28 of reprogramming. MLV is completely silenced in all the colonies stained positive for Tra-1-60 (Figure 9B, data not shown) which proved that MLV is well silenced by reprogramming dH1f, while HIV-SIN-hUbC is highly expressed in the colonies (Figure 9B). For HIV-Δenv and HIV-Δenv/nef reprogramming experiments, we randomly picked 50 colonies from each and screened for HIV viral positivity via genomic DNA qPCR and obtained colonies that are inserted with either HIV-Δenv or HIV-Δenv/nef alone (Figure 9D). We found that all the HIV-Δenv/nef colonies highly expressed the HIV gene except for a few that has tat-mutation in tat-exon1 (generating a premature stop codon in the middle of exon 1) bearing a silenced HIV-Δenv/nef, while all colonies silenced the MLV. However HIV-Δenv is completely silenced in the integrated colonies. Colonies that has a single infection of HIV-Δenv/nef have similar HIV expression phenotypes (data not shown). We further sequenced the proviral gene and confirmed that except for the tat
mutated colony (iPS C3), the other integrated HIV proviruses in these colonies are intact without mutations (data not shown). Our result showed that HIV-Δenv can be silenced via somatic reprogramming while HIV-Δenv/nef cannot.

Figure 9. HIV-Δenv is silenced upon somatic reprogramming while HIV-Δenv/nef is not
A. Experiment design of reprogramming. B. Tra-1-60 stained positive colonies infected with four vectors. Both the HIV-Δenv and MLV are well silenced upon reprogramming, yet HIV-Δenv/nef and HIV-SIN-hUbC are not. C. FACS of the doubly infected dH1f cells before reprogramming. dH1f cells were infected with MLV and HIV before reprogramming. These cells were subjected to infection with SIN-vector encoding OSKM genes later. D. Viral integrated iPS colonies. To confirm the presence of viral genes in the iPS colonies containing HIV gene, qPCR of HIV genes were performed.

f. ChIP-PCR revealed the epigenetic signature of silenced HIV proviral genome in pluripotent stem cells.

To investigate the mechanism of HIV proviral gene silencing in human pluripotent stem cells, we collected the chromatins of iPS cells with a single integration of silenced HIV-Δenv (iPS C15, Figure 9D) or actively expressing HIV- Δenv/nef (iPS lateinfC3, Figure
9D) and performed ChIP-PCR assay to investigate the epigenetic features on HIV proviral gene. We used primers annealing to LTR region of HIV proviral genes, as well as \textit{env} region of HIV proviral genome to detect the histone marker enrichment on HIV genome. Two human genes, GAPDH and Pax3, were used as control. We found that compared to HIV-\textDelta env, HIV-\textDelta env/\textDelta nef is approximately two-fold more enriched for H3K4me3 (Figure 10B), which are associated with active gene expression in human pluripotent stem cells (Pan et al., 2007; Garcia-Perez JL, 2010; Hawkins et al., 2010). H3K9me3 is 1.5 fold more enriched on HIV-\textDelta env compared to HIV-\textDelta env/\textDelta nef, while H3K27me3 and H4Ac are similarly enriched on HIV genome in either HIV-\textDelta env/\textDelta nef or HIV-\textDelta env colonies (Figure 10B). We obtained similar result using primers annealing to HIV \textit{env} region on proviral gene (Figure 10A) indicating the histone modification is presence throughout HIV genome rather than solely on LTR. We conclude that HIV-\textDelta env is silenced in human pluripotent stem cells via enriched H3K9me3, and lack of H3K4me3.

![Figure 10](image)

**Figure 10.** Epigenetic signature difference between HIV-\textDelta env and HIV-\textDelta env/\textDelta nef proviral LTR in human pluripotent stem cells.

ChIP-based measurement of H3K9me3, H4Ac, H3K27me3 and H3K4me3 on (A) HIV \textit{env} gene, (B) HIV LTR, (C) GAPDH, (D) Pax3. The signal was normalized to Input (Y axis shows the percentage of signal relative to input). GAPDH gene should be enriched for H3K9me3, H3K4me3 and H4Ac, but not H3K27me3 (Hawkins et al., 2010). Pax3 should be enriched for H3K27me3 and H3K9me3 but limited
enrichment for H4Ac (Hawkins et al., 2010).

g. Nef suppressing HIV expression is not through genome-wide integration sites selection

Our data suggested that HIV-Δenv is silenced in human pluripotent stem cells epigenetically. However, due to the fact that HIV-Δenv is more cytotoxic than HIV-Δenv/Δenv/nef, it is also possible that Nef induce HIV silencing indirectly via selection for proviral integration sites that have less permissive chromatin structures. To investigate whether Nef regulates HIV expression via chromatin modification indirectly by inducing cell death, we sequenced the integration sites of HIV-Δenv and HIV-Δenv/Δenv/nef MOI=0.1 infection on day 2, 5 and 12 post infection (Figure 6A) using MiSeq (Illumina) deep sequencing technology. The strategy of recovering proviral integration sites and pooling samples for deep-sequencing is showing in Figure 11.
Figure 11. High-throughput integration sites sequencing strategy.
The genomic DNA were digested using DraI enzyme overnight. The fragments were purified and ligated with an adaptor. Two rounds of PCR were conducted using one primer that anneal to the adaptor and the other one primer annealing to viral LTR sequence. The second round of PCR used primers that have a 5-bp barcode on it. The samples were then mixed and sent for MiSeq.

We found 95% of the provirus are integrated into genes, and HIV is more prone to integrate into highly expressed genes compared to background (Figure 12A). Compared to HIV, the distribution of expression level of the genes integrated by MLV is very similar to background distribution. Comparing HIV-Δenv with HIV-Δenv/nef, there was no significant difference in the orientation of the proviruses inserted into host genes. Across all the HIV vectors, we found the genes integrated by HIV have similar expression distribution (Figure 12A). Therefore, we conclude that Nef suppresses HIV
proviral gene expression not through cytotoxic effects induced by integration site selection.

Figure 12. HIV-Δenv and HIV-Δenv/nef are integrated into similar regions of human genome. The histogram of the expression score of genes that contain HIV provirus. Score is based on microarray data of H9 cells (GSE45036). Dark blue/green: the HIV/MLV genome orientation is same as the host genes. Light blue/green: the HIV/MLV genome orientation is opposite to the host genes.

h. Over-expressing Nef imposes silencing to HIV proviral gene in human pluripotent stem cells

To investigate the role of Nef in silencing HIV proviral gene and to confirm that Nef along is able to contribute the phenotype, we made the Nef expression construct (collaterally with an EGFP expressing construct) from HIV-SIN based lenti-vector (Ma et al., 2003; Yu et al., 2007) with an internal promoter of EF1α (Figure 13B) that can be
expressed in human pluripotent stem cells (Norrman et al., 2010) in trans. We performed the Nef supplementing experiment shown in Figure 13A and confirmed Nef expression by qPCR (data not shown). Basically, we co-infect H9 with a Nef donor HIV based vector (pSIN-Nef, HIV-Δenv, or negative control of pSIN-EGFP), together with HIV-Δenv/nef. By co-infecting the H9 with pSIN-Nef and HIV-Δenv/nef, and surveying the HIV expression via FACS on Day 5 post infection, we observed reduced HIV expression by three-fold compared to infected with EGFP (Figure 13C). Co-infecting HIV-Δenv/nef with HIV-Δenv can also silence the HIV-Δenv/nef (Figure 13C), indicating Nef can silencing the proviral genome in trans.

Since we observed less lenti-viral expression efficiency at high MOI infection with multiple copies in one cell (Figure 6 and Figure 7), the copy numbers of both the HIV-Δenv/nef and the Nef donor lenti-vector can potentially affect the expression level of the viruses. Therefore, we measured the copy number of HIV-Δenv/nef and the Nef-donor virus to make sure the cells have a similar level of HIV integration (Figure 13E) as well as Nef-donor viruses.

By infecting the iPS colonies that stably expressing HIV (iPS lateinfC4, Figure 9D) with pSIN-Nef and surveying the HIV expression via FACS on Day 12 post infection, we observed a 6-fold decrease in viral expression (Figure 13D). Our data strongly suggest that over-expressing Nef is able to silence the actively expressing HIV-Δenv/nef in human pluripotent stem cells.
Figure 13. Supplementing nef into HIV-Δenv/ nef infection can silence the virus in human pluripotent stem cells.
A. Schematic presentation of experiment design. B. Nef and EGFP construct made from pSIN-EF1α-Lin28A-Puro (3). C. FACS result showing fraction of positive cells in pSIN-Nef and HIV-Δenv/ nef co-infection in H9 cells, Day5 post infection. D. FACS result showing fraction of positive cells in pSIN-Nef infected iPS cells bearing single HIV-Δenv/ nef integration (iPS C4, Figure 7D). E. Copy numbers of HIV-Δenv/ nef and the lenti-Nef in H9 infection.
i. HIV expression and Nef-induced proviral silencing is linked to cell pluripotency

To further answer the question that whether Nef-linked HIV latency occurs only in pluripotent cells, we infected several cell types (H9, 293T, dH1f and Jurkat) with MOI=1 HIV-Δenv and performed cell sorting on Day2 post infection. We surveyed the genomic DNA integration level of the cells in that doesn't express fluorescence proteins integrated in HIV genome (FACS negative fraction). HIV-Δenv gene is present in all cells studied(Figure 14A), meaning that the HIV-Δenv latency is universal in all four cell types surveyed. However, comparing the amount of viruses in positive fraction (represented by percent positive cells) and viruses in negative fraction (represented by copies of integrated viruses in negative fraction), HIV-Δenv is silenced to different extent in cells with different pluripotency, especially H9 cells showed the strongest degree of silencing the virus while the Jurkat cells showed the least (Figure 14A). To better understand the relationship between cell pluripotency and viral silencing, we induced differentiation of iPS cells inserted with single copy of HIV-Δenv (iPS C15, Figure 9D) into fibroblast. We found a 6.7% of viral reactivation after inducing differentiation to fibroblast cells (Figure 14C). Our results indicate that human pluripotent stem cells have a stronger ability to maintain the HIV-Δenv proviral gene in silence compared to dH1f, 293T and Jurkat.
Figure 14. HIV-Δenv latency is linked to cell pluripotency.
A. Cell sorting experiment of HIV-Δenv infecting cells with different pluripotency. H9, dH1f, 293T and Jurkat cells are infected with HIV-Δenv at MOI=1 (titrated by 293T), and sorted for negative fraction on Day2 post infection. Genomic DNA is collected from the cells in the negative fraction. HIV integration copy number was obtained by qPCR compared to single copy 8E-5 cells. Blue bar: genomic DNA integration level in the negative fraction, representing amount of viruses in negative fraction. Red bar: percent of positive cells, representing amount of viruses in positive fraction. B. FACS result of inducing differentiation to fibroblast cells from HIV-Δenv integrated iPS cells.

As macrophages and multiple cells in blood lineages are nature host of HIV, and since Jurkat cells has least ability to silence HIV-Δenv, we asked whether inducing differentiation to blood lineage can also reactivate the silenced HIV proviral gene. We induced differentiation of HIV-Δenv integrated iPS colony (iPS C15, Figure 9D) to blood lineage by a two step process. We first successfully obtained CD34+CD45+ hematopoietic stem cells or CD34+ endothelial progenitor cells from iPS cells via embryoid body formation assay. The silenced HIV-Δenv is not reactivated in both population(data not shown). We then performed colony formation assay to both CD34+ and CD34+CD45 population and obtained terminally differentiated blood cells including granulocytes, erythrocytes and macrophages (Figure 15A and B). Interestingly, 2.1% reactivation (Figure 15C) was observed after inducing differentiation to terminal blood lineages. We traced down this 2.1% reactivation via staining for lineage specific markers of CD14 (monocytes specific marker) and CD15 (neutrophil specific markers).
Specifically, the viral reactivation is in the CD14+CD15- monocytes fraction (Figure 15C). As a gate control, we performed similar assay to iPS cells without HIV integration.

Figure 15. Induce differentiation of iPS colonies bearing HIV-Δenv latent proviral gene to blood lineage (A). Multiple blood lineages after colony formation assay 14 days from the CD34+/CD45+ double positive cells generated by inducing differentiation of HIV-Δenv integrated iPS cells. G: Granulocytes. M: Macrophages. GM: Granulocytes and Macrophages mixed. E: Erythrocytes. GEMM: Granulocytes, macrophages and erythrocyte mix. (B). Colony counts and types observed after 14 days of CFU assay. Four independent CFU were performed to CD34+/CD45+ population while three independent CFU were performed to CD34+ only population. (C). FACS of multiple blood lineage markers and HIV-Δenv EGFP after collecting the cells from CFU assay. Left: HIV-Δenv integrated iPS induce differentiation with CD45 (myeloid lineage marker), CD33 (myeloid lineage marker), CD14 (monocytes markers), CD15 (neutrophil markers) stained. Right: control iPS line without HIV integration unstained.

There was evidence that CD34+ hematopoietic stem cells is able to silence the full-length HIV and potentially serves as a HIV reservoir (Carter CC, 2010). In contrast, since HIV-Δenv/nef is highly expressed in iPS colonies, we asked whether the latency phenotype of HIV in CD34+ population is also nef dependent. We induced differentiation of iPS
colonies with single insertion of HIV-Δenv/Δ nef (iPS lateinfC4, Figure 9D) to CD34+ hematopoietic progenitor cells. We failed to obtain CD34+CD45+ population from this iPS line. However, we found that HIV-Δenv/Δ nef expression has a 2.3-fold reduction upon inducing differentiation to CD34+ lineage (Figure 16), suggesting that CD34+ cells suppress HIV proviral expression partially via a Nef-independent mechanism.

![Figure 16](image)

**Figure 16. Inducing differentiation into CD34+ hematopoietic progenitor cells by EB differentiation.**
iPS colony with single integration of HIV-Δenv/Δ nef (iPS late C3, Figure 9D) were differentiated into CD34+ cells. The iPS colony with a tat-defective HIV-Δenv/Δ nef is used as a control (iPS C3).

Together, our result provides evidence Nef-induced HIV latency is more specific to human pluripotent stem cells, while inducing differentiation of iPS cells harboring a latent virus to both fibroblast and blood lineage can moderately reactive the proviral expression.

**j. Identifying host factors that interact with Nef**

As we observed a strong phenotype of Nef inducing viral latency in human pluripotent stem cells, to further identify the host factors involved in Nef's silencing of viral genes, we knockdown the host protein of N-myristoyltransferase 1 and 2 (NMT1 and NMT2)
that both are reported to interact with Nef to perform myristoylation (Seaton and Smith, 2008; Morgan et al., 2011; Jager et al., 2012). Other Nef partners, including ACOT8 (Jager et al., 2012), PAK2 (Rauch et al., 2008; Foster et al., 2011), Hck were also tested. After confirming successful knockdown of the host proteins in H9 cells, we co-infect the cells with HIV-Δenv/nef with pSIN-Nef or pSIN-EGFP (Figure 17) and checked the genomic copy number of HIV-Δenv/nef in both infections (Figure 18B).

We found that compared to random shRNA control, all the Nef partners reduced the effect of Nef in inducing viral silencing except PAK2 (Figure 18A). NMT1 and ACOT8,
which almost completely abolish Nef's silencing effect, were also reported to strongly interact with Nef in both 293T and Jurkat cells (Jager et al., 2012).

Figure 18. Knocking down Nef partners in pluripotent stem cells impairs Nef role of silencing HIV proviral genome in trans.
A. Percent of viral expressing cells measured by FACS in different shRNA groups after co-infection of HIV-Δenv/nef with pSIN-Nef (red) or pSIN-EGFP (green) B. The copy number measurements of HIV-Δenv/nef in pSIN-nef coinfection (red) or pSIN-EGFP coinfection (green).

II.5 Discussion
a. Model of HIV latency in human pluripotent stem cells

Previous studies reported the mechanism of mouse embryonic stem cells silencing MLV by forming histone modification machinery at the PBS region of MLV genome (Wolf D, 2007; Wolf D, 2008; Wolf D, 2009). However, in human pluripotent stem cells, after changing the PBS of MLV, we didn't observe reactivation of the MLV transcription. These results indicate that human pluripotent stem cells utilize a different system for retroviral silencing.

To further confirm the pluripotent state is needed to silence the virus and nef is responsible for inducing the latency, we performed reprogramming experiment and obtained single integrated HIV iPS colonies. The result is consistent with our previous observations. After supplementing Nef into the HIV-Δenv/nef infected human pluripotent stem cells, we observed silencing of HIV-Δenv/nef genome. These results suggest that the Nef is responsible for inducing the viral silencing.
To further analyze the mechanism of HIV latency in human-pluripotent stem cells, we performed ChIP-PCR assay and found that the HIV-Δenv genome is enriched for H3K9me3 marker while the HIV-Δenv/nef genome is enriched for H3K4me3. This result shows that the Nef induced HIV latency in human pluripotent stem cells is at epigenetic level.

We also performed differentiation induction assay towards blood or fibroblast lineages and showed limited reactivation of HIV- Δenv. However, when we infected the differentiated cells directly, the differentiated cells showed much less viral silencing compared to human pluripotent stem cells. These results suggest the model that pluripotency associated host factors are responsible for silencing the HIV but after it established long-term silencing in human pluripotent stem cells, the epigenetic modification of viral genome became irreversible so that the removal of the host pluripotent factors cannot reactivate the virus.

To try to identify the host factors responsible for viral latency, we knockdown five host genes that closely interact with Nef. The result showed that after knocking down these genes, the HIV silencing is compromised, indicating they play roles in inducing HIV latency in human pluripotent stem cells.

To summarize, our study suggested the model of HIV latency in human pluripotent stem cells shown below:
b. Function of Nef protein

In our study, we found that HIV-1 Nef played important role in regulating HIV transcription in human pluripotent stem cells. This role of Nef has not been reported before.

Nef is a HIV-1 accessory protein that has 206 amino acids. In 1991 Kestler and colleagues reported that infecting rhesus monkeys with SIVmac239, a SIV virus strain that has a premature stop codon of Nef, results in attenuated viral replication in vivo (Kestler et al., 1991). It has also been reported that patients infected with HIV harboring defective Nef remained asymptomatic 14-18 years after HIV-1 infection without any HAART (Learmont et al., 1999). These in vivo observations indicated that nef played important role in viral replication and maintaining high viral load in vivo. However the molecular mechanism is poorly understood.
It is worth noticing that because *nef* has no catalytic activity, its pathogenetic function is mediated through cellular machinery or binding to HIV genome directly (Geyer et al., 2001; Malim and Emerman, 2008). In vitro experiments showed that Nef can bind to the LTR region of HIV genome and negatively regulates the transcription (Ahmad and Venkatesan, 1988). This observation was made in a cancer cell line SW480. People named this protein Nef due to the fact that it served as "Negative regulatory factor". Further study showed that *nef* can induced cell death in infected cells (Zauli G, 1999; Simmons et al., 2001) and induce apoptosis in bystander cells (Lenassi et al., 2010).

However, people also observed that in T cells, HIV without Nef has reduced viral replication rate (Terwilliger E, 1986; Chowers MY, 1994; Spina CA, 1994; Pizzato M, 2008; Emert-Sedlak L, 2009; Ptak RG, 2010). Other study showed that Nef can prevent the cell death of infected cells by down-regulating cell surface MHCI, CD4, CD8, and CD28 and avoid visibility from immune cells (Leonard JA, 2011). These are conflicting observations, and our study showed that Nef reduced HIV transcription and induce cell death, which was consistent with the observation of Ahmad and Simmons. It is possible that Nef's function is largely dependent on cell type due to the fact that it has to collaborate with certain host factors. Those host factors' expression level and functions may differ in different cell types.

Nef function is largely dependent on myristoylation on the protein. N-myristoyltransferase-1 (NMT1) and N-myristoyltransferase-2 (NMT2) have been reported to directly myristoylate Nef protein though Nef was preferentially myristoylated
by NMT2 \textit{in vitro} (Seaton and Smith, 2008; Jager et al., 2012). The myristoylation of Nef is required to activate cell signaling pathways including enhancement of p21-activated protein kinase (PAK2) autophosphorylation and interacting with Hck to Perturb the function of the Golgi complexes Hck (Hiyoshi et al., 2012). It is possible that in human pluripotent stem cells, players downstream of PAK2 and Hck interact with HIV transcriptional machinery to regulate HIV expression level, and Nef indirectly serve as transcriptional regulator by interacting with PAK2 and Hck.

c. \textbf{HIV cytotoxicity in human pluripotent stem cells}

We performed Nef supplementing experiments and observed suppression of HIV-\(\Delta\text{env/nef}\) expression level in both co-infection and super infecting HIV-\(\Delta\text{env/nef}\) iPS colonies (Figure 13), which is consistent with our previous observation of Nef's role in silencing the proviral genome in human pluripotent stem cells. However, we failed to observe the cytotoxicity effect after supplementing Nef (data not shown). We suspect the cytotoxicity that we observed in HIV-\(\Delta\text{env}\) infection is likely not to be induced by Nef. Though previous arts reported Nef's role in inducing cell apoptosis via inducing the up-regulation of CD95/CD95L pathway (Zauli G, 1999), we suspect the in our experiment setting the cytotoxicity effect is not induced by Nef but the residual envelope expression (Westendorp MO, 1995).

Apoptosis, pyroptosis and necroptosis (non-apoptotic cell death pathway) can be induced by viral infection and serve as natural immunity to viral replication (Upton and Chan,
Necroptosis can also be induced by NLRP1 and RIP3 in hematopoietic stem cells and embryonic stem cells (Kaiser et al., 2011; Masters et al., 2012). Pyroptosis is induced by NLRP3 and Caspase-1 (Rathinam et al., 2012). Apoptosis is mediated by an extrinsic signal of FADD activating caspase-8 (Kagi et al., 1994) or intrinsic signal of Apaf-1 activating caspase-9 (Cain et al., 2000).

To further investigate the mechanism of HIV cytotoxicity in human pluripotent stem cells, we propose to use cell death inhibitors to further define how the cells died after HIV infection. Caspase-1 inhibitors (Millipore 400010) can block the pyroptosis (Thornberry et al., 1992). RIPK1 inhibitor (necrostatin-1, SIGMA N9037) and RIPK3 inhibitor are able to block necroptosis (Degterev et al., 2005). The Q-VD-OPh (SIGMA SML0063) (Caserta et al., 2003; Lavallard et al., 2009) and Z-VAD-FMK (SIGMA V116) (Hara et al., 1997) are broader caspase inhibitors thus can effectively block apoptosis and necrosis. We propose to screen the above chemicals and investigate which chemical is able to block the cell death upon HIV-Δenv infection in human pluripotent stem cells.

d. HIV long-term latency in human pluripotent stem cells

Though we identified multiple host proteins that can reduce the effect of Nef's role of silencing HIV provirus, we cannot reactivate the silenced HIV-Δenv in pluripotent stem cells via infecting shRNAs into iPS cells already bearing a HIV-Δenv latent gene. This result indicates that Nef functions at early stage after viral integration. The fact that infecting HIV-Δenv into H9 cells already selected for shRNAs also cannot reactivate the
HIV-Δenv led us to hypothesize that there are mechanism other than Nef that are responsive for the silencing of HIV-Δenv. We exclude the possibility of integrational sites selection bias by sequencing the viral integrational sites. We further hypothesize that cell death induced by HIV-Δenv probably exert additional effort on viral genome suppression through other intracellular pathways.

Another possibility is that HIV is silenced in human pluripotent stem cells via epigenetic modification such as CpG methylation. HIV long-term silencing via hyper-methylation on HIV proviral LTR has been reported to be associated with HIV long-term elite controller (Palacios et al., 2012). Other chromatin modification including histone deacetylation and histone methylation are both highly involved in the HIV latency (du Chéné I, 2007; Tyagi M, 2007; Pearson R, 2008; Archin NM, 2009; Friedman J, 2011). To further investigate the mechanism of HIV long-term silencing, we propose to conduct a bisulfite-sequencing to map the DNA methylation site on HIV genome. Furthermore, reactivation of the virus under extreme circumstances in vivo are important to restore viral replication restoring, and reactivating long-term reservoir by drugs can potentially lead to cure of AIDS. Here we propose to screen a library of DNA methylation inhibitors, HDAC inhibitors (such as TSA), 5-Azacytidine and Prostratin.

e. HIV latency in other cell lineages

We observed reactivation of HIV-Δenv in different cell lines after inducing differentiation. Since the expression level of nef partners in fibroblast and H9 cells are similar, according to public microarray data (GDS3892[ACCN]), the mechanism of viral
expression associated with cell pluripotency may not be regulated via expression control of Nef partners. We observed a bivalent histone status of HIV proviral genome in human pluripotent stem cells, which is enriched for both H3K9me3 and H3K4me3. Multiple genes in human stem cells maintain the bivalent histone modifications (Pan et al., 2007). We hypothesize that the difference of HIV expression in regards to cell pluripotency is regulated by global histone platform changes in cells when inducing differentiation. However, our observation shows that the reactivation effect is moderate by inducing differentiation, while directly infecting HIV-\textit{nef} virus into cells with different pluripotency shows more pronounced differences in viral expression. This phenomenon indicates that after initial histone modification performed by players downstream \textit{nef}, the virus genome is maintained in a bivalent status. This process is subjective to cell pluripotency. Further epigenetic modifiers are recruited afterwards that probably perform DNA modification to silence the virus in long-term, and the second modification process is not subject to cell pluripotency. Therefore, by inducing differentiation, the viruses cannot be reactivated due to a long-term silencing on HIV genome.

To better test whether the mechanism of HIV latency discovered using our model of human pluripotent stem cells can be transferable to human multipotent stem cells such as hematopoietic stem cells (HSC), the natural reservoir of HIV, more experiments such as infecting HSC with HIV-\textit{Δenv/nef} and HIV-\textit{Δenv} will be necessary.
Chapter III  A Genome-wide association study reveals multiple new genes for host control of HIV-1 Subtype-C AIDS disease progression in an African population
III.1 Abstract

The Sub-Saharan African population infected with HIV-1C makes up the largest AIDS-patient population in the world and exhibits large heterogeneity in disease progression before initiating HAART. Previously, HLA-B*5701, other HLA alleles (Fellay et al., 2007; Colombo et al., 2008), Ring Finger Protein 39 (RFN39) and Zinc Ribbon Domain-Containing 1 (ZNRD1){Fellay, 2007 #1055} were identified as associated with VL set point or disease progression by genome wide association studies (GWAS) in HIV-subtype B infected European patients. However, little research has been done on control of HIV Subtype-C in the African population. Additionally, most previous studies to examine issues of progression in HIV/AIDS were based on single points in time for CD4 or viral load (VL) over time in treatment-naïve HIV infected patients. While searching for associations with trajectories of CD4 or VL are important, such indicators do not always follow linear trajectories. In our study, we performed an initial GWAS on 317 Tswana treatment-naive AIDS patients and replicated the result in a second independent cohort of 239. By applying a method of functional principal component analysis (fPCA) to characterize variability in CD4 and VL trajectories, we gained power with a relatively small data set. We identified 3 independent SNPs that are significantly associated with AIDS disease progression measured by combined CD4 and VL trajectories. The genes involved are HCG22, CCNG1 and ZBTB7C. The result is consistent after adjusting for sex and age. We also found several additional promising loci below significance level after using both fPCA and Linear mixed approximation for phenotypes: PRKAA2, ARHGEF12, TRIM26, ARL8B, AK095365, PCDH11X, and ZFYVE28. Our finding
raises the possibility that more SNPs linked to AIDS disease pathogenesis may yet be discovered if the approach used in this study is adopted and suggests that mechanistic studies of how CCNG1 and ZBTB7C interact with HIV-1 subtype C may uncover previously unknown pathways of HIV-host interaction. Collectively, though further replication studies are required with larger sample size to validate the finding, the genes we identified suggest new possibilities for examining biological mechanisms in HIV1-Subtype-C host control *in vivo*.

### III.2 Introduction

Currently HIV-1C accounts for more than half of all global infections and several-fold more than any other subtype, including HIV-1B which predominates in Caucasian populations (Geretti, 2006). In 2012, of the 34 million people living with HIV globally, 23.5 million resided in Sub-Saharan Africa, where HIV-1 Subtype-C dominates (2012). In Botswana, the 2012 prevalence of HIV in adults (15-49 years) is estimated at 23.4% with 68.7% women (UNAIDS, 2013a). Given the diversity of the southern Africa population (Schuster et al., 2010), the fact that treatment-naive patients display significant heterogeneity in natural disease progression in our HIV-infected Botswana cohorts suggests that the host genetic makeup of individuals may contribute to their varying ability to maintain CD4 counts and viral load.

Studying the disease progression of AIDS is important for estimating optimal time to initiate Highly Active Antiretroviral Therapy (HAART) and understanding host restriction of viral replication, which may suggest avenues for development of drugs and vaccines. Despite the worldwide predominance of HIV-1 Subtype-C (HIV-1C), most
Genome wide association studies (GWAS) have been conducted primarily on males of European ancestry infected by HIV-1 subtype B (Fellay et al., 2007; Fellay et al., 2009; van Manen et al., 2009; Troyer et al., 2011; van Manen et al., 2011; Bartha et al., 2013; McLaren et al., 2013). Previous GWAS for disease progression rate on HIV-subtype B infected European patients showed that most hits associated with viral load set points are on HLA-B*5701 and other HLA alleles (Migueles et al., 2000; Fellay et al., 2007; Dalmasso et al., 2008; van Manen et al., 2009; An and Winkler, 2010; O'Brien and Hendrickson, 2013). Other hits associated with disease progression measured by CD4 declining slope include ring finger protein 39 (RFN39) and zinc ribbon domain-containing 1 (ZNRD1) which encodes an RNA PolII subunit (Fellay et al., 2007). Notch4, C6orf12 and Trim10 were identified using the permutation procedure (Fellay et al., 2009).

However all these findings can only explain less than 10% of the differences in disease progression (Fellay et al., 2007), which calls for both a more accurate method for defining disease progression and a model combining the viral load with the CD4 trajectory. Using a single value of viral load set point or early CD4 counts cannot fully describe the course of measurements' change over time. Since we do not know in what manner SNPs might affect viral load or CD4 count, we employed functional principal component analysis (fPCA) (Gareth M. James, 2000; Yao F, 2005; Yao, 2007; Wu S, 2011) to characterize the variability in the trajectories of CD4 and VL. This fPCA method allows us to identify major modes of variation in the trajectories without any assumptions on the shape of the trajectories. It furthermore easily identifies typical modes of variation like mean and slope while also accommodating and identifying nonlinearity
in the trajectories. Compared to linear mixed simulation, our result showed that by adopting fPCA we can increase the study power and reveal additional SNPs that are associated with characteristics of CD4-VL combined trajectory other than slope.

In this study, we identified three significant host genetic restriction factors for viral replication using a novel definition of disease progression by fPCA. Our study is the first GWAS limited to HIV-1C and the southern African population. This is also the first study modeling AIDS disease progression trajectory using (fPCA) based on longitudinal observations of CD4 counts and viral load, and the first GWAS on HIV using Illumina Omni2.5 chips.

III.3 Material and Methods
a. Cohorts and patients eligibility

The BHP010 (Botsogo and Botsogo Extension Study) was a natural history observational prospective cohort study recruited from clinics in Gaborone, which was composed of HIV-1C infected individuals with starting CD4 cell counts above 400 cells per µl and not yet qualified for the Botswana HAART treatment program. Any one below 18, those with active AIDS defining illness requiring the initiation of HAART, the presence of an AIDS-related malignancy, or pregnant or breast feeding women who have been previously exposed to HAART were excluded. Participants returned to the clinic at 3-month intervals with an additional visit 1 month after enrollment and were followed up for up to 255 weeks (Figure 20A). The mean sampling time for CD4 was 91 weeks (Figure 20A). The end state of the study was censored (216/456), or enrollment on HAART treatment (120/456). Participants who could not be contacted after they have
missed three consecutive visits were considered “lost to follow-up” (90/456). Of the 456 patients enrolled, 370 were women and age distribution of BHP010 is shown in Figure 20B. DNA samples were collected at around three years post enrollment and 294 samples were available (Table 1). In this subset, the age and gender distribution are similar to the whole cohort, indicating a non-biased sampling (Table 2).

![Figure 20. Population and sampling comparison between BHP010 and BHP011.](image)

A. The histogram showing the distribution of CD4 sampling time point (weeks after enrollment) of BHP010 (blue) and BHP011(pink). BHP010 has a longer follow-up time with the longest of 255 weeks post enrollment and mean of 91 weeks. BHP011 has a shorter follow-up time with the longest of 169 weeks and mean of 56 weeks. B. Age distribution of BHP010 (blue) and BHP011 (pink). This figure involves all the patients in BHP010 and BHP011. C. EIGENSTRATE using genotypes of BHP010 and BHP011 after standard QC(Table 5). D. EIGENSTRATE using genotypes of BHP011, showing homogenous distribution of different treatment arms.

The BHP011 (Dikotlana Study) was a randomized, multifactorial, double-blind placebo-controlled trial to determine the efficacy of micronutrient supplementation (supplementation of multivitamin, selenium or both) in improving immune function and preventing early mortality of HIV-1C infection (Baum et al., 2013). It was composed of 875 treatment naive patients with CD4 higher than 350/µl. 638 out of 875 patients were women and age distribution of BHP011 is shown in Figure 20B. DNA samples were
collected at around two years post enrollment and 364 samples were available (Table 1).

In this subset, the age and gender distribution were similar to the whole cohort, indicating
a non-biased sampling (Table 3). Patients were followed up for maximum 169 weeks post
enrollment (Figure 20A). They returned to clinics every 6 months to get CD4 samples.

The mean CD4 sampling time was 56 weeks post enrollment (Figure 20A).

Table 1 Information of cohorts

<table>
<thead>
<tr>
<th></th>
<th>BHP010, natural history observational cohort</th>
<th>BHP011, randomized multivitamin treatment cohort</th>
<th>Combined (number with DNA)</th>
</tr>
</thead>
<tbody>
<tr>
<td>participants</td>
<td>456 (294)</td>
<td>876 (364)</td>
<td>1332 (658)</td>
</tr>
<tr>
<td>Intervention</td>
<td>No</td>
<td>Four arms:</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Multivitamines alone: 219 (93)</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Selenium alone: 220 (88)</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Multivitamines+Selenium: 217 (94)</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Placebo: 219 (88)</td>
<td></td>
</tr>
<tr>
<td>Women percentage</td>
<td>81.0% (84.7%)</td>
<td>72.9% (71.8%)</td>
<td>75.7% (77.6%)</td>
</tr>
<tr>
<td>Median baseline CD4 (cells/mm3)</td>
<td>468.5 (484.5)</td>
<td>422.2 (424.8)</td>
<td>439.1 (456.0)</td>
</tr>
<tr>
<td>Median baseline VL (Copies/ml)</td>
<td>12.9K (10.1K)</td>
<td>14.0K (14.5K)</td>
<td>13.6K (12.1K)</td>
</tr>
</tbody>
</table>

* Numbers in parentheses indicate result of patients with DNA available

Table 2 Average baseline and median CD4 and VL (BHP010)

<table>
<thead>
<tr>
<th>BHP010</th>
<th>whole cohort</th>
<th>Has DNA</th>
<th>qualified for GWAS</th>
<th>call rate pass through</th>
<th>Involved in GWAS analysis</th>
</tr>
</thead>
<tbody>
<tr>
<td>participants</td>
<td>456</td>
<td>294</td>
<td>258</td>
<td>250</td>
<td>240</td>
</tr>
<tr>
<td>average baseline CD4</td>
<td>513.6</td>
<td>520.9</td>
<td>532.1</td>
<td>534.5</td>
<td>533.1</td>
</tr>
<tr>
<td>Median baseline CD4</td>
<td>468.5</td>
<td>484.5</td>
<td>497</td>
<td>497.5</td>
<td>497.5</td>
</tr>
<tr>
<td>average baseline VL</td>
<td>56577</td>
<td>38743</td>
<td>37258</td>
<td>36952</td>
<td>35787</td>
</tr>
<tr>
<td>Median baseline VL</td>
<td>12900</td>
<td>10090</td>
<td>9565</td>
<td>8815</td>
<td>8255</td>
</tr>
<tr>
<td>average age</td>
<td>33.9</td>
<td>34.2</td>
<td>34.2</td>
<td>34.2</td>
<td>34.3</td>
</tr>
<tr>
<td>Women percentage</td>
<td>81%</td>
<td>85%</td>
<td>84%</td>
<td>83%</td>
<td>83%</td>
</tr>
</tbody>
</table>

Table 3 Average baseline and median CD4 and VL (BHP011)

<table>
<thead>
<tr>
<th>BHP011</th>
<th>whole cohort</th>
<th>Has DNA</th>
<th>qualified for GWAS</th>
<th>call rate pass through</th>
<th>Involved in GWAS analysis</th>
</tr>
</thead>
<tbody>
<tr>
<td>participants</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>average baseline CD4</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Median baseline CD4</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>average baseline VL</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Median baseline VL</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>average age</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Women percentage</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Table 3 Average baseline and median CD4 and VL (BHP011) (Continued)

<table>
<thead>
<tr>
<th>participants</th>
<th>875</th>
<th>364</th>
<th>334</th>
<th>326</th>
<th>321</th>
</tr>
</thead>
<tbody>
<tr>
<td>average baseline CD4</td>
<td>462.3</td>
<td>477.0</td>
<td>475.3</td>
<td>476.6</td>
<td>475.6</td>
</tr>
<tr>
<td>Median baseline CD4</td>
<td>422.2</td>
<td>424.8</td>
<td>427.8</td>
<td>432.4</td>
<td>429.7</td>
</tr>
<tr>
<td>average baseline VL</td>
<td>67600</td>
<td>72269</td>
<td>64972</td>
<td>66380</td>
<td>66189</td>
</tr>
<tr>
<td>Median baseline VL</td>
<td>14000</td>
<td>14500</td>
<td>14600</td>
<td>14900</td>
<td>15200</td>
</tr>
<tr>
<td>average age</td>
<td>34.0</td>
<td>35.0</td>
<td>35.3</td>
<td>35.2</td>
<td>35.5</td>
</tr>
<tr>
<td>Women percentage</td>
<td>73%</td>
<td>72%</td>
<td>71%</td>
<td>72%</td>
<td>71%</td>
</tr>
</tbody>
</table>

b. Genotyping process and quality control

The genotyping was performed using the protocol of Illumina® Infinium® Omni2.5 Multi-Use Assay. Due to discontinuation of manufacturer's product, two kinds of chips were used: Illumina HD BeadsChip and Illumina LCG BeadsChip. The numbers of patients genotyped by each kind of Illumina BeadsChip are shown in Table 4.

Table 4 Genotyping platforms used for BHP010 and BHP011

<table>
<thead>
<tr>
<th>BHP010 (Botsogo)</th>
<th>112</th>
<th>142</th>
</tr>
</thead>
<tbody>
<tr>
<td>BHP011 (Dikotlana)</td>
<td>0</td>
<td>335</td>
</tr>
</tbody>
</table>

Note: One patient from BHP010 was genotyped twice: one by HD Chip and the other by LCG Chip.

Samples with DNA concentration higher than 38ng/µl were qualified for genotyping. As shown in Table 5, of the qualified 592 samples, 4 samples didn't get genotyped due to discontinuation of manufacturer's product, so we genotyped 588 patients. Among them, 7 patients' call rate didn't pass 0.99. We also checked the specification of the gender to verify that patients with gender recorded in the clinical records as male should have heterozygous X genotypes (>=1%) or who if recorded as females should have
homozygous X genotypes. Six patients who didn't match the criteria were excluded. We also only include patients that have at least two clinical laboratory observations for both CD4 and VL; 19 patients were excluded by this criteria. After the initial QC, 556 patients were selected for further analysis, 239 of them are from BHP010 and 317 are from BHP011. Detailed characteristics of the patients before QC and after each step of initial QC are shown in Table 5. The gender ratio, median baseline VL, median baseline CD4, average baseline VL, average baseline CD4 and average age after each step of QC were all similar to the initial cohort for both BHP010 and BHP011, indicating unbiased sampling (Table 2, Table 3).

To avoid artifacts associated with either chemistry or clustering errors, we performed the quality control to the SNPs included in the association study as follows:

(1). Infinium BeadStudio Raw Data clustering:

Due to the incompatibility of LCG and HD chip data processing through Illumina Genome Studio, BHP010 on HD, BHP010 on LCG and BHP011 on LCG were processed separately. We used the reclustering file based on the previous unpublished Malawi genotyping data we generated. We applied the Quality control consistent with previous report (Fellay et al., 2007). After the initial reclustering, any sample that had very low intensity or a call rate lower than 95% is deleted. All SNPs on Y chromosomes were individually evaluated to make sure they are only called in male samples. All SNPs on X chromosomes were also individually evaluated to make sure no male was heterozygous. The autosomal SNPs fulfilling any one of the five criteria were screened within BeadStudio and reclustered manually: call frequencies of less than 0.62; the mean of the normalized r-values for the AB genotypes of less than 0.28; the mean of the normalized
theta angles for the AB genotype is smaller than 0.2 or larger than 0.8; the Het Excess value is between -1.0 and -0.3 or between 0.2 and 1.0; and a Minor Allele Frequency less than 10%. If a SNP fails the reclustering, the SNP was zeroed manually. This process led to deletion of 3739, 6727, 6263 SNPs in BHP010 genotyped by Illumina HD BeadsChip, BHP010 genotyped by Illumina LCG BeadsChip and BHP011 genotyped by Illumina LCG BeadsChip, respectively.

(2). Minor Allele Frequency (MAF) check, genotyping missingness and Hardy-Weinberg Equilibrium (HWE):

After the reclustering, we further excluded SNPs that had Minor Allele Frequency lower than 5% in both BHP010 and BHP011. The rationale behind this quality control is that at least 28 people should bear the SNP in the cohort. We excluded SNPs with call rates smaller than 0.95 after reclustering. We further tested the Hardy-Weinberg Equilibrium with α=0.0001 and only included the SNPs that passed the test. The detailed result of the QC is shown in Table 5.

<table>
<thead>
<tr>
<th>Filtering</th>
<th>Dropped*</th>
<th>Included*</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total subjects</td>
<td>1332 (456, 876)</td>
<td>752 (294, 458)</td>
</tr>
<tr>
<td>Blood sample collected</td>
<td>580 (162, 418)</td>
<td>658 (294, 364)</td>
</tr>
<tr>
<td>DNA sample available</td>
<td>94 (0, 94)</td>
<td>592 (258, 334)</td>
</tr>
<tr>
<td>Qualify for Genotyping</td>
<td>66 (36, 30)</td>
<td>588 (255, 333)</td>
</tr>
<tr>
<td>Genotyped</td>
<td>4 (3,1)</td>
<td>581 (250, 331)</td>
</tr>
<tr>
<td>Call rate pass 0.99</td>
<td>7 (5,2)</td>
<td>575 (249, 326)</td>
</tr>
<tr>
<td>Gender matched</td>
<td>6 (1,5)</td>
<td>556 (239, 317)</td>
</tr>
<tr>
<td>More than two clinical observations</td>
<td>19 (10,9)</td>
<td>556 (239, 317)</td>
</tr>
<tr>
<td><strong>Total used</strong></td>
<td><strong>556 (239, 317)</strong></td>
<td><strong>556 (239, 317)</strong></td>
</tr>
</tbody>
</table>

* Numbers in parentheses indicate result of BHP010 (left) and BHP011 (right)
<table>
<thead>
<tr>
<th>Filtering</th>
<th>Dropped</th>
<th>Included</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total number of SNPs on Illumina Omni2.5 BeadsChip HD and LCG</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Consistent phenotype direction between primary and secondary cohorts for both CD4 and VL</td>
<td>680905</td>
<td>1773227</td>
</tr>
<tr>
<td>Call rate pass 0.99 in either cohorts</td>
<td>17799</td>
<td>1755428</td>
</tr>
<tr>
<td>MAF above cut off in either cohort</td>
<td>859517</td>
<td>895911</td>
</tr>
<tr>
<td>Pass HWE $\alpha=0.05$</td>
<td>43022</td>
<td>852889</td>
</tr>
<tr>
<td>Autosomal</td>
<td>33857</td>
<td>819032</td>
</tr>
<tr>
<td>Total used (if including X and Y chromosomes)</td>
<td></td>
<td>852889</td>
</tr>
</tbody>
</table>

**c. EIGENSTRAT method to perform population stratification**

We applied the EIGENSTRAT as described before (Fellay et al., 2007). Basically we selected EIGENSTRAT axes to analyze the population ancestry. We selected SNPs based on the QC described above.

**d. Functional principal component analysis steps**

We performed the fPCA methods as described (Gareth M. James, 2000; James, 2002; Yao F, 2005) for log CD4 and log viral load in BHP010 and BHP011 separately. For each cohort and measure, we decomposed each person’s trajectory into a mean function (shared among all patients) and a linear combination of Eigenfunctions, $\phi_k(t), k = 1, 2, 3$. Because the Eigenfunctions are constructed such that the first Eigenfunction explains the most variability in the data, the second Eigenfunction explains the second
most variability, and so on, only 3 Eigenfunctions are necessary to explain more than 99% of the variation in the data. The coefficients, or scores, $\xi_{ik}$, associated with the Eigenfunctions succinctly summarize the patient-specific trajectories.

$$\log \text{CD4}_{ij} = \mu^{(\text{CD4})}(t_{ij}) + \sum_{k=1}^{3} \xi^{(\text{CD4})}_{ik} \phi^{(\text{CD4})}_k(t_{ij}) + \epsilon^{(\text{CD4})}_{ij}$$

$$\log \text{VL}_{ij} = \mu^{(\text{VL})}(t_{ij}) + \sum_{k=1}^{3} \xi^{(\text{VL})}_{ik} \phi^{(\text{VL})}_k(t_{ij}) + \epsilon^{(\text{VL})}_{ij}$$

where for the $i$th patient at time $t_{ij}$, $\text{CD4}_{ij}$ represents their observed CD4 count; $\mu^{(\text{CD4})}(t_{ij})$ represents the population mean function for log-$\text{CD4}$ evaluated at $t_{ij}$; $\phi^{(\text{CD4})}_k(t_{ij})$ represents the $k$th Eigenfunction evaluated at $t_{ij}$; $\xi^{(\text{CD4})}_{ik}$ represents the random score for the $k$th Eigenfunction, which can also be seen as a random effect; and $\epsilon^{(\text{CD4})}_{ij}$ represents random measurement error. Quantities are defined analogously for viral load.

The population means $\mu^{(\text{CD4})}$ and $\mu^{(\text{VL})}$ are estimated via nonparametric smoothing, and the Eigenfunctions are obtained via an eigendecomposition of the smoothed covariance matrices of log CD4 and log VL (Rice, 1991; Hall, 2006). To estimate the scores, we used the best linear unbiased predictor (Robinson, 1991) because the data on CD4 and viral load were not densely observed. Thus, to estimate the scores for person $i$, we borrow information from all other subjects. Only three Eigenfunctions were needed to explain more than 99% of the variability in the trajectories.

e. Linear regression model to characterize CD4 and VL trajectories
We also fit a more familiar linear mixed effects model (Laird, 1982; Smith et al., 2010; Zhang et al., 2010) to bolster the evidence from the FPCA model

\[
\log \text{CD}_4_{ij} = \beta_0^{(\text{CD}_4)} + \beta_1^{(\text{CD}_4)} t_{ij} + \xi_{i0}^{(\text{CD}_4)} + \xi_{i1}^{(\text{CD}_4)} t_{ij} + \epsilon_{ij}^{(\text{CD}_4)}
\]

\[
\log \text{VL}_{ij} = \beta_0^{(\text{VL})} + \beta_1^{(\text{VL})} t_{ij} + \xi_{i0}^{(\text{VL})} + \xi_{i1}^{(\text{VL})} t_{ij} + \epsilon_{ij}^{(\text{VL})}
\]

This model corresponds to pre-specifying the model above to have \( \mu^{(\text{CD}_4)}(t_{ij}) = \beta_0^{(\text{CD}_4)} + \beta_1^{(\text{CD}_4)} t_{ij}, \phi_1^{(\text{CD}_4)}(t_{ij}) = 1, \phi_2^{(\text{CD}_4)}(t_{ij}) = t_{ij}, \) and \( \phi_3^{(\text{CD}_4)}(t_{ij}) = 0 \) (and likewise for viral load).

\section{f. Testing}

To characterize the strength of association between a SNP, \( G_i \), and the trajectories of \( \log \text{CD}_4 \) and \( \log \text{viral load} \), we obtained variance component score statistics, similar to those used previously (Lin, 1997; Wu et al., 2011), that correspond to the working models

\[
\xi_{ik}^{(\text{CD}_4)} \sim N(y_k^{(\text{CD}_4)} G_i, \tau^2), \xi_{ik}^{(\text{VL})} \sim N(y_k^{(\text{VL})} G_i, \tau^2)
\]

where \( y_k^{(\text{CD}_4)} \) represents the effect of the SNP on the trajectory of CD4. For the null hypothesis of no association between the SNP and the longitudinal trajectories of CD4 and viral load, \( H_0: y_k^{(\text{CD}_4)} = y_k^{(\text{VL})} = 0 \), the form of the test statistic is given by

\[
\sum_{k=1}^{3} \sum_{l=\text{CD}_4, \text{VL}} (n^{-1/2} \sum_{i=1}^{n} \hat{\zeta}_{ik}^{(l)} G_i)^2
\]

where \( \hat{\zeta}_{ik}^{(l)} \) is the best linear unbiased predictor of \( \zeta_{ik}^{(l)} \). To adjust for age and sex, the test statistic becomes

\[
\sum_{k=1}^{3} \sum_{l=\text{CD}_4, \text{VL}} (n^{-1/2} \sum_{i=1}^{n} (\hat{\zeta}_{ik}^{(l)} - \alpha_{1k} \text{ age}_i + \alpha_{2k} \text{ sex}_i) G_i)^2
\]
where $z_{1k}(t)$ and $z_{2k}(t)$ are obtained under the null hypothesis.

Under the null hypothesis, the test statistic is distributed asymptotically as a mixture of $\chi^2$'s. We obtain p-values by estimating its null distribution and comparing it to the observed test statistic. The overall error rate was controlled using the Benjamini-Hochberg false discovery rate (FDR) controlling procedure (Benjamini, 1995). The FDR was controlled at the level 0.10.

III.4 Results

a. Population characteristics and genetic stratification

Two independently recruited Tswana cohorts were involved in our study: a natural history observational prospective cohort (BHP010) observed prior to drug intervention, and a randomized multivitamin treatment cohort (BHP011) with four treatment arms (Table 1, also see Methods). Both cohorts consist of AIDS patients from Botswana that are infected with subtype-C and have not been exposed to HAART. Age distribution was similar for the two cohorts (Figure 20), despite BHP010 having a slightly higher proportion of women (Table 1). Based on EIGENSTRAT population stratification analysis (Price et al., 2006), the genetic makeup of BHP010 and BHP011 is not separated and the two genotyping platforms showed similar results (Figure 20C). The four treatment arms of BHP011 are also genetically similar (Figure 20D). After going through a stringent exclusion criteria (See methods, Table 5), 556 patients were finally involved in our association analysis.
b. Phenotype definition using Functional Principal Component Analysis (fPCA) to reconstitute of CD4 and VL trajectory

The natural history of treatment naive patients from BHP010 and BHP011 cohorts shows various disease progression patterns (Figure 21). We performed fPCA {Holte, 2012 #3311} separately on BHP010 and BHP011 due to the fact that the length of follow-up differed between the two cohorts. The most important modes of variation in the trajectories were selected such that 99% of the observed variability in the data was explained. This is analogous to regular principal component analysis where the most important directions of variation of a matrix are obtained via Eigen decomposition. These three modes highlight three typical trajectories of maintaining, declining and varying CD4/VL counts (Figure 21A). Extremity of the first function scores is able to separate patients maintaining high or low CD4/VL, while extremity of second function scores can separate flat versus declining CD4/VL scores (Figure 21B). The third Eigenfunction captures the variation of longitudinal trajectories. We then used the score of the first Eigenfunction to approximate the level of baseline CD4/ VL, and the score of the second Eigenfunction to approximate the progressiveness of their diseases.
Figure 21. Functional principle component analysis to estimate Eigenfunctions of CD4 and VL trajectories.

A. Estimated Eigenfunctions used to fit CD4 and viral load longitudinal data for BHP010 and BHP011 separately. Three functions were used for viral load and CD4 separately in each cohort. B. Extremity of the first Eigenfunction coefficients (Left) separate patients with high or low CD4 counts (Upper) or Viral load (lower). Extremity of the second Eigenfunction coefficients (Right) separate patients with fast declining CD4 or VL versus flat or increased CD4 or VL. C. Examples of Eigenfunction fit for six patients. With similar viral load counts (red), patients showed different patterns of CD4 trajectory (black) over time. Solid lines represent approximation of CD4 or VL trajectories by fPCA. Dash lines represent predicted trajectories of VL by fPCA. D. Proportion of variation explained by FPCA (light blue) and linear model (dark blue) as measured by pseudo-R^2. Larger values indicate better model fit to the data. E. Power curve of association study using FPCA (red) or linear regression (blue).

We found that by fitting curves via fPCA Eigenfunctions, we can largely reconstitute the CD4 and viral load trajectories for patients (Figure 21C) with a lower variation represented by higher pseudo-R^2 compared to linear regression (Figure 21D). FPCA provides relevant phenotypes by separating fast progressors from slow progressors.
(Figure 21B, Figure 22) and it does not require the restrictive linearity assumption of linear mixed effects models, potentially gaining power to detect effects on nonlinearity in CD4 and VL trajectories. Additionally, in order to account for the profound interaction between viral load and CD4, we combined VL and CD4 fPCA scores (coefficients) to define the phenotype of patients.

Figure 22. Extremity of intercept and slope values using linear model
Green: high intercept value or low absolute value of slope. Red: low intercept value or high absolute value of slope. The intercept extremities separate patients with high viral load or CD4 (green) versus patients with low viral load or CD4 (red). The slope extremities separate patients that are fast progressors (red) versus slow progressors (green).

c. GWAS using BHP011 as primary cohort and BHP010 as replication cohort

We conducted an initial GWAS on BHP011 sized 317 as primary cohort with 819032 SNPs included (Table 5) and selected 61 SNPs with P-value lower than $6 \times 10^{-4}$ to replicate in the second cohort of BHP010 sized 239. After false discovery rate (FDR) adjustment with the significance level of 0.1 we identified 39 SNPs that stand
significantly associated with HIV disease progression with a consistent association direction in two cohorts. We tested the direction of the effect and eliminated SNPs that were significant after replication but have opposite direction of effect in the primary cohort and the secondary cohort (Figure 23, Table 6). After Linkage disequilibrium analysis, we found 8 SNPs lying in independent regions on the human genome (Table 7, Figure 24) that have same association direction in both cohorts. Among them, 3 reached genome-wide significance level with a P-value lower than $2.38 \times 10^{-6}$ in a pooled analysis of the two cohorts (Table 8, Figure 24 A and B). Three genes are involved in these regions which include HCG22 (peak variant rs2535307, $P_{\text{combined}}=3.72 \times 10^{-7}$), CCNG1 (peak variant kgp22385164, $P_{\text{combined}}=1.88 \times 10^{-6}$) and ZBTB7C (peak variant kgp7815029, $P_{\text{combined}}=1.12 \times 10^{-6}$). We also found 5 more SNPs below genome-wide significant level in pooled analysis that passed the false discovery rate adjustment in replication. The genes involved are PRKAA2, ARHGEF12, AK095365, ZFYVE28 (Table 7). Additionally, we found 2 more SNPs (close to genes of IPO8 and CYP1B1-AS1) reached genome-wide significance level in combined analysis but they are only significant in one cohort and one SNP lying in the intron of DPYD that is significant in the primary cohort but didn't reach MAF 5% in the replication cohort (Figure 24B, Table 9).
Figure 23. Average of trajectory of significant SNPs that are eliminated due to opposite direction in primary and secondary cohorts generated by fPCA model.
Yellow: The average trajectory of patients with 0 minor allele. Green: The average trajectory of patients with 1 minor allele.

Figure 24. GWAS results at 852889 SNPs with using combined VL and CD4 fPCA scores as phenotype under a dominant model.
A. Genes involved in the top 100 SNPs (smallest genome-wide $P_{\text{combined}}$) identified using fPCA model (red) or linear mixed model (Blue).
B. The Manhattan plot before adjusting for age and sex shown in combined $p$-value of BHP010 and BHP011. Six SNPs in independent regions reached genome wide significance level. Three SNPs (in purple circle) are significant in primary cohort of BHP011 and remain significant after replication in BHP010. For the other three SNPs (above significance level but not circled), though their combined $p$-value reached significance level, they are significant in only one cohort but not the other.
C-E. Linkage disequilibrium around significant SNP loci identified by fPCA model.
Table 6 Significant SNPs that have opposite effects on BHP010 and BHP011

<table>
<thead>
<tr>
<th>SNP</th>
<th>locus</th>
<th>Annotation</th>
<th>Nearest Gene</th>
<th>Model</th>
<th>MAF (primary)</th>
<th>P-value (replicate)</th>
<th>P-value (primary)</th>
<th>P-value (combined)</th>
</tr>
</thead>
<tbody>
<tr>
<td>rs1434384</td>
<td>4q32.1</td>
<td>14kb upstream intergenic</td>
<td>LOC340017</td>
<td>Linear/fPCA</td>
<td>0.446875</td>
<td>0.00047</td>
<td>2.43E-05</td>
<td>2.21E-07</td>
</tr>
<tr>
<td>kgp5554701</td>
<td>3q27.3</td>
<td>14kb upstream intergenic</td>
<td>LOC340017</td>
<td>Linear/fPCA</td>
<td>0.315625</td>
<td>0.008044</td>
<td>4.02E-06</td>
<td>5.90E-07</td>
</tr>
<tr>
<td>kgp7674952</td>
<td>4q22.1</td>
<td>intron</td>
<td>CCSER1</td>
<td>Linear/fPCA</td>
<td>0.06875</td>
<td>0.002022</td>
<td>5.75E-05</td>
<td>1.97E-06</td>
</tr>
<tr>
<td>rs6956267</td>
<td>7p21.3</td>
<td>25kb downstream intron</td>
<td>ARL4A</td>
<td>fPCA</td>
<td>0.159375</td>
<td>0.012766</td>
<td>2.21E-07</td>
<td>5.90E-07</td>
</tr>
<tr>
<td>kgp652371</td>
<td>4q11.2</td>
<td>intron</td>
<td>LOC340017</td>
<td>Linear/fPCA</td>
<td>0.082813</td>
<td>0.006201</td>
<td>1.78E-05</td>
<td>1.88E-06</td>
</tr>
<tr>
<td>rs13165594</td>
<td>5q34</td>
<td>intergenic</td>
<td>LOC340017</td>
<td>Linear/fPCA</td>
<td>0.0625</td>
<td>0.014882</td>
<td>9.68E-05</td>
<td>3.10E-05</td>
</tr>
<tr>
<td>kgp2018975</td>
<td>8q11.2</td>
<td>intergenic</td>
<td>LOC340017</td>
<td>Linear/fPCA</td>
<td>0.075</td>
<td>0.039468</td>
<td>4.79E-05</td>
<td>2.68E-05</td>
</tr>
<tr>
<td>kgp1078004</td>
<td>3p25.2</td>
<td>intron</td>
<td>SYN2</td>
<td>fPCA</td>
<td>0.107813</td>
<td>0.006795</td>
<td>0.00029</td>
<td>2.79E-05</td>
</tr>
<tr>
<td>kgp2275257</td>
<td>16p13.3</td>
<td>intron</td>
<td>SRL</td>
<td>fPCA</td>
<td>0.21875</td>
<td>0.056438</td>
<td>3.55E-05</td>
<td>1.77E-05</td>
</tr>
<tr>
<td>kgp6123710</td>
<td>4p11.2</td>
<td>intron</td>
<td>SRL</td>
<td>fPCA</td>
<td>0.06875</td>
<td>0.005555</td>
<td>2.55E-05</td>
<td>2.83E-05</td>
</tr>
<tr>
<td>rs6956267</td>
<td>7p21.3</td>
<td>25kb downstream intron</td>
<td>ARL4A</td>
<td>fPCA</td>
<td>0.159375</td>
<td>0.012766</td>
<td>2.21E-07</td>
<td>5.90E-07</td>
</tr>
<tr>
<td>kgp652371</td>
<td>4q11.2</td>
<td>intron</td>
<td>LOC340017</td>
<td>Linear/fPCA</td>
<td>0.082813</td>
<td>0.006201</td>
<td>1.78E-05</td>
<td>1.88E-06</td>
</tr>
<tr>
<td>rs13165594</td>
<td>5q34</td>
<td>intergenic</td>
<td>LOC340017</td>
<td>Linear/fPCA</td>
<td>0.0625</td>
<td>0.014882</td>
<td>9.68E-05</td>
<td>3.10E-05</td>
</tr>
<tr>
<td>kgp2018975</td>
<td>8q11.2</td>
<td>intergenic</td>
<td>LOC340017</td>
<td>Linear/fPCA</td>
<td>0.075</td>
<td>0.039468</td>
<td>4.79E-05</td>
<td>2.68E-05</td>
</tr>
<tr>
<td>kgp1078004</td>
<td>3p25.2</td>
<td>intron</td>
<td>SYN2</td>
<td>fPCA</td>
<td>0.107813</td>
<td>0.006795</td>
<td>0.00029</td>
<td>2.79E-05</td>
</tr>
<tr>
<td>kgp2275257</td>
<td>16p13.3</td>
<td>intron</td>
<td>SRL</td>
<td>fPCA</td>
<td>0.21875</td>
<td>0.056438</td>
<td>3.55E-05</td>
<td>1.77E-05</td>
</tr>
</tbody>
</table>

Table 7 SNPs significant* after replication using fPCA model (excluding X and Y chromosomes)

<table>
<thead>
<tr>
<th>SNP</th>
<th>locus</th>
<th>Annotation</th>
<th>Nearest Gene</th>
<th>Allele (ref/alt)</th>
<th>MAF (primary)</th>
<th>P-value (replicate)</th>
<th>P-value (primary)</th>
<th>P-value (combined)</th>
</tr>
</thead>
<tbody>
<tr>
<td>rs2535307</td>
<td>6p21.33</td>
<td>25kb downstream 20kb</td>
<td>HCG22</td>
<td>T/C</td>
<td>0.480456</td>
<td>2.39E-05</td>
<td>3.72E-07</td>
<td></td>
</tr>
<tr>
<td>kgp7815029</td>
<td>18q21.1</td>
<td>upstream</td>
<td>ZBTB7C</td>
<td>T/C</td>
<td>0.154688</td>
<td>0.064199</td>
<td>9.90E-07</td>
<td>1.12E-06</td>
</tr>
<tr>
<td>kgp2238516</td>
<td>5q34</td>
<td>5kb upstream</td>
<td>CCNG1</td>
<td>T/G</td>
<td>0.082813</td>
<td>0.006201</td>
<td>1.78E-05</td>
<td>1.88E-06</td>
</tr>
<tr>
<td>rs4912408</td>
<td>5p32.2</td>
<td>intron</td>
<td>PRKAA2</td>
<td>T/C</td>
<td>0.209375</td>
<td>0.00583</td>
<td>0.000109</td>
<td>9.73E-06</td>
</tr>
<tr>
<td>rs10892573</td>
<td>11q23.3</td>
<td>intron</td>
<td>ARHGEF12</td>
<td>T/C</td>
<td>0.141509</td>
<td>0.003182</td>
<td>0.000303</td>
<td>1.43E-05</td>
</tr>
<tr>
<td>kgp1882816</td>
<td>12p13.3</td>
<td>2</td>
<td>AK095365</td>
<td>T/C</td>
<td>0.126563</td>
<td>0.037386</td>
<td>5.62E-05</td>
<td>2.96E-05</td>
</tr>
<tr>
<td>kgp1101858</td>
<td>3</td>
<td>intergenic</td>
<td>ZFYVE28</td>
<td>A/G</td>
<td>0.063091</td>
<td>0.006051</td>
<td>0.000525</td>
<td>4.34E-05</td>
</tr>
</tbody>
</table>

* SNP p-value lower than 6x10^-4 in BHP011 (primary cohort) and passed FDR (α=0.1) in BHP010 (second independent cohort)
Table 8 Genome-wide significant loci that are significant in both cohorts by fPCA and Linear mixed phenotype simulation

<table>
<thead>
<tr>
<th>SNP</th>
<th>locus</th>
<th>Annotation</th>
<th>Nearest Gene</th>
<th>Allele (ref/alt)</th>
<th>MAF (primary)</th>
<th>P-value (primary)</th>
<th>P-value (replicate)</th>
<th>P-value (combined)</th>
<th>Phenotype simulation model</th>
</tr>
</thead>
<tbody>
<tr>
<td>rs2535307</td>
<td>6p21.33</td>
<td>25kb down</td>
<td>HCG22</td>
<td>T/C</td>
<td>0.48045</td>
<td>2.39E-05</td>
<td>0.00083</td>
<td>3.72E-07</td>
<td>fPCA</td>
</tr>
<tr>
<td></td>
<td></td>
<td>stream 5kb</td>
<td></td>
<td></td>
<td>0.14568</td>
<td>9</td>
<td>9.90E-07</td>
<td>1.13E-06</td>
<td></td>
</tr>
<tr>
<td>kgp7815029</td>
<td>18q21.1</td>
<td>downstream</td>
<td>ZBTB7C</td>
<td>T/C</td>
<td>0.08281</td>
<td>0.00620</td>
<td>1.78E-05</td>
<td>1.88E-06</td>
<td>fPCA</td>
</tr>
<tr>
<td></td>
<td></td>
<td>5kb down</td>
<td></td>
<td></td>
<td>0.47805</td>
<td>1.26E-04</td>
<td>3.28E-06</td>
<td>5.37E-09</td>
<td></td>
</tr>
<tr>
<td>kgp22385164</td>
<td>5q34</td>
<td>upstream</td>
<td>CCNG1</td>
<td>T/G</td>
<td>0.15468</td>
<td>0.06558</td>
<td>6.02E-06</td>
<td>6.22E-06</td>
<td></td>
</tr>
<tr>
<td>rs6921557</td>
<td>6p21.33</td>
<td>upstream</td>
<td>HCG22</td>
<td>T/C</td>
<td>0.15468</td>
<td>1</td>
<td>1.78E-05</td>
<td>1.3E-06</td>
<td></td>
</tr>
<tr>
<td>kgp7815029</td>
<td>18q21.1</td>
<td>downstream</td>
<td>ZBTB7C</td>
<td>T/C</td>
<td>0.48045</td>
<td>2.39E-05</td>
<td>0.00083</td>
<td>3.72E-07</td>
<td>fPCA</td>
</tr>
<tr>
<td></td>
<td></td>
<td>5kb down</td>
<td></td>
<td></td>
<td>0.14568</td>
<td>9</td>
<td>9.90E-07</td>
<td>1.13E-06</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>stream 5kb</td>
<td></td>
<td></td>
<td>0.47805</td>
<td>1.26E-04</td>
<td>3.28E-06</td>
<td>5.37E-09</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>upstream</td>
<td>CCNG1</td>
<td>T/G</td>
<td>0.15468</td>
<td>0.06558</td>
<td>6.02E-06</td>
<td>6.22E-06</td>
<td></td>
</tr>
</tbody>
</table>

Table 9 SNPs significant in one cohort that fail to replicate as significant in second cohort

<table>
<thead>
<tr>
<th>SNP</th>
<th>locus</th>
<th>Annotation</th>
<th>Nearest Gene</th>
<th>Model</th>
<th>MAF (BHP011)</th>
<th>Significant cohort</th>
<th>P-value in significant cohort</th>
<th>P-value in the other cohort</th>
<th>P-value (combined)</th>
</tr>
</thead>
<tbody>
<tr>
<td>rs11050931</td>
<td>12p11.2</td>
<td>100kb down</td>
<td>IPO8</td>
<td>FPCA/Linear</td>
<td>0.3843</td>
<td>BHP01</td>
<td>7.13E-08</td>
<td>0.5551</td>
<td>7.15E-07</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>25kb down</td>
<td>CYP1B1-AS1</td>
<td>FPCA/Linear</td>
<td>0.4593</td>
<td>BHP01</td>
<td>2.90E-07</td>
<td>0.4348</td>
<td>2.13E-06</td>
</tr>
<tr>
<td>rs7606492</td>
<td>2p22.2</td>
<td>downstream</td>
<td></td>
<td>FPCA/Linear</td>
<td>0.3156</td>
<td>BHP01</td>
<td>3.62E-05</td>
<td>0.0036</td>
<td>2.20E-06</td>
</tr>
<tr>
<td>rs6703307</td>
<td>1p21.3</td>
<td>intron</td>
<td>DPYD</td>
<td>FPCA</td>
<td>0.3156</td>
<td>BHP01</td>
<td>3.62E-05</td>
<td>0.0036</td>
<td>2.20E-06</td>
</tr>
<tr>
<td>kgp1867220</td>
<td>5p15.33</td>
<td>intergenic</td>
<td></td>
<td>Linear</td>
<td>0.3156</td>
<td>BHP01</td>
<td>3.62E-05</td>
<td>0.0036</td>
<td>2.20E-06</td>
</tr>
</tbody>
</table>

d. Adjusting by age and sex

We also adjusted the result by age and sex. All the SNPs discussed above still remain in the top 100 lowest by p-value after adjusting for sex and age (Figure 25). We excluded sex chromosome in all the analysis to avoid potential bias after adjusting for age and sex.
e. Using the dominant model

Our results are based on the dominant model (Bush and Moore, 2012). We chose to use the dominant model because most autosomal genetic traits are inherited in a dominant pattern. However there are also autosomal traits inherited by non-dominant patterns. So we replicated our study using a non-dominant model to see whether we can gain power for some loci. We found that using the non-dominant model decreased power significantly, but SNPs with the lowest p-values remain the same as with the regular model (Figure 26).
Figure 26. QQ plots and Manhattan plots using non-dominant model before stratifying by age and sex.
A. QQ plot using FPCA model before stratifying by age and sex. B. QQ plot using linear model before stratifying by age and sex. C. Manhattan plot using FPCA regression model before stratifying by age and sex. No SNP is significant in the non-dominant model.

f. Linear model for phenotype approximation

We also replicated the above association study using a traditional linear mixed model to estimate the CD4 and VL trajectories. Without adjusting for age and sex, we obtained 80 significant SNPs after FDR (a=0.1) adjustment lying in 5 independent regions of the human genome. These SNPs were in HCG22 (peak variant rs6921557, Pcombined=5.37×10^{-9}), ZBTB7C (peak variant kgp7815029, Pcombined=6.22×10^{-6}),
CCNG1 (peak variant kgp22385164, Pcombined=9.14×10^{-6}), TRIM26 (peak variant kgp10188320, Pcombined=1.45×10^{-5}) and ARL8B (peak variant kgp8910014, Pcombined=1.99×10^{-5}) (Figure 27). Only HCG22 and ZBTB7C reached genome-wide significance level after FDR(a=0.1, p-value lower than 7.48×10^{-6}) (Table 8). It is worth noticing that though genes with the smallest p-values largely overlap in results obtained using the fPCA model and linear mixed model (Figure 24A), the linear mixed model has less power reflected by larger SNPs p-values for the same SNPs compared to the fPCA model, except for HCG22 (Table 7, Table 10). The result using the linear mixed model further confirms the association we observed using the fPCA model. Though the linear mixed model gives us a much lower p-value for HCG22, we generally gained power to detect other significant SNPs using the fPCA model.
Figure 27. QQ plots and Manhattan plots using different phenotype simulation models before adjusting for age and sex (excluding sexual chromosome).
All of the plots used the dominant model. A. QQ plot using fPCA model before adjusting for age and sex. B. QQ plot using linear model before adjusting for age and sex. C. Manhattan plot using linear mixed model before adjusting for age and sex.

Table 10 SNPs significant* after replication using linear mixed model (excluding X and Y chromosomes)

<table>
<thead>
<tr>
<th>SNP</th>
<th>locus</th>
<th>Annotation</th>
<th>Nearest Gene</th>
<th>Allele (ref/alt)</th>
<th>MAF (primary)</th>
<th>P-value (primary)</th>
<th>P-value (replicate)</th>
<th>P-value (combined)</th>
</tr>
</thead>
<tbody>
<tr>
<td>rs6921557</td>
<td>6p21.33</td>
<td>20kb upstream</td>
<td>HCG22</td>
<td>T/C</td>
<td>0.480456</td>
<td>1.26E-04</td>
<td>3.28E-06</td>
<td>5.37E-09</td>
</tr>
<tr>
<td>kgp78150</td>
<td>18q21.1</td>
<td>5kb upstream</td>
<td>ZBTB7C</td>
<td>T/C</td>
<td>0.154688</td>
<td>0.070688</td>
<td>6.22E-06</td>
<td></td>
</tr>
<tr>
<td>kgp22385</td>
<td>5q34.3</td>
<td>intron</td>
<td>CCNG1</td>
<td>T/G</td>
<td>0.082813</td>
<td>0.011275</td>
<td>9.14E-06</td>
<td></td>
</tr>
<tr>
<td>kgp10188</td>
<td>6p22.1</td>
<td>intron</td>
<td>TRIM26</td>
<td>T/C</td>
<td>0.440625</td>
<td>0.065684</td>
<td>1.45E-05</td>
<td></td>
</tr>
<tr>
<td>kgp89100</td>
<td>3p21.1</td>
<td>50kb upstream</td>
<td>ARL8B</td>
<td>A/G</td>
<td>0.1125</td>
<td>3.22E-03</td>
<td>0.000173</td>
<td>1.99E-05</td>
</tr>
</tbody>
</table>
 SNPs p-value lower than $6 \times 10^{-4}$ in BHP011 (primary cohort) and passed FDR ($\alpha=0.1$) in BHP010 (second independent cohort)

**g. Phenotype of patients with and without minor alleles**

We further examined the CD4 and VL trajectories in patients with different variations of the significant SNPs. Of the three significant SNPs identified by the fPCA model, the trajectories of patients without minor alleles and with at least one minor allele display significant differences (Figure 28, Figure 29, Figure 30). This effect is reflected by the most significant difference in their first Eigenfunction score and moderate differences in their second Eigenfunction score (Figure 29, Figure 30). Using the linear mixed model can also differentiate phenotypic difference but with lower resolution (Figure 31). All the three SNPs affect CD4 and VL in an opposite way: minor alleles of rs2535307 and kgp22385164 are associated with lower CD4 and higher VL suggesting they are risk alleles, and minor allele of kgp7815029 (chr18, pos45548347, MAF 15%, $P_{fPCA}=6.22E-06$) is associated with higher CD4 and lower VL suggesting it is a protective allele (Figure 28). This result is biologically meaningful.

![Figure 28](image-url)

**Figure 28.** Average trajectories of VL and CD4 are highly corrected with the top significant SNP using dominant model.
Average trajectory of patients with 0 minor alleles (red) or at least 1 minor allele (blue) of significant loci of A.
rs2535307 (chr6, pos31053741, MAF 48%, P_{pca}=3.72E-07), B. kgp7815029(chr18, pos45548347, MAF 15%, P_{pca}=6.22E-06), C. kgp22385164(chr5, pos162859382, MAF=8%, P_{pca}=9.14E-06). Primary cohort of BHP011 and replication cohort of BHP010 are shown separately. Having at least one copy of minor allele of rs2535307 and kgp22385164 is associated with lower CD4 and higher VL (A and C), reflected by lower value of Eigenfunction 1 of CD4 and higher Eigenfunction 1 of VL compared to patients with zero minor alleles in both BHP010 and BHP011 (Figure 29A, Figure 29C). Minor allele of kgp7815029 is also associated with higher CD4 and lower VL.

**Figure 29. Coefficients measuring the trajectory of VL and CD4 are highly corrected with the top significant SNP using dominant model.**

Clinical presentation of patients with alleles of significant loci of A. rs2535307 (chr6, pos31053741, MAF 48%, P_{pca}=3.72E-07), B. kgp7815029(chr18, pos45548347, MAF 15%, P_{pca}=6.22E-06), C. kgp22385164 (chr5, pos162859382, MAF=8%, P_{pca}=9.14E-06). Average fPCA scores Eigenfunction 1 (represents levels of maintaining CD4 or VL), Eigenfunction 2 (represents CD4 or VL decline rate if negative and increase rate if positive), Eigenfunction 3 (variation of the CD4 or VL trajectories). Primary cohort of BHP011 and replication cohort of BHP010 are shown separately. Having at least one copy of minor allele of rs2535307 and kgp22385164 is associated with lower CD4 and higher VL (Figure 28A and Figure 28C). This is also reflected by lower value of Eigenfunction 1 of CD4 and higher Eigenfunction 1 of VL compared to patients with zero minor allele in both BHP010 and BHP011. Minor Allele of rs2535307 and kgp22385164 is also associated with faster CD4 decline and smaller VL changing rate reflected by a more negative Eigenfunction 2 of CD4 and the VL Eigenfunction 2 is closer to zero compared to patients with zero minor allele in both BHP010 and BHP011. The magnitude of effect on Eigenfunction 1 is much larger than Eigenfunction 2, indicating the variants of rs2535307 and kgp22385164 affect the level of CD4 and VL more than changing rate. Similarly, minor allele of kgp7815029 is associated with higher CD4 level and lower VL level indicating their protective role.

**Figure 30. Coefficients measuring the trajectory of VL and CD4 are highly corrected with the top significant SNP using dominant model.**
A. The first coefficient of Eigenfunction measuring the average level of CD4 overtime. B. The second coefficient of Eigenfunction measuring the declining rate of CD4 over time. C. The first coefficient of Eigenfunction measuring the average level of VL overtime. B. The second coefficient of Eigenfunction measuring the increase rate of VL over time.

Green: Patients with at least one minor allele of “G”. Red: Patients with no minor allele.

Figure 31. Phenotype presentation of patients with different alleles of top SNPs identified by linear mixed.

Clinical presentation of patients with alleles of significant loci of A. rs6921557 (chr6, pos31049022, MAF 48%, P_{lin}=5.37E-09) B. kgp7815029(chr18, pos45548347, MAF 15%, P_{fpca}=6.22E-06), C. kgp22385164(chr5, pos162859382, MAF=8%, P_{fpca}=9.14E-06), D. kgp10188320, E. kgp8910014. Top: Average VL scores of Intercept (represents baseline), Slope (represents decline rate if negative and increase rate if positive) of patients with 0 minor allele (yellow) or at least 1 minor allele (green); Bottom: Average VL scores. Primary cohort of BHP011 (right) and replication cohort of BHP010 (left) are shown separately.
III.5 Discussion

In this study, we applied both functional principal component analysis and the traditional linear mixed model to define AIDS disease trajectory using longitudinal CD4 and VL. We discovered several new SNPs significantly associated with AIDS disease progression in HIV-Subtype-C infected Botswana treatment naive patients. Using functional principal component analysis, we were able to gain power and identify SNPs that may not be identified using the traditional linear mixed model alone.

a. HCG22 involved in HIV restriction in African population but not others

In both the linear mixed model as well as the fPCA, HCG22 is significant. Previously the SNP rs9262632 in the intron region of HCG22 has been shown to be associated with HIV disease progression in the African American population (International et al., 2010). Our study suggested significant loci of rs2535307(chr6, pos31053741, MAF 48%, \( P_{\text{fPCA}}=3.72E^{-07} \)) and rs6921557 (chr6, pos31049022, MAF 48%, \( P_{\text{lin}}=5.37E^{-09} \)) near HCG22 which have not been shown before. Patients without the minor allele displayed a significantly higher level of CD4 and lower level of VL (Figure 28A, Figure 31A). It seems the SNP is more related to the level of CD4 and VL more than a decline/increase rate over time. We found that our peak SNP near HCG22 is in high LD with rs9262632 and also the genes within 200kb (Figure 24C) including PSORS1C1, CDSN, CCHCR1, TCF19, POU5F1, PSORS1C3, MUC22 and MUC21. Despite previous GWAS studies identifying HLA-B, HLA-C and HCP5 as strongly associated with AIDS progression in European men (Fellay et al., 2007; Dalmasso et al., 2008; Fellay et al., 2009; Limou et al., 2009; International et al., 2010), it is likely that HCG22 is more significantly related to AIDS progression in African populations. Additionally, the gene cluster of HCG22 has
been reported to be associated total white blood cell counts (Nalls et al., 2011), susceptibility to follicular lymphoma (Skibola et al., 2009) and multiple myeloma (Chubb et al., 2013) in European individuals, suggesting its role in immune cell cycle regulation. This further supported our finding that HCG22 affects CD4 level more than VL. Another GWAS study showed that HCG9, a gene located approximately 1MB upstream of HCG22 on chromosome 6, was a proxy for HLA- A*11:01(Tang et al., 2012). Exon sequencing is needed in the surrounding regions to further identify the causal SNPs for AIDS disease progression in African populations.

b. P21/WAF1 pathway may be highly involved in HIV host restriction in the African population

Interestingly, the other two of our top hits were both on the p53 and p21 pathway that affects cell cycle. Kgp7815029(chr18, pos45548347, MAF 15%, P_{fpca}=6.22E-06) is 5kb downstream of the ZBTB7C gene and is in LD with the intron and coding region of ZBTB7C (Figure 24D). Patients without a minor allele displayed lower CD4 and higher VL compared to patients with at least one minor allele (Figure 28B, Figure 31B). ZBTB7C (APM-1, KR-POK) encodes a zinc finger transcriptional factor that was first described as cotranscribed with HPV DNA in the cervical carcinoma cell line ME180 (Reuter et al., 1998). It has been reported to interact with p53 and regulate the transcription of CDKN1A and p21WAF1, both of which control for cell proliferation(Jeon et al., 2012a). Both p53(Duan et al., 1994) and p21 WAF1 have been shown to be regulated by tat, vpr and SP1 collaboratively (Sawaya et al., 1998; Amini et al., 2004b) and affect cell cycle (Cherrier et al., 2009). ZBTB7C also interacts with SP1
directly (Jeon et al., 2012b), a transcriptional regulator that directly regulates HIV transcription via interaction with tat (Kamine et al., 1991; Sune and Garcia-Blanco, 1995; Zhang et al., 2000; Yedavalli et al., 2003; Gee et al., 2007), vpr (Wang et al., 1995; Deshmene et al., 2009) and gag (Zhang et al., 2000). Kgp22385164 (chr5, pos162859382, MAF=8%, P_{pca}=9.14E-06) is 5kb upstream of the CCNG1 gene and in LD with the coding region of CCNG1 (Figure 24E). Patients without the minor allele had lower VL and higher CD4 (Figure 28C, Figure 29C, Figure 31C). CCNG1 is a member of cyclin family and shares 53% sequence identity with its closest family member CCNG2, which was shown to promote HIV replication by shRNA screen in Jurkat cells (Yeung et al., 2009). Despite CCNG2, multiple other cyclin family members such as CDKN1A (Ryu et al., 2004; Kapasi et al., 2006; Thakur et al., 2012), CDK7 (Nekhai et al., 2002), CCNH (Kino et al., 2002), CCND1(Nelson et al., 2002) have been reported to play roles in HIV replication. HIV vpr protein has been shown to arrest the cell cycle by interacting with the cyclins and affecting the function of p53/p21 pathway (Chowdhury et al., 2003; Amini et al., 2004a). It is likely that variation in SNPs close to ZBTB7C and CCNG1 affect CD4 and VL trajectories of AIDS patients via regulation of HIV transcription and the cell cycle of infected cells through the SP1, p53 and p21 pathway. Our association study further raised the possibility of an important role of multiple cross talk between HIV proteins, the p53/p21 pathway and cyclins to regulate HIV replication as well as cell cycle in vivo reflected by both viral load and CD4 counts.

c. Multiple genes below significance level may still be associated with host control of HIV/AIDS progression
We also detected several SNPs with combined P-values below a significant level that consistently have small P-values in both cohorts ($P_{\text{combined}}<4.4\times10^{-5}$, with both cohort’s P value $<0.05$) (Table 11). PRKAA2 can phosphorylate HIV tat (Zhang and Wu, 2009; Zhang et al., 2011), vif (Shirakawa et al., 2008) and vpr (Barnitz et al., 2010; Francis et al., 2011) and regulate their activity directly. ARL8B is a member of the ADP-ribosylation factor family of GTP-binding proteins in which ARL1(Wonderlich et al., 2011) and ARL6 (Blagoveshchenskaya et al., 2002) have been reported to collaborate with HIV- nef to disrupt MHC-1 surface level in HIV infected T cells. ARHGEF12 is a Rho GTPase which has also been shown to interact with HIV pol in both Jurkat and 293T cells (Jager et al., 2012). ALK is a tyrosine kinase protein that has been reported to interact with HIV- nef and played roles in HIV associated dementia (Bergonzini et al., 2009). TRIM family proteins have been reported to be highly involved in innate immunity modulation (Ozato et al., 2008) and anti-retroviral defenses (Nisole et al., 2005) and in our list there are two TRIM proteins that were identified: TRIM26 and TRIM16. TRIM26 is located in the MHC region on chromosome 6 and has been suggested to have roles in inhibiting HIV release. Importantly, TRIM26 was shown as an anti-HIV restriction factor both in vitro and in vivo such that expression level of TRIM26 has been shown to be elevated in HIV elite controllers compared to HAART suppressed individuals (Abdel-Mohsen et al., 2013) and expression of TRIM26 is correlated with decreased HIV infectivity in PBMC (Raposo et al., 2013).

<table>
<thead>
<tr>
<th>SNP</th>
<th>locus</th>
<th>Annotation</th>
<th>Nearest Gene</th>
<th>Model</th>
<th>MAF (primary)</th>
<th>P-value (replicate)</th>
<th>P-value (primary)</th>
<th>P-value (combined)</th>
</tr>
</thead>
<tbody>
<tr>
<td>rs4912408</td>
<td>1p32.2</td>
<td>intron</td>
<td>PRKAA2</td>
<td>fPCA</td>
<td>0.209375</td>
<td>0.00583</td>
<td>0.000109</td>
<td>9.73E-06</td>
</tr>
<tr>
<td>kgp14514681</td>
<td>2q22.1</td>
<td>100kb upstr am</td>
<td>HNMT</td>
<td>fPCA</td>
<td>0.095313</td>
<td>2.34E-05</td>
<td>0.029334</td>
<td>1.04E-05</td>
</tr>
<tr>
<td>rs680168</td>
<td>10q26. 2</td>
<td>20kb upstream</td>
<td>CLRN3</td>
<td>Linear</td>
<td>0.420313</td>
<td>0.000607</td>
<td>0.001146</td>
<td>1.17E-05</td>
</tr>
<tr>
<td>rs1089257</td>
<td>11q23. 3</td>
<td>intron</td>
<td>ARHGEF 12</td>
<td>fPCA</td>
<td>0.141509</td>
<td>0.003182</td>
<td>0.000303</td>
<td>1.43E-05</td>
</tr>
<tr>
<td>kgp10188320</td>
<td>6p22.1</td>
<td>intron</td>
<td>TRIM26</td>
<td>Linear</td>
<td>0.440625</td>
<td>0.065684</td>
<td>0.000302</td>
<td>1.74E-05</td>
</tr>
<tr>
<td>kgp11221083</td>
<td>16q23. 2</td>
<td>intergenic</td>
<td>50kb upstream</td>
<td>ARL8B</td>
<td>Linear</td>
<td>0.1125</td>
<td>0.003224</td>
<td>0.000173</td>
</tr>
<tr>
<td>kgp8910014</td>
<td>3p26.1</td>
<td>intron</td>
<td>TRIM16</td>
<td>fPCA/Linear</td>
<td>0.121875</td>
<td>0.000228</td>
<td>0.006107</td>
<td>2.02E-05</td>
</tr>
<tr>
<td>rs1563011</td>
<td>17p12</td>
<td>3p26.1</td>
<td>intron</td>
<td>TRIM16</td>
<td>fPCA/Linear</td>
<td>0.126563</td>
<td>0.037386</td>
<td>5.62E-05</td>
</tr>
<tr>
<td>kgp18828163</td>
<td>12p13. 32</td>
<td>6p24.1</td>
<td>intergenic</td>
<td>AK095365</td>
<td>fPCA</td>
<td>0.126563</td>
<td>0.037386</td>
<td>5.62E-05</td>
</tr>
<tr>
<td>rs12111444</td>
<td>Xq21. 32</td>
<td>4p16.3</td>
<td>exon</td>
<td>PCDH11 X</td>
<td>fPCA</td>
<td>0.184375</td>
<td>0.018558</td>
<td>0.000167</td>
</tr>
</tbody>
</table>

### Table 11 The promising SNPs below genomic significant level (Continued)

### d. PCDH11X on X chromosome may be associated with HIV restriction in vivo

We excluded X and Y chromosomes in our association analysis to avoid potential bias if adjusting for sex and age. We also replicated the study including the sex chromosomes on 852889 SNPs and before adjusting for age and sex, we found an additional interesting SNP lying in the intron of PCDH11X (peak variant kgp22829958, $P_{combined}=4.24\times10^{-5}$) (Table 11). PCDH11X is a member of protocadherin family that knock down PCDH11X via siRNA and has been shown to affect HIV transcription in TZM-bl cells (Brass et al., 2008). After adjusting for age and sex, we got an additional 3 new SNPs with low P-
value (错误！未找到引用源。). The top SNPs before adjustment were unaffected by adjustment (Figure 32C).

**Figure 32.** The QQ plot and Manhattan plot after adjusted by age and sex if including sex chromosome in association analysis (dominant model) for fPCA. A. QQ plot using fPCA model after adjusting for age and sex, including X and Y chromosomes. B. QQ plot using linear model after adjusting for age and sex, including X and Y chromosomes. C. Manhattan plot using fPCA model after adjusting for age and sex, including X and Y chromosomes. Top SNPs identified before stratification still remain top. Four independent SNPs passed significant thresholds and HCG22 is the only one with a consistent small p-value in both cohorts. All sex chromosome SNPs failed to reach the genome-wide significant level after gender and age adjustment.

e. **New SNPs with small p-value after adjusting for age and sex**

Previous studies of GWAS for both the "Centre for HIV/AIDS Vaccine Immunology study "(CHAVI) Study (Fellay J, 2007) and the microsatellite-based GWAS of SIV infected monkeys showed that an unknown X chromosome variation is associated with slow AIDS progression defined by a lower viral load at set point and slower CD4 decline.
Strikingly, this effect is only in HIV infected females but not males (Siddiqui RA, 2009). Aging is also associated with alterations in T-cell function and decreasing ability to reconstitute CD4 T-cells in HIV patients after antiviral treatment (Rickabaugh TM, 2010; Appay V, 2011). These pieces of evidence suggest gender and age are potential effect modifiers for host genes associations with CD4 trajectory.

We also found three interesting SNPs with small p-values after adjusting for age and sex, indicating these SNPs may affect HIV host control in specific gender or age groups. The genes involved are mir4500HG (peak variant kgp16854268, Pcombined=4.35×10^{-6}), CHIAP2 (peak variant rs3122595, Pcombined=1.77×10^{-5}) and FHit (peak variant kgp5164069, Pcombined=2.02×10^{-5}).

mir4500HG is a ncRNA that hasn't been linked to a particular function. CHIAP2 is a pseudogene of chitinase and also has no report of function. Chitinases have been reported to have important roles in plants and insects (Nagpure et al., 2013). Humans have 8 Chitinases (GH18) but most bear essential mutations on a glutamic acid residue at the catalytic domain so the protein is not functional (Guan et al., 2009). FHit (fragile histidine triad) encodes a diadenosine 5',5''-P1,P3-triphosphate hydrolase involved in purine metabolism. It has been identified as a tumor suppressor gene in multiple cancer types (Wu et al., 2003; Cecener et al., 2010; Al-Temaimi et al., 2013). Yu GR et.al used a HIV-tat fused FHit to deliver the pro-apoptotic protein to hepatocytes and showed inhibition of hepatocellular carcinoma(Yu et al., 2012). However there hasn't been any report of FHit's role in HIV restriction.
f. Using fPCA allows better simulation of small variations in longitudinal AIDS clinical data and identifies new SNPs

We noticed that except for HCG22, all other novel SNPs we discovered haven't been reported to be associated with AIDS disease phenotypes in earlier reports. We thought this may be due to two reasons. Firstly, we applied a new mathematical method (fPCA) which can better adapt the variation of trajectory. Previous GWAS studies applied linear mixed method to obtain individual viral loads or CD4 count slopes which assumed linear trajectory shape and usually excluded certain CD4 and VL counts manually to best fit for the assumed linear trajectory (Fellay et al., 2007). Linear shape displays the lowest degree of freedom to capture the variation. However the pathogenesis of HIV is complicated in that both CD4 and VL trajectories can be non-linear. Consequently the linear mixed method loses resolution and scores by the linear mixed method are more subject to misclassification of patients' real phenotypes (SE, 2012). Secondly, given that the SNPs reported before were mostly shown by GWAS conducted on Caucasian men infected with HIV-1-subtype B, it is possible that host genome uses some different mechanisms in controlling HIV-1-Subtype-C and the larger variation within the African population enables us to discover new SNPs. We raised this hypothesis because even though we
replicated the previous study using their reported linear mixed method (Fellay et al., 2007) on our cohorts, the p-value of SNPs suggested by previous GWAS that strongly associated with AIDS VL or disease progression (rs9264942, rs9261174 and many other SNPs identified to be associated with AIDS progression or CD4/VL levels) (Fellay et al., 2007; van Manen et al., 2012) were not among the top100000 smallest p-values in our study (Table 13). We failed to examine the top hit in previous study of rs2395029 due to the fact that this SNP's MAF is lower than 0.05 (Table 13) in both of our cohorts.
Table 13 P-value of SNPs suggested before by other GWAS

<table>
<thead>
<tr>
<th>SNP</th>
<th>Phenotype</th>
<th>Ethnicity</th>
<th>Nearest Gene</th>
<th>Linear model Combined p value</th>
<th>fPCA model Combined p value</th>
<th>Cohort</th>
</tr>
</thead>
<tbody>
<tr>
<td>rs2395029</td>
<td>RNA VL set point, plasma HIV RNA primary infection, cellular HIV DNA primary infection, long-term non-progression, VL controllers</td>
<td>Caucasian</td>
<td>HCP5</td>
<td>Minor allele frequency &lt;0.05</td>
<td></td>
<td>CHAVI, ANRS PRIMO, GRIV</td>
</tr>
<tr>
<td>rs2649442</td>
<td>RNA VL set point, VL controllers</td>
<td></td>
<td>HLA-C</td>
<td>0.856069</td>
<td>0.869542</td>
<td>CHAVI</td>
</tr>
<tr>
<td>rs9261174</td>
<td>CD4 T cell decline</td>
<td>ZNRD1</td>
<td>0.076274</td>
<td>0.070016</td>
<td></td>
<td>CHAVI</td>
</tr>
<tr>
<td>rs9368699</td>
<td>long-term non-progression</td>
<td>C6orf48</td>
<td>0.328242</td>
<td>0.288011</td>
<td></td>
<td>GRIV</td>
</tr>
<tr>
<td>rs4118325</td>
<td>rapid progression</td>
<td></td>
<td>PRMT6</td>
<td>0.441716</td>
<td></td>
<td>GRIV</td>
</tr>
<tr>
<td>rs1522232</td>
<td>rapid progression</td>
<td></td>
<td>SOX5</td>
<td>0.616375</td>
<td></td>
<td>GRIV</td>
</tr>
<tr>
<td>rs17762192</td>
<td>progression to AIDS</td>
<td></td>
<td>PROX1</td>
<td>0.649821$</td>
<td>0.261922$</td>
<td>MACS</td>
</tr>
<tr>
<td>rs2234358</td>
<td>long-term non-progression</td>
<td></td>
<td>CXCR6</td>
<td>0.441716</td>
<td></td>
<td>GRIV, MACS, ACS</td>
</tr>
<tr>
<td>rs152363</td>
<td>progression to AIDS, or AIDS-related death</td>
<td></td>
<td>AGR3</td>
<td>0.616375</td>
<td>0.641990</td>
<td>ACS</td>
</tr>
<tr>
<td>rs4418214</td>
<td>VL controllers</td>
<td></td>
<td>MICA</td>
<td>Minor allele frequency &lt;0.05</td>
<td></td>
<td>International HIV controllers study</td>
</tr>
<tr>
<td>rs3131018</td>
<td>VL controllers</td>
<td></td>
<td>PSORS1C3</td>
<td>0.817738</td>
<td></td>
<td>International HIV controllers study</td>
</tr>
<tr>
<td>rs2523608</td>
<td>RNA VL set point</td>
<td>African</td>
<td>HLA-B*5703</td>
<td>0.221724</td>
<td>0.239658</td>
<td>DoD HIV NHS and</td>
</tr>
<tr>
<td>rs11884476</td>
<td>progression to AIDS</td>
<td>African</td>
<td>PARD3B</td>
<td>0.105197$</td>
<td>0.153524$</td>
<td>MACS, ACS</td>
</tr>
</tbody>
</table>

$^3$The original SNP rs17762192 is not on the Chip. We used the closest SNP kgp15424027 which is 103 bp upstream of rs17762192.

$^6$The original SNPs rs1522232 is not on the Chip. We used the closest SNP kgp1244933 which is 673 bp upstream of rs11884476.
Chapter IV  Discussion
IV.1 Using human pluripotent stem cell as an in vitro model of HIV latency

The concept of a latent reservoir for HIV arose in the 1990s from clinical observations of residual cells harboring the HIV proviral genome and the treatment of AIDS patients that could maintain low viral load for a period of time. The natural HIV reservoir was first identified as resting CD4+ T cells (Chun et al., 1997). These cells have a very slow decay rate so that they serve as a major barrier to viral eradication by HAART (Finzi et al., 1999; Zhang et al., 1999a; Ramratnam et al., 2000; Siliciano et al., 2003; Strain et al., 2003; Han Y, 2007b). Besides resting CD4+ T cells, it has been suggested that other cell types can also serve as HIV proviral reservoirs including hematopoietic stem cells (Carter CC, 2010), macrophages (Zhang et al., 1999b; Igarashi et al., 2001), dendritic cells (McIlroy et al., 1995; Donaghy et al., 2001) and monocytes (Lambotte et al., 2000; Sonza et al., 2001; Zhu et al., 2002; Ellery et al., 2007; Jaworowski et al., 2007; Hasegawa et al., 2009). However latently infected cells represent a very low fraction of all blood cells (for CD4+ T cells about 1 integrated copy of HIV in 10000 CD4+ T cells) (Chun et al., 1995). Additionally, the definition of "latently infected cells" states that the infected cells are reversibly infected. Such cells produce little HIV-1 mRNA, and the mRNA produced are usually either prematurely terminated (Brooks et al., 2003; Hermankova et al., 2003) or mislocalized (Kao et al., 1987; Adams et al., 1994; Lassen et al., 2004). Therefore the latent cells have no viral proteins as markers and can only be distinguished from uninfected cells by integrated proviral DNA. The latent reservoirs may get reactivated under certain conditions and produce infectious viral particles.
Because of the importance of HIV reservoirs as an obstacle to curing AIDS, it is crucial to have more in vitro models that allow scientists to study the mechanism of HIV latency. The exploration of reactivation pathways for drug screening may lead to new strategies for treatment. The most ideal in vitro model of HIV latency might be cells from patients that harbor latent HIV genomes, especially if the cells were available for culturing and proliferation in vitro to perform intervention experiments. However, due to the fact that we still don't have an effective tool to survey the proviral HIV integrated at a single cell level without killing the cells, that the frequency of latent cells are extremely low in vivo, and that latent cells cannot be effectively propagated in vitro, new approaches are needed to provide systems for analysis.

Surrogate cell models, such as Jurkat cells, Hela cells and 293T cells, have been widely used to study HIV latency. In fact, many important discoveries of HIV latency (mostly at epigenetic level) were made using those cell types (reviewed in II.2). However it has been argued that the active machinery in transformed cell types are different from untransformed cells (Chen and Dent, 2014). Therefore the relative value of findings derived from the in vitro cancer cell models remains to be decided.

Human pluripotent stem cells are not a natural host for HIV due to the lack of CD4 HIV entry receptors and the co-receptors. There has been no evidence showing HIV is able to infect embryonic stem cells during natural infection. It has also been shown that sperm are also not targets of HIV infection (Kim et al., 1999), though washing sperm has been a standard practice to decrease the chance of having HIV positive babies born to
serodiscordant-couples (Pasquier et al., 2000; Garrido et al., 2006). Therefore our rationale of using human embryonic stem cells was not to mimic natural infection.

What makes human pluripotent stem cells appealing as a new model for HIV latency are the following characteristics:

i). Theoretically human pluripotent stem cells are able to differentiate into all the lineages of adult cells (Murry and Keller, 2008; Hanna et al., 2010; Yabut and Bernstein, 2011), including all the natural host cell types of that can be infected by HIV. Due to the limited knowledge of developmental biology and cell differentiation techniques, we cannot derive all the host cell types that HIV infects in vitro, but the current cell types that we can derive from human pluripotent stem cells include hematopoietic progenitor cells (Vodyanik et al., 2005; Doulatov et al., 2013), T cells (Galic et al., 2006), macrophages (Senju et al., 2011; van Wilgenburg et al., 2013) and dendritic cells (Li et al., 2014).

Potentially we will be able to derive some natural reservoirs of HIV using the iPS cell model we generated that harbors latent HIV. Such a model could perhaps also be used for drug screening in vitro.

ii). The pluripotent stem cells' epigenetic signature is similar to that of hematopoietic stem cells (Spivakov and Fisher, 2007). The epigenetic signature on the genome of human pluripotent stem cells displays a bivalent profile that many markers such as H3K9ac, H3K4me and H3K27me tend to be more enriched (Efroni et al., 2008; Cui et al., 2009). Many essential HIV transcription activators or repressors such as SUV39H1 (du Chene et al., 2007) and HDACs (Zhang et al., 2011) are also actively participating in establishing
the chromatin signature of pluripotent stem cells or hematopoietic stem cells. Due to the fact that hematopoietic stem cells cannot effectively be cultured ex vivo, the embryonic stem cells used represent the best approximation to hematopoietic stem cells.

To better adapt the model to studying patient-specific mechanisms for HIV latency, we proposed to use patient cell-derived iPS cells generated from PBMC that harbor HIV viruses. Based on our study, we found that the embryonic stem cells were able to silence the full-length HIV virus. After reprogramming the blood cells from patients into the iPS stage, we can further re-differentiate the patient-derived iPS towards blood cell lines again (Figure 33). Those cells will harbor HIV proviral genes, and patient specific genetic mutations. According to our finding of long-term HIV latency, those proviruses seem likely to be silenced in the re-differentiated blood cells. By doing so we will be able to generate patient-specific HIV reservoirs that harbor latent HIV. Importantly this strategy may allow us to expand patient-specific iPS cells \textit{in vitro} infinitely and derive the HIV blood cell reservoirs for research and drug screen usage (Figure 33).

Theoretically, we will be able to perform personalized drug screens on molecules that will potentially be able to reactivate the HIV latent viruses on these patient-derived HIV reservoirs. However, many technical difficulties may still need to be overcome with this strategy and more research is necessary.
IV.2 Implication of host restriction to HIV replication by transcription factors

We suggested that ZBTB7C and CCNG1 are associated with HIV disease progression in treatment naïve African patients infected with HIV-1C. As discussed in III.5b, these two genes are both in the p53 and p21 pathway that affect cell cycle. Though there hasn't been direct reports showing these two genes are transcriptional factors, they have been shown to interact with p53 and p21, two very important transcriptional factors regulating the cell cycle.
Despite the fact that most of the GWAS on HIV identified hits around HLAs, previous GWAS studies have also identified several transcriptional factors including ZNRD1 (Fellay et al., 2007), PARD3B (Chinn et al., 2010), SOX5 (Le Clerc et al., 2009) and PROX1 (Herbeck et al., 2010). In vitro, many other transcriptional factor pathways including the NF-KB pathway (Antoni BA, 1994), Sp1 (Kamine et al., 1991), activator protein 1 (AP-1) proteins (Gonzalez et al., 2001), nuclear factor of activated T cells (NFAT) (Lahti et al., 2003), and CCAAT enhancer binding protein (C/EBP) (Dahiya et al., 2014) were identified to be key regulators of HIV transcription. However in this in vivo association study we only identified two genes involved in transcriptional regulation, both in the same p21/p53 pathway. Our result raises the importance of p21/p53 pathway in regulating HIV transcription in vivo. However, other pathways that were shown to be important in vitro were not among the top hits in vivo, despite the sophisticated cross-talk among these pathways. We think this may be due to two reasons: First, many of these genes may be crucial for cell function so that variation in the genes that affect protein functions will be very rare in the population. Second, HIV genome transcription is governed by multiple transcriptional pathways as discussed above. Because of the cross-talk among these pathways and the redundancy of the players, the effect of mutations in one pathway may be rescued by another so that the effect of transcriptional factor variation may be subtle and hard to detect at the level of CD4 and VL. Our novel application of FPCA is able to capture minor variations in CD4/VL trajectories that the linear mixed model cannot, yet some effects may still be too subtle for detection and more parameters need to be measured in order to better define the phenotype. We,
however, haven't looked at proviral load that may be one parameter we can measure to study the effect of HIV host control at the transcriptional level.

Additionally, though there has been a report showing that ZBTB7C is cotranscribed with HPV DNA in the cervical carcinoma cell line ME180 (Reuter et al., 1998) and it may interact with p53 to regulate p21WAF1 which in turn interacts with multiple HIV accessory proteins (reviewed in III.5b), there hasn't been a report on ZBTB7C and HIV replication in vitro. Similarly, for CCNG1, though its close family member CCNG2 was shown to promote HIV replication by shRNA screen in Jurkat cells, there has been no report of CCNG1 function in HIV transcriptional control. It will be interesting to check these two proteins in vitro to test their functionality in HIV transcription restriction.

In II.4j, using in vitro experiments, we identified that the HIV accessory protein Nef serves upstream of PAK2 to act as a transcriptional regulator. HIV Nef, activated by myristoylation by NMT1 and NMT2, activates PAK2 and other host factors such as Hck. Full length PAK2 is autophosphorylated when activated by p21 (Pirruccello et al., 2006) and Nef. It is possible that downstream transcriptional factors regulated by PAK2 interact with the HIV genome and modulate HIV transcription in human pluripotent stem cells (Figure 34).

Combining our in vivo and in vitro studies, we raised the following model in Figure 34. Our in vivo study raises the possibility that host factor of ZBTB7C may act upstream of p53 and p21 to inhibit multiple cyclins including CCNG1, to arrest the cell cycle. This
could possibly affect CD4 counts. The activation or inhibition of p53 is regulated by HIV accessory proteins tat and vpr (Figure 34). Furthermore, ZBTB7C interacts directly with SP1 to activate the HIV transcription, which will directly affect viral load. CCNG1 interacts with Vpr to repress HIV transcription. The opposite function of ZBTB7C and CCNG1 is also consistent with our in vivo observation that patients having minor variation of ZBTB7C or CCNG1 have opposite phenotypes regarding CD4 and VL levels.

Overall both our in vitro and in vivo studies suggest that the p21 pathway played a central role in HIV pathogenesis.
Figure 34. In vitro and In vivo combined finding of HIV viral factors collaborating with host factors to regulate HIV replication and host cell cycle.

Purple: Host factors (Dark Purple: factors identified in the study; Light Purple: factors suggested by literature); Orange: Viral factors; Green arrows: activating function; Red: inhibiting function; Blue: interaction that hasn't been characterized. Blue background: pathways identified by in vivo GWAS study; Pink background: pathways identified by using pluripotent stem cells in vitro.
IV.3 The collaboration of host genes and viral genes in controlling HIV replication and affecting CD4 and viral RNA

Our research showed that three genes are associated with CD4/VL trajectories over time as a combined phenotype. From an epidemiology perspective, HIV infection and progression to AIDS is a complicated process where host genes, viral properties and environmental exposures are all involved in determining CD4 and VL trajectories (Figure 35A). Previous GWAS showed that the variation within HLAs, ZNRD1 and RFN39 can only explain less than 10% of the phenotypic variation in outcomes (Fellay et al., 2007). Environmental factors such as previous exposure to drugs, nutrition, education, and viral genome variation may also affect the phenotypes. Host genes, viruses and environmental factors serve as the causal factors, interacting with each other in the human body. By using the methodology of GWAS, we can only infer a fraction of the complicated interaction among these causal factors by establishing an association between genetic alleles and measurable outcomes such as CD4 and VL (Figure 35A).

Host factors exert their role of viral restriction by interacting with viral elements, while viral factors facilitate viral replication by interacting with host genes. Viral and host genes are always interacting with each other to play roles in viral replication restriction. Identifying the viral target of host genes and stage of viral restriction in the HIV life cycle is very important if we want to design HIV interventions.
Figure 35. Epidemiology modeling of HIV-1 infection.
A. Models depicting factors that could affect the read-outs of HIV infection. B. causal DAG model showing the causal relationship of factors/exposure and the GWAS study read out. Blue: The candidate mechanism of casual inference; Red: candidate effect modifiers to the causal inference.

Before virus gets into the cells, host cell surface receptors CD4 and CCR5/CXCR4 interact with gp120 to facilitate entry of HIV into host cells. Then host genes such as APOBEC3G and LEDGF/p75 interact with vif and integrase respectively to control viral replication before integration (Anderson et al., 2009). To isolate the host viral interaction before integration, GWAS studies based on populations with high exposure to HIV infection (such as sex workers) use HIV positivity as phenotype and tested the association between host genes and HIV susceptibility (An et al., 2014). After integration, multiple host genes are involved in viral transcription, or regulating HIV transcription indirectly by chromatin modification. It is very hard to isolate the host genetic restriction mechanisms at a transcriptional level by GWAS. Here we suggest that by surveying HIV proviral copy number, it may be possible to detect extreme phenotypes.
of patients with heavy proviral loads but low viral RNA or the opposite (Figure 35B). Though previous studies have shown some positive relationship between HIV proviral load and viral load, the association is not strong (Novitsky et al., 2006) (Figure 36). We suggest this new defined phenotype combining proviral load and viral load may help capture the subtle effect of transcription factors and other post integration restriction mechanisms in vivo.

![Figure 36 Association between proviral load and plasma viral load in HIV-1C infection](image)

The CD4 level of patients can be affected by factors other than those that directly affect replication. Immune regulation factors such as multiple interleukins that affect cell proliferation, and factors associated with control of cell death, may also be possibly involved in affecting the CD4 level of patients. On the other hand, viral factors that determine viral virulence can interact with host cell death pathways to influence the cell cycle. Host and viral factors that affect cell cycles should have stronger associations with CD4 level as compared to VL (Figure 35B). There is another interesting observation in
our study, that if we use only CD4 phenotype, though the top few SNPs are still similar, their p values are much larger except for region around rs2535307(chr6, pos31053741, MAF 48%, P_{pca}=3.72E-07) (data not shown). This observation suggested that the HCG22 region is more associated with CD4 counts while ZBTB7C or CCNG1 may be more associated with VL. However, since we haven't tested the role of ZBTB7C and CCNG1 in viral restriction \textit{in vitro}, and since these two genes are highly involved in cell cycle pathways (Figure 34), it is also possible that ZBTB7C and CCNG1 are affecting both cell cycle and viral load. To isolate the host genes' effect in regulating cell proliferation, an analysis to examine the association between host genes and CD4 or VL alone, rather than a combined phenotype of CD4 and VL, may be needed.

IV.4 Using genetic markers as a public health indication for AIDS treatment prioritization

From a public health stand point, treatment prioritization is crucial for improving quality of life and averting premature death in adults and children living with HIV. In Southern Africa, due to the high rate of AIDS and economic constraint, not all patients are able to get HAART upon detection of HIV positivity. Instead, an approach was adopted based on WHO guidelines: in Botswana, only patients with CD4 counts lower than 350 cells/$\mu$l on clinical illness are eligible for HAART treatment. Patients with CD4 higher than 350 cells/$\mu$l and lower than 500 cells/$\mu$l are eligible for periodic follow up for every three months(Government-of-Botswana, 2012). Patients with CD4 higher than 500 cells/$\mu$l are followed every six months. However, due to the disease progression variability of AIDS patients, the patients not immediately eligible for HAART may have different rates of
progression to severe disease development. For some patients (Figure 37 upper panel), it took less than half a year for CD4 to decline from above 500 cells/µl to less than 200 cells/µl, while other patients can maintain their CD4 around 500 cells/µl for at least 4 years (Figure 37 lower panel). For patients with low CD4 counts, even short delays (less than one month) in initiating treatment are associated with increased mortality. Therefore eligibility of initiating HAART based on arbitrary CD4 levels is not an optimal approach.

However, it is often difficult to identify patients that are predisposed to abrupt CD4 decline. We established an association between host genes and disease progression using both CD4 and VL as phenotype. We found three alleles that are strongly associated with disease progression. Using the combination of the three alleles, it may be possible to develop algorithms to help predict CD4 trajectory of each patient, and to use this prediction as an indicator for HAART eligibility.

For example, we randomly picked three patients from BHP010 that showed fast decline of CD4 within 6 months, and three patients with stable CD4 trajectories from BHP010 (Figure 37). Their genotypes of the three SNPs, which we identified to be strongly associated with disease progression (Table 10), are listed below their phenotype plots. The patients that are fast-progressors all carry two risk alleles of rs2535307(chr6, pos31053741, MAF 48%, P_{fpca}=3.72E-07), and some carry another risk allele kgp22385164(chr5, pos162859382, MAF=8%, P_{fpca}=9.14E-06), though all of them carry protective alleles at kgp7815029(chr18, pos45548347, MAF 15%, P_{fpca}=6.22E-06). For the three slow-progressors, none of them carry risk alleles rs2535307 and kgp22385164.
It is worth noticing that though this is not a study designed for phenotypic prediction, by random sampling the cohort, it is possible to use the three SNPs to predict the disease progression for these patients. And since these three alleles are significantly associated with the phenotype, it is possible to build statistical simulation models to predict patients' AIDS progression using the three genes identified. It also worth studying if rs2535307 alone is sufficient in predicting the disease phenotype as it has the strongest association in our result. Ideally if the model works well, we can potentially design a genotyping kit to survey these SNPs among patients that are HIV positive but not yet eligible for HAART due to higher CD4 levels. Such a result could help inform HAART prioritization in people carrying risk alleles that are linked to faster AIDS progression.
Figure 37. Extreme phenotypes of CD4 decline and genotype of the three significant SNPs.
U: non-risk allele, R: risk allele, P: protective allele, N: non-protective allele. Using dominant model, one risk allele or one protective allele will result in dominant effect. Upper panel: patients that experience fast CD4 decline from above 500 cells/µl to <200 cells/ml within half a year. Lower panel: patients that are able to maintain CD4 around 500 cells/µl for at least four years.
Bibliography


pathway by HIV-1 Vpr leads to induction of hypoxia-inducible factor 1alpha expression. The Journal of biological chemistry 284, 11364-11373.


the DNA binding of SREBP-1c and Sp1 at the proximal promoter. Journal of lipid research 53, 755-766.


NIAID (2014). “Mississippi baby” now has detectable HIV, researchers find. (NIH News).


Tyagi M, K.J. (2007). CBF-1 promotes transcriptional silencing during the establishment of HIV-1 latency. EMBO J 26, 10.


