Emergence of simian immunodeficiency virus in rhesus macaques is characterized by changes in structural and accessory genes that enhance fitness in the new host species

Emergence of simian immunodeficiency virus in rhesus macaques is characterized by changes in structural and accessory genes that enhance fitness in the new host species

A dissertation presented

by

Alison Kimberly Hill

to

The Division of Medical Sciences

in partial fulfillment of the requirements

for the degree of

Doctor of Philosophy

in the subject of

Virology

Harvard University

Cambridge, Massachusetts

January 2016

Emergence of simian immunodeficiency virus in rhesus macaques is characterized by changes in structural and accessory genes that enhance fitness in the new host species

Abstract

The distribution of lentiviruses among primates reflects a history of interspecies transmission and emergence of new virus-host relationships. The degree to which viruses must adapt to the genetic environment of new host species, and how adaptations to the new host initially affect viral fitness are two understudied elements of emergence. The simian immunodeficiency virus (SIV) of rhesus macaques (SIVmac) emerged as the result of a cross-species transmission of SIV from the sooty mangabey monkey (SIVsm) into rhesus macaques, and comparing cohorts of SIVmac- and SIVsm-infected macaques provides an opportunity to examine a lentivirus evolving during the early stages of emergence. Using archived samples from four cohorts of macaques, we compared evolution of "established" macaque-adapted viruses (SIVmac239, SIVmac251) to incompletely-adapted, "emerging" viruses (SIVsmE543, SIVsmE660). Longitudinal samples included the inoculum for each cohort, as well as acute and chronic plasma samples for each animal. Samples were processed for deep sequencing, and consensus sequences of complete viral coding regions were assembled *de novo*. Computational and manual analysis of the sequences revealed a set of loci that diverged considerably only in the SIVsm-infected animals, suggesting that adaptations at these loci are important for emergence of SIVsm in rhesus macaques. These candidate adaptations included known adaptations to overcome restriction by macaque TRIM5α. In order to quantify the impact of these candidate adaptations on viral replication, each mutation was introduced into SIVsmE543 (forward mutations, reflecting adaptation to the macaque host) and SIVmac239 (reversions to the ancestral residue). These were then tested in a deep sequencing-based fitness assay, in which changes in

the frequencies of mutant and parental sequences replicating in cell culture were used to calculate differences in relative fitness. Substitutions in the coding sequences for the Matrix, Capsid, and Vif proteins were found to enhance fitness of SIVsm in rhesus cells, confirming the hypothesis that they represent species-specific adaptations. Together, these studies represent a novel approach to the identification and functional characterization of viral determinants of cross-species transmission.

Acknowledgments

First, I would like to thank Welkin Johnson for his endless support throughout my graduate career. I chose Welkin as an advisor because of his enthusiastic yet equable attitude about research, and I never regretted that decision. In addition to having all of the typical qualities of a good mentor (supportiveness, creativity, expertise, and integrity), Welkin has several qualities that set him apart. First, he treats everyone, from undergraduates to senior scientists, as a colleague, which creates an inclusive environment in his lab. Second, he has a unique ability to deliver criticism in an exclusively constructive manner, which allows his trainees the room they need to make mistakes and learn from them. I am grateful to him for his scientific guidance and for his calming perspective when experiments did not go as planned.

I am also grateful for the kind and cooperative group of people I've had the pleasure of working with in the Johnson lab over the past five and half years. There are too many past and present lab members who helped me in some way or another to list here, but I would especially like to thank Sergio Ita, Laura Hall, Jennifer Morgan, and Max Mangano. I am grateful to Sergio, who was both a lab mate and a classmate, for his collaboration, support, and friendship, without which my dissertation would have been much less enjoyable. Laura and Jen have been such consistently helpful presences during my time in graduate school that I doubt any of my research would have been possible without them. Max was perhaps the most eager and motivated undergraduate I've ever encountered, and it was a delight to work with him for the two years he assisted me in the lab.

I would also like to thank the faculty, students, and administrators of the Virology Program for the ideal environment they created for young researchers to learn and grow. I especially want to thank my classmates for both scientific and moral support throughout my time

at Harvard. The friendships we formed have heartened me through the most difficult academic and scientific challenges. I'm also particularly grateful to Todd Allen, Miti Kaur, and Lee Gerkhe for their support and advice as my dissertation advisory committee members.

Finally, none of my accomplishments would have been possible without my wonderful family. My parents, Margianne and George Hill, have put my education (both personal and professional) above their own interests for as long as I can remember. They and my brother, Brendan, have encouraged and supported me through any challenge I have undertaken. My husband, Travis Watters, has gone above and beyond the duties of a partner since I met him at the beginning of my second year of graduate school. He has helped me with data analysis, writing, and presentations, in addition to providing an abundance of emotional and moral support. My family never doubted my ability to complete my PhD, even when I felt most doubtful, and it's their belief in me that sustained me throughout all the highs and lows of graduate school.

Table of Contents

List of Figures

List of Tables

1.0 Introduction

1.1 Cross-species transmission and emergence

1.1.1 Stages of emergence

New viruses often enter the human population from animal reservoirs, in a process called zoonosis. Zoonotic infections are a major public health concern, as they present a significant threat to human health. Viruses such as severe acute respiratory syndrome (SARS) and Middle East respiratory syndrome (MERS) coronaviruses, Ebola virus, and human immunodeficiency viruses (HIV) type 1 and type 2 are the results of cross-species transmissions from animal populations (1-3). There are many more examples of cross-species transmissions that have resulted in the emergence of new viruses that cause human disease; in fact, the majority of new human pathogens arise in this manner (4). However, there are also many documented (and likely many more undocumented) instances of failed cross-species transmissions. In these cases, at least one human has been exposed to and infected by a virus from another species, but the virus failed to spread beyond the originally infected individual, or those in immediate contact with that individual (5). For example, of the nine known subtypes of HIV-2, each of which resulted from an independent cross-species transmission event, seven have only been found in one or two infected individuals, indicative of "dead-end" transmissions (6-9). Why some cross-species transmissions lead to the emergence of new viruses and diseases, while others lead to dead-end infections remains largely unknown, and the molecular mechanisms that govern emergence have yet to be understood. This dissertation will address the importance of adaptations to the new host species in facilitating the emergence of viruses in new species, specifically focusing on the Primate Lentiviruses (PLVs), defined and discussed in the next section.

In order for a virus to spread in a new species, several major steps must occur: First, the new host species must be exposed to the virus; second, the virus must infect the cells of the new host; third, the virus must replicate and spread within the new host and to other individual hosts

within the same species; and fourth, the virus must adapt to the environment of the new host species. The third and fourth steps most likely progress concurrently, with greater spread facilitating more adaptations, and more adaptations facilitating further spread (5, 10).

In the exposure stage, the new species must come into contact with the virus, typically through direct contact with the reservoir or donor host, or through indirect contact by way of a vector, such as a mosquito. Whether or not an exposure occurs largely depends on ecological factors and host behavior. For example, humans can be exposed to animal viruses through domestication and hunting of animals, as well as by sharing territory and natural resources with animal populations (4, 5, 10).

After an exposure, the virus must infect the cells of its new host for zoonosis to occur. Whether or not this happens depends on the extent of virus-host compatibility, such as receptor usage and evolutionary relationships between the original and new host species. One hypothesis is that viruses are more likely to infect a new species if the donor species and the recipient species are closely related evolutionarily (11). Indeed, there are more examples of zoonoses from primates and mammals than from non-mammalian species. There is also the possibility of indirect zoonoses via intermediate hosts, where a virus is first transmitted from the reservoir host to a species that is more similar to humans and then goes on to infect human populations (4, 10-13).

If the virus is able to infect cells of the new host, it must also replicate to a large enough population size to accumulate sufficient genetic variation upon which natural selection can act. RNA viruses in particular are likely to generate ample diversity for adaptation because they tend to have error-prone replication enzymes. Indeed, more documented zoonoses have occurred with RNA viruses than DNA viruses (11). The high mutation rate coupled with large population size

enables rapid adaptation to the new host (11). However, host cells typically express innate and intrinsic antiviral factors with which an emerging virus population must contend. Viruses are particularly susceptible to intrinsic, constitutively expressed antiviral proteins or restriction factors, which are the host cell's first line of defense against cross-species transmissions (14). Specific restriction factors are described in detail in later sections of the introduction.

The potential for a virus to spread in a new host species is a function of its reproductive rate, R_0, defined as the number of secondary transmissions from an infected individual in a susceptible population (15). If the virus can adapt sufficiently to spread to secondary, tertiary, and further hosts of the new species, particularly so that it reaches an R_0 value of 1 or higher, it is likely to emerge in that species and sustain transmission chains in the new host (15). However, mathematical modeling studies have suggested that even in cases where the initial R_0 is below 1, adaptation can still occur; in fact, it may be at this stage that the most essential adaptations to the new host arise (16, 17).

1.1.2 Studies of emergence: Difficulties and practical considerations

The early stages of emergence are mysterious in part because of the difficulties in obtaining relevant samples to analyze. In order to closely examine the molecular events leading to emergence, one would ideally have access to samples from multiple phases of the process: from the donor immediately prior to transmission to the recipient; longitudinal samples from the recipient; and longitudinal samples from secondary and tertiary hosts in subsequent transmission chains. However, accessing such samples is unlikely for the majority of cross-species transmissions given how much surveillance of donor and recipient species would be required. Surveillance of at-risk individuals is typically not extensive enough for researchers and public health officials to obtain samples from individuals at the beginning of a transmission chain. For

example, during the SARS outbreak of 2002-2003, the coronavirus responsible was not identified until it had already spread to over 300 people (18, 19). This lack of surveillance precludes the study of samples close to the original transmission from the donor species. Even though SARS coronavirus had not spread far by comparison to HIV-1 upon its identification as the cause of Acquired Immune Deficiency Syndrome (AIDS), the precise donor animal that infected the original human host was still not likely to be found (19). In the case of HIV-1 and -2, the long asymptomatic phase of infection meant that AIDS as a disease was not recognized, nor its viral origin elucidated, until it had been circulating in multiple countries for close to a century (20).

Despite these difficulties, several groups have undertaken intensive efforts to uncover the natural history of HIVs by analyzing clinical samples predating their discovery as well as surveying non-human primate populations in the geographical regions where the earliest cases of AIDS were documented. Such studies have led to accurate geographical and temporal data about the original cross-species transmissions that led to the multiple groups and subtypes of HIV-1 and -2 (21-26). However, there is a limit to what we can learn about the adaptation process from studying extant viruses that share common ancestors with HIV-1 and -2, due to the amount of divergence that has occurred in the century or so since these lineages diverged after the original transmissions to humans.

1.2 Introduction to primate lentiviruses

1.2.1 Classification

HIV-1 and HIV-2 are members of the *Lentivirus* genus within the family *Retroviridae*. Lentiviruses are characterized by their distinct particle morphology and tendency to cause long-lasting infections. Lentiviruses have been identified as agents of disease in a variety of

mammalian species, including sheep, cows, cats, horses, goats, and multiple primate species (27). This section will focus specifically on the primate lentiviruses (PLVs), a subgroup of the genus. There are fourteen phylogenetically distinct groupings of PLVs, which in many cases are further divided into subspecies, each infecting a particular species of primate. For example, the simian immunodeficiency virus (SIV) that infects African green monkeys (SIVagm) is subdivided into groups infecting the sabeus, vervet, tantalus, and grivet subspecies ($SIVagm_{sab, ver, tan,}$ and $_{gri}$) (27). Species-specific PLVs have been identified in humans, apes (including chimpanzees and gorillas), more than forty different African monkey species, and one Asian monkey species (27, 28).

1.2.2 Natural histories of cross-species transmission

PLVs have a long history of interspecies transmission and emergence of new virus-host relationships (Figure 1). Most notably, HIV-1 and HIV-2 both emerged in humans after cross-species transmissions from chimpanzees, gorillas, and sooty mangabey monkeys (3, 6, 23). There are 4 documented groups of HIV-1: M, N, O, and P. Groups M and N originated in cross-species transmissions of the SIV of chimpanzees (SIVcpz) (21, 29). Groups O and P are the result of zoonoses of the SIV of gorillas (SIVgor) (23, 30). Finally, the 9 subtypes of HIV-2 (A-I) emerged in humans following transmissions of the SIV of the sooty mangabey monkey (SIVsm) (31, 32). Each group of HIV-1 and subtype of HIV-2 came from independent cross-species transmission events, most likely through contamination of human blood with blood or other bodily fluids from non-human primates during hunting or butchering (3). In addition to hunting or butchering them for consumption, it is also not uncommon for people to keep smaller primates such as sooty mangabeys as pets (33).

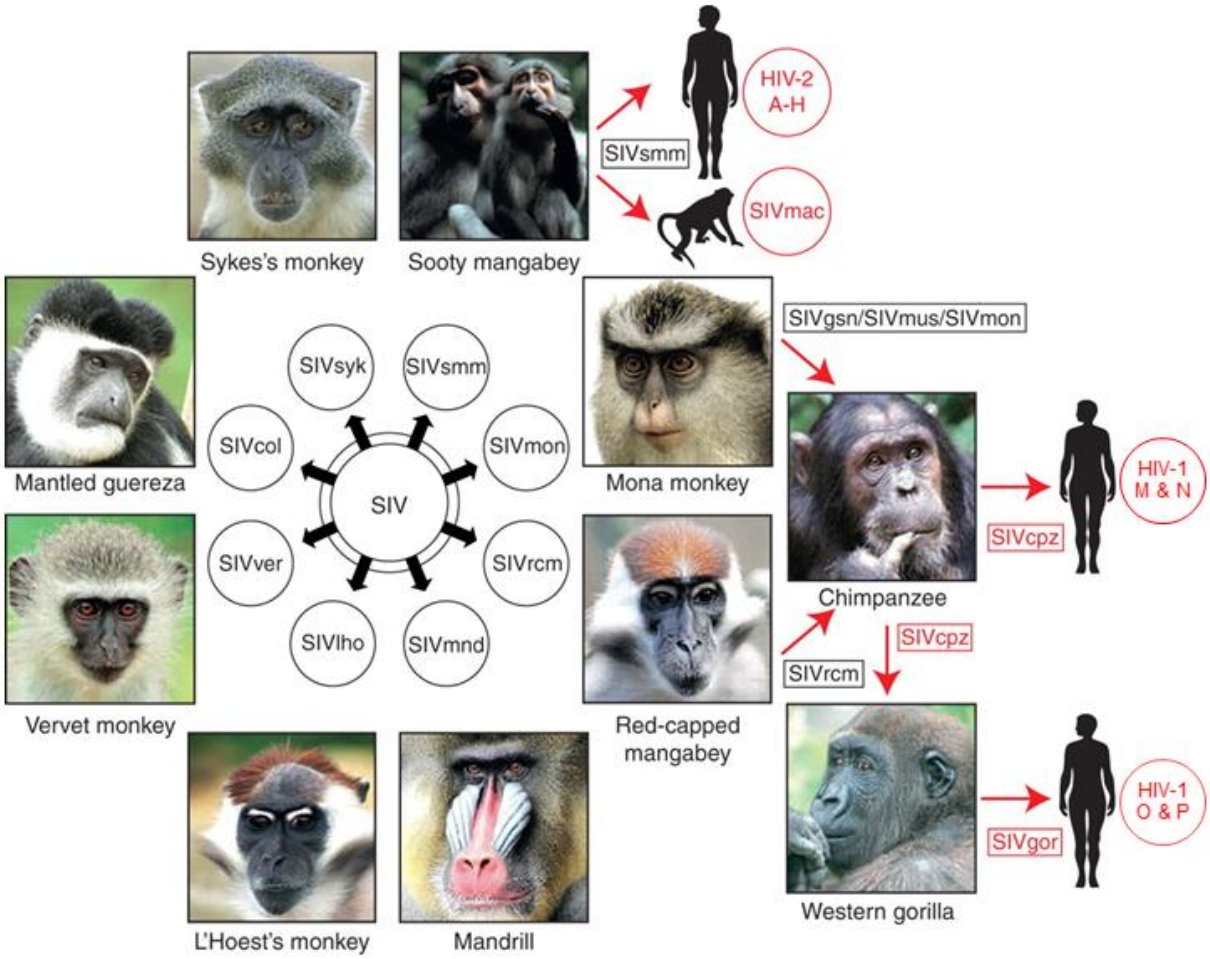


Figure 1. SIVs in natural and non-natural hosts.

SIVs naturally infect many African monkey species, but several have also crossed species barriers into apes and humans, leading to the emergence of new viruses. Black circles and lines indicate longstanding natural infection, while red lines indicate evolutionarily recent cross-species transmissions. Red circles are emergent viruses. Adapted with permission from (6).

These interactions between humans and other primate species likely facilitated the exposure and subsequent infection of the first infected humans (3, 20, 33). Despite somewhat frequent exposures (13 documented exposures and undoubtedly more undocumented examples), only HIV-1 Group M has led to a pandemic. HIV-1 Group O and HIV-2 subtypes A and B have caused epidemics largely restricted to West Africa, while all of the remaining groups and subtypes have failed to spread substantially in human populations (6, 34). For example, HIV-1 N has been reported in fewer than twenty people (35), while Group P has been detected in only two individuals (29, 36). The amplification and spread of HIV in West Africa and eventually around the world has been attributed to the use of unsterilized needles in vaccination campaigns which, if true, would have effectively resulted in serial passage of the virus through a series of individuals, facilitating the adaptation of HIV to humans (20, 37).

SIVcpz and SIVgor also resulted from cross-species transmissions. The ancestor of SIVcpz is likely a recombinant virus originating from co-infection of a chimpanzee with ancestral SIVs related to those currently found in red-capped mangabeys (SIVrcm) and greater spot-nosed monkeys (SIVgsn) (22). Both the red-capped mangabey and the greater spot-nosed monkey are targets of chimpanzee predation, which is believed to be the mechanism for these cross-species transmissions (22). Subsequently, SIVcpz was transmitted to gorillas, leading to the emergence of SIVgor (38, 39).

In addition to infecting humans, SIVsm was transmitted from sooty mangabeys to rhesus macaques, leading to the emergence of SIVmac. This transmission was likely triggered by experimental transfer of bodily fluids from SIVsm-infected sooty mangabeys to rhesus macaqaues in captivity at the California National Primate Research Center (CNPRC) in the 1960s. At the time, neither HIV nor SIV had been discovered, and research was being done to

elucidate the transmissibility of Kuru, a prion disease that causes severe neurological symptoms and death. It has been suggested that these experiments led to the original transmission of SIVsm to rhesus macaques, and to the subsequent spread of the virus through colonies of captive macaques at the CNPRC and eventually to other primate research centers (40). In the 1970s, staff at various centers began reporting outbreaks of a wasting disease in macaque colonies, but it was not until 1985 that SIVmac was identified as the cause of this disease, and its origins were eventually elucidated through an analysis of archived veterinary records and tissue samples at several primate research centers (40, 41).

1.2.3 Pathogenesis in natural and non-natural hosts

The PLVs are of significant interest to the HIV/AIDS field for several reasons beyond their various cross-species transmissions. In particular, the study of SIV pathogenesis in so-called "natural" as opposed to "non-natural" host species has been an area of intense research for many years. Many species of African non-human primates have been co-evolving with species-specific SIVs for thousands of years and are considered natural hosts of SIV (42). The primate species discussed above that were the recipients of cross-species transmissions are considered non-natural hosts (Figure 1). Generally, natural hosts remain asymptomatic throughout their lives, and do not develop the immune deficiencies and opportunistic infections that manifest in non-natural hosts. This phenotype is often referred to as "AIDS resistance" (43). Natural hosts maintain their health in spite of high levels of viral replication comparable to levels in non-natural hosts. This is likely due to the extensive co-evolution that is presumed to have occurred between PLVs and their natural hosts. It is possible that ancestors of extant PLV species caused disease when they were first introduced into their host species, but that over millennia of the virus-host arms race, an evolutionary détente was reached, in which the virus is able to maintain

high levels of replication without causing disease (42). Wild-living natural hosts typically die of other causes before the virus damages the immune system enough to cause disease. Indeed, there are examples of natural hosts in captivity (where animals tend to live longer than their feral counterparts) that developed simian AIDS (44). However, natural hosts in captivity still do not progress to disease as rapidly as non-natural hosts. A typical SIV-infected macaque will develop symptoms of simian AIDS within one year of infection, while SIV-infected sooty mangabeys live for years without exhibiting any symptoms (44, 45).

There is also a great deal of evidence to suggest that there are differences in the way PLVs affect natural hosts compared to non-natural hosts. In particular, the immune responses during chronic PLV infection are markedly different between the two host types. During acute infection, both types of hosts experience uncontrolled viral replication (with characteristically high viral loads), activation of innate immune responses leading to production of inflammatory cytokines, and depletion of mucosal CD4+ T cells (43, 46, 47). However, in the chronic phase, non-natural host infection is characterized by sustained inflammation from persistent activation of interferon (IFN), loss of central memory T cells, and translocation of microbes from the gut to circulation, all of which are absent in chronic infection of natural hosts. The precise mechanism for AIDS resistance of natural hosts is unknown, and several potential explanations are supported by experimental evidence. One generality that is agreed upon is that the resolution of the innate immune response and the cessation of inflammatory cytokine production (despite persistent viral replication) in natural hosts is the key to the disease-resistant phenotype they exhibit (43, 46, 47).

1.3 Primate lentivirus biology

1.3.1 Genome and particle organization

As members of the retrovirus family, PLV replication is characterized by several unique life cycle stages, including reverse transcription of the single-stranded RNA genome into double-stranded DNA (48, 49), as well as the subsequent integration of that DNA into the host cell genome. All retroviral genomes encode several key gene products that carry out the major stages of replication, flanked by long terminal repeats (LTRs) at both the 5' and 3' ends (Figure 2, top). The genes *gag*, *pol*, *pro*, and *env* are found in all members of the retrovirus family. Gag is a structural protein that forms a core around the viral genomic RNA. It is translated as a polyprotein that is then cleaved into the Matrix (MA), Capsid (CA), Nucleocapsid (NC), and p6 proteins (50). PLV Gag proteins have additional minor cleavage products, called spacer peptides 1 and 2 (SP1 and SP2) (51). CA forms the main core of the virus, while MA forms a structured lattice on the inside of the viral membrane (Figure 2, bottom). NC remains bound to the viral RNA and is required for its inclusion in the budding virion. P6 binds to several host cellular factors that facilitate particle budding and release, while SP1 and SP2 play a role in the conformational changes that take place during maturation. In PLVs, *pro* is a part of *pol,* which is translated as the Gag-Pol polyprotein, then cleaved into Reverse Transciptase (RT), Integrase (IN), and Protease (PR) enzymes, which conduct the reverse transcription, integration, and cleavage of viral proteins, respectively (50, 52). Finally, the Env protein facilitates entry into the host cell via receptor-binding and fusion with the cell's plasma membrane. The two subunits of PLV Env, gp120 (the surface subunit, or SU) and gp41 (the transmembrane subunit, or TM) are initially translated as the gp160 precursor before being cleaved by cellular proteases (27, 50-52).

The PLVs also encode several additional gene products known as auxiliary and accessory genes. The auxiliary genes include *tat* and *rev*, which are involved in RNA transcription and

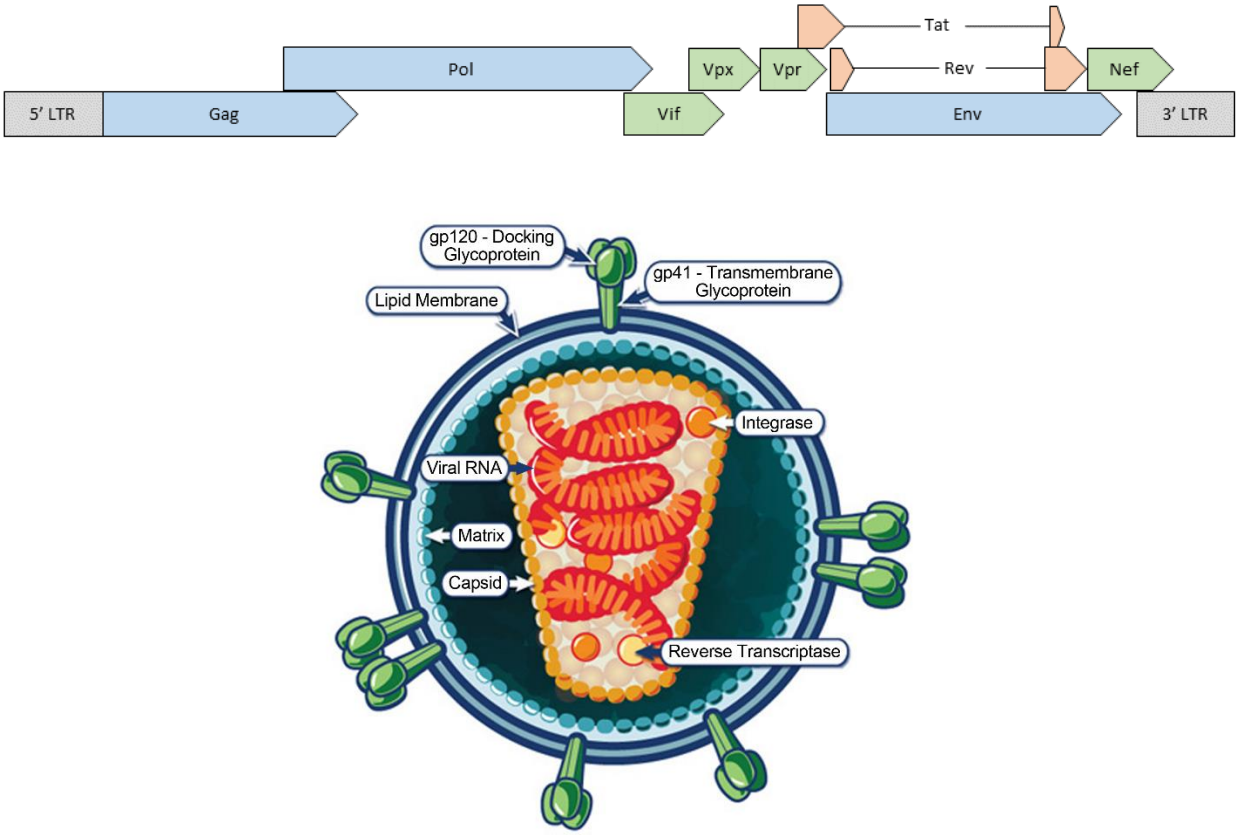


Figure 2. Genome and particle organization of PLVs.

Top: open reading frames (ORFs) are represented by colored block arrows (blue, structural genes; orange, auxiliary genes; green, accessory genes). LTRs are represented by gray blocks. In some PLV lineages, Vpx is replaced with Vpu. Bottom: mature lentivirus particle structure. Reproduced from (53).

export (50, 52). The accessory genes include *vif*, *vpr*, either *vpx* or *vpu*, and *nef* (50, 52, 54, 55). These genes encode proteins that are typically dispensable for replication, but can dramatically enhance replicative capacity of the virus *in vivo* (and under some *in vitro* conditions) (27, 56-58). They have a variety of functions including suppression of host intrinsic, innate, and adaptive immune responses (27, 50, 52, 59).

The mature retrovirus particle (Figure 2, bottom) contains two copies of the genomic RNA bound by NC, and surrounded by the CA core. The core is in turn enveloped by a roughly spherical lipid bilayer derived from the infected host cell, which is lined with MA proteins. Incorporated into the bilayer are trimeric complexes of Env, with the SU subunit on the outside of the membrane and the TM subunit incorporated into the membrane. A C-terminal stretch of residues known as the cytoplasmic tail (CT) extends into the interior of the particle. The particle is initially released in an immature form, with Gag and Pol both still in their uncleaved state. Maturation occurs when PR initiates cleavage and a dramatic rearrangement of the structural proteins occurs. (50, 51)

1.3.2 Replication

PLV infection of a target cell begins when the gp120 subunit of the Env trimer attaches and binds to its primary receptor, CD4 (Figure 3) (52, 60, 61). CD4 binding causes a conformational change in Env, which reveals the binding site for the co-receptor, typically chemokine receptors CCR5 (the primary co-receptor) or CXCR4, although many PLVs are able to use additional alternative co-receptors (52, 62-67). Engagement of the co-receptor enables the gp41 subunit to initiate fusion of the viral membrane with the cellular membrane, a complex process involving further conformational changes and stages of partial fusion before completion, when the capsid core is released into the host cell cytoplasm (52, 68).

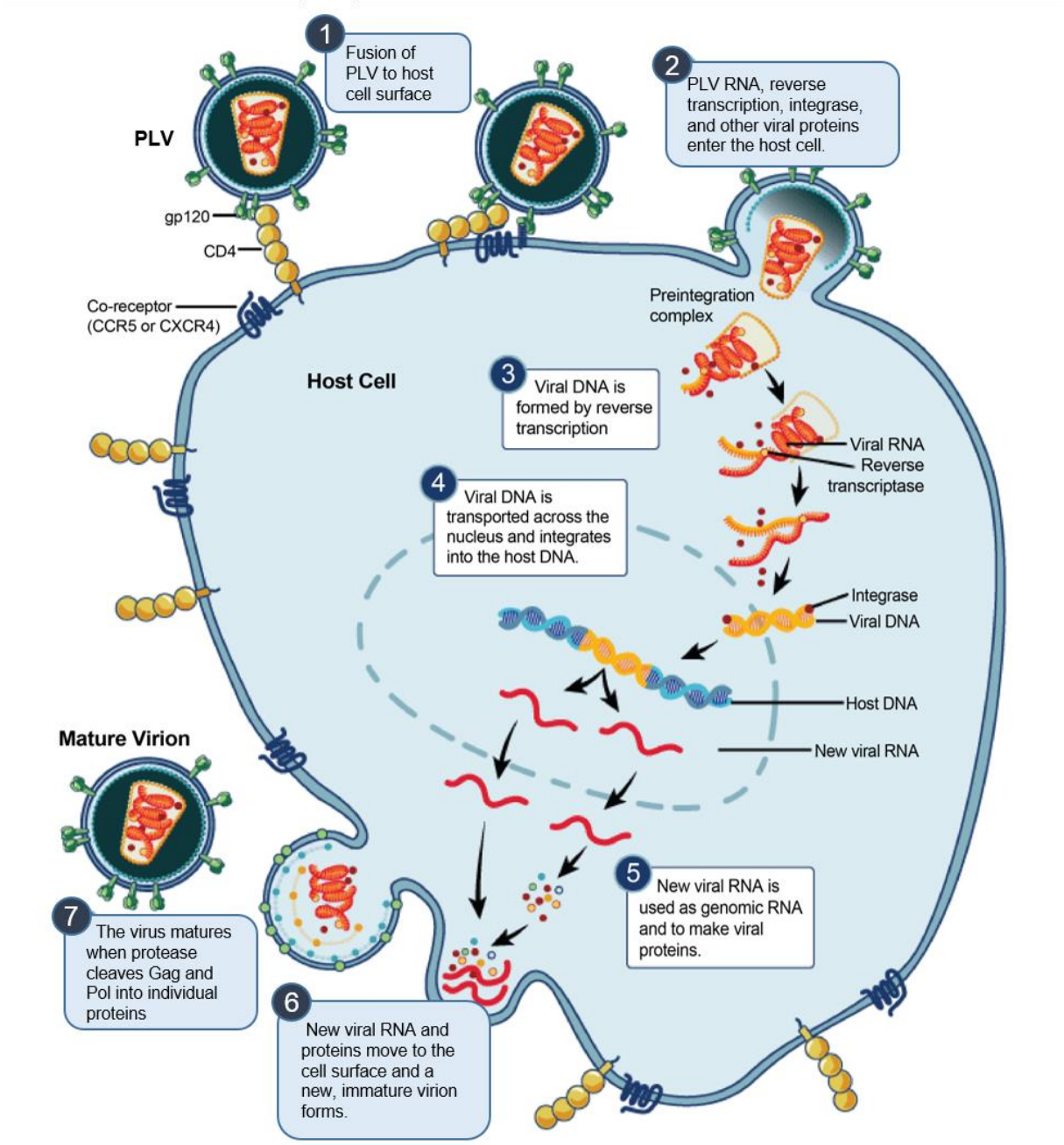


Figure 3. PLV Life Cycle

(1) Receptor binding and membrane fusion, (2) deposition of virion contents into host cell cytoplasm, (3) reverse transcription of viral genomic RNA (4) migration of pre-integration complex and integration into host cell DNA, (5) transcription of viral mRNA and translation of viral gene products by cellular machinery, (6) assembly of viral proteins at the plasma membrane and budding of the immature viral particle, (7) virion maturation. Adapted from (69).

Once inside the cell membrane, the capsid core disassembles in a regulated manner, though the precise timing of this process remains to be determined (70). Reverse transcription of the viral genomic RNA is also a highly regulated process that leads to the formation of a linear, double-stranded DNA product (50, 71), which is transported to the nuclear pore in complex with viral and cellular proteins (72, 73). This complex is deposited into the nucleus where IN, in association with cellular co-factors, inserts the viral DNA into the host cell chromosomal DNA (50, 52). This form of the viral genome is called the provirus. The integration step allows viral mRNA to be produced using cellular transcription machinery, with additional transcriptional regulation from the viral LTRs and Tat. Rev facilitates the nuclear export of unspliced and singly spliced viral mRNAs, which—with the exception of those encoding *env*—are then translated on free ribosomes in the cytoplasm. Env is translated on the rough endoplasmic reticulum and co-translationally glycosylated before travelling to the Golgi, where it is cleaved by a cellular protease into gp120 and gp41. The processed subunits then move to the plasma membrane, where the other viral proteins have also been transported to begin the assembly process (50, 52).

Interaction between the CT of Env and the MA domain of Gag ensures inclusion of Env trimers into the nascent particles. Two copies of the viral genomic RNA are incorporated into the virion through their interactions with NC. Gag multimerization leads to bulging and eventually budding of the plasma membrane, forming the immature particle. Autocatalysis of Gag-Pol by the PR domain leads to the separation of structural cleavage products, which rearrange to form the mature particle structure. Cell-free virus particles may go on to infect new target cells, but the fusogenic nature of Env at the plasma membrane can also lead to cell-cell fusion, and viral products and particles can be transferred to new target cells in this manner (50, 52).

1.3.3 Cellular and species tropism

PLVs have a highly specific cellular tropism due to their requirement for the CD4 receptor and CCR5/CXCR4 co-receptor. They specifically infect CD4+ T lymphocytes and macrophages (27, 52). As discussed above, most PLVs also have a highly species-specific tropism, primarily infecting a single host species, with limited ability to infect other types of primates (27). While not fully understood, species tropism likely depends on differences between host factor homologs among species. For example, PLVs exploit a number of host proteins in order to replicate, including the aforementioned CD4 and CCR5, which the viruses uses to enter cells (52). Other examples include host transcription and translation machinery, which make viral mRNA and proteins (50, 52), as well as vesicle budding proteins, which the virus co-opts to bud from the infected cell at the end of the replication cycle (51). In general, PLVs are able to use such host factors from the particular species to which they are adapted, but may not necessarily be able to use homologous host proteins in different species (74). The same is true for antiviral proteins which act to block PLV replication. PLVs have generally evolved to evade or counteract these antiviral factors in their natural hosts, but when a cross-species transmission occurs, the virus is often blocked by the homologous protein in the new species, and must adapt in order to avoid restriction (75, 76). A subset of antiviral proteins known as restriction factors, which are intrinsically expressed antiviral proteins that are also often upregulated by the innate immune response, heavily influence species tropism (14). While there are many known restriction factors, and likely more that remain to be identified, three factors have been shown to play particularly important roles in both the prevention of interspecies transmission and selection of emergent PLVs: TRIM5α, APOBEC3G, and Tetherin (77).

1.3.3.1 TRIM5α

TRIM5α (tripartite motif-containing protein-5α) is a cytoplasmic protein that recognizes the incoming viral capsid core, and restricts infection by interfering with capsid uncoating (78, 79). Numerous studies have revealed the restriction potential of TRIM5α from different species against various PLVs, but the most extensively studied relationships are between rhesus macaque TRIM5α (rhTRIM5α) and HIV-1 and SIVsm. In fact, TRIM5α was initially identified as a restriction factor because of the ability of rhTRIM5α to restrict HIV-1 replication (78, 80, 81). Since then, the site of interaction between HIV-1 and rhTRIM5α has been identified as a patch of surface-exposed residues on the HIV-1 CA N-terminal domain that is dramatically different from its counterpart on the rhTRIM5α-resistant SIVmac CA (82, 83).

RhTRIM5α also played a major role in the emergence of SIVmac after the transmission of SIVsm to macaques. There are multiple alleles of rhesus TRIM5α that can be divided into several functional classes based on their ability to restrict SIVsm. The presence of either a TFP or a ΔΔQ at positions 339-340 in the C-terminal SPRY/B30.2 domain of rhTRIM5α determines that allele's restrictive capacity against SIVsm (84, 85). Specifically, restriction assays involving exogenous expression of different alleles of rhTRIM5α in permissive cells have shown that SIVsm may be resistant or perhaps mildly restricted by Q alleles, but is potently restricted by TFP alleles (85, 86). During adaptation to rhesus macaques, particular changes in the SIVsm CA were selected (specific adaptations will be discussed in sections 2.0 and 3.0), and the emergent SIVmac is now capable of evading all rhTRIM5α alleles (84, 85).

1.3.3.2 APOBEC3 proteins

Similar stories have arisen with APOBEC3G, or A3G (apolipoprotein B mRNA editing enzyme, catalytic polypeptide-like 3G), which incorporates into virions in infected producer cells

and causes cytidine deamination of minus strand, single-stranded DNA during reverse transcription, thereby restricting replication in the target cell (87). A3G is part of a larger family of cytidine deaminase proteins, including A3A through A3H, the majority of which have some antiviral activity (88, 89). PLVs have evolved to counteract restriction by A3 proteins in their natural hosts by using the accessory protein Vif. Vif links A3 proteins to ubiquitination machinery, leading to their proteasomal degradation (59), and may also counteract A3G by down-modulating its transcription (90). Despite the adaptability of Vif proteins, A3G is thought to prevent cross-species transmission of various PLVs to gorillas, whose A3G carries a mutation that renders it resistant to degradation by Vif proteins from PLVs other than SIVgor (23, 91). Differences in rhesus macaque and sooty mangabey A3G also presented an initial barrier to the emergence of SIVsm in rhesus macaques. Like TRIM5α, A3G is polymorphic in macaques, and SIVsm Vif was initially unable to counteract all of the rhesus alleles. But over time, changes in Vif were selected that allowed SIVsm/SIVmac to evade restriction by all rhesus A3G alleles (92).

1.3.3.3 Tetherin

Tetherin (also known as bone marrow stromal antigen 2, or BST-2) prevents the spread of virions to target cells by physically tethering them as they bud from the producer cell membrane. In their natural hosts, both SIVcpz and SIVsm counteract tetherin using Nef. However, human tetherin lacks the essential Nef-interacting motif, and so upon transmission to humans, SIVcpz and SIVsm would have been susceptible to restriction by tetherin. In the case of SIVsm, Env eventually evolved the ability to counteract human tetherin, and emergent HIV-2 subtype A is able to counteract human tetherin (6). In the case of SIVcpz, the Vpu protein acquired the ability to counteract human tetherin, but only in the case of HIV-1 Group M viruses.

HIV-1 Group N viruses remain susceptible to restriction by human tetherin (93, 94). These findings suggest that group M, which has infected well over sixty million people worldwide (6), may have been able to spread much further than N (diagnosed in fewer than fifteen people as of 2011 (35)) at least in part because it acquired adaptations that allowed it to overcome restriction by tetherin (6, 94).

1.3.3.4 Other factors influencing species tropism

Cross-species transmissions of SIVsm to both humans and rhesus macaques have involved key mutations that allowed the viral populations to overcome species-specific barriers to infection and successfully emerge in their new hosts (6, 85, 86, 92, 95). The three antiviral factors discussed above are among the best studied as they relate to restriction of PLVs. However, there are still many proteins involved in antiviral responses that could act as important selection pressures during cross-species transmission. For example, the impact of the recently identified MxB/Mx2 restriction factor on cross-species transmission of SIVs remains to be determined (96, 97), and yet-to-be-identified IFN-inducible factors have been shown to restrict cross-species infection of SIVmac and HIV-1 (98). Furthermore, if restriction factors that inhibit infection select for resistant PLVs, it follows that host proteins that are necessary for viral replication may also select for viral adaptations that enhance virus-host factor interaction. In order to better understand viral evolution after cross-species transmission, we sought to identify novel adaptations in SIVsm that may have contributed to its adaptation to and emergence in new hosts.

1.4 This study

1.4.1 SIVsm transmission to rhesus macaques as a model to study emergence

The natural history of PLVs has involved myriad cross-species transmissions, many with significant ecological and public health consequences, making them relevant models for studying adaptation. In particular, the historical transmission of SIVsm from sooty mangabeys to rhesus macaques in captivity provides an apt, albeit unintentional, animal model of PLV emergence. Since the discovery and isolation of SIVmac in the 1980s, numerous studies have been conducted in which SIVsm strains were used to experimentally infect rhesus macaques (99-102). Although only a small number of these studies were conducted with the goal of studying cross-species transmission and emergence—indeed, the majority have been to investigate vaccine strategies using SIVsm as a heterologous challenge strain for SIVmac-based vaccines—samples of SIVsm-infected macaques from both types of studies contain a body of viral genomic information that can further our understanding of SIVsm emergence. One key benefit of this model is that the precise source inoculum is known and is the same for all animals in a particular cohort. Access to the original source inoculum as well as longitudinal samples post-infection makes it possible to track evolution of the viral population throughout the process of adapting to a new host, allowing investigation of fundamental questions about cross-species transmission.

1.4.2 Approach and Innovation

The few studies that have aimed to examine adaptation of SIVsm during the early stages of emergence in rhesus macaques have been limited to small regions of the genome, and involved cloning and/or bulk sequencing for analysis of viral populations. For example, in a study where a primary SIVsm isolate was used to infect rhesus macaques and sooty mangabeys in parallel, only a small portion of the Gag gene and the Env variable loops 1 and 2 (V1/V2)

were analyzed by PCR-amplification and sequencing of individual clones (103, 104). Other studies, including several from our own group, focused on regions of the SIVsm genome that were known to interact with specific host restriction factors (85, 86, 92). While informative, such targeted studies can yield biased variant frequencies and may fail to detect low frequency variants. Thus, we turned to deep sequencing as an alternative approach to identifying adaptations and defining variation within viral populations (section 2.0). Specifically, we conducted whole-genome, deep sequence comparisons of SIVsm and SIVmac populations replicating in rhesus macaques. We developed novel algorithms to identify hotspots of divergence in emerging viral populations (SIVsm in macaques) compared to established populations (SIVmac in rhesus macaques). Additionally, we created a set of criteria to identify potential rhesus-specific adaptations, with the hypothesis that the adaptations would inform us of selection pressures that the virus must overcome to achieve sufficient replication and spread in the new host species.

We also sought to expand on previous studies in which species-specific adaptations were validated in restriction assays involving over-expression of restriction factors. Such studies tend to rely on testing replication-defective viruses and only examine the effects of adaptations during a portion of the viral replication cycle (23, 85, 92, 105). Observing only one portion of the replication cycle precludes the detection of any fitness tradeoffs that a putative adaptation might cause, in which a mutation is beneficial in one part of the life cycle but detrimental in another. In order to assess the full effect of adaptations on the viral life cycle, and to do so in a quantitative manner, we developed a fitness assay called FitSeq that measures the relative fitness of adaptations in the context of full-length, replication-competent viruses over the course of a full (or multiple) replication cycle(s) (section 3.0). Overall, the work presented here increases

understanding of SIVsm emergence by utilizing novel approaches to comparative viral genomics, adaptation-identification, and fitness evaluation.

2.0 Deep sequencing of SIVsm populations emerging in rhesus macaques reveals adaptations in both structural and accessory genes

2.1 Attributions

The data in this chapter are part of a manuscript in preparation by Alison K. Hill, Sivan Leviyang, Ruchi Newman, Michael Zody, Xiao Yang, Vanessa Hirsch, and Welkin E. Johnson.

AKH, SL, and WEJ designed research and wrote the manuscript. AKH performed experimental procedures and analyses except: RN, MZ, and XY conducted Illumina sequencing and computational read processing as outlined in section 2.4.1.3. SL developed novel algorithms and performed computational analyses detailed in sections 2.4.2 and 2.5.4.2. VH provided all emerging virus samples.

2.2 Abstract

Human Immunodeficiency Virus types 1 and 2 emerged in humans after cross-species transmissions from chimpanzees (type 1), gorillas (type 1), and sooty mangabey monkeys (type 2). Little is known about how the simian immunodeficiency viruses (SIVs) from non-human primate hosts adapted to and subsequently spread within human populations. Using the transmission of SIV from sooty mangabeys (SIVsm) to rhesus macaques as a model system for emergence, we employed deep sequencing to analyze viral populations from infected macaques. We obtained archived plasma samples from four different cohorts, representing both "emerging" (SIVsm-infected) and "established" (SIVmac-infected) viral populations. Computational analysis of the sequences revealed regions throughout the genome that diverged dramatically only in the emerging virus populations, suggesting a role in adaptation following cross-species transmission. Manual comparison of sequences revealed a set of potentially adaptive point mutations, including a previously reported adaptation in Capsid selected by the host restriction factor TRIM5α.

2.3 Introduction

Human Immunodeficiency Virus types 1 and 2 (HIV-1 and HIV-2) emerged in humans following the transmissions of the SIVs of chimpanzees (HIV-1 groups M and N), gorillas (HIV-1, groups O and P), and sooty mangabey monkeys (HIV-2, subtypes A-I), to humans (6, 23). SIVcpz, SIVgor and SIVsm were each transmitted to humans on multiple independent occasions, giving rise to the multiple groups and subtypes of HIV-1 and HIV-2 (3). SIVsm led to the emergence of a second virus in addition to HIV-2, when it was unintentionally transmitted from sooty mangabeys to rhesus macaques in captivity at U.S. primate centers (40, 77). SIVsm adapted to macaques, and eventually emerged as the highly pathogenic species SIVmac (40, 41). These events, along with other evidence (22), suggest that cross-species transmissions of SIVs have been frequent throughout evolutionary history, and that similar species jumps may continue to occur in the future (6, 32). What distinguishes a "successful" cross-species transmission from so-called "dead-end" transmissions, in which the virus is unable to spread to secondary hosts of the new species, remains an open question, but it likely involves rapid adaptation of the virus to the genetic environment of the new host species, and the acquisition of mutations that facilitate further infection (11).

The identification of such mutations for the zoonotic transmissions of SIVs is difficult, and in most cases must be inferred indirectly by comparing sequences of extant viruses, sampled decades or even centuries after the initial cross-species transmission events (94, 106). For example, comparing the ability of HIV-1 Group M and HIV-1 Group N viruses to counteract the antiviral activity of the restriction factor bone marrow stromal cell antigen 2 (BST-2)/Tetherin revealed that only Group M viruses were able to evade restriction, suggesting a correlation between successful spread in human populations and adaptation to counteract human tetherin (93, 94). While successful, such methods of identifying adaptations critical to emergence may

overlook key mutations that occur shortly after transmission, and will likely fail to detect any amino acid "sampling" that the virus population goes through in the process of adaptation. For example, studies investigating viral adaptation to the cytotoxic T lymphocyte (CTL) response within individual hosts often find that multiple primary and compensatory mutations are sampled before the fixation of an optimal escape variant (107, 108). Moreover, without sampling of the viral population from the donor and recipient species, it is not possible to determine whether critical adaptations are present at low-frequency prior to initial transmission, or arise *de novo* during replication in the new host. Generally, for many documented cross-species transmissions, including those involving SIVs, examining sequences of viral populations closer to the initial transmission is not possible due to the lack of appropriate samples available for analysis.

The transmission of SIVsm to rhesus macaques (and the subsequent emergence of SIVmac) provides a uniquely tractable model for investigating the early stages of cross-species transmission and emergence. Since the discovery of SIVmac, strains of SIVsm have often been used for experimental infection of macaques, either as heterologous challenge strains for SIVmac-based vaccinations, or as part of comparative pathogenesis studies (99, 109, 110). Plasma samples from such animals (particularly infected, unvaccinated or untreated control animals) contain a wealth of viral genomic information that can be used to study the evolution of SIV in the context of a recent cross-species transmission. In addition, the precise source inoculum is known and is identical for all animals in a particular study, removing the need to infer estimated ancestral sequences. Access to the source inoculum as well as longitudinal samples post-infection makes it possible to track evolution of the viral population throughout the process of adapting to a new host. This facilitates the investigation of both general questions

about the overall evolution of virus populations in different host contexts, as well as specific questions about adaptations at particular loci and in response to known selection pressures.

Several key adaptations have been identified using this model system, including adaptations at position 98 in the capsid (CA) protein and position 17 in the Vif protein, which allow SIVsm to escape restriction by rhesus TRIM5α and A3G, respectively (85, 92). Other adaptations have also been identified in CA, as well as in the V1/V2 region of the SIVsm envelope glycoprotein (86, 103, 104, 111). A drawback of these studies is that they focused on small regions of the viral genome, and relied on either sequencing of bulk PCR products or multiple clones from the same sample. These strategies can be laborious and can fail to detect minority variants. Here, we take advantage of next-generation sequencing technology to produce whole-genome, population-level sequences of "emerging" SIVsm strains in the process of adapting to the rhesus macaque host, as well as control "established" SIVmac strains. We use this method to evaluate overall evolutionary trends between emerging and established virus populations, and to identify putative rhesus macaque-specific adaptations that may have contributed to the successful emergence of SIVmac.

2.4 Materials and Methods

2.4.1 Deep Sequencing and Analysis of in vivo viral populations

2.4.1.1 Plasma Samples

Plasma samples for emerging virus (SIVsm) cohorts were obtained from Vanessa Hirsch. Plasma samples from SIVmac251-infected animals were obtained from David T. Evans (112), and SIVmac239 samples were obtained from Ronald C. Desrosiers.

2.4.1.2 Sample Processing

Viral RNA was extracted directly from 200ul plasma using the High Pure Viral RNA kit (Roche), and RNA was eluted in 50ul RNase-free water. 8ul of RNA was treated with DNase (Invitrogen), then divided into 4 aliquots to use in one-step RT-PCR (Qiagen OneStep RT-PCR kit) of four amplicons spanning the SIV genome. See Supplemental Table 1 for table of primers used. All primers were synthesized with the 5' amino modifier (Integrated DNA Technologies). Thermocycling conditions were: 45°C for 2 hours, 95°C for 15 minutes, followed by 40-50 cycles of 94°C for 15 seconds, 50°C for 30 seconds, and 68°C for 6 minutes; followed by a final extension at 68°C for 6 minutes. Amplification success was confirmed by gel electrophoresis. Multiple controls were included in order to determine whether cross-contamination of the samples had occurred (water and uninfected plasma controls), as well as to check for DNA contamination in the RNA samples (PCR without reverse transcriptase). None of the negative control samples resulted in visible PCR products after gel electrophoresis. PCR products were purified using the Qiagen PCR Cleanup kit, and then the 4 amplicons from each sample were pooled at 50 ng per amplicon and prepared for sequencing.

2.4.1.3 Illumina Sequencing and variant analysis

Illumina library construction was performed using the NexteraXT (Illumina) kit according to the manufacturer's protocol. Sequencing was performed on the Illumina MiSeq platform. Reads for each sample were assembled de novo using *VICUNA* (113). Variant calling was then performed using *V-Phaser2* (114, 115) and codon frequencies displayed using *V-Profiler* (115).

2.4.1.4 Divergence and Diversity

Divergence and diversity were determined from the *V-Profiler* output, by calculating the percentage of reads at each site that differed from either the consensus of the viral stock corresponding to each sample (divergence), or from each sample's own consensus (diversity). Diversity plots were generated using Microsoft Excel. Divergence due to putative APOBEC3-protein-mediated hypermutation was inferred using the Hypermut tool using the default settings, available at http://www.lanl.gov.

2.4.2 Computational analysis

2.4.2.1 Identification of adaptive loci

For each cohort, we used the dN/dS and dN hotspot methods (described below) to identify codons or groups of codons that rejected the null hypothesis of neutral evolution, for dN/dS, and uniform mutation rates, for dN hotspot. All confidence intervals (CIs) were at 80% confidence level. In neither case did we account for correlations between codons due to an underlying shared phylogeny, but within a cohort the independence across animals lowers the resultant bias and estimates calculated for the different cohorts were independent. With this in mind, we labeled a locus as adaptive when the two emerging cohorts had significantly high values for dN/dS or the dN hotspot estimates, and the two established cohorts did not. Overall, correlations between codons within an animal should raise false positive rates, since some correlated mutations are taken as independent, but this is acceptable given our goals.

2.4.2.2 dN/dS method

We calculated dN/dS using a pairwise sequence comparison method similar to that introduced in (116). A HKY85 mutation model was used. Nucleotide equilibrium frequencies were based on the distribution in the consensus inoculum (π_T=.22, πC=.18, π_A=.35 π_G=.25) and

the transition/transversion ratio was set to 3 following results in (117) for HIV. Results were essentially unchanged for different transition/transversion ratios.

For a given codon, N was defined as the rate of non-synonymous mutations according to the HKY85 model. N' was defined as the summed frequency of observed non-synonymous mutations at the codon over all animals in the cohort. For our dataset, codon mutations had only single nucleotide substitutions, but different mutations had different frequencies within animals and across animals. When we considered a grouping of codons, N and N' were summed across codons in the grouping. To lessen the impact of synonymous mutations, we calculated S' and S analogously to N' and N, but we only considered S' and S across the whole gene. In this way, S'/S served as a baseline rate for synonymous mutations.

We set dN/dS = (N'/N)/(S'/S) and calculated CI assuming N' and S' were Poisson distributed. CI considered each codon or grouping of codons individually (i.e. no multiple test correction was performed).

Using a gene wide estimate for S'/S means that we underestimated dN/dS for regions with low synonymous mutation rate but relatively high non-synonymous mutation rates. But in those cases, S' and S specific to the region have high variance and our estimates would not have been significant even if we used locus specific S' and S.

2.4.2.3 dN Hotspot Method

To explain this method, let M be the number of codons in a certain gene and Ni', Ni for i=1,2,..,M be the observed counts and model rates of non-synonymous mutation as described above for the dN/dS method. Then, after rounding the Ni', we model the Ni' as a multinomial distribution with N' = N1'+N2'+...+NM' trials and Ni' successes in category i. The estimate for the multinomial probability associated with category (codon) i is simply Ni'/N' and we set

pi=(Ni'/N')/(Ni/N) where N = N1+N2+...+N. The denominator, Ni/N, is a scaling factor that makes pi=1 if all categories have equal non-synonymous mutation rates. When we are dealing with M groups of codons, everything generalizes by redefining Ni' and Ni over the M groupings.

We form a confidence set for all pi, i=1,2,..,M as follows. If pi < 1, then we choose $[0,\infty)$ as the CI range for that category. Since pi < 1, we do not want to waste power on this category, for which we cannot reject the null model. For the pi > 1, we build a one sided interval of the form $[x,\infty)$ since we do not want to waste power on developing upper bounds. These one-sided intervals are built simultaneously with a Bonferroni multiple test correction.

2.5 Results and Discussion

2.5.1 Cohort Composition and Sequencing

We gathered archived plasma samples from SIV-infected macaques used in previous studies representing both "emerging" virus populations (macaques infected with SIVsm), as well as control "established" virus populations (macaques infected with SIVmac), (Table 1). We selected samples from two emerging cohorts and two established cohorts. Each group (emerging or established) contained samples from one cohort of macaques that had been infected with a clonal stock, and one cohort that had been infected with an uncloned isolate, or "swarm." This was done in order to control for differences that may be due to the level of heterogeneity of the stock, as has been suggested previously (118). For each cohort, we obtained a sample of the source inoculum/stock, as well as an acute sample (between 2-6 weeks post-infection) and a chronic sample (between 40-72 weeks post-infection) for each animal.

We chose SIVsm strains that were partially adapted to rhesus macaques: SIVsmE543, and the related SIVsmE660. The passage history of the two viruses is as follows: biological

Table 1. Cohorts and samples processed for deep sequencing.

					Genotypes		Acute		Chronic	
Adaptation Level	**Cohort**	**Stock Type**	**Animal**	**Inoculation**	**TRIM5α**	**MHC class I**	**Weeks p.i.**	**Viral Load**	**Weeks p.i.**	**Viral Load**
Established	SIVmac239	Clone	MM001	Intravenous	N.D.	N.D.	2	1.4E+07	40	7.0E+05
			MM002	Intravenous	N.D.	N.D.	2	1.2E+07	40	4.6E+05
			MM003	Intravenous	N.D.	N.D.	2	7.3E+07	40	1.6E+06
	SIVmac251	Swarm	MM004	Vaginal	TFP/TFP	A*01	6	1.3E+06	41	7.3E+05
			MM005	Vaginal	TFP/TFP	A*01, A*08, B*01	5	4.8E+05	41	1.3E+04
			MM006	Vaginal	Q/Cyp	A*08, B*17, B*29	4	1.3E+06	40	6.5E+05
			MM007	Vaginal	TFP/TFP	A*01, A*02	5	3.3E+06	42	1.2E+06
Emerging	SIVsmE543	Clone	MM008	Intra-rectal	Q/Cyp	A*02, A*08	2	5.7E+05	41	1.8E+05
			MM009	Intra-rectal	TFP/Q	B*01	2	1.8E+04	48	3.5E+05
			MM010	Intra-rectal	TFP/TFP	-	3	1.5E+04	72	2.3E+04
			MM011	Intra-rectal	TFP/Q	B*01	3	1.2E+05	56	2.0E+06
	SIVsmE660	Swarm	MM012	Intra-rectal	TFP/Q	-	2	5.7E+05	44	1.4E+05
			MM013	Intra-rectal	TFP/TFP	A*08	2	1.5E+06	44	6.0E+04
			MM014	Intra-rectal	TFP/TFP	A*02	2	6.5E+05	44	8.0E+03
			MM015	Intra-rectal	TFP/Q	A*01	2	1.7E+06	44	1.5E+04

N.D., not done. -, negative for all alleles tested

material from an SIV-infected sooty mangabey, CaE038, was used to infect the rhesus macaque MmF236. Material from MmF236 was then used to inoculate MmE543 (77, 119). A terminal PBMC sample from MmE543 was briefly co-cultured with the human B/T hybrid cell line CEMx174, and genomic DNA isolated from the co-culture was used to clone SIVsmE543 into a lambda bacteriophage vector (77, 120, 121). Inguinal lymph node cells also isolated from MmE543 were used to inoculate MmE660 (119, 122). SIVsmE660 was subsequently isolated by co-culture of spleen tissue from MmE660 with CEMx174 cells (119). Despite having been passaged in macaques, SIVsmE543 and E660 produce variable viral loads when introduced into macaques *in vivo* (85, 120), suggesting they are not yet fully adapted to this host. Importantly, these strains are capable of sustaining relatively high levels of viremia through the chronic phase and can lead to the development of simian AIDS (86, 99). In contrast, when primary, unpassaged strains of SIVsm are introduced into macaques, viral loads often drop precipitously following peak viremia. Chronic samples from such animals often have undetectable viral loads, making their analysis difficult (123, 124). This difference in chronic viremia between unpassaged and minimally passaged SIVsm strains may reflect some adaptation to macaques that occurred during the passaging of SIVsmE543 and SIVsmE660.

In order to compare these emerging SIVsm viruses to established, macaque-adapted viruses, we also assembled samples from cohorts of rhesus macaques infected with SIVmac239 and SIVmac251 (Table 1). SIVmac251 is an uncloned isolate from rhesus macaque Mm251-79, which had been inoculated with minced tissue from a macaque with spontaneous lymphoma (125, 126). Splenocytes from this animal were frozen and later co-cultured with the human T-cell line HuT-78, from which the SIVmac251 strain was isolated (126). Minced lymphoma tissue from Mm251-79 was also used to inoculate several macaques, and pooled blood samples from

three such animals were used to infect Mm61-82, along with other macaques (126). Filtered plasma from Mm61-82 was used to inoculate Mm239-82, from which a terminal serum sample was isolated and subsequently cultured with HuT-78 cells (126, 127). Genomic DNA from these cells was extracted, digested, and cloned into a lambda bacteriophage vector, producing the SIVmac239 infectious molecular clone (127, 128). SIVmac239 and 251 reliably and uniformly maintain high viral loads throughout the chronic phase of infection and lead to simian AIDS within approximately one year, suggesting that they are well adapted to rhesus macaques as a host species (129, 130).

Conditions were optimized for the amplification of whole genomes of SIV via RT-PCR of viral RNA (vRNA) extracted directly from the plasma samples. Primers were designed to amplify vRNA in four overlapping amplicons (Figure 4A, Supplemental Table 1). Primers for the SIVsm samples were designed based on the published SIVsmE543 molecular clone sequence (NCBI accession number U72748), and SIVmac primers were based on the published SIVmac239 molecular clone sequence (NCBI accession number M33262). Primers were designed to target conserved regions of the genome and were validated using multiple strains of SIVsm and SIVmac prior to amplification of samples from the cohorts described above (data not shown).

Plasma sample processing proceeded as follows: vRNA was extracted from plasma, DNase-treated, and used for one-step RT-PCR amplification. Source, acute, and chronic samples from each cohort were processed separately in order to prevent cross-contamination prior to amplification. Amplified products were purified and pooled by animal and time point, totaling 34 genomes. Samples were fragmented, indexed, and sequenced on the Illumina MiSeq platform. Reads were processed using a suite of software developed by the Broad Institute. Briefly, the

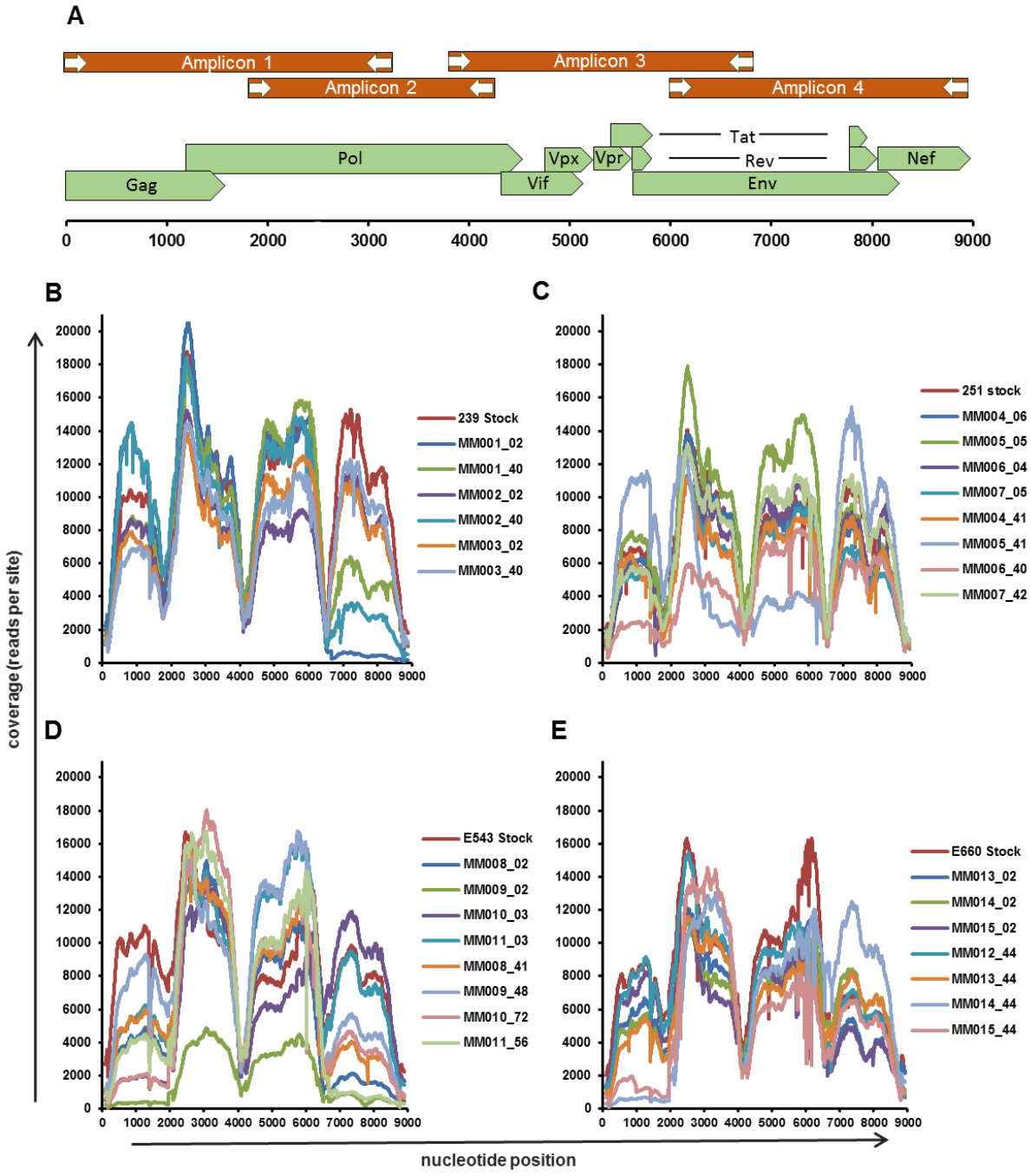


Figure 4. Amplification strategy and coverage.

(A) Schematic of SIV genome with locations of overlapping amplicons used for RT-PCR. (B-E) Sequencing coverage (number of reads per codon) for SIVmac239, SIVmac251, SIVsmE543, and SIVsmE660 cohort samples, respectively.

VICUNA program was used to assemble the reads *de novo*. Consensus sequences from the *de novo* assemblies were then used to align the reads (*Mosaik2*), and *VPhaser2* was used to calculate variant frequencies at each codon in the genome (114, 115).

2.5.2 Coverage and Diversity

To validate the sequencing results, we compared the depth of coverage of each sample, and found that average coverage per sample was 9,213X (sequencing reads per codon, Figure 4B-E), with a range of 2,233-13,306X. We next confirmed that we could observe expected patterns of evolution in the virus populations. For example, infections with uncloned stocks (swarms) typically experience a bottleneck at the point of transmission to a new host, so we expected that in the SIVsmE660 and SIVmac251 samples we would see a substantial loss of diversity at the acute time point compared to the diversity in the stocks (131, 132). Methods of measuring diversity typically involve calculating pairwise genetic distances at individual sites and/or averaged across a region of interest (133). At the time of this analysis, suitable applications/modifications of these methods that could accommodate next-generation sequencing data—particularly datasets with short read lengths (150bp for the Illumina platform) and large number of reads covering each nucleotide position—did not exist ((134) and Todd Allen, personal communication). We therefore sought to simply visually display the diversity of each sample at-a-glance, by calculating diversity at each codon as the percentage of sequencing reads that contained a different codon than the consensus of that sample (118). This analysis did reveal a steep drop in diversity at the acute time point for all animals infected with swarms (Figure 5, Supplemental Figure 1-4). Diversity was predictably low to nonexistent in the clonal stocks, and remained low at the acute time point. In the majority of animals, diversity had increased considerably by the chronic time point (Figure 5, Supplemental Figure 1-4).

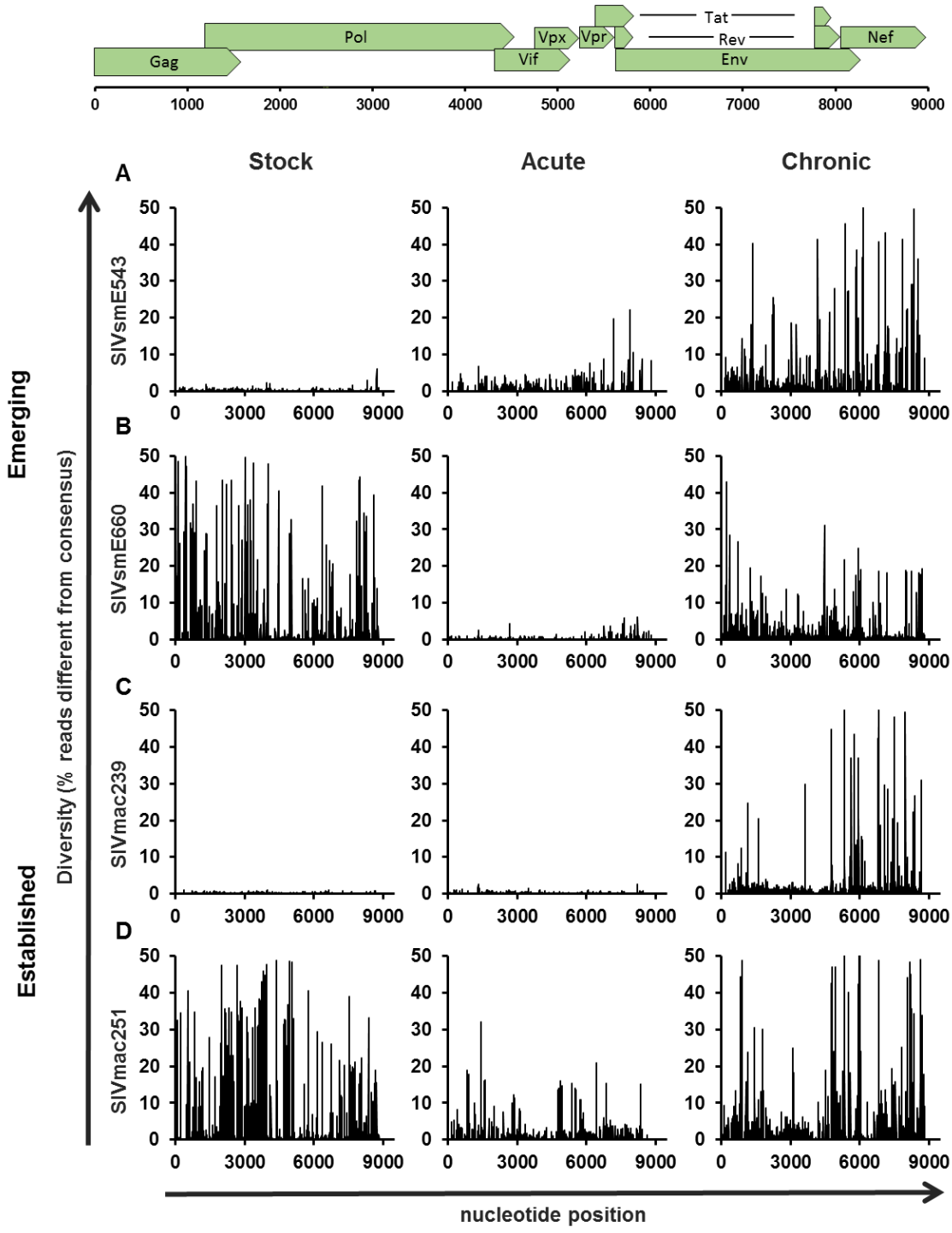


Figure 5. **Diversity of representative samples**.

Diversity, displayed as the percentage of sequencing reads at each site differing from the consensus codon, for a representative set of samples from each cohort (MM003, MM006, MM008, and MM013). A and B, emerging cohorts; C and D, established cohorts. A and C, cohorts infected with clonal stocks; B and D, cohorts infected with swarms.

We used consensus sequences covering the full coding regions of the viruses to construct a neighbor-joining tree, incorporating several sequences of naturally occurring SIVsm strains (135), as well as an HIV-2 outgroup (Figure 6). Each cohort formed a distinct cluster in the tree. We noted a difference between genomes arising from the cloned stocks compared to those from the uncloned stocks. For the cloned stocks, the acute time points tended to cluster with the stock, illustrating the lack of divergence from the source inoculum by the acute time point. In contrast, for the uncloned stocks, the acute and chronic time points tended to cluster by animal, suggesting that the acute genomes were more closely related to the chronic genomes from the same animal than they were to the source inoculum. This observation highlights the impact of the transmission bottleneck on the evolution of virus populations from uncloned stocks.

2.5.3 Early APOBEC3-mediated hypermutation

We hypothesized that the emerging virus populations would be more sensitive to restriction by rhesus macaque APOBEC3 proteins. To investigate this hypothesis, we used the Hypermut tool from the Los Alamos National Laboratories website (www.hiv.lanl.gov) to look for evidence of APOBEC3-mediated hypermutation in the consensus sequences of each sample. Hypermut analyzes sequences for specific nucleotide changes from a reference sequence to determine if there is an enrichment for G to A substitutions (136). Alignments of sequences from each cohort relative to the stock were generated using Geneious and uploaded to the Hypermut web interface. We detected a strong signature in one SIVsmE543-infected animal (MM009) at the acute time point (Figure 7, top). Examination of the sequences from animal MM009 revealed that hypermutation resulted in the introduction of one premature stop codon, located in the CT of

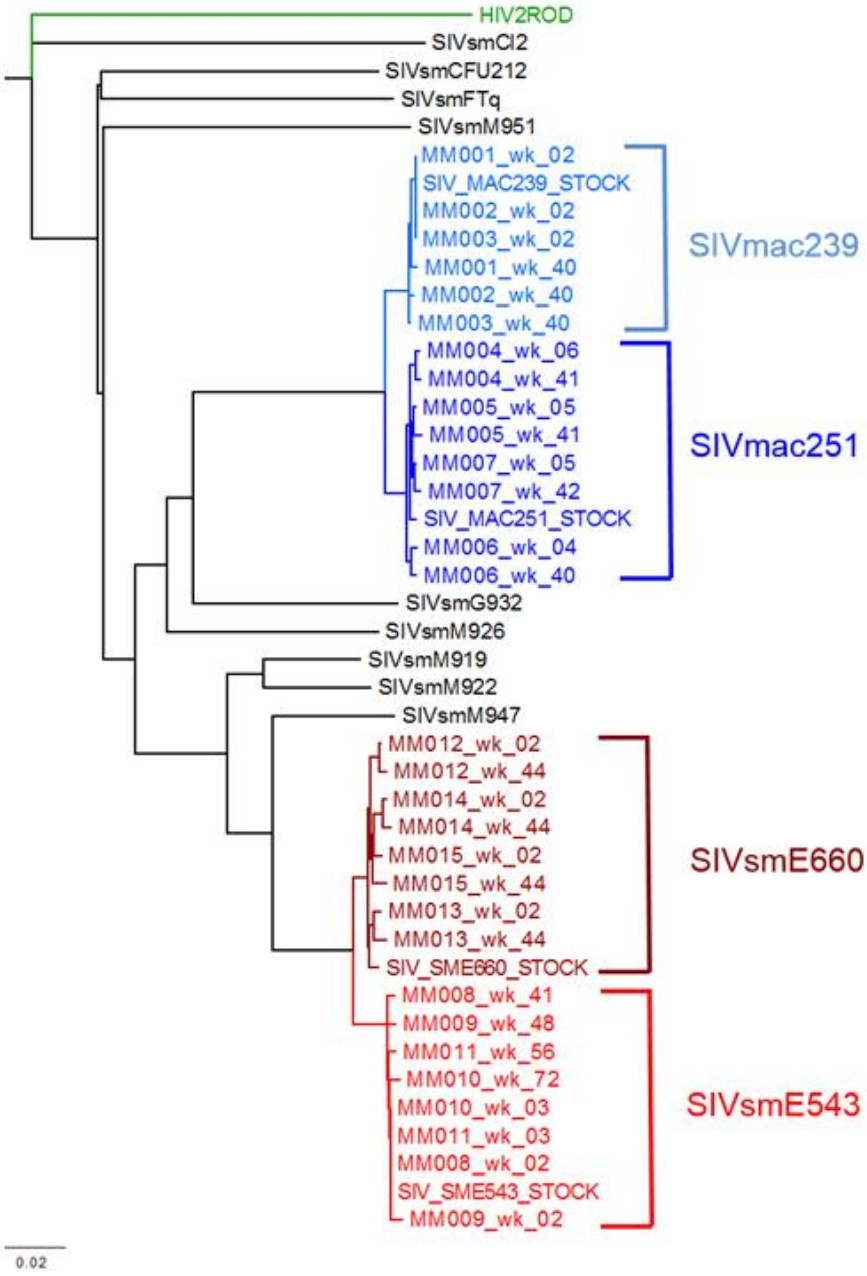


Figure 6. **Whole genome neighbor-joining tree of all samples**.
Interleaved with samples from this study are naturally-occurring SIVsm sequences (135). The tree is rooted on the HIV-2 ROD strain.

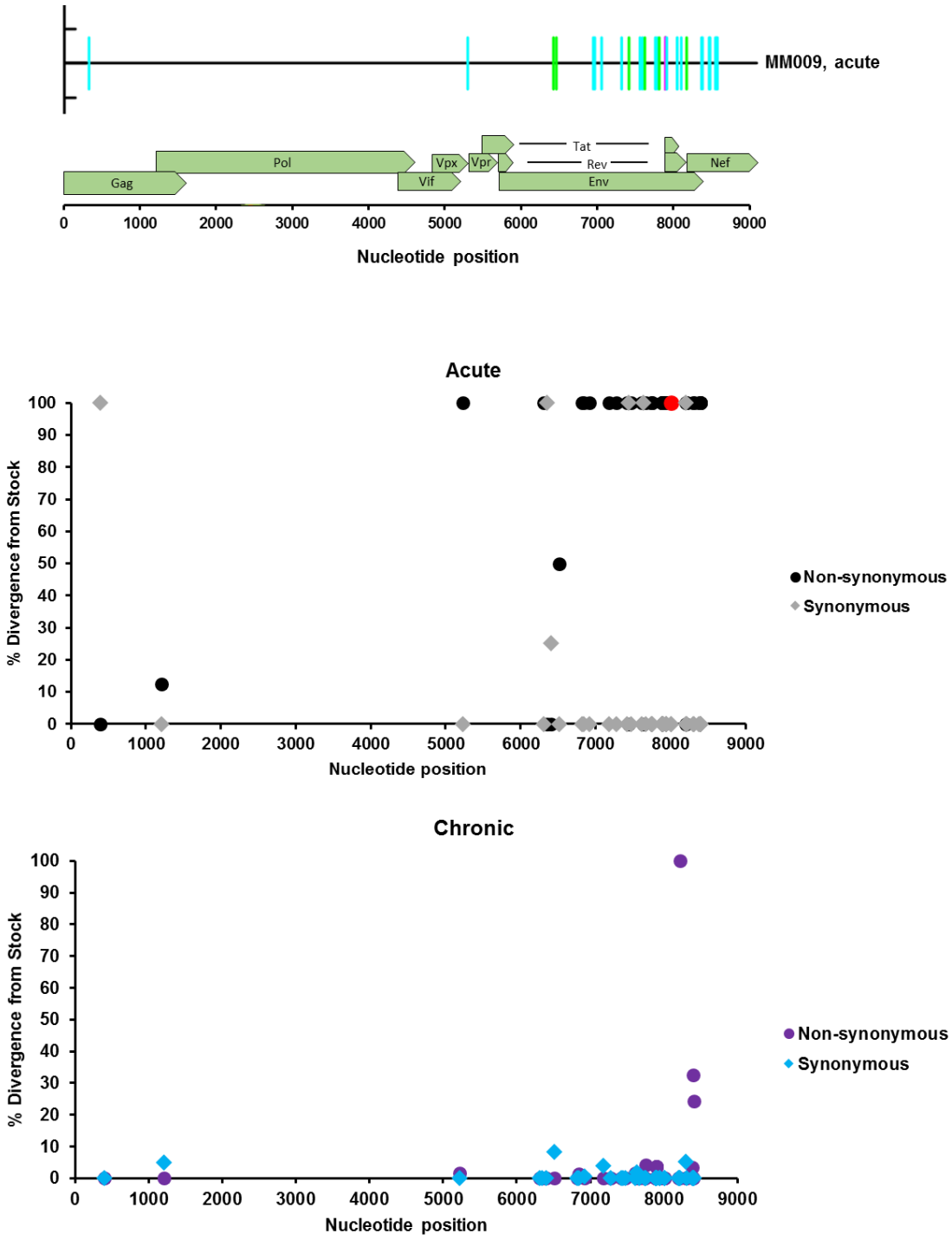


Figure 7. APOBEC3 hypermutation.

Top panel: Hypermut output for SIVsmE543-infected animal MM009, acute time point. Cyan lines, GA to AA mutations; green lines, GC to AC mutations; magenta, GT to AT mutations. Middle panel: sites in animal MM009 where G to A changes occurred at the acute time point (red dot indicates a premature stop codon). Bottom panel: divergence at acutely hypermutated sites at the chronic time point**.**

Env (Figure 7, middle panel and Supplemental Table 2). The majority of the changes in MM009 presumably introduced by APOBEC3 proteins early in infection reverted to the stock consensus sequence by the chronic time point, suggesting that these mutations were deleterious (Figure 7, bottom and Supplemental Table 2).

2.5.4 Candidate adaptations

We next wanted to determine if there were any overall differences in sequence evolution between the emerging virus cohorts and the established virus cohorts. We reasoned that, while each virus population would have been subjected to animal-specific selection pressures such as CTL and antibody responses, the emerging virus populations would have experienced additional selective pressures imposed by a novel host environment. Genetic differences between sooty mangabeys and rhesus macaques, particularly in loci that have an impact on viral replication, could drive additional evolution that would not occur in the established virus populations. We therefore hypothesized that there might be different patterns of divergence from stock in the emerging virus cohorts due to the additional species-specific selection pressure.

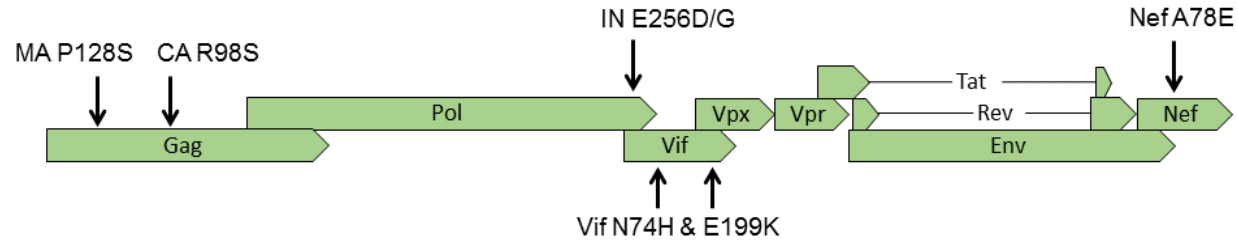
To address this, we took two approaches to identify substitutions of potential importance for the adaptation of SIVsm to the rhesus macaque host. First, we manually inspected sequences for sites that conform to a set of predictions (criteria). These are based on reports that have observed convergent evolution producing the same or similar changes in independent cross-species transmissions of PLVs (85, 86, 92, 106). Second, to cast a wider net and to avoid *a priori* assumptions, we also took a computational approach to identify sites of potential positive selection unique to the SIVsm cohorts.

2.5.4.1 Criteria-based candidate identification

For the first approach, we generated an alignment of all consensus sequences and manually searched for sites that met a majority of the following criteria: (1) Did multiple SIVsm-infected animals acquire mutations at that site? (2) Does the substitution match the residue at the equivalent position in SIVmac? (3) Is the substitution rarely found in natural SIVsm sequences and commonly found in SIVmac sequences? These questions became useful for distinguishing substitutions that were potentially relevant for species-specific adaptation to rhesus macaques, from animal-specific or random mutations. Based on these criteria, six putative adaptations were selected for further study (Figure 8). The candidates were spread throughout the SIV genome, appearing in both structural regions (*gag, pol*) as well as accessory genes (*vif*, *nef*).

Among the candidates was a substitution that had been previously identified as an adaptation that enables SIVsm to escape restriction by certain alleles of rhesus TRIM5α *in vivo* (85). TRIM5α is polymorphic in rhesus macaques, and the presence of either a TFP or a ΔΔQ at positions 339-340 in the C-terminal SPRY/B30.2 domain of rhTRIM5α determines that allele's restrictive capacity against SIVsm (84, 85). Specifically, SIVsm is potently restricted by TFP alleles of rhTRIM5α (85). This restriction is partially explained by the residue at position 98 in the SIV CA protein. In SIVsm sequences from naturally infected sooty mangabeys, an arginine (R) is almost always found at this position, while a serine (S) is almost always found in SIVmac sequences (data not shown). Restriction assays demonstrated that the residue at CA98 had a profound impact on the ability of rhTRIM5α alleles to restrict different SIV strains (85, 86). We predicted that we would see this R to S change in our SIVsm-infected cohorts, because the majority of the animals had at least one copy of the restrictive TFP allele (Table 1). Indeed, in the SIVsm cohorts, this position changed in all of the SIVsm-infected animals possessing a TRIM5α-TFP allele. In 5 of those animals, we saw the exact R to S change we predicted

A



B

			Candidate adaptations					
Time Point	**Cohort**	**Animal**	**MA 128**	**CA 98**	**IN 256**	**Vif 74**	**Vif 199**	**Nef 78**
Inoculum	SIVsmE543	STOCK	P	R	E	N	E	A
	SIVsmE660	STOCK	P	R	E	N	E	A
Acute	SIVsmE543	MM008	P	R	E	N	E	A
		MM009	P	R	E	N	E	A
		MM010	P	R	E	N	E	A
		MM011	P	R	E	N	E	A
	SIVsmE660	MM012	P	R	E	N	E	**E**
		MM013	P	R	E	N	E	**E**
		MM014	P	R	E	N	E	**T**
		MM015	P	R	E	N	E	A
Chronic	SIVsmE543	MM008	P	R	E	**H**	E	A
		MM009	**S**	**S**	**G**	**H**	E	A
		MM010	P	**G**	E	N	**K**	A
		MM011	P	**T**	E	**H**	E	A
	SIVsmE660	MM012	P	**S**	**D**	N	E	**E**
		MM013	P	**S**	**D**	N	E	A
		MM014	P	**S**	**G**	N	E	**T**
		MM015	P	**S**	E	**H**	**K**	**T**

Figure 8. Candidate adaptations.

(A) Schematic of genomic locations of candidate adaptations. (B) Majority variant present at each putative adaptation site in all SIVsm samples. Numbering based on SIVsmE543. Residues highlighted in red and bolded have diverged from the stock majority variant. Variant frequencies at candidate loci for all SIVsm and SIVmac samples are shown in Supplemental Table 3.

(Figure 8, Supplemental Table 3). In the two remaining animals, we observed sampling of potentially sub-optimal residues, with the original R changing to a threonine (T) in one animal and to a glycine (G) in another. In animal MM008, which was the only SIVsm-infected animal without a copy of the rhTRIM5α TFP allele, the R residue remained unchanged throughout the time points sampled, and all SIVmac-infected animals maintained a serine residue throughout the course of the infection (Supplemental Table 3). All of these observations support the conclusion that the observed changes at CA98 were selected by rhTRIM5α.

We could not draw such direct conclusions about the selection pressures that led to the emergence of the other candidate adaptations we identified, although we speculate that the two changes in Vif may have been selected by rhesus-specific alleles of APOBEC-family proteins. SIVsm/mac Vif position 74 is homologous to position 71 in HIV-1 Vif, which falls within a region previously shown to govern interaction of HIV-1 Vif with both human A3G and A3F (137, 138). This position is located in the middle of a YXXL motif that is conserved in many PLV Vifs (137, 138). The YXXL pattern is also classified as one of several functionally interchangeable "late domain" motifs, which are essential for the replication of many lentiviruses and are important for the interactions of viral proteins with host proteins (139). The second adaptation is at SIV Vif position 199, which has no direct equivalent in HIV-1 (HIV-1 Vif is only 192 amino acids in length based on the HXB2 reference sequence).

SIVmac/sm Nef is known to counteract the antiviral activity of Tetherin, and so it is tempting to speculate that this restriction factor selected for the change we observed at position 78 in Nef. However, in previous studies, restriction assays did not reveal any differential restriction of SIVsm and SIVmac by rhesus Tetherin, which would suggest that it was not the selective force responsible for this substitution (105). Nevertheless, restriction assays are

typically conducted in the context of both Nef and Tetherin overexpression, and so a subtle restriction effect could have been overlooked in previous experiments. Identifying a possible functional role of residue 78 by homology with HIV-1 did not provide any alternative hypotheses, as the corresponding position in HIV-1 (S46) is not known to have any functional role in HIV-1 replication (140).

The substitution in MA is particularly intriguing because it directly follows a distinctive PTAP motif in SIVsm MA. PTAP is also a late domain motif, and viruses in the SIVsm/SIVmac/HIV-2 lineage are unique among PLVs in having a late domain in MA (141). All retroviruses have at least one late domain in Gag, which is essential for proper assembly and budding of nascent virions. As mentioned above, the motif YXXL is a late domain, as is PPPY (139). Other lentiviruses tend to have PTAP motifs at the very C-terminus of the Gag polyprotein, which is p6 in the case of HIV-1 (139). The function and binding partners of the MA late domain have not been investigated to our knowledge, but late domains in p6 are essential for replication of HIV-1 and mediate interaction with the host protein apoptosis-linked gene 2-interacting protein, or ALIX, which is important for budding (51, 142). HTLV-1 also possesses a late domain at the C-terminus of its MA protein, which plays a role in Gag interaction with the host assembly factor Tumor susceptibility gene 101 (Tsg101) (141). It is possible that this substitution could have been selected to enhance an interaction of MA with a late-domain interacting host protein.

The substitution at position 256 in IN is located in the C-terminal domain, which is relatively divergent across PLVs (143). This lack of conservation could suggest an interaction of the C-terminal domain with host proteins that differ between PLV hosts, or could simply indicate mutational flexibility in this region. To our knowledge, this position has not specifically been

reported to have a functional role in either SIV or HIV, however it falls within a region of HIV-1 IN that is required for interaction with Ku70, a host DNA repair protein that shields IN from proteasomal degradation during infection (144).

2.5.4.2 Computational candidate identification

dN/dS analysis is the most common method of inferring and quantifying selection from sequence data. This analysis is most often applied to diverse/divergent sequences, such as sequences of gene homologs from different animal species (145, 146). In the case of viruses, dN/dS analysis is most often conducted using either sequence alignments of different virus lineages (147), or alignments of viruses from the same lineage but across many different individuals (135). In the context of HIV/SIV evolution, Bonhoeffer, et al. calculated dN/dS for variable loop 3 (V3) in Env using pairwise sequence comparison (i.e., by counting the number of synonymous and non-synonymous differences between pairs of sequences and then averaging over all sequences) and found $dN/dS > 1$ for sequences collected in year 3 of infection but not later (148). Subsequent studies considered dN/dS on the codon scale and used phylogenetic methods rather than pairwise sequence comparison to estimate the number of synonymous and non-synonymous mutations (104, 116, 117, 149-151). Such methods applied to Env revealed codons with $dN/dS > 1$ at times when $dN/dS < 1$ for Env as a whole (117). Other studies have used a generalization of the McDonald-Kreitman test to estimate the fraction of mutations that were selected in Env (104, 152-154).

While the later dN/dS studies demonstrate the advantage of considering selection at individual codons, dN/dS analysis on the codon level has little statistical power to reject neutral evolution when sequence diversity is low (155). In particular, when diversity is low, estimates of dS have high variance and confidence intervals (CIs) for dN/dS become wide (116, 149, 155).

Correspondingly, most previous HIV/SIV analyses have either focused on Env, which displays significant sequence diversity relatively early in infection, or considered data years into infection or across hosts. In this vein, Vanderford, et al. considered Env early in infection but for a highly diverse inoculum and large inoculum doses, thereby purposely, as the authors note, creating high levels of diversity even during acute infection (104). McDonald-Kreitman methods share similar problems, but further, the assumption of neutrality for polymorphic sites in such methods is likely not true for intrahost HIV/SIV evolution.

In contrast to the studies mentioned above, our dataset has low levels of diversity, with sequences collected during the first year of infection differing from the inoculum consensus at 1-4% of nucleotides and with codon mutations almost exclusively (>99%) associated with single nucleotide substitutions. Further, in our dataset approximately 99% of codons have no synonymous mutation or no non-synonymous mutation, placing our dataset well within a regime for which codon scale dN/dS analysis has low statistical power. Moreover, dN/dS analysis on the gene level does us little good in pinpointing specific cross-species adaptations and, in fact, for every cohort over all genes we calculated $dN/dS < 1$.

Another distinguishing feature of our dataset is lack of linkage information. Previous dN/dS studies based on deep sequencing have considered small genomic regions, where reads do provide linkage information (e.g. (156)), but our aim was to find loci of cross-species adaptation across the full genome. One possible approach would be to build phylogenies in a windowed manner across the genome, similar in spirit to existing methods for dN/dS analysis in the context of recombination (157). However, given read lengths of roughly 150 base pairs, windows would be roughly 50 codons, leading to significant computational and statistical challenges in analyzing data across a roughly 3000 codon genome. Using consensus sequences of the full genome, while

computationally feasible, fails to take advantage of the depth of our study by leaving out non-consensus variants.

While low diversity and lack of linkage information make dN/dS analysis on our dataset problematic, the experimental nature of our dataset and our desire to identify candidate adaptations for *in vitro* fitness assays presented some simplifications. First, for each cohort we know the ancestral sequence, i.e. the inoculum. Second, since codon mutations are almost exclusively associated with a single nucleotide substitution from the inoculum consensus, there is no uncertainty in the mutational pathway at each codon. Third, we know that evolution across animals is independent (i.e. has a star phylogeny). As a result, we can use simple statistical methods when considering mutations across animals, even while mutations within animals may reflect interdependencies flowing from an unknown underlying phylogeny. Finally, our goal is to identify loci associated with cross-species adaptation for further *in vitro* experimentation. With this in mind, we are not overly concerned with false positives (loci labeled as adaptive which are not), but prefer to lessen false negatives (loci unlabeled as adaptive which are).

With these observations in mind, we used two methods to identify candidate adaptations: a dN/dS method and a method to detect non-synonymous mutation hotspots, which we refer to as the dN hotspot method. For each method we considered single codons as well as grouping codons together through a windowing approach. Grouping codons reduced the stochasticity through averaging and served as a middle ground between single codon and whole gene analysis. As detailed below, dN/dS estimates had wide CIs and were not particularly informative, as expected given the low diversity level of our dataset. dN hotspot estimates were more informative.

In the dN hotspot method we looked for loci (either codons or codon groupings) in which non-synonymous mutations are enriched relative to other loci in the gene. Put another way, while dN/dS at a locus compares dN against dS at that locus, our dN hotspot approach compares dN at that locus against dN across the gene. For each locus we calculate an estimate p that is the rate of non-synonymous mutation at the locus relative to the gene as a whole. So, $p > 1$ is a sign of enriched non-synonymous mutation.

The advantages of the dN hotspot method are that we ignore synonymous mutations which have high variance in our dataset and that standard multinomial statistical methods allow us to form estimates and CI. The disadvantage is that we are inferring mutation rates rather than selection rates, so we may miss selected loci with low mutation rates. However, given the low diversity of our dataset, loci for which mutation rates are low will either have no mutations or too few mutations for meaningful inference.

2.5.4.2.1 Single Codon Results

We calculated dN/dS and dN hotspot estimates for single codons. In general, dN/dS values had wide CI's. We found $dN/dS > 1$ with confidence level exceeding 80% for 116 codons over the 4 cohorts (27 in the SIVsmE543 cohort, 45 in SIVsmE660, 26 in SIVmac239, and 18 in SIVmac251). Of those codons, only 3 had significant $dN/dS > 1$ for the emerging cohorts while not being significant for the established cohorts, which were our criteria for labeling a codon as a candidate adaptation: Gag 233/CA 98, Env 132, and Env 421.

In contrast, the dN hotspot method identified more candidate adaptations. We found $p > 1$ with confidence level exceeding 80% for 212 codons (47 in the SIVsmE543 cohort, 71 in SIVsmE660, 29 in SIVmac239, and 65 in SIVmac251). Of those codons, 8 codons had

significant p > 1 for the SIVsm cohorts but not for the SIVmac cohorts: Gag 120/MA 120, Gag 233/CA 98, Tat 90, Tat 96, Env 103, Env 421, Nef 165, and Nef 74.

Table 2 shows the dN/dS and dN hotspot estimates along with CIs for the 9 codons identified. One codon, Env 132, was identified through dN/dS but not the dN hotspot method. Two codons, Env 421and Gag 233/CA 98, were identified through both dN/dS and the dNhotspot method. Six codons; Env 103, Gag 120/MA 120, Nef 165, Nef 74, Tat 90, andTat 96; were identified solely through the dN hotspot method. For both methods, the codon with the highest estimates was Gag 233/CA 98, which is associated with escape from rhTRIM5α described in the previous section.

2.5.4.2.2 Grouped Codon Results

Figure 9 shows the data and dN/dS results for Gag and the SIVsmE543 cohort. For Gag, we considered codons grouped in windows of 25 codons, so each window is roughly 5% of the gene. The bottom panel shows the location and frequency of synonymous and non-synonymous mutations relative to the consensus as well as the location of the windows. The middle panel shows the dN/dS estimates and associated CI at a confidence level of 80% for each window. This middle panel is analogous to CI presented in Table 2. The top panel provides annotation information. As can be seen in the middle panel, even with an aggressive confidence level of 80%, the dN/dS CI are wide and no estimate of dN/dS > 1 is significant, meaning we are unable to reject the null hypothesis of neutral evolution. For example, for the window Gag 100-125, dN/dS is estimated as 2.2, but the 80% CI falls far below 1. The window Gag 225-250, which contains the adaptive Gag 233/CA 98 codon, has a dN/dS estimate of 5.5, but here the 80% CI falls just below 1. Analysis for all genes and cohorts is shown in Supplemental Figure 5-13.

Figure 10 extends Figure 9, but now the dN hotspot results take the place of the dN/dS

Table 2. Candidate adaptive codons identified through dN/dS and dN hotspot analysis. 80% CI for dN/dS and dN hotspot estimates are shown in parentheses. For each codon, CIs that allow rejection of the null hypothesis of neutral evolution for dN/dS and uniform mutation for dN hotspot are highlighted: SIVsm in purple (dN/dS) and yellow (dN hotspots), SIVmac in green. Codons identified as candidate adaptations by dN/dS (dN hotspot) are visualized by purple (yellow) highlight and no green highlighting in the dN/dS (dN hotspot) column. Two codons were identified by both methods (Env 421, Gag 233/CA 98). One codon was identified only by dN/dS (Env 132). Six codons were identified only by dN hotspot (Env103, Gag 120/MA 120, Nef 165, Nef 74, Tat 90, Tat 96).

Codon	Cohort	dN/dS	dN Hotspot
Env 103	SIVmac239	4.9 (0,169.7)	6.8 (0,Inf)
	SIVmac251	0 (0,56.4)	0 (0,Inf)
	SIVsmE543	15.5 (0.7,146.3)	25.5 (7,Inf)
	SIVsmE660	7.8 (0.3,41.3)	15.2 (9.4,Inf)
Env 132	SIVmac239	14.6 (0,191)	18.1 (2.3,Inf)
	SIVmac251	7.7 (0.5,89)	14.3 (2.4,Inf)
	SIVsmE543	33.6 (4.5,193.5)	55.6 (25.1,Inf)
	SIVsmE660	26.7 (8,89.2)	46.4 (47.2,Inf)
Env 421	SIVmac239	0 (0,158)	0 (0,Inf)
	SIVmac251	0 (0,63.2)	0 (0,Inf)
	SIVsmE543	35.9 (4.2,180.8)	63.5 (34.8,Inf)
	SIVsmE660	11.2 (1.6,46.2)	19.4 (18.5,Inf)
Gag 120 / MA 120	SIVmac239	0 (0,1554.1)	0 (0,Inf)
	SIVmac251	0 (0,51)	0 (0,Inf)
	SIVsmE543	22.2 (0.9,436.3)	45.8 (15.1,Inf)
	SIVsmE660	9.4 (0.3,54.5)	52.3 (39.4,Inf)
Gag233 / CA98	SIVmac239	0 (0,1748.4)	0 (0,Inf)
	SIVmac251	0 (0,57.4)	0 (0,Inf)
	SIVsmE543	106.9 (13.5,685.3)	229 (174.9,Inf)
	SIVsmE660	29.8 (7.3,101.1)	172.1 (186.1,Inf)
Nef 165	SIVmac239	0 (0,5795.8)	0 (0,Inf)
	SIVmac251	0 (0,218.4)	0 (0,Inf)
	SIVsmE543	29.5 (2.7,195)	34.7 (21.1,Inf)
	SIVsmE660	5.4 (0.2,34.1)	12 (7.4,Inf)
Nef 74	SIVmac239	0 (0,5432.2)	0 (0,Inf)
	SIVmac251	0 (0,204.7)	0 (0,Inf)
	SIVsmE543	22.4 (2.7,195)	26.3 (14.9,Inf)
	SIVsmE660	3.8 (0.2,30.3)	9.6 (5.2,Inf)
Tat 90	SIVmac239	0 (0,175.4)	0 (0,Inf)
	SIVmac251	0 (0,105.3)	0 (0,Inf)
	SIVsmE543	24.8 (0.3,306.1)	43 (16.6,Inf)
	SIVsmE660	19.7 (0.4,704.4)	11.4 (4.4,Inf)
Tat 96	SIVmac239	0 (0,198.1)	0 (0,Inf)
	SIVmac251	0 (0,118.9)	0 (0,Inf)
	SIVsmE543	14.4 (0.3,292.8)	26.1 (8.2,Inf)
	SIVsmE660	10.8 (0,425.4)	6.5 (2.2,Inf)

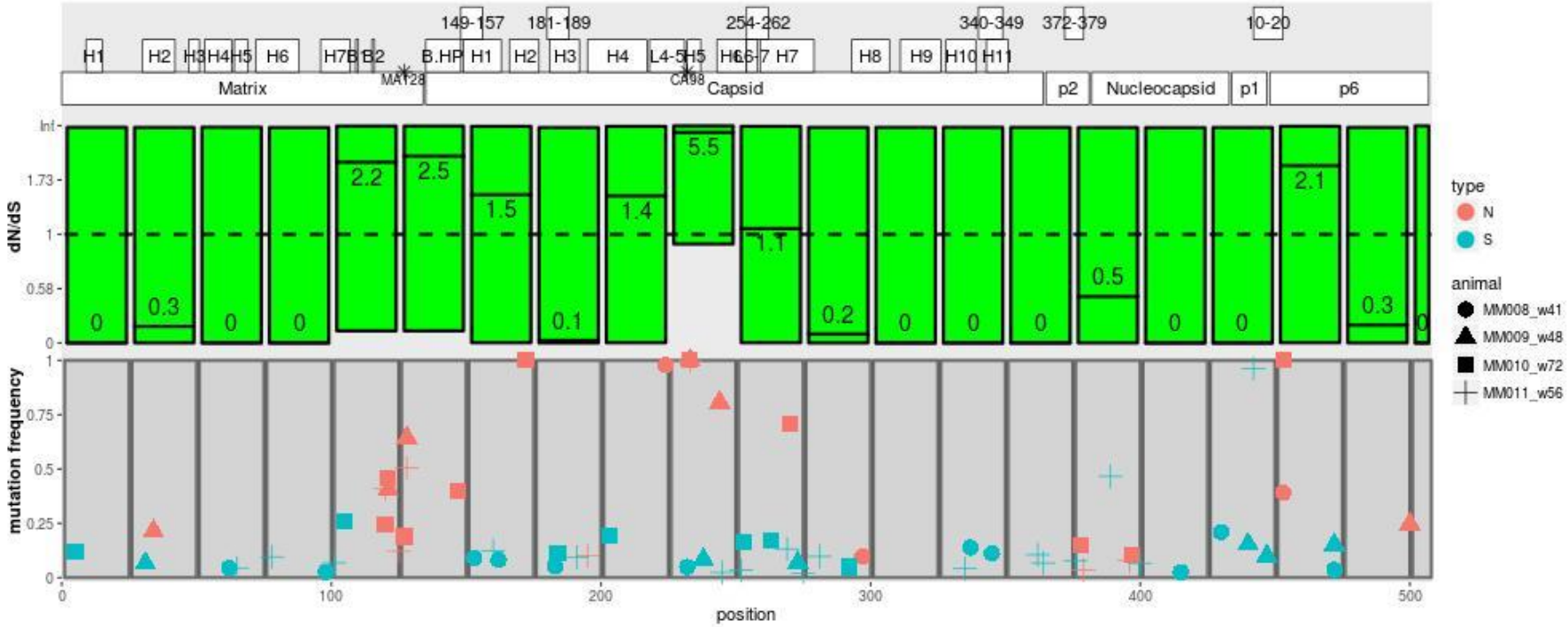


Figure 9. dN/dS estimates for Gag evolution in the SIV smE543 cohort.

The bottom panel (mutation frequency) shows the frequency of non-synonymous (red) and synonymous (blue) mutations across the four animals in the cohort, with the x-axis representing the nucleotide position at the start of each codon. In the legend, the animal number is followed by the week post infection. Codons are grouped in windows according to the rectangles shown. The middle panel (dN/dS) shows the 80% confidence interval (CI) for codons grouped by windows. Numbers shown represent the estimated dN/dS value. Note the y-axis is not on a linear scale, for orientation the dashed black line shows dN/dS=1. The top panel provides Gag annotation information, with the individual proteins at the bottom, structural features in the middle (H = helix, B = beta strand, B.HP = beta hairpin, L = loop), and known CTL epitopes at the top. Asterisks mark locations of candidate adaptations identified in section 2.5.4.1.

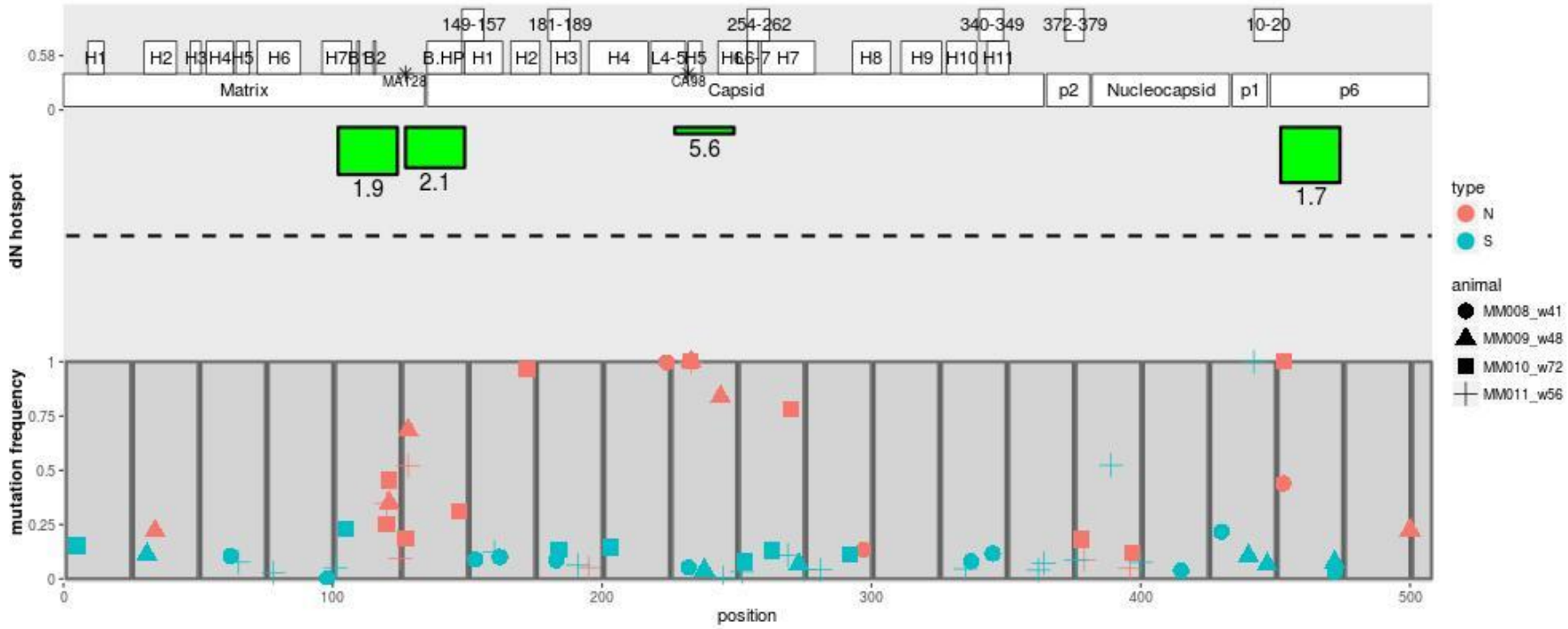


Figure 10. dN hotspot estimates for Gag evolution in the SIV smE543 cohort.

This figure is analogous to Figure 9, but the middle panel now shows dN hotspot estimates rather than dN/dS estimates. Numbers shown represent the lower end of the CI for p, the dN hotspot estimate, rather than the p itself. Note the y-axis is not on a linear scale, for orientation the dashed black lines shows p=1. Four windows have p > 1 with greater than 80% confidence, in contrast to the dN/dS estimates for which no window had dN/dS > 1 with 90% confidence.

results in the middle panel. For the dN hotspot results, the numbers shown are not the estimates of p but rather the lower bounds of the CI. For example, in the window Gag 450-475, $p > 5.6$, meaning that the non-synonymous mutation rate in the window is greater than 5.6 times the rate seen over the whole gene. While no window had $dN/dS > 1$ with 80% confidence for dN/dS, we find $p > 1$ with confidence 80% for 4 windows: Gag 100-125, Gag 125-150, Gag 225-250, and Gag 450-475. Across all genes we find 6 windows with significant $p > 1$ in the SIVsm cohorts but not in the SIVmac cohorts: Gag 100-125, Gag 125-150, Gag 225-250, Nef 160-180, Pol 909-959, and Tat 80-100 (Supplemental Figure 5-13). Nef 160-180 includes a locus for a known CTL epitope and is likely not associated with cross-species adaptation.

The results of the deep sequencing generated several new hypotheses to investigate, particularly regarding the candidate adaptations discussed above. In order to begin to assess the potential functionally adaptive roles of the substitutions we observed, we first sought to determine whether these putative adaptations actually provided a replicative benefit to SIVsm in the rhesus host cell environment. We addressed this by systematically evaluating each candidate for its impact on relative fitness, which will be discussed in detail in section 3.0.

3.0 Fitness effects of SIVsm adaptations to the rhesus macaque host

3.1 Attributions

The data in this chapter are part of a manuscript in preparation by Alison K. Hill, Sergio Ita, Max Mangano, Jennifer Morgan, and Welkin E. Johnson.

AKH and WEJ designed research and wrote the manuscript. AKH, MM, and JM engineered recombinant viral plasmids and produced viral stocks. AKH performed infections and data analysis for FitSeq assays. AKH, SI, and MM processed samples for FitSeq assays.

This research was carried out in part with resources from the Emory Center for AIDS Research, NIH Grant # P30-AI-50409.

3.2 Abstract

Cross-species transmission of viruses into new hosts is often characterized by extensive adaptation of the virus to the novel host species. The primate lentiviruses (PLVs) are particularly well-studied models of adaptation to new hosts due to their many documented cross-species transmissions, including the transmissions of SIVcpz, SIVgor, and SIVsm to humans, leading to the emergence of HIV-1 and HIV-2. SIVsm was also transmitted to rhesus macaques in captivity, causing the emergence of SIVmac. Many studies have pointed to host restriction factors as significant selection pressures that both suppressed the emergence of new viruses and selected for resistant emergent variants, but the effect of acquiring these escape mutations on the fitness of the emerging virus population has yet to be addressed. Here, we report the development of FitSeq, a quantitative assay that measures the relative fitness of viral variants. While the flexibility of the FitSeq assay can be applied to answer many questions about viral fitness, we have used it to assess the fitness effects of rhesus macaque-specific adaptations acquired by SIVsm after cross-species transmission. Applying the FitSeq assay to this model system revealed that mutations in the Capsid, Matrix, and Vif proteins provide subtle yet

significant fitness advantages to SIVsm in rhesus macaque T cells, supporting the hypothesis that these adaptations facilitated the emergence of SIVmac.

3.3 Introduction

Since the discovery of the zoonotic origins of HIV-1 and HIV-2, numerous studies have been conducted in order to understand the molecular mechanisms of these zoonoses (6, 23). In particular, studies have sought to identify important selection pressures that viruses ancestral to HIV-1 and HIV-2 (SIVcpz, SIVgor and SIVsm) faced upon first infecting humans. Our lab and others have also investigated the emergence of SIVmac after cross-species transmission of SIVsm to rhesus macaques in order to better understand how emerging virus populations adapt to new host species. A preponderance of evidence suggests that host restriction factors, including TRIM5α, APOBEC3G, and Tetherin (described in detail in section 1.3.3), presented significant selection pressure to emerging PLV populations, and also selected for escape variants that eventually facilitated emergence (23, 85, 86, 92, 95, 105, 158, 159).

Despite what is known about the importance of these selection pressures, little has been done to quantify their effects on viral fitness. Adaptations that facilitate emergence after cross-species transmission are typically validated using restriction assays, in which a known antiviral host factor and/or viral antagonist is overexpressed in a permissive cell line (85, 92, 94, 105, 159). While powerful, these approaches may overlook subtle differences that cannot be detected in conditions of host and/or viral protein overexpression. Furthermore, they are typically limited to only a portion of the viral replication cycle and/or use viruses that are replication-defective. These caveats preclude the comparison of the relative effectiveness of different restriction factors against emerging viruses, which would enhance our understanding of how viruses can emerge in new species despite formidable host defenses.

We therefore developed FitSeq: a fitness assay in which full-length, replication-competent wild-type and mutant viruses (containing potentially adaptive mutations) dually infect cells. Viral RNA from infected supernatants is used as a template for targeted RT-PCR and products are sequenced using massively parallel sequencing technology, enabling multiplexing of numerous samples in a single sequencing run. The frequencies of mutant and wild-type are determined by their relative counts in the resulting sequencing reads. Fitness values are then derived from linear regression analysis of the change in relative variant frequencies over time. FitSeq has the potential to be applied to numerous virus strains and cell types, making it more flexible than conventional methods of testing adaptations, in addition to providing a sensitive determination of their fitness effects. It is also more flexible than methods that utilize allele-specific PCR probes, which require optimization each time a different region of the genome is targeted, or a different strain of virus or a different point mutation is used. FitSeq requires only target-specific RT-PCR primers and so can be applied easily to different viral strains and target regions. FitSeq is also scalable, and can be applied to cell culture infections in a variety of formats, though the work presented here has been done primarily in 96-well plates.

Our previous work using the SIVsm transmission to rhesus macaques as a model for primate lentiviral cross-species transmission led to the identification of several potentially adaptive mutations that may have facilitated the emergence of SIVmac (section 2.0). We tested four of those mutations in the FitSeq assay: MA-S128P, CA-R98S, IN-E256D, and Vif-N74H. We found that the mutations in MA, CA, and Vif all provided a fitness advantage over wild type SIVsm in rhesus cells, while the mutation in IN was neutral. This represents the first *in vitro* fitness assessment of adaptive mutations in SIVsm that may have contributed to the emergence of SIVmac.

3.4 Materials and Methods

3.4.1 Viral mutant production

Mutations were introduced into parental SIVsmE543 or SIVmac239 either by site-directed mutagenesis, or by subcloning synthetic gBlocks (Integrated DNA Technologies) directly into the parental viral plasmid. Mutant virus stocks were produced by transfecting full-length viral plasmids into 293T cells and harvesting virus-containing supernatant. Stock concentrations were determined by p27 antigen capture (Advanced Bioscience Laboratories), and infectivity was determined using the standard TZM-bl assay (160).

3.4.2 Cell Culture

The adherent cell lines 293T and TZM-bl were maintained in Dulbecco's modified Eagle's medium (DMEM) supplemented with 10% fetal bovine serum (FBS), 1% Penicillin/Streptomycin, and 1% L-glutamine. The immortalized rhesus 221 and 444 T-cell lines (161, 162) were maintained in RPMI medium 1640 supplemented with 20% FBS, 1% Penicillin/Streptomycin, 1% L-glutamine, and 0.001% interleukin-2 (IL-2, obtained from NIH AIDS Reagent Program). Rhesus 444 cells were kindly provided by Vanessa Hirsch (NIAID).

3.4.3 Viral load measurement

Levels of viral RNA in tissue culture supernatants were determined by quantitative RT-PCR (with a detection limit of 160 RNA copies/ml) by the Emory Center for AIDS Research (163, 164).

3.4.4 FitSeq assay

Rhesus 221 and 444 T cell lines (161, 162) were dually infected with wild-type and mutant viruses according to conditions indicated in figure legends (Figure 11-Figure 17). Viral RNA was extracted using the either the High Pure Viral RNA kit (Roche) or the ZR-96 Viral

RNA kit (Zymo Research). RT-PCRs of short, targeted amplicons were performed using the OneStep RT-PCR Kit (Qiagen). A table of primers used can be found in Supplemental Table 4. Samples were then indexed with Nextera XT Index Kits (Illumina), normalized to approximately 1 ng/ul using the SequalPrep Normalization Kit (Invitrogen), and pooled prior to a single run on an Illumina MiSeq.

Reads were assembled to a reference/parental genome sequence and variant frequencies were calculated over the course of the infection using Geneious (Biomatters). After assembly to the reference sequence, the frequencies of mutant and wild type were calculated either by highlighting the mutated locus and using the Geneious statistics tool to determine the frequency of each nucleotide at that locus, or by using the Find Variations/SNPs tool. Mutant or wild-type allele frequencies were plotted over time in Microsoft Excel. Three consecutive time points representing exponential growth were chosen for linear regression analysis. Slope values were taken to represent the relative fitness (w) of each allele measured. R^2 values were also calculated and a one-sample T test was used to determine *p*-values for the null hypothesis that the slopes of the regression lines were equal to zero, which would indicate a neutral fitness impact of the allele being measured.

3.5 Results and Discussion

3.5.1 FitSeq overview

In order to test our hypothesis that the mutations we identified in section 2.0 were indeed adaptations acquired by SIVsmE543 and/or E660 to improve replication in rhesus macaque cells, we sought to compare the fitness of these variants in the rhesus T cell lines rh221 and rh444 (161, 162). We wanted to address two questions about our candidate adaptations. First, do they increase the relative fitness of SIVsm in rhesus cells, and second, are they necessary for SIVmac

to remain fit in rhesus cells? To address these questions, we engineered each of the mutations in Figure 8 into the parental SIVsmE543 (forward mutations) and SIVmac239 (reverse mutations).

While many methods exist for evaluating viral fitness (defined here as the ability of a virus to survive and reproduce under the selection pressures present in a given environment), those in which a reference strain and a mutant strain are grown together (in the same well, plate, or flask) offer the benefit of eliminating variation arising from independent measurements, allowing for more sensitive detection and quantification of small differences in relative fitness (165). Many commonly used methods of measuring viral fitness focus on a small region of the genome, necessitating development and optimization of allele-specific, quantitative PCR primers and conditions (166), or replacement of a non-essential viral gene with a reporter (165). However, because of the various locations of our candidate mutations in the SIV genome (some of which are in non-essential genes), and because we intended to study each mutation in the context of two different genetic backgrounds (SIVsm and SIVmac), we required a fitness assay that could be easily applied to compare the relative fitness of mutations in disparate regions of the genome, as well as in multiple parental strains. In addition, we wanted a fitness assay with multiplex capability allowing higher throughput. We therefore developed a deep sequencing-based approach, similar to one used previously to assess the fitness of drug-resistance mutations in HIV-1 RT (167).

In the FitSeq assay, two or more replication-competent viral variants are combined into a single inoculum used to infect cells. Over the course of several days, supernatants are harvested from which viral RNA is then extracted. The RNA is used for targeted RT-PCR, in which a small portion of the viral genome containing the mutation of interest is amplified. The resulting amplicons are then indexed and pooled for sequencing on the Illumina MiSeq platform. The

target amplicons are designed to be approximately the same length as a typical sequencing read (150 base pairs in the case of the MiSeq platform), and so that they can be directly aligned to the wild type (WT) reference sequence. The frequency of the mutation relative to WT is then calculated directly from its representation in sequencing reads using conventional sequence analysis tools such as those found in Geneious, without the need for more sophisticated software. Relative fitness is then calculated as the slope of a regression line determined from plotting the change in the mutant frequency (versus a WT reference strain) over time.

3.5.2 Application and optimization of FitSeq

We chose to first test the mutation in the Capsid protein at position 98. In the SIVsm cohorts, this position changed in multiple animals from an arginine to a serine (CA98), while the SIVmac-infected animals had a serine that did not change throughout the course of the infection (Figure 8 and Supplemental Table 3). This mutation is known to facilitate evasion of SIVsm from restrictive rhesus TRIM5α alleles (85, 86). We hypothesized that a virus containing this mutation (SIVsmE543-CA-R98S) would replicate better than WT in the presence of restrictive rhesus TRIM5α alleles. To test this hypothesis, and to pilot the FitSeq assay, we infected the TRIM5α-TFP/Q heterozygous cell line rh221 with a mixture of SIVsmE543-WT and mutant virus. In parallel, we infected rh221 cultures singly with either WT or CA-R98S. We collected supernatants daily for eight days. Supernatants from the individual infections were measured for p27 (Capsid protein) content and supernatants from dual infections were processed for Illumina sequencing. We conducted both individual and dual infections in order to compare the FitSeq results to standard methods that compare the growth of individual viruses in separate cultures.

We found that the SIVsmE543-CA-R98S mutant did indeed appear to replicate better than WT in both the individual and dual infections (Figure 11). Two out of the three replicates of

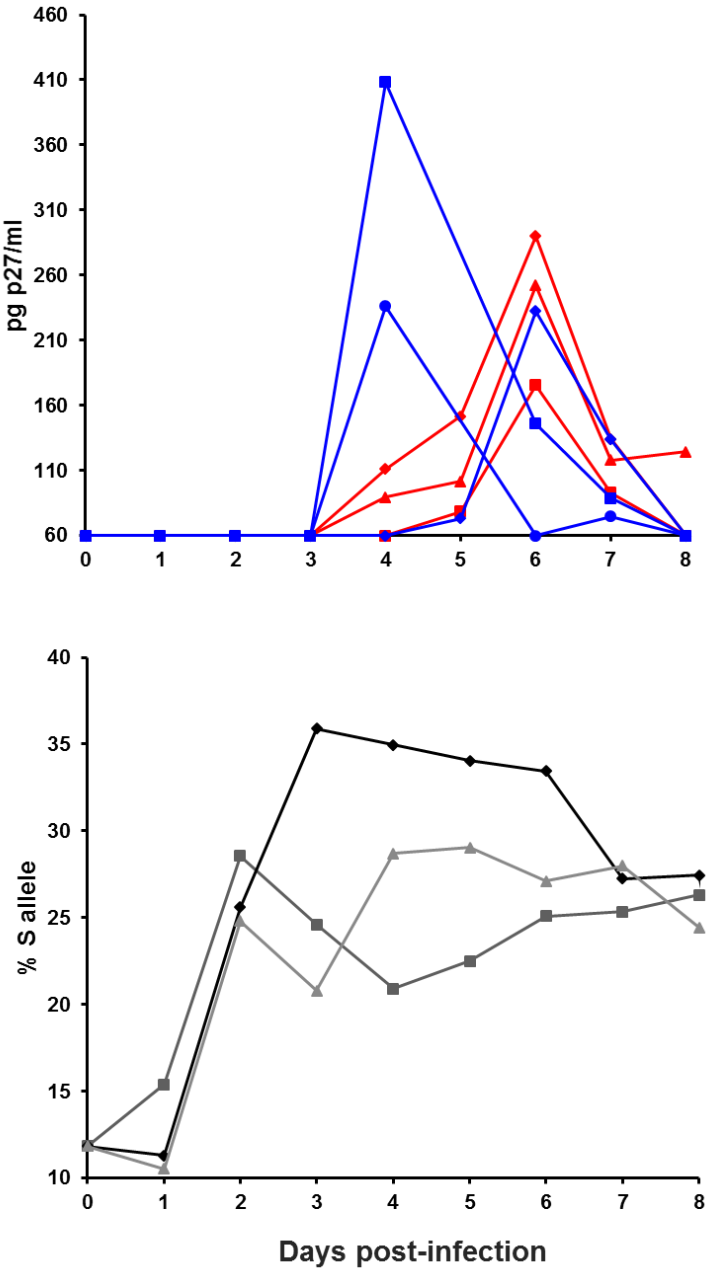


Figure 11. FitSeq pilot

Rh221 cells were infected at a density of 200,000 cells/ml in 12-well plates. Top, three replicates each were infected with SIVsmE543-WT (11.1 ng p27 per replicate, red lines) or –CA-R98S (1.5 ng p27 per replicate, blue lines). Cells were washed to remove input virus after 4 hours. Supernatants were collected daily and p27 content was determined. *(Note: the y-axis starts at the limit of detection of the assay, 62.5 pg p27/ml)*. Bottom, a mixture of both SIVsmE543-WT and –CA-R98S (totaling 25.1 ng p27) was used to infect three replicate wells. Supernatants were processed for Illumina sequencing and mutant frequencies were determined as outlined in section 3.4.4.

SIVsmE543-CA-R98S-infected cells produced detectable p27 levels earlier than the WT-infected cells (Figure 11, top panel). In the dually infected cells, the frequency of the mutant S allele increased in all three replicates, demonstrating its fitness advantage over WT (Figure 11, bottom panel). We noticed that this positive fitness effect seemed to be saturated in the dual infections by three days post-infection while p27 remained undetectable until day four. This suggested to us that, under the conditions tested, FitSeq was the more sensitive of the two methods, as it measured changes in the relative frequencies of the two viruses prior to detection of growth by p27 antigen capture.

In this pilot experiment, we had attempted to use an equal amount of infectious units (IU) of each virus. IU per nanogram of p27 (IU/ng) for each virus was determined using a standard TZM-bl infectivity assay. TZM-bl is a HeLa-derived cell line that constitutively expresses human CD4, CCR5, and CXCR4. These cells also encode a β-galactosidase reporter gene that is expressed in the presence of HIV or SIV Tat (160, 168). IU/ng can then be determined by counting the number of blue cells after seventy-two hours of infection. We found that SIVsmE543-CA-R98S had a higher IU/ng than the WT (data not shown), so we attempted to normalize for this when preparing the virus input. As shown in both the legend and bottom panel of Figure 11, this meant that we used considerably more WT virus in both the dual and individual infections.

In our follow-up experiment, we wanted to determine whether we would see the same fitness benefit of the S allele over the WT R allele if we used a different IU/ng input and a different starting ratio of SIVsmE543-CA-R98S to WT. We also wanted to ensure that thviruses were indeed growing in the dually infected cell cultures during the first three days of infection. We therefore repeated the dual infection experiment with a slightly lower total virus input, at an

approximately 50/50 ratio of the two viruses. We also sent a portion of the collected supernatants to the Emory Center for AIDS Research for measurement of viral load (viral RNA copy number/ml supernatant), which we anticipated would be more sensitive than the p27 antigen capture. We again saw a similar fitness benefit to the S allele in all replicates as in the first experiment (Figure 12, bottom panel). We also found that the viral load analysis did indeed detect viral growth during the first seventy-two hours post-infection (Figure 12, top panel). We concluded that the viruses were indeed replicating during first three days of infection, and that collecting time points within those first three days was sufficient to determine relative fitness going forward.

While the fitness benefit of the S allele relative to SIVsmE543-WT was clear from these two pilot experiments (Figure 11 and Figure 12), it appeared that the effect was saturating at approximately sixty hours post-infection in some replicates. We speculated that the cell cultures might be overgrowing during the course of the experiment, and that this may have caused them to become refractory to infection. We wanted to investigate this idea, and to further optimize the FitSeq assay to reduce the variability between replicates, which we hypothesized was a result of not using enough total virus to initiate the infections. To address these issues, we conducted an experiment in which both the virus input and starting cell density were varied systematically (Figure 13). In this experiment, we included a second rhesus T-cell line, rh444. We wanted to use two cell lines going forward to test our candidate adaptations because we know from previous studies that restriction factors can be polymorphic within rhesus macaques, and that this can have a profound effect on the ability of SIVsm strains to replicate in animals with different genotypes (85, 86, 92). The rh444 cell line has different TRIM5α and A3G genotypes from the

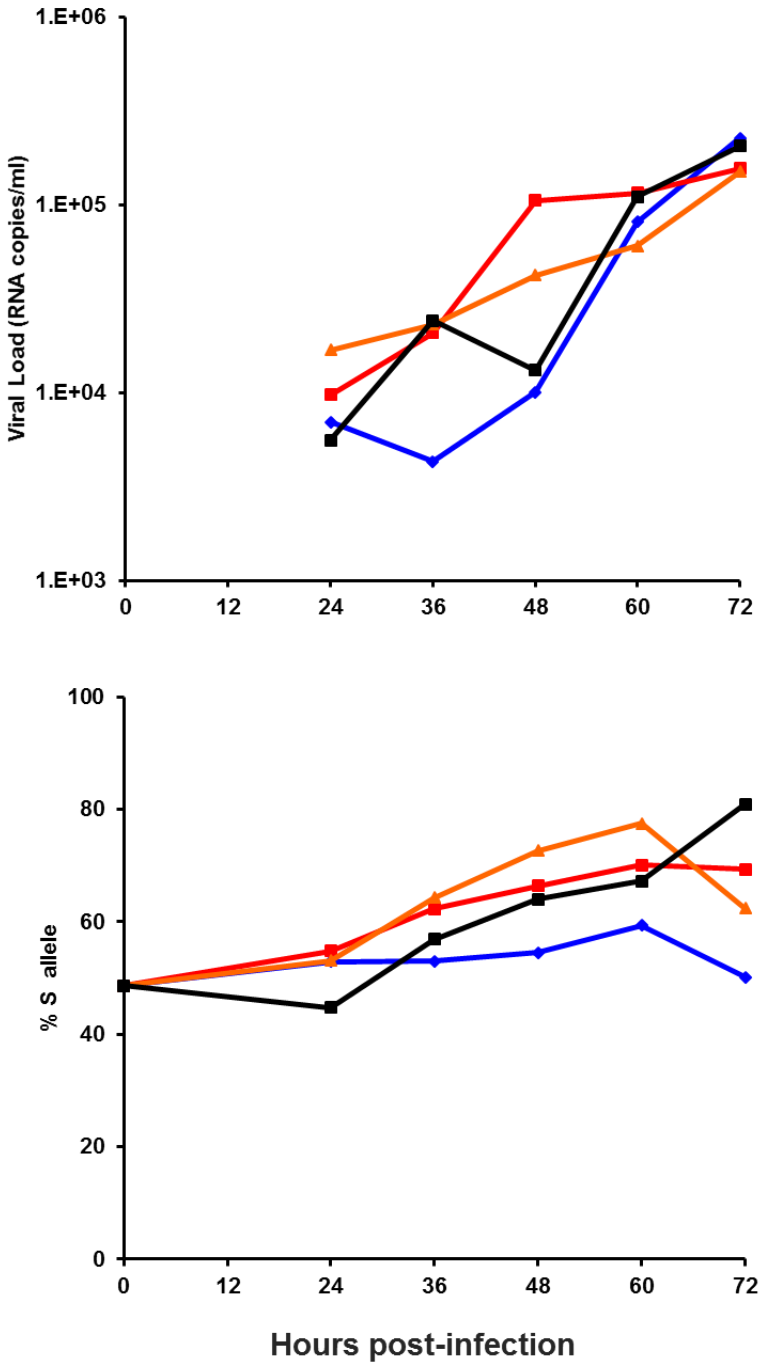


Figure 12. Growth of dual virus populations during the first three days of infection. Rh221 cells were infected at a density of 200,000 cells/ml in 12-well plates. Four replicates each were infected with a mixture of SIVsmE543-WT and –CA-R98S (approximately 1.5 ng of each virus per replicate). Supernatants were collected at 24, 36, 48, 60 and 72 hours post-infection and split into two aliquots. One aliquot was sent for viral load measurement (top) and the other aliquot was processed for Illumina sequencing (bottom). Mutant frequencies were determined as outlined in section 3.4.3. *Note: the viral load of the input (zero hours post-infection) was not measured in this experiment.)*

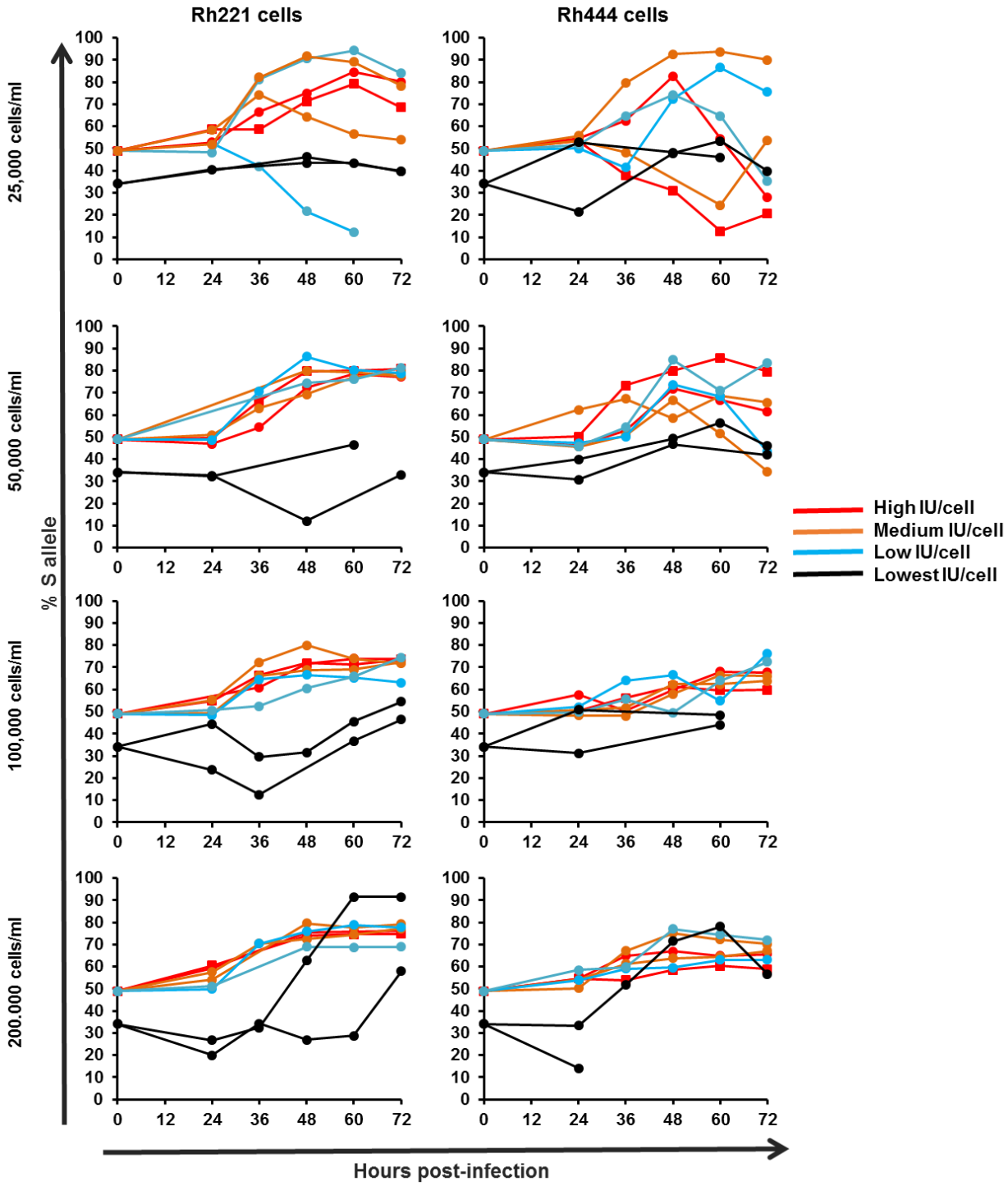


Figure 13. Higher cell density and virus input yield more reproducible results.

Rh221 (left panels) and rh444 cells (right panels) were infected in duplicate at the indicated densities with a mixture of SIVsmE543-WT and –CA-R98S. Input for each infection was normalized to the total number of cells, so that each category (high, medium, low, lowest) could be compared across cell density conditions. The bottom two panels use the same conditions as the previous experiment (Figure 12).

rh221 cell line (data not shown), which is why we chose it for this and subsequent experiments. We found that, in general, higher IU/cell and higher cell density yielded the most reproducible results (Figure 13).

We realized after this experiment that we had likely been using far too low a cell density for T-cell lines such as rh221 and rh444. Looking at the literature, primary peripheral blood mononuclear cell (PBMC) cultures are typically kept at much higher densities (86, 166). We reasoned that further increasing the cell density could lead to even better reproducibility between replicates. We therefore conducted an experiment where we used a high IU/cell input, and further increased the cell density in both the rh221 and rh444 cells. We again found that the mutant S allele increased in proportion relative to the wild type in both cell types, and that the variability between replicates was substantially reduced compared to previous experiments (Figure 14, top panels). We also tested the impact of the reverse mutation in the SIVmac239 genetic background using the same conditions (Figure 14, bottom panels), which revealed an apparently neutral relative fitness between SIVmac239-WT and SIVmac239-CA-S97R.

Now that we had optimized the infection culture conditions, we were concerned about potential variation introduced into the assay by technical differences between individual sequencing runs and index primer pairs. To measure this variation, we chose two samples and divided each into three identical aliquots (Table 3, samples A and B). For each sample, two aliquots were indexed with the same indexing primer set and one aliquot was indexed with a distinct primer set. The aliquots indexed with different primers were sequenced together in the same run to compare results for identical samples with different indices. In addition, the identically indexed aliquots were sequenced in separate runs in order to compare results for

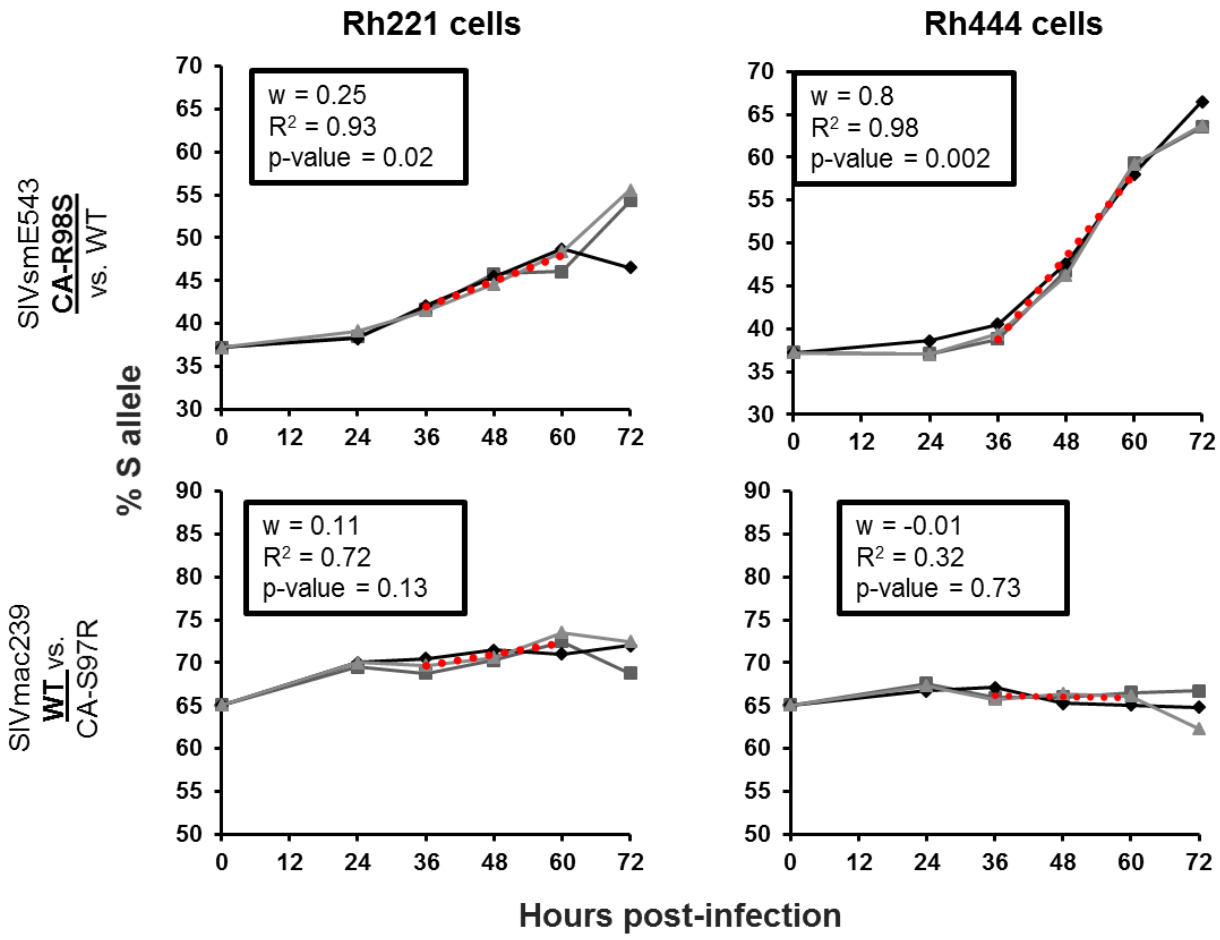


Figure 14. R to S adaptation at CA position 98 improves SIVsm fitness in rhesus cells. rh221 cells (left panels) or rh444 cells (right panels)—at a density of 10^6 cells/ml in 96-well plates—were infected with a mixture of SIVsmE543-WT and –CA-R98S (top panels) or SIVmac239-WT and –CA-S97R (bottom panels), totaling 40 ng p27 input per infection. Supernatant was collected at 24, 36, 48, 60, and 72 hours post infection. Viral RNA was extracted and used as template for targeted RT-PCRs amplifying small regions surrounding CA position 98. The percentages of reads mapping to the S variant were calculated using Geneious. The frequency of the S allele (y-axis) is plotted over time (x-axis), with values for zero hours post-infection indicating the frequency of the input stock. The red dotted line is a trendline representing the average of the three replicates over three consecutive time points representing the exponential growth phase of infection. Linear regression analysis was performed separately on each of the three replicates per condition. Insets: average slope (relative fitness, w), R-squared, and p-values.

Table 3. Variation between sequencing runs and indices is negligible.

Samples A and B, SIVsmE543-WT and –CA-R98S dual infection samples were split into three aliquots and indexed with the indicated primers, and sequenced in one of two separate Illumina sequencing runs. Samples C-H, data were generated by Sergio Ita and reproduced here with permission: SIVmac239-WT and –Env-V67M dual infection samples were split into two aliquots, indexed with the indicated primers, and sequenced in independent sequencing runs.

Sample	Sequencing Run	Index primer pair	% Mutant
A	1	S504, N704	24.8
	2	S505, N722	21.8
	2	S504, N704	22.3
B	1	S504, N707	28.7
	2	S506, N722	30
	2	S504, N707	30.4
C	1	S504, N701	74
	2	S516, N718	74.2
D	1	S505, N702	60.2
	2	S521, N718	58.7
E	1	S507, N701	70.1
	2	S517, N718	69.8
F	1	S517, N701	67.6
	2	S518, N718	68.6
G	1	S503, N702	61.3
	2	S520, N718	59.7
H	1	S507, N702	53.5
	2	S522, N718	51.7

identical samples in different sequencing runs. We also took six additional samples amplifying a different target region and divided them each into two aliquots (Table 3, samples C-H). Each aliquot was indexed with separate primers and then sequenced in separate runs, to measure the effect of the combination of different indices and sequencing runs. We found that the average difference between aliquots was 1.3%. We concluded from these experiments that variation due to different sequencing runs and indices was negligible.

We were surprised to find that SIVsmE543-CA-R98S had a greater fitness benefit in the rh444 cells compared to the rh221 cells. Although SIVsm is restricted by both the Q and TFP alleles of rhTRIM5α, TFP is the more restrictive of the two, and so we hypothesized that there would be a greater benefit of having the escape mutation at position 98 in TFP-expressing cells (rh221). However, this result highlights the sensitivity of the FitSeq assay compared to traditional restriction assays. One could conclude from restriction assay data that the Q allele is "permissive" for SIVsm infection, based on the small differences seen in previous studies (85, 86), but the FitSeq results suggest that the presence of an R at position 98 makes SIVsm susceptible to restriction by the Q allele.

Previous work has also demonstrated that the R to S change is not entirely sufficient to provide SIVsm with resistance to the TFP allele of rhTRIM5α in restriction assays (86). A particular compensatory mutation has been identified that confers TFP-resistance to SIVsm, at position 37 in CA, where the ancestral proline changed to a serine (86). This change was not one of our candidate adaptations because it was only found in one animal, and because SIVsm did not change to match the residue typically found in SIVmac sequences, which is also a proline. This finding suggests that the serine is a suboptimal TFP-resistance variant, and that SIVsm may have sampled other compensatory mutations during its emergence in rhesus macaques. These

results from a known, well-documented adaptation (CA-R98S), validate the utility of the FitSeq assay for analyzing the impact of candidate adaptations on relative fitness.

3.5.3 Fitness effects of candidate adaptations in Matrix, Vif, and Integrase

We next tested the forward and reverse mutations at MA128 (Figure 15), IN256 (Figure 16), and Vif74 (Figure 17) in the FitSeq assay. The calculated relative fitness values (w) are summarized in Figure 18. The mutation at MA128 from a P to an S provided a significant fitness advantage to SIVsmE543 over wild-type in both rh221 and rh444 cells (Figure 15). As we saw in the CA98 mutation, the advantage was more substantial in the rh444 cells (w = 0.54) than the rh221 cells (w = 0.14). When we reverted the S to the ancestral P in SIVmac, however, we unexpectedly found that the wild-type S residue was disadvantageous in both cell types (rh221, w = -0.17; rh444, w = -0.10, Figure 15). MA128 is within the Gag 125-150 window that we identified as a dN hotspot (section 2.5.4.2.2, Figure 10), and there are other loci within that window that showed evidence of adaptation in the SIVsm cohorts. We speculate that a number of variants could have been sampled during adaptation of SIVsm to rhesus macaques, and that compensatory mutations elsewhere at the C-terminal end of MA may have rendered the S adaptation unnecessary and even detrimental to SIVmac.

We found that, in the context of both SIVsmE543 and SIVmac239, both the E and the D residues at IN256 were approximately neutral (Figure 16). We noted that, in the rh444 cells, SIVmac239-D256 had a statistically significant p-value (w = -0.05), allowing us to reject the null hypothesis that the D allele was neutral in this context (Figure 16 and Figure 18). However, we noted that the R2 value was low for this condition, indicating a lack of correlation between the time point and the frequency of the D allele. We therefore speculate that the D allele may be neutral compared to E in SIVmac239 replicating in rh444 cells. The overall neutrality of the D

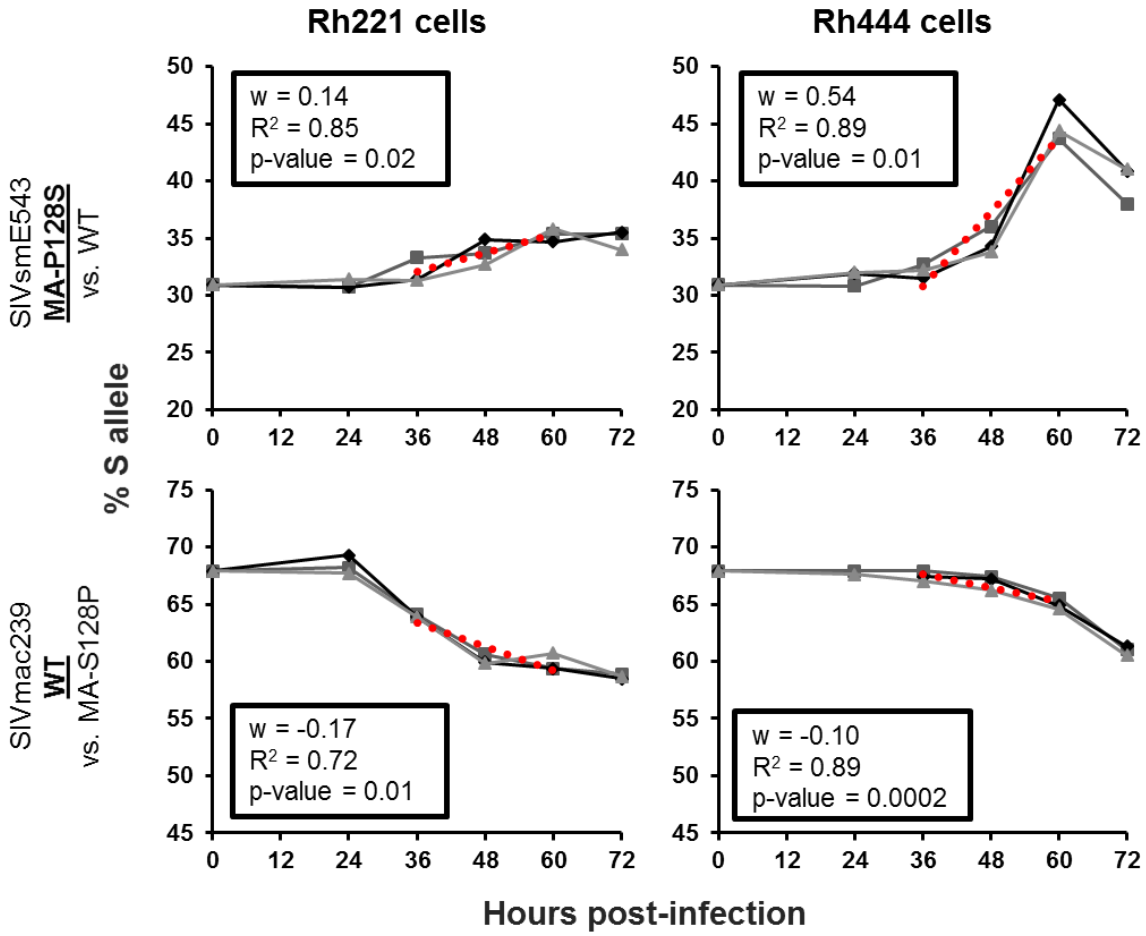


Figure 15. S at position 128 enhances fitness of SIVsm but not SIVmac in rhesus cells. rh221 (left panels) or rh444 cells (right panels)—at a density of 10^6 cells/ml in 96-well plates—were infected with a mixture of SIVsmE543-WT and –MA-P128S (top panels) or SIVmac239-WT and –MA-S128P (bottom panels), totaling 80 ng p27 virus input per infection. Supernatant was collected at 24, 36, 48, 60, and 72 hours post infection. The frequency of the S allele (y-axis) is plotted over time (x-axis), with values for zero hours post-infection indicating the frequency of the input stock. The red dotted line is a trendline representing the average S frequency of the three replicates over three consecutive time points representing the exponential growth phase of infection. Linear regression analysis was performed separately on each of 3 replicates per condition. Insets: average slope (relative fitness, w), R-squared, and *p*-values.

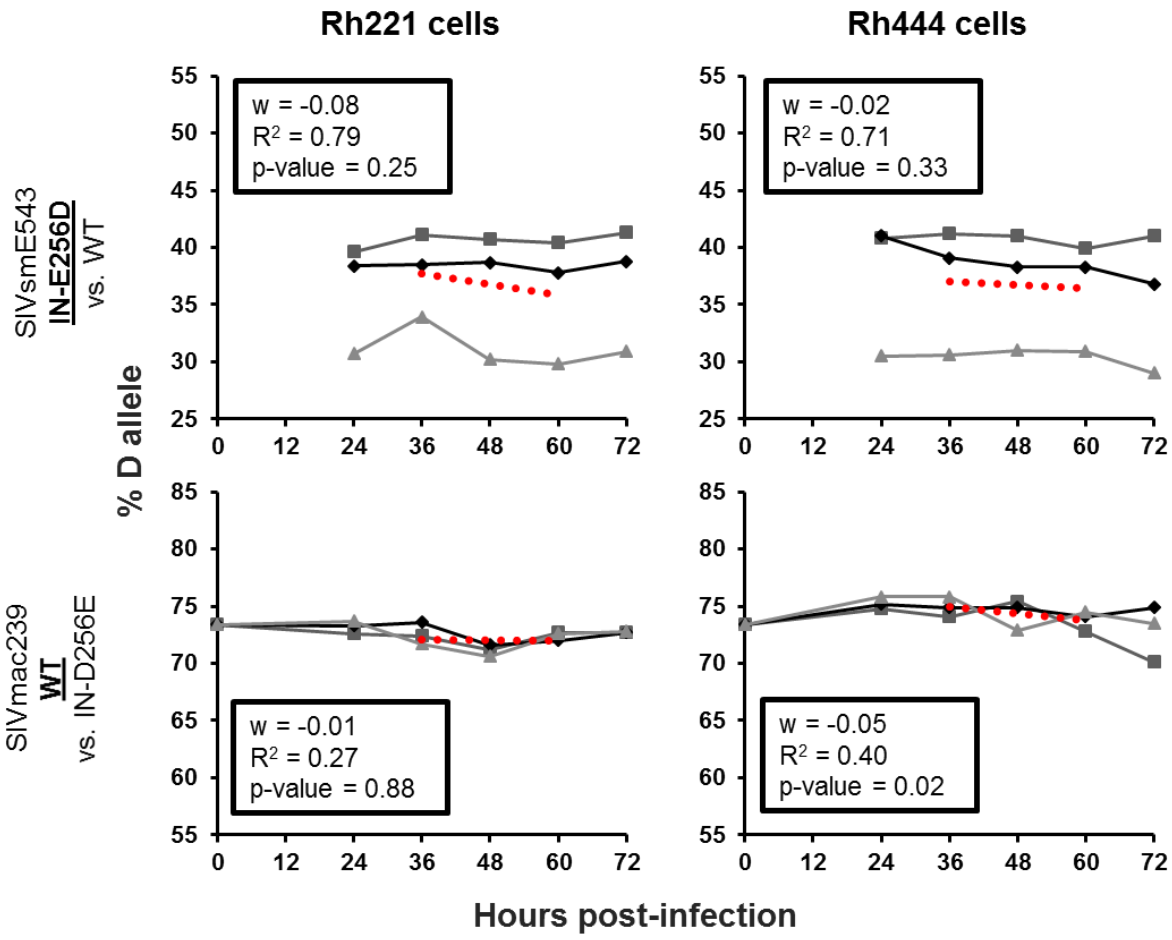


Figure 16. E and D have neutral impacts on SIVsm and SIVmac fitness at IN256. rh221 (left panels) or rh444 cells (right panels)—at a density of 10^6 cells/ml in 96-well plates—were infected with a mixture of SIVsmE543-WT and –IN-E256D (top panels) or SIVmac239-WT and –IN-D256E (bottom panels) totaling 32 and 40 ng p27 virus input per infection, respectively. Supernatant was collected at 24, 36, 48, 60, and 72 hours post infection. The frequency of the D allele (y-axis) is plotted over time (x-axis), with values for zero hours post-infection indicating the frequency of the input stock (not shown in top panels because stock samples were contaminated during processing). Top panels, one replicate per cell line was infected with a different ratio of WT to mutant virus. The red dotted line is a trendline representing the average D frequency of the three replicates over three consecutive time points representing the exponential growth phase of infection. Linear regression analysis was performed separately on each of 3 replicates per condition. Insets: average slope (relative fitness, w), R-squared, and *p*-values.

allele is not entirely surprising because both residues are negatively charged, and differ only in that D has a shorter side chain than E, making the change from E to D in the SIVsm cohorts relatively conservative. However, IN256 is located close to a dN hotspot window (Pol 909-959) identified in section 2.5.4.2.2, indicating that other potential adaptations did arise in the SIVsm cohorts in the C-terminal domain of Integrase. It could be the case that the E to D change in SIVsm is a compensatory mutation for another adaptation that is neutral on its own.

A replacement of asparagine with histidine at position 74 of Vif was found to be beneficial to SIVsmE543 in both rh221 and rh444 cell lines (Figure 17, top panels). It was also beneficial to SIVmac239 in rh221 cells, but neutral (with a non-significant downward trend) in rh444 cells (Figure 17 and Figure 18). In section 2.0, we speculated that an adaptation at this position may facilitate interaction between the emerging SIVsm Vif protein and the rhesus host APOBEC3 proteins, specifically A3G and/or A3F. Certain alleles of rhesus A3G are already known to have selected for a substitution at Vif position 17 during SIVsm emergence in macaques (92). However, it's likely this mutation (Vif-E17G) would have occurred very early after cross-species transmission, i.e., in the first or second infected macaque. Both SIVsmE543 and SIVmacE660, which were passaged through two and three macaques prior to isolation, respectively, have the macaque-adapted G residue at Vif17. In restriction assays where both A3G and Vif are overexpressed, this adaptation appears to be sufficient for SIVsm escape from restriction by rhesus A3G (92), but such assays are unlikely to detect a subtle restriction effect. We speculate that either rhesus A3G or other APOBEC3-family restriction factors could account for the fitness benefit of the H allele at Vif74.

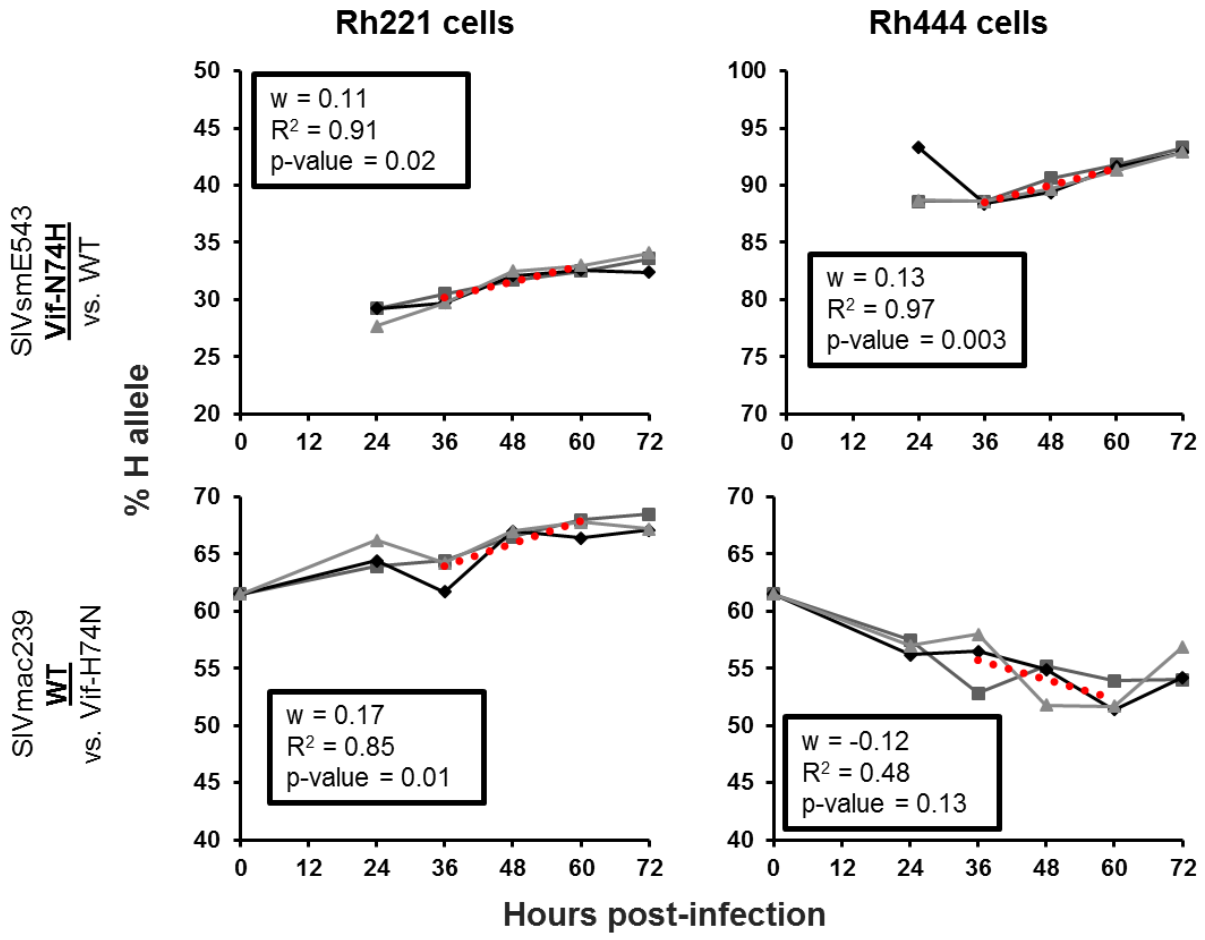


Figure 17. H at Vif74 is beneficial to SIVsm in rhesus cells, and to SIVmac in rh221 cells. rh221 (left panels) or rh444 cells (right panels)—at a density of 10^6 cells/ml in 96-well plates—were infected with a mixture of SIVsmE543-WT and –Vif-N74H (top panels) or SIVmac239-WT and –Vif-H74N (bottom panels) totaling 80 ng p27 virus input per infection. Supernatant was collected at 24, 36, 48, 60, and 72 hours post infection. The frequency of the H allele (y-axis) is plotted over time (x-axis), with values for zero hours post-infection indicating the frequency of the input stock (not shown in top panels because stock sample was contaminated during processing). The red dotted line is a trendline representing the average H frequency of the three replicates over three consecutive time points representing the exponential growth phase of infection. Linear regression analysis was performed separately on each of 3 replicates per condition. Insets: average slope (relative fitness, w), R-squared, and *p*-values.

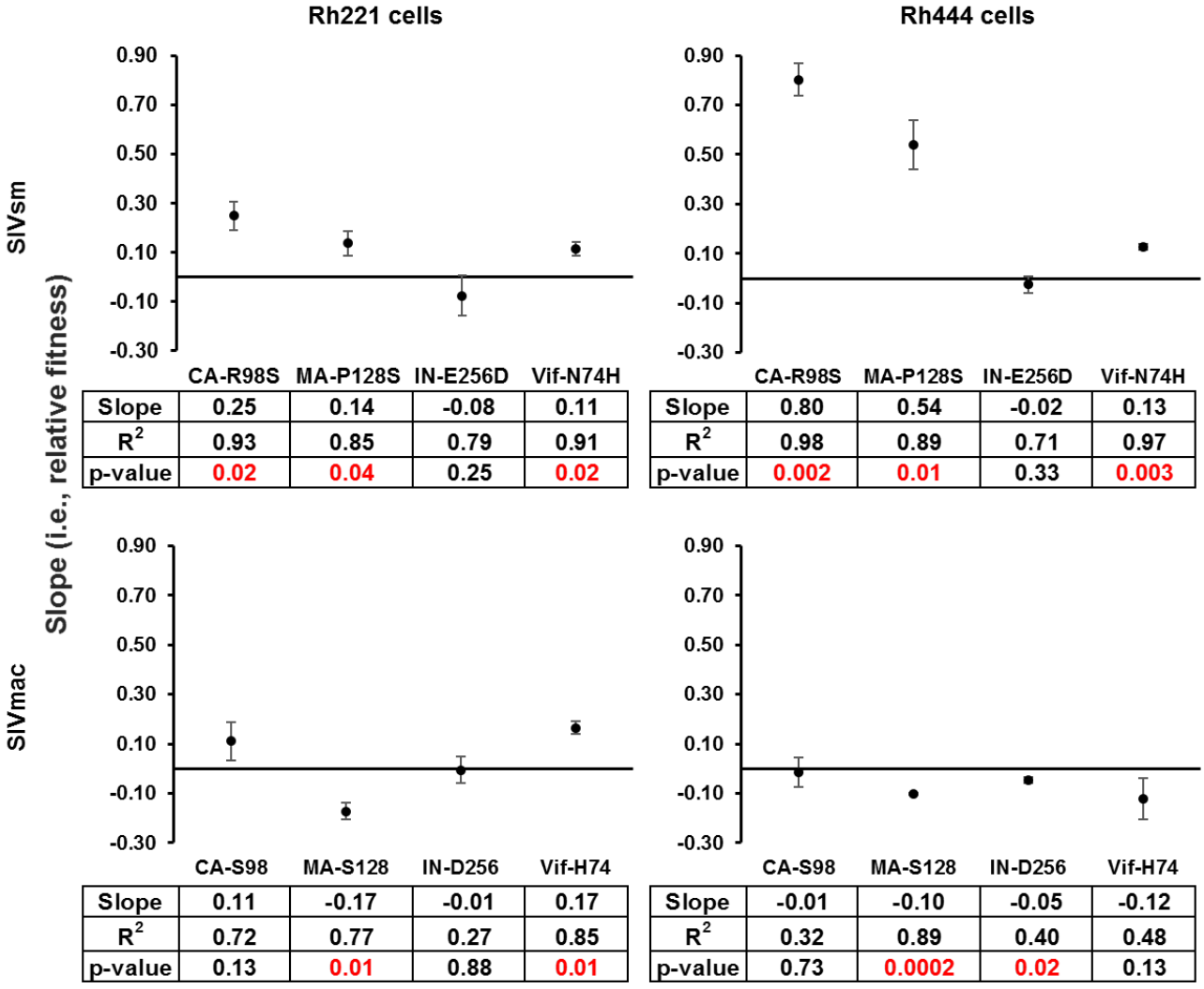


Figure 18. Fitness Summary.

The average slope for each condition was taken to represent the fitness of (top panels) the mutant relative to the WT or (bottom panels) the WT relative to the mutant in rh221 (left panels) or rh444 cells (right panels). Error bars represent standard deviations from the mean slope. A one-sample T-test was used to test the null hypothesis that the slope was equal to zero, indicating no impact of the mutant on relative fitness. *p*-values less than 0.05 are highlighted in red.

4.0 Summary and Discussion

Cross-species transmission and emergence of new viruses in human populations is a significant problem in virology and public health (4). While the ecological and behavioral factors leading to the exposure and infection of humans with zoonotic viruses have been studied extensively, less attention has been paid to the molecular evolution of virus populations immediately following cross-species transmission. The primate lentiviruses have undergone numerous cross-species transmissions among primate species in Africa (6). These include at least four independent transmissions to humans from great apes leading to the emergence of the four groups of HIV-1, and at least nine transmissions from the sooty mangabey monkey leading to the emergence of the nine subtypes of HIV-2 (3, 7, 23, 32). As a model for understanding the early adaptive events in PLV emergence, we have investigated the evolution of SIVsm in rhesus macaques shortly after cross-species transmission. SIVsm was accidentally transferred to captive macaques at primate centers in the United States, which led to outbreaks of simian AIDS later revealed to be caused by the emergent virus SIVmac (40, 41). Since SIVmac's discovery, a number of *in vivo* studies have been conducted in which SIVsm strains were used to infect cohorts of rhesus macaques (86, 99, 101, 109, 169). We considered samples from such cohorts to be ideally suited for the study of viral evolution after cross-species transmission, due to the controlled nature of the infections and the accessibility of key samples including the stock/source inoculum, as well as longitudinal samples after infection.

We sought to compare evolution of emerging virus populations (SIVsm) to established virus populations (SIVmac, which is adapted to the rhesus macaque host). We obtained plasma samples from 15 rhesus macaques (eight SIVsm-infected and seven SIVmac-infected) at acute (2-6 weeks p.i.) and chronic (40-72 weeks p.i.) time points, as well as the stock virus used to inoculate them (Table 1). We then processed those samples in order to sequence the entire

coding regions of each sample using next generation sequencing technology. With the resulting data, we identified one animal in which substantial hypermutation occurred during acute infection and demonstrated that those mutations likely introduced by APOBEC3 family proteins were largely removed by the chronic time point. This was interesting to us, as several studies have suggested that APOBEC3-mediated hypermutation introduces beneficial as well as detrimental mutations (170). Our data agree more with a model in which APOBEC3 hypermutation acts primarily as an antiviral force. However, our sample size is too small to draw any general conclusions.

We also identified several candidate adaptations as well as hotspots of potential adaptation using two approaches: manual identification of convergent evolution based on biological criteria, and novel computational methods. For the first approach, we applied strict criteria to select candidate adaptations that would be worthy of follow-up biological assays. We looked for sites where the emerging virus sequences changed to match what was found in the established virus sequences, reasoning that these were likely to be examples of convergent evolution, where SIVmac strains and SIVsm strains ended up with the same residue at a particular locus after passaging in rhesus macaques. Each of our candidate adaptations also occurred in at least two animals in the emerging cohorts. We reasoned that we would be less likely to select animal-specific adaptations due to individual immune responses if we included this criterion. We also knew the MHC class I genotypes for the majority of our animals (Table 1), so we could rule out many putative CTL-induced substitutions. Finally, we looked in alignments of naturally occurring SIVsm sequences and of a variety of published SIVmac sequences (pre-assembled HIV-2/SIVsm/SIVmac alignments from http://www.lanl.gov) to check that the adaptations we identified reflected a departure from typical SIVsm sequences. For

example, there were several candidates that met the first two criteria, but when we looked at the LANL alignments we found that natural SIVsm strains already shared the same residue with SIVmac strains, indicating that SIVsmE543/E660 simply had a suboptimal residue at that particular locus, rather than adapting specifically to rhesus macaques.

We next used computational methods to identify sites under selection that we might have missed using the first method. We found it challenging to select an appropriate method, because the majority of those available tend to be designed for a very different kind of data than we had (see section 2.5.4.2). Traditionally, dN/dS analysis is used to detect sites under selection from an alignment of diverse sequences, such as from virus strains from different geographic locations or even viruses from different species. The comparatively low diversity of our dataset made it likely that this type of analysis would yield almost exclusively false negatives. Because we were interested in finding relatively rare cross-species adaptations, we needed a method that favored false positives over false negatives. In particular, because we planned to test candidate adaptations in biological assays to confirm their relevance, we were less concerned with identifying false positives. The development of the novel algorithms described in sections 2.4.2 and 2.5.4.2 provided us (and the field) with new tools to identify adaptive loci using temporally and evolutionarily similar sequences, such as those sampled from the first 1-2 years of infection.

The candidate adaptations we identified by both methods were located in structural (MA, CA, IN, and Env), auxiliary (Tat) as well as accessory proteins (Vif and Nef). When we designed the study, we had hypothesized that we might identify a greater proportion of adaptations in accessory genes rather than structural ones. We reasoned that changes in accessory genes might be more tolerable for the virus (e.g., having a lesser fitness cost) compared to changes in structural genes. However, a majority of the changes we identified were in structural proteins.

We speculate that this finding could reflect a priority of adapting "essential" rather than "non-essential" proteins early in emergence. Another possibility is that the most tolerable changes for the virus to make are made as rapidly as possible upon replication in a new host species. Our set of samples would therefore have already acquired such adaptations due to prior, limited passage in one or two macaques.

Our study has several characteristics that set it apart from previous work. First, with our unique set of samples, we were able to investigate very early molecular events of emergence, which has not been possible for the other PLV cross-species transmissions. Key adaptations have been identified for the emergence of HIV-1 and HIV-2 by comparing reference or consensus HIV sequences to currently circulating strains of their ancestral SIVs (6, 106). Unfortunately, both the HIV and SIV populations have evolved for decades beyond the common ancestor that first infected humans, making it impossible to study the earliest events of emergence for these cross-species transmissions. Second, we are also looking at early events of intra-host *in vivo* infections. These two characteristics put us in a unique position to observe potential sampling of different adaptations, in order to gain insight into the molecular constraints of emergence. Third, we sequenced the entire coding region of the virus populations in each sample, which meant that we could look for adaptations in areas of the genome not currently known to be involved in emergence and contribute new information to the field. Finally, we had access to the stocks that were used to infect the rhesus macaques, which allowed us to trace the fates of low-frequency variants in the stock.

One open question in the field is whether adaptations to new species are present as rare variants in the source inoculum and then selected as the majority variant after transmission; or whether emerging viruses must adapt rapidly to the genetic environment of the new host (11).

Using next-generation sequencing technology (and Illumina technology in particular) provided a higher depth of coverage than conventional methods of variant identification, such as cloning and single-genome amplification (115, 171, 172). We identified a panel of putative adaptations that emerged in SIVsm-infected macaques, and did not find evidence that they were present in the stock. However, we cannot determine conclusively whether a rare variant was present below the limit of detection of our assay.

One caveat of using Illumina sequencing technology is the lack of linkage information (171). Different next generation sequencing platforms face a tradeoff between depth of coverage and read length. By choosing the Illumina platform for our study, we privileged depth of coverage and thus had short read lengths of approximately 150 base pairs in length. The short read lengths meant that we could only determine linkage of adaptations that were very close together in the genome. While computational methods exist for determining linkage from short read data, there did not seem to be a consensus on an approach for doing this reliably. We therefore deemed haplotyping beyond the scope of this project, but are interested in investigating this in the future. As an alternative to computational methods of determining linkage, future studies would use PacBio (or similar) technology, which can cover an entire viral genome in a single sequencing read (173).

Expanding both the number of animals and the number of longitudinal time points would also be a worthwhile follow-up to our study. Obtaining whole-genome sequencing data from more animals would enhance the statistical power of our computational analyses, and help to identify more putative rhesus-specific adaptations. The criteria used to select the six substitutions discussed in section 2.5.4.1 were strict and likely caused us to rule out adaptations because they appeared only in one animal. Data from more animals would allow us to re-examine candidates

that were ruled out due to the scope of the study. Expanding the number of longitudinal time points would enable an investigation of the *in vivo* kinetics of the emergence of adaptations.

As a follow-up to our analysis of emerging and established viral populations, we sought to test the candidate adaptations we identified in a fitness assay. We reasoned that if our candidates were truly adaptations to the rhesus macaque host, they would improve the fitness of wild-type SIVsm in rhesus T cells. We therefore engineered each candidate adaptation into wild-type SIVsmE543. We also engineered the reverse mutations into SIVmac239. We chose to do both forward and reverse mutations because of the possibility that the adaptations could require additional compensatory substitutions in order to confer a fitness benefit to the virus. This possibility could result in a false negative conclusion that a candidate was not adaptive. Testing the reverse mutation in SIVmac239, which presumably has acquired all necessary compensatory substitutions, would help us to determine if this was indeed happening.

We felt that it was important to evaluate the fitness of our candidates because fitness is not typically quantified in studies of cross-species adaptation. Rather, for PLVs in particular, restriction assays are the norm, where host restriction factors are overexpressed in permissive cell types and the susceptibility of different viruses is measured and compared (85, 92, 94, 105, 159). These assays do not assess the impact of restrictions on the full replication cycle, as they typically use replication-defective viruses or only look at a portion of the replicative cycle. This precludes the detection of fitness tradeoffs that may exist, where an adaptation benefits the virus during one portion of the life cycle (such as a substitution in CA enhancing infectivity), but is detrimental in another portion of the life cycle (such as assembly). Since we were interested in examining the full effects of adaptations on viral fitness, we required an assay that would use replication-competent viruses, and take into account the full viral replication cycle. Furthermore,

restriction assays are unlikely to detect a subtle impact of a restriction factor on fitness because it is overexpressed. We therefore chose to conduct our assays with native host factors. We developed the FitSeq assay to address all of these requirements, as well as a desire to create a protocol that could be flexible and also scalable to enable testing of viral variants in a high-throughput manner. Illumina sequencing is ideal for such experiments because of the ability to multiplex hundreds of samples in a single sequencing run, which cuts down on both labor and financial costs.

We tested four of our candidate adaptations in the FitSeq assay (section 3.5), and found that the results we obtained were highly reproducible (Table 3). We found that viruses containing adaptations at MA128, CA98, and Vif74 were all significantly more fit than wild-type SIVsmE543 in the immortalized rhesus T cell lines rh221 and rh444. We also found that an IN256 mutant virus was equally as fit as wild-type, with both mutant and wild-type maintaining their original frequency throughout the assay. We had hypothesized that, in the context of SIVmac239, the wild type would always be more fit than the mutant virus, given that SIVmac is adapted to replicate in rhesus cells specifically. However, we found that this was not the case in the several of our FitSeq experiments. We found that wild-type SIVmac239 was equally as fit as SIVmac239 variants containing reverse mutations at CA98 in both rh221 and rh444 cells, as well as IN256 in rh221 cells and Vif74 in rh444 cells. We also found that SIVmac239 was actually less fit than variants containing reversions at MA128 in both cell types, and potentially IN256 in rh444 cells, although we speculate due to the low R^2 value that this result reflected a neutral fitness impact. We hypothesize that these results may reflect rapid fixation of the original adaptations, followed by acquisition of additional, more optimal neighboring substitutions. Thus,

returning to the ancestral residue at the original adapted site is actually beneficial now that the emergent virus has become more adapted to its host.

It would be ideal to test all of the candidate adaptations (both those identified by manual and computational methods) in the FitSeq assay. Doing so would allow us to evaluate the efficacy of the two adaptation-identification methods we used. We would also like to use the assay in the context of primary rhesus macaque CD4+T cells, to increase the physiological relevance of our findings. While the rh221 and rh444 cells are CD4+CCR5+ T cells, they have been artificially immortalized and thus have the usual caveats of working with cell lines over primary cells (161, 162). Another reason to turn to primary T cells is that we are interested in investigating whether the fitness of our mutants is altered by the IFN-induced antiviral state. As discussed in section 1.3.3, antiviral proteins of various functions are upregulated as part of the IFN response (98). We hypothesize that the fitness impacts of our adaptations might increase in magnitude in the presence of higher levels of antiviral host factors, since wild-type SIVsm is likely to be even more restricted in the presence of the IFN-induced antiviral state than in non-IFN-stimulated cells. We have conducted a small pilot experiment using rhesus IFNα treatment of rh221 and rh444 cells, and found that the cells may not be IFN-responsive (data not shown).

Finally, it would be interesting to conduct the FitSeq assays in sooty mangabey CD4+ T cells. Such experiments would enable us to investigate any fitness tradeoffs of the candidate adaptations more thoroughly. In particular, it would be interesting to determine whether they incur a fitness cost to SIVsm in the absence of the selective benefit they may confer in rhesus cells. Presumably, the rhesus-specific selection pressures are not present in sooty mangabey cells, making them an appropriate cell type to test the effects of the candidate adaptations on purely replicative fitness, in the absence of the selection pressure that drove their emergence.

Overall, the work presented in this dissertation supports a model of emergence in which virus populations must adapt to the new host species, and amino acid substitutions that enhance viral fitness in the new host environment are selected. We have developed an experimental framework with which to further investigate the relative strength of the selective forces that drive this adaptation, and to identify new adaptations that will enhance our understanding of cross-species transmission and emergence.

5.0 References

1. Lu G, Wang Q, Gao GF. Bat-to-human: spike features determining 'host jump' of coronaviruses SARS-CoV, MERS-CoV, and beyond. Trends Microbiol. 2015.

2. Baize S, Pannetier D, Oestereich L, Rieger T, Koivogui L, Magassouba N, et al. Emergence of Zaire Ebola virus disease in Guinea. N Engl J Med. 2014;371(15):1418-25.

3. Hahn BH, Shaw GM, De Cock KM, Sharp PM. AIDS as a Zoonosis: Scientific and Public Health Implications. Science. 2000;287:607-14.

4. Cutler SJ, Fooks AR, van der Poel WH. Public health threat of new, reemerging, and neglected zoonoses in the industrialized world. Emerg Infect Dis. 2010;16(1):1-7.

5. Parrish CR, Holmes EC, Morens DM, Park EC, Burke DS, Calisher CH, et al. Cross-species virus transmission and the emergence of new epidemic diseases. Microbiol Mol Biol Rev. 2008;72(3):457-70.

6. Sharp PM, Hahn BH. Origins of HIV and the AIDS Pandemic. Cold Spring Harb Perspect Med. 2011;1(1):a006841.

7. Smith SM, Christian D, de Lame V, Shah U, Austin L, Gautam R, et al. Isolation of a new HIV-2 group in the US. Retrovirology. 2008;5:103.

8. Gao F, Yue L, Robertson DL, Hill SC, Hui H, Biggar RJ, et al. Genetic diversity of human immunodeficiency virus type 2: evidence for distinct sequence subtypes with differences in virus biology. J Virol. 1994;68(11):7433-47.

9. Chen Z, Luckay A, Sodora DL, Telfer P, Reed P, Gettie A, et al. Human immunodeficiency virus type 2 (HIV-2) seroprevalence and characterization of a distinct HIV-2 genetic subtype from the natural range of simian immunodeficiency virus-infected sooty mangabeys. J Virol. 1997;71(5):3953-60.

10. Wolfe ND, Dunavan CP, Diamond J. Origins of major human infectious diseases. Nature. 2007;447(7142):279-83.

11. Holmes EC, Drummond AJ. The Evolutionary Genetics of Viral Emergence. CTMI. 2007;315:51-66.

12. Wang LF, Shi Z, Zhang S, Field H, Daszak P, Eaton BT. Review of bats and SARS. Emerg Infect Dis. 2006;12(12):1834-40.

13. Stein RA. Lessons from outbreaks of H1N1 influenza. Ann Intern Med. 2009;151(1):59-62.

14. Johnson WE. Rapid adversarial co-evolution of viruses and cellular restriction factors. Curr Top Microbiol Immunol. 2013;371:123-51.

15. May RM, Gupta S, McLean AR. Infectious disease dynamics: What characterizes a successful invader? Philos Trans R Soc Lond B Biol Sci. 2001;356(1410):901-10.

16. Antia R, Regoes RR, Koella JC, Bergstrom CT. The role of evolution in the emergence of infectious diseases. Nature. 2003;426(6967):658-61.

17. Holmes EC. The molecular epidemiology, phylogeography, and emergence of RNA viruses. The Evolution and Emergence of RNA Viruses. Oxford Series in Ecology and Evolution: Oxford University Press; 2009. p. 131-55.

18. Update 95 - SARS: Chronology of a serial killer: World Health Organization; 2003 [updated July 2003; cited 2015]. Available from: http://www.who.int/csr/don/2003_07_04/en/.

19. CDC SARS Response Timeline: Centers for Disease Control; [updated April 26, 2013; cited 2015]. Available from: http://www.cdc.gov/about/history/sars/timeline.htm.

20. Pepin J. The Origins of AIDS: Cambridge University Press; 2011.

21. Gao F, Bailes E, Robertson DL, Chen Y, Rodenburg CM, Michael SF, et al. Origin of HIV-1 in the chimpanzee Pan troglodytes troglodytes. Nature. 1999;397(6718):436-41.

22. Bailes E, Gao F, Bibollet-Ruche F, Courgnaud V, Peeters M, Marx PA, et al. Hybrid origin of SIV in chimpanzees. Science. 2003;300(5626):1713.

23. D'arc M, Ayouba A, Esteban A, Learn GH, Boué V, Liegeois F, et al. Origin of the HIV-1 group O epidemic in western lowland gorillas. Proc Natl Acad Sci U S A. 2015;112(11):E1343-52.

24. Gao F, Yue L, White AT, Pappas PG, Barchue J, Hanson AP, et al. Human Infection by genetically diverse SIVsm-related HIV-2 in West Africa. Nature. 1992;358:495-9.

25. Hirsch VM, Olmsted RA, Murphey-Corb M, Purcell RH, Johnson PR. An African primate lentivirus (SIVsm) closely related to HIV-2. Nature. 1989;339:389-92.

26. Lemey P, Pybus OG, Wang B, Saksena NK, Salemi M, Vandamme AM. Tracing the origin and history of the HIV-2 epidemic. Proc Natl Acad Sci U S A. 2003;100(11):6588-92.

27. Desrosiers RC. Nonhuman Lentiviruses. In: Knipe DM, Howley PM, editors. Field's Virology. 5th Ed.2006. p. 2215-44.

28. Locatelli S, Peeters M. Cross-species transmission of simian retroviruses: how and why they could lead to the emergence of new diseases in the human population. AIDS. 2012;26(6):659-73.

29. Simon F, Mauclere P, Roques P, Loussert-Ajaka I, Muller-Trutwin MC, Saragosti S, et al. Identification of a new human immunodeficiency virus type 1 distinct from group M and group O. Nat Med. 1998;4(9):1032-7.

30. Plantier JC, Leoz M, Dickerson JE, De Oliveira F, Cordonnier F, Lemee V, et al. A new human immunodeficiency virus derived from gorillas. Nat Med. 2009;15(8):871-2.

31. Sharp PM, Bailes E, Chaudhuri RR, Rodenburg CM, Santiago MO, Hahn BH. The origins of acquired immune deficiency syndrome viruses: where and when? Philos Trans R Soc Lond B Biol Sci. 2001;356(1410):867-76.

32. Ayouba A, Akoua-Koffi C, Calvignac-Spencer S, Esteban A, Locatelli S, Li H, et al. Evidence for continuing cross-species transmission of SIVsmm to humans: characterization of a new HIV-2 lineage in rural Cote d'Ivoire. AIDS. 2013.

33. Peeters M, Courgnaud V, Abela B, Auzel P, Pourrut X, Bibollet-Ruche F, et al. Risk to human health from a plethora of simian immunodeficiency viruses in primate bushmeat. Emerging infectious diseases. 2002;8(5):451-7.

34. Damond F, Worobey M, Campa P, Farfara I, Colin G, Matheron S, et al. Identification of a highly divergent HIV type 2 and proposal for a change in HIV type 2 classification. AIDS Res Hum Retroviruses. 2004;20(6):666-72.

35. Delaugerre C, De Oliveira F, Lascoux-Combe C, Plantier JC, Simon F. HIV-1 group N: travelling beyond Cameroon. Lancet. 2011;378(9806):1894.

36. Vallari A, Holzmayer V, Harris B, Yamaguchi J, Ngansop C, Makamche F, et al. Confirmation of Putative HIV-1 Group P in Cameroon. Journal of Virology. 2010;85(3):1403-7.

37. Marx PA, Alcabes PG, Drucker E. Serial human passage of simian immunodeficiency virus by unsterile injections and the emergence of epidemic human immunodeficiency virus in Africa. Philos Trans R Soc Lond B Biol Sci. 2001;356(1410):911-20.

38. Van Heuverswyn F, Li Y, Neel C, Bailes E, Keele BF, Liu W, et al. Human immunodeficiency viruses: SIV infection in wild gorillas. Nature. 2006;444(7116):164.

39. Takehisa J, Kraus MH, Ayouba A, Bailes E, Van Heuverswyn F, Decker JM, et al. Origin and biology of simian immunodeficiency virus in wild-living western gorillas. J Virol. 2009;83(4):1635-48.

40. Apetrei C, Lerche NW, Pandrea I, Gormus B, Silvestri G, Kaur A, et al. Kuru experiments triggered the emergence of pathogenic SIVmac. AIDS. 2006;20(3):317-21.

41. Apetrei C, Kaur A, Lerche NW, Metzger M, Pandrea I, Hardcastle J, et al. Molecular Epidemiology of Simian Immunodeficiency Virus SIVsm in U.S. Primate Centers Unravels the Origin of SIVmac and SIVstm. Journal of Virology. 2005;79(14):8991-9005.

42. Apetrei C, Robertson DL, Marx PA. The History of SIVs and AIDS: Epidemiology, Phylogeny and Biology of Isolates from Naturally SIV Infected Non-Human Primates (NHP) in Africa. Frontiers in Bioscience. 2004;9:225-54.

43. Chahroudi A, Bosinger SE, Vanderford TH, Paiardini M, Silvestri G. Natural SIV hosts: showing AIDS the door. Science. 2012;335(6073):1188-93.

44. Ling B, Apetrei C, Pandrea I, Veazey RS, Lackner AA, Gormus B, et al. Classic AIDS in a Sooty Mangabey after an 18-Year Natural Infection. Journal of Virology. 2004;78(16):8902-8.

45. Kaur A, Grant RM, R.E. M, H. M, Feinberg M, R.P. J. Diverse Host responses and outcomes following SIVmac239 infection in Sooty Mangabeys and Rhesus Macaques. Journal of Virology. 1998;72(12):9597-611.

46. Paiardini M, Pandrea I, Apetrei C, Silvestri G. Lessons Learned from the Natural Hosts of HIV-Related Viruses. Annual Review of Medicine. 2009;60(1):485-95.

47. Sodora DL, Allan JS, Apetrei C, Brenchley JM, Douek DC, Else JG, et al. Toward an AIDS vaccine: lessons from natural simian immunodeficiency virus infections of African nonhuman primate hosts. Nature Medicine. 2009;15(8):861-5.

48. Temin HM, Mizutani S. RNA-dependent DNA polymerase in virions of Rous sarcoma virus. Nature. 1970;226(5252):1211-3.

49. Baltimore D. RNA-dependent DNA polymerase in virions of RNA tumour viruses. Nature. 1970;226(5252):1209-11.

50. Goff SP. Retroviridae: The Retroviruses and Their Replication. In: Knipe DM, Howley PM, editors. Field's Virology. 5th Ed.2006. p. 1999-2070.

51. Sundquist WI, Krausslich HG. HIV-1 Assembly, Budding, and Maturation. Cold Spring Harb Perspect Med. 2012;2(7):a006924.

52. Freed EO, Martin MA. HIVs and Their Replication. In: Knipe DM, Howley PM, editors. Field's Virology. 5th Ed.2006. p. 2107-86.

53. NIAID. Image # 18163 Public Heath Image Library (PHIL)2010 [cited 2015 10 November]. Available from: http://phil.cdc.gov/phil/.

54. Kappes JC, Morrow CD, Lee SW, Jameson BA, Kent SB, Hood LE, et al. Identification of a novel retroviral gene unique to human immunodeficiency virus type 2 and simian immunodeficiency virus SIVMAC. J Virol. 1988;62(9):3501-5.

55. Cohen EA, Terwilliger EF, Sodroski JG, Haseltine WA. Identification of a protein encoded by the vpu gene of HIV-1. Nature. 1988;334(6182):532-4.

56. Strebel K, Klimkait T, Martin MA. A novel gene of HIV-1, vpu, and its 16-kilodalton product. Science. 1988;241(4870):1221-3.

57. Gibbs JS, Regier DA, Desrosiers RC. Construction and *In Vitro* Properties of SIVmac Mutants with Deletions in "Nonessential" Genes. AIDS Research and Human Retroviruses. 1994;10(5):607-16.

58. Desrosiers RC, Lifson JD, Gibbs JS, Czajak SC, Howe AYM, Arthur LO, et al. Identification of Highly Attenuated Mutants of Simian Immunodeficiency Virus. Journal of Virology. 1998;72(2):1431-7.

59. Malim M, Emerman M. HIV-1 Accessory Proteins—Ensuring Viral Survival in a Hostile Environment. Cell Host & Microbe. 2008;3(6):388-98.

60. Klatzmann D, Champagne E, Chamaret S, Gruest J, Guetard D, Hercend T, et al. T-lymphocyte T4 molecule behaves as the receptor for human retrovirus LAV. Nature. 1984;312(5996):767-8.

61. Dalgleish AG, Beverley PC, Clapham PR, Crawford DH, Greaves MF, Weiss RA. The CD4 (T4) antigen is an essential component of the receptor for the AIDS retrovirus. Nature. 1984;312(5996):763-7.

62. Elliott ST, Riddick NE, Francella N, Paiardini M, Vanderford TH, Li B, et al. Cloning and Analysis of Sooty Mangabey Alternative Coreceptors That Support Simian Immunodeficiency Virus SIVsmm Entry Independently of CCR5. J Virol. 2012;86(2):898-908.

63. Feng Y, Broder CC, Kennedy PE, Berger EA. HIV-1 entry cofactor: functional cDNA cloning of a seven-transmembrane, G protein-coupled receptor. Science. 1996;272(5263):872-7.

64. Deng H, Liu R, Ellmeier W, Choe S, Unutmaz D, Burkhart M, et al. Identification of a major co-receptor for primary isolates of HIV-1. Nature. 1996;381(6584):661-6.

65. Choe H, Farzan M, Sun Y, Sullivan N, Rollins B, Ponath PD, et al. The beta-chemokine receptors CCR3 and CCR5 facilitate infection by primary HIV-1 isolates. Cell. 1996;85(7):1135-48.

66. Alkhatib G, Combadiere C, Broder CC, Feng Y, Kennedy PE, Murphy PM, et al. CC CKR5: a RANTES, MIP-1alpha, MIP-1beta receptor as a fusion cofactor for macrophage-tropic HIV-1. Science. 1996;272(5270):1955-8.

67. Riddick NE, Hermann EA, Loftin LM, Elliott ST, Wey WC, Cervasi B, et al. A novel CCR5 mutation common in sooty mangabeys reveals SIVsmm infection of CCR5-null natural hosts and efficient alternative coreceptor use in vivo. PLoS Pathog. 2010;6(8):e1001064.

68. Klasse PJ. The molecular basis of HIV entry. Cell Microbiol. 2012;14(8):1183-92.

69. NIAID. Image # 18162 Public Heath Image Library (PHIL)2010 [cited 2015 10 November]. Available from: http://phil.cdc.gov/phil/.

70. Campbell EM, Hope TJ. HIV-1 capsid: the multifaceted key player in HIV-1 infection. Nat Rev Microbiol. 2015;13(8):471-83.

71. Gilboa E, Mitra SW, Goff S, Baltimore D. A detailed model of reverse transcription and tests of crucial aspects. Cell. 1979;18(1):93-100.

72. Bukrinsky MI, Sharova N, Dempsey MP, Stanwick TL, Bukrinskaya AG, Haggerty S, et al. Active nuclear import of human immunodeficiency virus type 1 preintegration complexes. Proc Natl Acad Sci U S A. 1992;89(14):6580-4.

73. Bowerman B, Brown PO, Bishop JM, Varmus HE. A nucleoprotein complex mediates the integration of retroviral DNA. Genes Dev. 1989;3(4):469-78.

74. Humes D, Emery S, Laws E, Overbaugh J. A species-specific amino acid difference in the macaque CD4 receptor restricts replication by global circulating HIV-1 variants representing viruses from recent infection. Journal of Virology. 2012;86(23):12472-83.

75. McCarthy KR, Johnson WE. Plastic proteins and monkey blocks: how lentiviruses evolved to replicate in the presence of primate restriction factors. PLoS Pathog. 2014;10(4):e1004017.

76. Kirmaier A, Krupp A, Johnson WE. Understanding restriction factors and intrinsic immunity: insights and lessons from the primate lentiviruses. Future Virol. 2014;9(5):483-97.

77. Hatziioannou T, Evans DT. Animal models for HIV/AIDS research. Nat Rev Microbiol. 2012;10(12):852-67.

78. Stremlau M, Owens CM, Perron MJ, Kiessling M, Autissier P, Sodroski J. The cytoplasmic body component TRIM5a restricts HIV-1 infection in Old World monkeys. Nature. 2004;427:848-53.

79. Stremlau M. From the Cover: Specific recognition and accelerated uncoating of retroviral capsids by the TRIM5 restriction factor. Proceedings of the National Academy of Sciences. 2006;103(14):5514-9.

80. Shibata R, Sakai H, Kawamura M, Tokunaga K, Adachi A. Early replication block of human immunodeficiency virus type 1 in monkey cells. Journal of General Virology. 1995;76:2723-30.

81. Shibata R, Kawamura M, Sakai H, Hayami M, Ishimoto A, Adachi A. Generation of a chimeric human and simian immunodeficiency virus infectious to monkey peripheral blood mononuclear cells. Journal of Virology. 1991;65(7):3514-20.

82. McCarthy KR, Schmidt AG, Kirmaier A, Wyand AL, Newman RM, Johnson WE. Gain-of-sensitivity mutations in a Trim5-resistant primary isolate of pathogenic SIV identify two independent conserved determinants of Trim5alpha specificity. PLoS Pathog. 2013;9(5):e1003352.

83. Soll SJ, Wilson SJ, Kutluay SB, Hatziioannou T, Bieniasz PD. Assisted Evolution Enables HIV-1 to Overcome a High TRIM5alpha-Imposed Genetic Barrier to Rhesus Macaque Tropism. PLoS Pathogens. 2013;9(9):e1003667.

84. Newman RM, Hall L, Connole M, Chen GL, Sato S, Yuste E, et al. Balancing selection and the evolution of functional polymorphism in Old World monkey TRIM5 Proceedings of the National Academy of Sciences. 2006;103(50):19134-9.

85. Kirmaier A, Wu F, Newman RM, Hall LR, Morgan JS, O'Connor S, et al. TRIM5 Suppresses Cross-Species Transmission of a Primate Immunodeficiency Virus and Selects for Emergence of Resistant Variants in the New Species. PLoS Biology. 2010;8(8).

86. Wu F, Kirmaier A, Goeken R, Ourmanov I, Hall L, Morgan JS, et al. TRIM5 alpha Drives SIVsmm Evolution in Rhesus Macaques. PLoS Pathogens. 2013;9(8):e1003577.

87. Wissing S, Galloway NL, Greene WC. HIV-1 Vif versus the APOBEC3 cytidine deaminases: an intracellular duel between pathogen and host restriction factors. Molecular aspects of medicine. 2010;31(5):383-97.

88. Harris RS, Dudley JP. APOBECs and virus restriction. Virology. 2015;479-480:131-45.

89. Simon V, Bloch N, Landau NR. Intrinsic host restrictions to HIV-1 and mechanisms of viral escape. Nat Immunol. 2015;16(6):546-53.

90. Anderson BD, Harris RS. Transcriptional regulation of APOBEC3 antiviral immunity through the CBF-beta/RUNX axis. Sci Adv. 2015;1(8):e1500296.

91. Letko M, Silvestri G, Hahn BH, Bibollet-Ruche F, Gokcumen O, Simon V, et al. Vif proteins from diverse primate lentiviral lineages use the same binding site in APOBEC3G. J Virol. 2013;87(21):11861-71.

92. Krupp A, McCarthy KR, Ooms M, Letko M, Morgan JS, Simon V, et al. APOBEC3G polymorphism as a selective barrier to cross-species transmission and emergence of pathogenic SIV and AIDS in a primate host. PLoS Pathogens. 2013;9(10):e1003641.

93. Sauter D, Unterweger D, Vogl M, Usmani SM, Heigele A, Kluge SF, et al. Human Tetherin Exerts Strong Selection Pressure on the HIV-1 Group N Vpu Protein. PLoS Pathogens. 2012;8(12):e1003093.

94. Sauter D, Schindler M, Specht A, Landford WN, Münch J, Kim K-A, et al. Tetherin-Driven Adaptation of Vpu and Nef Function and the Evolution of Pandemic and Nonpandemic HIV-1 Strains. Cell Host & Microbe. 2009;6(5):409-21.

95. Le Tortorec A, Neil SJ. Antagonism to and intracellular sequestration of human tetherin by the human immunodeficiency virus type 2 envelope glycoprotein. Journal of Virology. 2009;83(22):11966-78.

96. Goujon C, Moncorge O, Bauby H, Doyle T, Ward CC, Schaller T, et al. Human MX2 is an interferon-induced post-entry inhibitor of HIV-1 infection. Nature. 2013;502(7472):559-62.

97. Matreyek KA, Wang W, Serrao E, Singh P, Levin HL, Engelman A. Host and viral determinants for MxB restriction of HIV-1 infection. Retrovirology. 2014;11(1):90.

98. Bitzegeio J, Sampias M, Bieniasz PD, Hatziioannou T. Adaptation to the interferon-induced antiviral state by human and simian immunodeficiency viruses. Journal of Virology. 2013;87(6):3549-60.

99. Wu F, Ourmanov I, Kuwata T, Goeken R, Brown CR, Buckler-White A, et al. Sequential Evolution and Escape from Neutralization of Simian Immunodeficiency Virus SIVsmE660 Clones in Rhesus Macaques. J Virol. 2012;86(16):8835-47.

100. Reynolds MR, Sacha JB, Weiler AM, Borchardt GJ, Glidden CE, Sheppard NC, et al. The TRIM5{alpha} genotype of rhesus macaques affects acquisition of simian immunodeficiency virus SIVsmE660 infection after repeated limiting-dose intrarectal challenge. J Virol. 2011;85(18):9637-40.

101. Keele BF, Li H, Learn GH, Hraber P, Giorgi EE, Grayson T, et al. Low-dose rectal inoculation of rhesus macaques by SIVsmE660 or SIVmac251 recapitulates human mucosal infection by HIV-1. J Exp Med. 2009;206(5):1117-34.

102. Manrique J, Piatak M, Lauer W, Johnson W, Mansfield K, Lifson J, et al. Influence of Mismatch of Env Sequences on Vaccine Protection by Live Attenuated Simian Immunodeficiency Virus. Journal of Virology. 2013.

103. Demma LJ, Logsdon JM, Vanderford TH, Feinberg M, Staprans S. SIVsm Quasispecies Adaptation to a New Simian Host. PLoS Pathogens. 2005;1(1):e3.

104. Vanderford TH, Demma LJ, Feinberg MB, Staprans SI, Logsdon JM. Adaptation of a Diverse Simian Immunodeficiency Virus Population to a New Host Is Revealed through a Systematic Approach to Identify Amino Acid Sites under Selection. Molecular Biology and Evolution. 2007;24(3):660-9.

105. Jia B, Serra-Moreno R, Neidermyer W, Rahmberg A, Mackey J, Fofana IB, et al. Species-Specific Activity of SIV Nef and HIV-1 Vpu in Overcoming Restriction by Tetherin/BST2. PLoS Pathogens. 2009;5(5):e1000429.

106. Wain LV, Bailes E, Bibollet-Ruche F, Decker JM, Keele BF, Van Heuverswyn F, et al. Adaptation of HIV-1 to Its Human Host. Molecular Biology and Evolution. 2007;24(8):1853-60.

107. Seki S, Matano T. CTL Escape and Viral Fitness in HIV/SIV Infection. Front Microbiol. 2011;2:267.

108. Friedrich TC, Frye CA, Yant LJ, O'Connor DH, Kriewaldt NA, Benson M, et al. Extraepitopic compensatory substitutions partially restore fitness to simian immunodeficiency virus variants that escape from an immunodominant cytotoxic-T-lymphocyte response. Journal of Virology. 2004;78(5):2581-5.

109. Sugimoto C, Watanabe S, Naruse T, Kajiwara E, Shiino T, Umano N, et al. Protection of macaques with diverse MHC genotypes against a heterologous SIV by vaccination with a deglycosylated live-attenuated SIV. PLoS ONE. 2010;5(7):e11678.

110. Patterson LJ, Daltabuit-Test M, Xiao P, Zhao J, Hu W, Wille-Reece U, et al. Rapid SIV Env-specific mucosal and serum antibody induction augments cellular immunity in protecting immunized, elite-controller macaques against high dose heterologous SIV challenge. Virology. 2011;411(1):87-102.

111. Wu F, Ourmanov I, Riddick N, Matsuda K, Whitted S, Plishka RJ, et al. TRIM5alpha Restriction Affects Clinical Outcome and Disease Progression in Simian Immunodeficiency Virus-Infected Rhesus Macaques. Journal of Virology. 2015;89(4):2233-40.

112. Alpert MD, Rahmberg AR, Neidermyer W, Ng SK, Carville A, Camp JV, et al. Envelope-modified single-cycle simian immunodeficiency virus selectively enhances antibody responses and partially protects against repeated, low-dose vaginal challenge. J Virol. 2010;84(20):10748-64.

113. Yang X, Charlebois P, Gnerre S, Coole MG, Lennon NJ, Levin JZ, et al. De novo assembly of highly diverse viral populations. BMC Genomics. 2012;13:475.

114. Macalalad AR, Zody MC, Charlebois P, Lennon NJ, Newman RM, Malboeuf CM, et al. Highly sensitive and specific detection of rare variants in mixed viral populations from massively parallel sequence data. PLoS Comput Biol. 2012;8(3):e1002417.

115. Henn MR, Boutwell CL, Charlebois P, Lennon NJ, Power KA, Macalalad AR, et al. Whole Genome Deep Sequencing of HIV-1 Reveals the Impact of Early Minor Variants Upon Immune Recognition During Acute Infection. PLoS Pathog. 2012;8(3):e1002529.

116. Yang Z, Nielsen R. Estimating synonymous and nonsynonymous substitution rates under realistic evolutionary models. Mol Biol Evol. 2000;17(1):32-43.

117. Nielsen R, Yang Z. Likelihood models for detecting positively selected amino acid sites and applications to the HIV-1 envelope gene. Genetics. 1998;148(3):929-36.

118. Del Prete GQ, Scarlotta M, Newman L, Reid C, Parodi LM, Roser JD, et al. Comparative Characterization of Transfection- and Infection-Derived SIV Challenge Stocks for In Vivo Non-Human Primate Studies. Journal of Virology. 2013.

119. Hirsch VM, Johnson PR. Pathogenic diversity of simian immunodeficiency viruses. Virus Res. 1994;32(2):183-203.

120. Hirsch V. The Impact of TRIM5 polymorphisms on viremia in rhesus macaques with SIV. Conference Proceedings, Conference on Retrovirology and Opportunistic Infections. 2011.

121. Hirsch VM, Adger-Johnson D, Campbell B, Goldstein S, Brown C, Elkins WR, et al. A Molecularly Cloned, Pathogenic, Neutralization-Resistant Simian Immunodeficiency Virus, SIVsmE543-3. Journal of Virology. 1997;71(2):1608-20.

122. Hirsch VM, Zack PM, Vogel AP, Johnson PR. Simian immunodeficiency virus infection of macaques: end-stage disease is characterized by widespread distribution of proviral DNA in tissues. J Infect Dis. 1991;163(5):976-88.

123. Meythaler M, Martinot A, Wang Z, Pryputniewicz S, Kasheta M, Ling B, et al. Differential CD4+ T-Lymphocyte Apoptosis and Bystander T-Cell Activation in Rhesus Macaques and Sooty Mangabeys during Acute Simian Immunodeficiency Virus Infection. Journal of Virology. 2008;83(2):572-83.

124. Silvestri G, Fedanov A, Germon S, Kozyr N, Kaiser WJ, Garber DA, et al. Divergent Host Responses during Primary Simian Immunodeficiency Virus SIVsm Infection of Natural Sooty Mangabey and Nonnatural Rhesus Macaque Hosts. Journal of Virology. 2005;79(7):4043-54.

125. Hunt RD, Blake BJ, Chalifoux LV, Sehgal PK, King NW, Letvin NL. Transmission of naturally occurring lymphoma in macaque monkeys. Proc Natl Acad Sci U S A. 1983;80(16):5085-9.

126. Daniel MD, Letvin NL, King NW, Kannagi M, Sehgal PK, Hunt RD, et al. Isolation of T-cell tropic HTLV-III-like retrovirus from macaques. Science. 1985;228(4704):1201-4.

127. Naidu YM, Kestler HW, 3rd, Li Y, Butler CV, Silva DP, Schmidt DK, et al. Characterization of infectious molecular clones of simian immunodeficiency virus (SIVmac) and human immunodeficiency virus type 2: persistent infection of rhesus monkeys with molecularly cloned SIVmac. J Virol. 1988;62(12):4691-6.

128. Letvin NL, Daniel MD, Sehgal PK, Desrosiers RC, Hunt RD, Waldron LM, et al. Induction of AIDS-like disease in macaque monkeys with T-cell tropic retrovirus STLV-III. Science. 1985;230(4721):71-3.

129. Pegu P, Vaccari M, Gordon S, Keele BF, Doster M, Guan Y, et al. Antibodies with high avidity to the gp120 envelope protein in protection from simian immunodeficiency virus SIV(mac251) acquisition in an immunization regimen that mimics the RV-144 Thai trial. Journal of Virology. 2013;87(3):1708-19.

130. Sato S, Yuste E, Lauer WA, Chang EH, Morgan JS, Bixby JG, et al. Potent Antibody-Mediated Neutralization and Evolution of Antigenic Escape Variants of Simian Immunodeficiency Virus Strain SIVmac239 In Vivo. Journal of Virology. 2008;82(19):9739-52.

131. Stone M, Keele BF, Ma ZM, Bailes E, Dutra J, Hahn BH, et al. A limited number of simian immunodeficiency virus (SIV) env variants are transmitted to rhesus macaques vaginally inoculated with SIVmac251. J Virol. 2010;84(14):7083-95.

132. Liu J, Keele BF, Li H, Keating S, Norris PJ, Carville A, et al. Low-dose mucosal simian immunodeficiency virus infection restricts early replication kinetics and transmitted virus variants in rhesus monkeys. J Virol. 2010;84(19):10406-12.

133. Kearney M, Spindler J, Shao W, Maldarelli F, Palmer S, Hu SL, et al. Genetic diversity of simian immunodeficiency virus encoding HIV-1 reverse transcriptase persists in macaques despite antiretroviral therapy. J Virol. 2011;85(2):1067-76.

134. Shao W, Kearney MF, Boltz VF, Spindler JE, Mellors JW, Maldarelli F, et al. PAPNC, a novel method to calculate nucleotide diversity from large scale next generation sequencing data. Journal of virological methods. 2014;203:73-80.

135. Fischer W, Apetrei C, Santiago ML, Li Y, Gautam R, Pandrea I, et al. Distinct evolutionary pressures underlie diversity in simian immunodeficiency virus and human immunodeficiency virus lineages. J Virol. 2012;86(24):13217-31.

136. Rose PP, Korber BT. Detecting hypermutations in viral sequences with an emphasis on G --> A hypermutation. Bioinformatics. 2000;16(4):400-1.

137. He Z, Zhang W, Chen G, Xu R, Yu XF. Characterization of conserved motifs in HIV-1 Vif required for APOBEC3G and APOBEC3F interaction. J Mol Biol. 2008;381(4):1000-11.

138. Pery E, Rajendran KS, Brazier AJ, Gabuzda D. Regulation of APOBEC3 proteins by a novel YXXL motif in human immunodeficiency virus type 1 Vif and simian immunodeficiency virus SIVagm Vif. J Virol. 2009;83(5):2374-81.

139. Freed EO. Viral late domains. J Virol. 2002;76(10):4679-87.

140. Roeth JF, Collins KL. Human immunodeficiency virus type 1 Nef: adapting to intracellular trafficking pathways. Microbiol Mol Biol Rev. 2006;70(2):548-63.

141. Bouamr F, Melillo JA, Wang MQ, Nagashima K, de Los Santos M, Rein A, et al. PPPYVEPTAP motif is the late domain of human T-cell leukemia virus type 1 Gag and mediates its functional interaction with cellular proteins Nedd4 and Tsg101 [corrected]. J Virol. 2003;77(22):11882-95.

142. Strack B, Calistri A, Craig S, Popova E, Gottlinger HG. AIP1/ALIX is a binding partner for HIV-1 p6 and EIAV p9 functioning in virus budding. Cell. 2003;114(6):689-99.

143. Cannon PM, Byles ED, Kingsman SM, Kingsman AJ. Conserved sequences in the carboxyl terminus of integrase that are essential for human immunodeficiency virus type 1 replication. J Virol. 1996;70(1):651-7.

144. Zheng Y, Ao Z, Wang B, Jayappa KD, Yao X. Host protein Ku70 binds and protects HIV-1 integrase from proteasomal degradation and is required for HIV replication. The Journal of biological chemistry. 2011;286(20):17722-35.

145. Meyerson NR, Rowley PA, Swan CH, Le DT, Wilkerson GK, Sawyer SL. Positive selection of primate genes that promote HIV-1 replication. Virology. 2014;454-455:291-8.

146. Sawyer SL, Wu LI, Emerman M, Malik HS. Positive selection of primate TRIM5alpha identifies a critical species-specific retroviral restriction domain. Proc Natl Acad Sci U S A. 2005;102(8):2832-7.

147. Streicker DG, Altizer SM, Velasco-Villa A, Rupprecht CE. Variable evolutionary routes to host establishment across repeated rabies virus host shifts among bats. Proc Natl Acad Sci U S A. 2012.

148. Bonhoeffer S, Holmes EC, Nowak MA. Causes of HIV diversity. Nature. 1995;376(6536):125.

149. Zanotto PM, Kallas EG, de Souza RF, Holmes EC. Genealogical evidence for positive selection in the nef gene of HIV-1. Genetics. 1999;153(3):1077-89.

150. Yamaguchi-Kabata Y, Gojobori T. Reevaluation of amino acid variability of the human immunodeficiency virus type 1 gp120 envelope glycoprotein and prediction of new discontinuous epitopes. J Virol. 2000;74(9):4335-50.

151. Ross HA, Rodrigo AG. Immune-mediated positive selection drives human immunodeficiency virus type 1 molecular variation and predicts disease duration. J Virol. 2002;76(22):11715-20.

152. Williamson S. Adaptation in the env gene of HIV-1 and evolutionary theories of disease progression. Mol Biol Evol. 2003;20(8):1318-25.

153. McDonald JH, Kreitman M. Adaptive protein evolution at the Adh locus in Drosophila. Nature. 1991;351(6328):652-4.

154. Smith NG, Eyre-Walker A. Adaptive protein evolution in Drosophila. Nature. 2002;415(6875):1022-4.

155. Kosakovsky Pond SL, Frost SD. Not so different after all: a comparison of methods for detecting amino acid sites under selection. Mol Biol Evol. 2005;22(5):1208-22.

156. Poon AF, Swenson LC, Dong WW, Deng W, Kosakovsky Pond SL, Brumme ZL, et al. Phylogenetic analysis of population-based and deep sequencing data to identify coevolving sites in the nef gene of HIV-1. Mol Biol Evol. 2010;27(4):819-32.

157. Scheffler K, Martin DP, Seoighe C. Robust inference of positive selection from recombining coding sequences. Bioinformatics. 2006;22(20):2493-9.

158. Neil SJ, Zang T, Bieniasz PD. Tetherin inhibits retrovirus release and is antagonized by HIV-1 Vpu. Nature. 2008;451(7177):425-30.

159. Kluge SF, Mack K, Iyer SS, Pujol FM, Heigele A, Learn GH, et al. Nef proteins of epidemic HIV-1 group O strains antagonize human tetherin. Cell Host Microbe. 2014;16(5):639-50.

160. Platt EJ, Wehrly K, Kuhmann SE, Chesebro B, Kabat D. Effects of CCR5 and CD4 cell surface concentrations on infections by macrophagetropic isolates of human immunodeficiency virus type 1. J Virol. 1998;72(4):2855-64.

161. Alexander L, Du Z, Rosenzweig M, Jung JU, Desrosiers RC. A role for natural simian immunodeficiency virus and human immunodeficiency virus type 1 nef alleles in lymphocyte activation. J Virol. 1997;71(8):6094-9.

162. Goldstein S, Brown CR, Dehghani H, Lifson JD, Hirsch VM. Intrinsic susceptibility of rhesus macaque peripheral CD4(+) T cells to simian immunodeficiency virus in vitro is predictive of in vivo viral replication. J Virol. 2000;74(20):9388-95.

163. Amara RR, Villinger F, Altman JD, Lydy SL, O'Neil SP, Staprans SI, et al. Control of a mucosal challenge and prevention of AIDS by a multiprotein DNA/MVA vaccine. Science. 2001;292(5514):69-74.

164. Velu V, Titanji K, Zhu B, Husain S, Pladevega A, Lai L, et al. Enhancing SIV-specific immunity in vivo by PD-1 blockade. Nature. 2009;458(7235):206-10.

165. Dykes C, Wang J, Jin X, Planelles V, An DS, Tallo A, et al. Evaluation of a multiple-cycle, recombinant virus, growth competition assay that uses flow cytometry to measure replication efficiency of human immunodeficiency virus type 1 in cell culture. Journal of clinical microbiology. 2006;44(6):1930-43.

166. Boutwell CL, Rowley CF, Essex M. Reduced viral replication capacity of human immunodeficiency virus type 1 subtype C caused by cytotoxic-T-lymphocyte escape mutations in HLA-B57 epitopes of capsid protein. Journal of Virology. 2009;83(6):2460-8.

167. Brumme CJ, Huber KD, Dong W, Poon AF, Harrigan PR, Sluis-Cremer N. Replication fitness of multiple nonnucleoside reverse transcriptase-resistant HIV-1 variants in the presence of etravirine measured by 454 deep sequencing. Journal of Virology. 2013;87(15):8805-7.

168. Derdeyn CA, Decker JM, Sfakianos JN, Wu X, O'Brien WA, Ratner L, et al. Sensitivity of human immunodeficiency virus type 1 to the fusion inhibitor T-20 is modulated by coreceptor specificity defined by the V3 loop of gp120. J Virol. 2000;74(18):8358-67.

169. Yeh WW, Rao SS, Lim SY, Zhang J, Hraber PT, Brassard LM, et al. The TRIM5 Gene Modulates Penile Mucosal Acquisition of Simian Immunodeficiency Virus in Rhesus Monkeys. J Virol. 2011;85(19):10389-98.

170. Kim EY, Lorenzo-Redondo R, Little SJ, Chung YS, Phalora PK, Maljkovic Berry I, et al. Human APOBEC3 induced mutation of human immunodeficiency virus type-1 contributes to adaptation and evolution in natural infection. PLoS Pathogens. 2014;10(7):e1004281.

171. Zagordi O, Daumer M, Beisel C, Beerenwinkel N. Read length versus depth of coverage for viral quasispecies reconstruction. PLoS One. 2012;7(10):e47046.

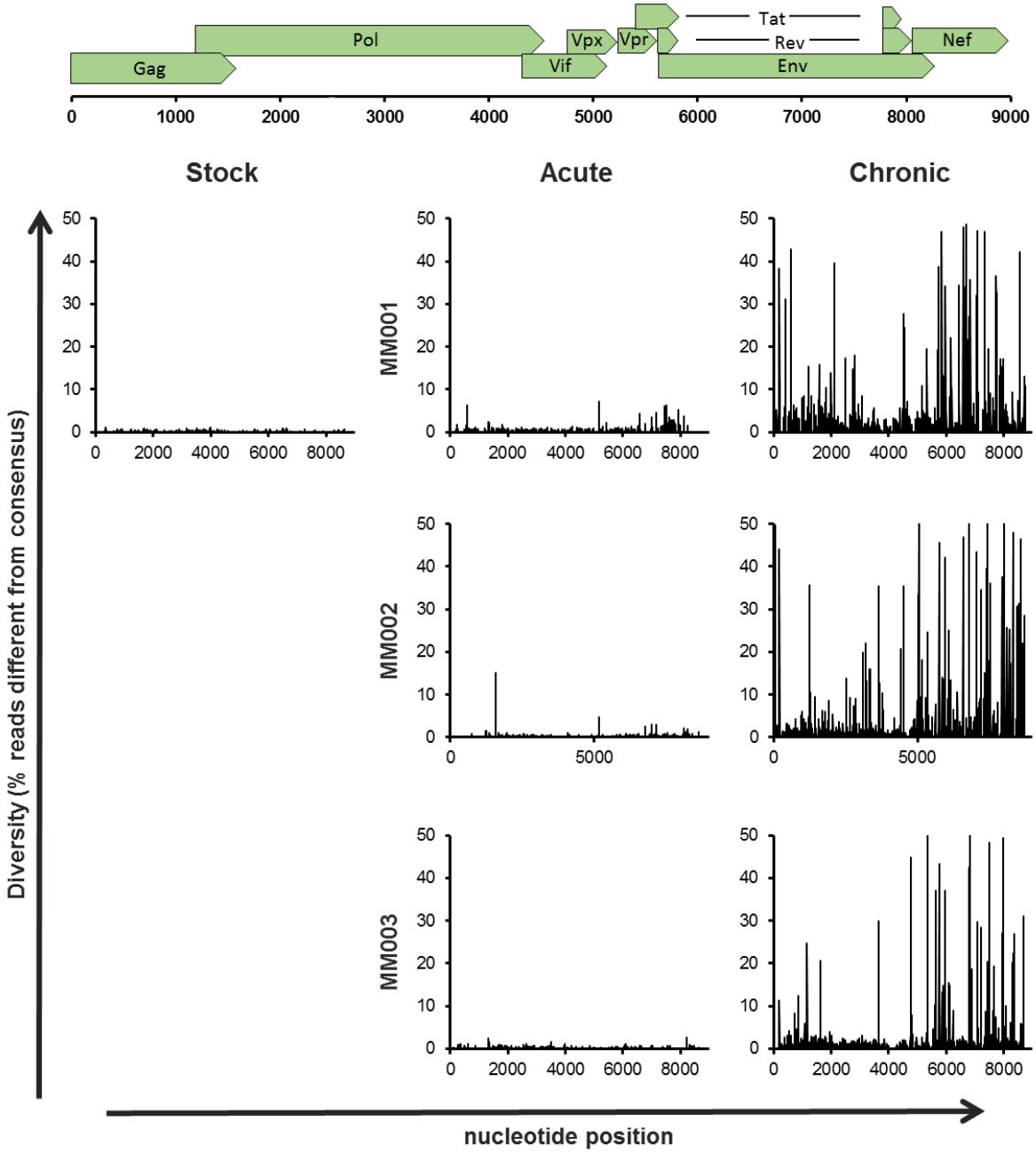
172. Li JZ, Chapman B, Charlebois P, Hofmann O, Weiner B, Porter AJ, et al. Comparison of illumina and 454 deep sequencing in participants failing raltegravir-based antiretroviral therapy. PLoS One. 2014;9(3):e90485.

173. Eid J, Fehr A, Gray J, Luong K, Lyle J, Otto G, et al. Real-time DNA sequencing from single polymerase molecules. Science. 2009;323(5910):133-8.

6.0 Appendix: Supplemental Tables and Figures

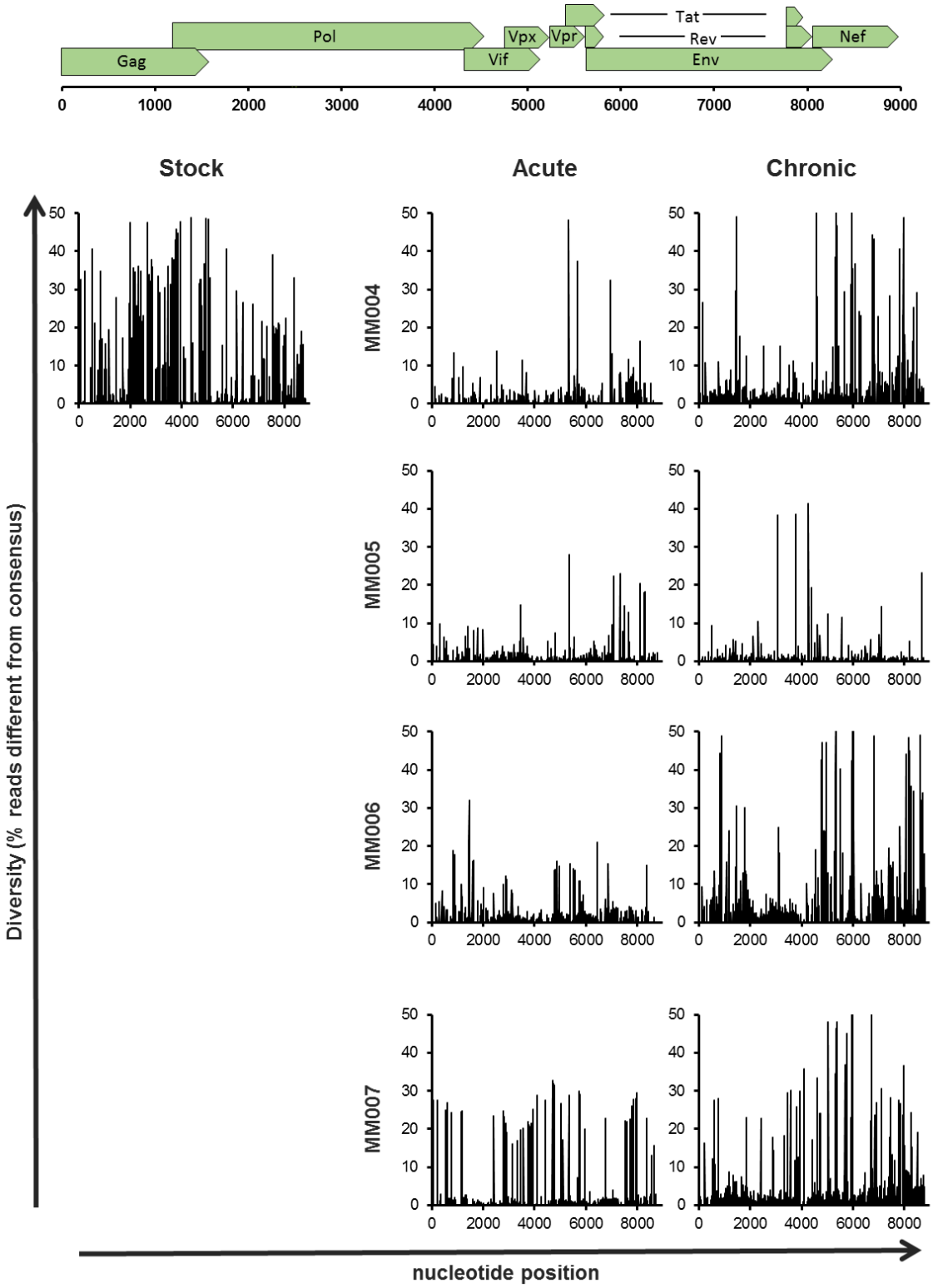
Supplemental Table 1. Virus-specific primer sequences for amplification of SIV genomes.
All primers were synthesized with the 5' amino modifier (IDT), and included Illumina adapter sequences at the 3' end to facilitate indexing**.**

Virus	Amplicon	Length (bp)	Forward Primer	Reverse Primer
SIVmac	**1**	3124	5'GAGGCCTCCGGTTGCAGGTAAG	5'TCCCATTCCGGTATCCAGGT
	2	2296	5'GCAGTGGCCATTATCAAAAG	5'CTCTGCCTTCTCTGTAATAGACC
	3	3764	5'TCAGATCTAGGGACTTGGCA	5'GTGGCAGACTTGTCTAAACG
	4	3207	5'TATGGTGTACCAGCTTGGAG	5'ACATCCCCTTGTGGAAAGTC
SIVsm	**1**	3114	5'GAGGCCTCCGGTTGCAGGTAAG	5'TCCCAATCTGGTATCCAGGT
	2	2306	5'ACAGTGGCCGCTATCAAAAG	5'CTCTGCCTTCTCTGTAATAGACC
	3	3783	5'GCAGAACTAGGGACTTGGCA	5'GTGACAGACTTGCCTAAAAG
	4	3231	5'TATGGTGTACCAGCATGGAA	5'AAAGTCCCTTGTGGAAAGTC



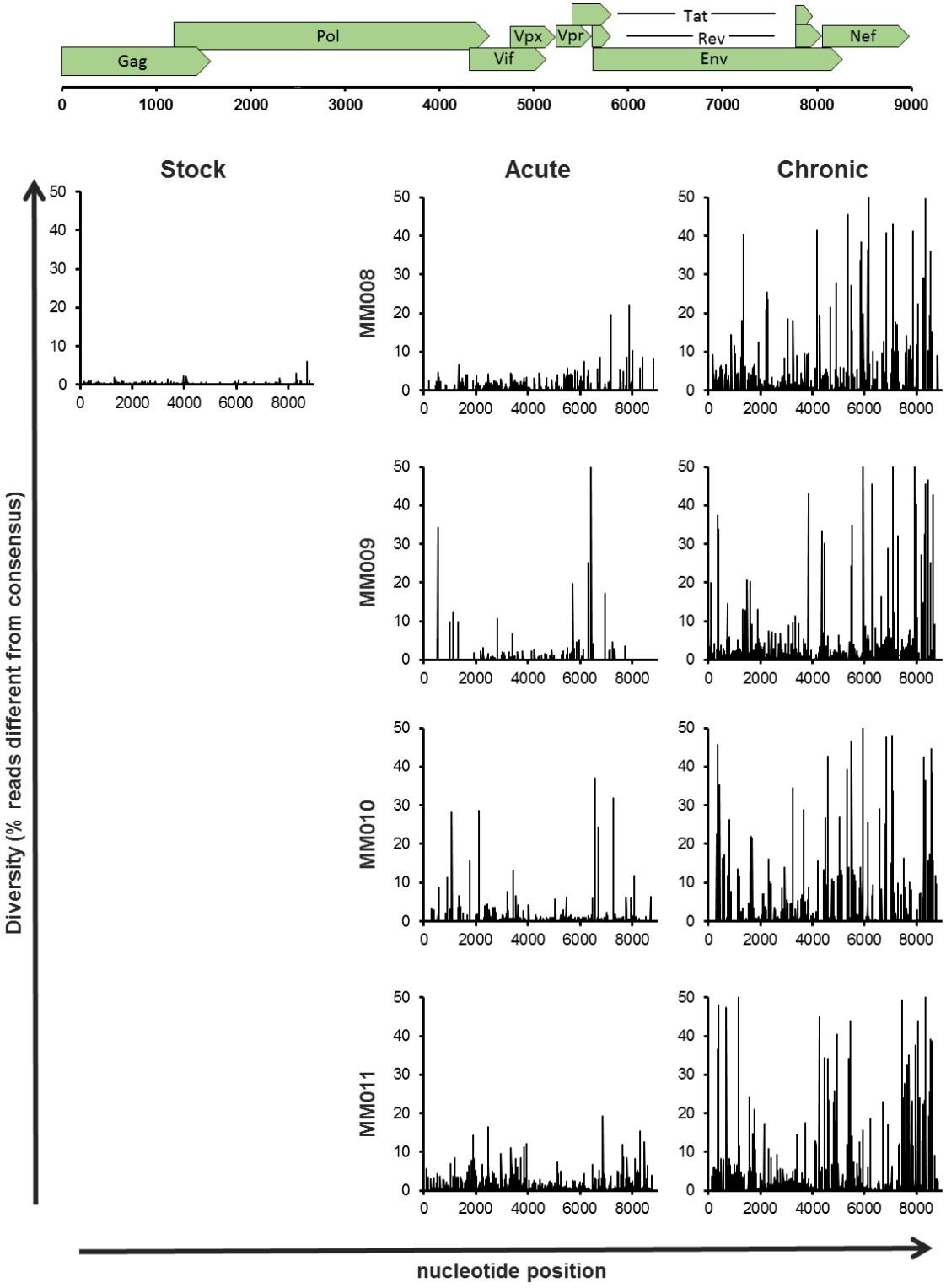
Supplemental Figure 1. Diversity of samples from SIVmac239 cohort.

Diversity, displayed as the percentage of sequencing reads at each site differing from the consensus codon, for the SIVmac239-infected cohort.



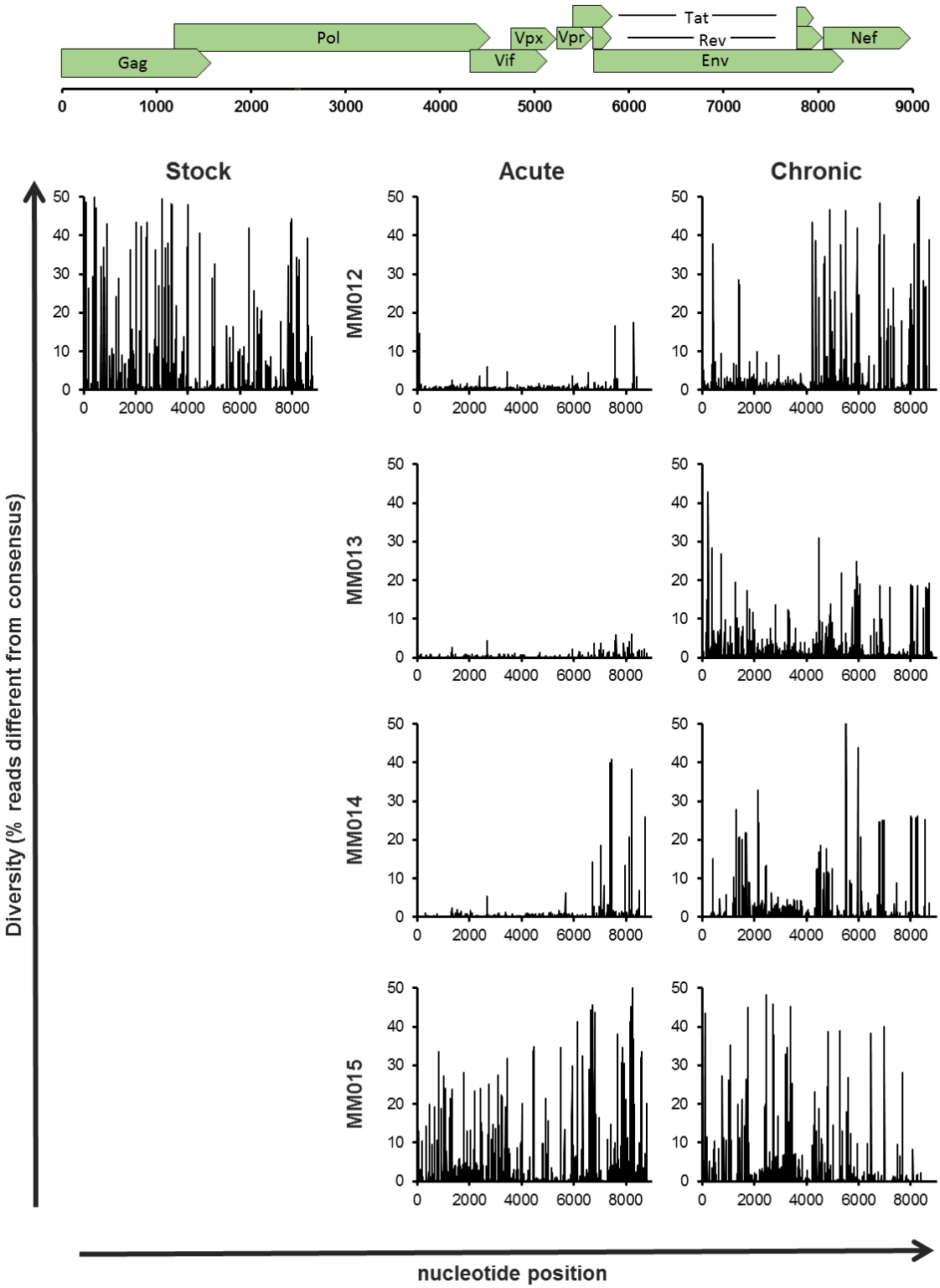
Supplemental Figure 2. Diversity of samples in SIVmac251 cohort.

Diversity, displayed as the percentage of sequencing reads at each site differing from the consensus codon, for the SIVmac251-infected cohort.



Supplemental Figure 3. Diversity of samples in SIVsmE543 cohort.

Diversity, displayed as sequencing reads at each site differing from the consensus codon, for the SIVsmE543-infected cohort.



Supplemental Figure 4. Diversity of samples in SIVsmE660 cohort.

Diversity, displayed as sequencing reads at each site differing from the consensus codon, for the SIVsmE660-infected cohort.

Supplemental Table 2. G to A changes in MM009

List of nucleotide positions where a G to A change occurred at the acute time point in animal MM009 (red text indicates a site where a premature stop codon was introduced). All hypermutants were replaced by the chronic time point. Nt, nucleotide, AA, amino acid, %, percentage of sequencing reads mapping to each codon.

Nt Position	Gene	E543 Stock		MM009, acute						MM009, chronic					
				1° Codon			2° Codon			1° Codon			2° Codon		
		AA	Codon	AA	Codon	%	AA	Codon	%	AA	Codon	%	AA	Codon	%
396	Gag	Q	CAG	Q	CAA	100				Q	CAG	100			
1209	Gag	D	GAC	D	GAC	87.53	N	AAC	12.47	D	GAC	95.05			
5225	Vpr	E	GAA	K	AAA	100				E	GAA	98.32			
6308	Env	A	GCT	T	ACT	100				A	GCT	100			
6347	Env	L	TTG	L	TTA	100				L	TTG	100			
6404	Env	V	GTG	V	GTG	74.91	V	GTA	25.09	V	GTG	100			
6515	Env	R	AGA	R	AGA	50.12	K	AAA	49.88	R	AGA	91.73	R	AGG	8.27
6821	Env	E	GAA	K	AAA	100				E	GAA	100			
6839	Env	M	ATG	I	ATA	100				M	ATG	98.6			
6923	Env	R	AGA	K	AAA	100				R	AGA	99.41			
7181	Env	R	AGA	K	AAA	100				R	AGA	96.12			
7277	Env	A	GCG	T	ACG	100				A	GCG	100			
7418	Env	R	AGA	K	AAA	100				R	AGA	100			
7442	Env	L	CTG	L	CTA	100				L	CTG	100			
7478	Env	A	GCT	T	ACT	100				A	GCT	100			
7613	Env	E	GAA	K	AAA	100				E	GAA	98.31			
7634	Env	K	AAG	K	AAA	100				K	AAG	98.02			
7664	Env	S	AGC	N	AAC	100				S	AGC	100			
7742	Env	G	GGA	K	AAA	100				G	GGA	100			
7745	Env	V	GTA	I	ATA	100				V	GTA	100			
7757	Env	R	AGA	K	AAA	100				R	AGA	95.82			
7886	Env	G	GGA	E	GAA	100				G	GGA	100			
7887	Rev	R	AGG	R	AGA	100				R	AGG	100			
7887	Tat	E	GAG	K	AAG	100				E	GAG	100			
7892	Env	E	GAA	K	AAA	100				E	GAA	96.42			
7892	Rev	R	AGA	K	AAA	100				R	AGA	96.42			
7892	Tat	K	AAG	K	AAA	100				K	AAG	96.42			
7895	Env	G	GGA	E	GAA	100				G	GGA	96.41			
7895	Rev	R	AGG	R	AGA	100				R	AGG	96.41			
7896	Tat	E	GAG	K	AAG	100				E	GAG	96.41			
7943	Env	E	GAA	K	AAA	100				E	GAA	100			
7943	Rev	R	AGA	K	AAA	100				R	AGA	100			
7943	Tat	*	TAG	*	TAA	100				*	TAG	100			
8014	**Env**	**W**	**TGG**	*****	**TGA**	**100**				**W**	**TGG**	**100**			
8014	Rev	A	GCT	T	ACT	100				A	GCT	100			
8202	Nef	G	GGG	G	GGA	100				G	GGG	100			
8204	Env	E	GAG	K	AAG	100				E	GAG	100			
8213	Env	G	GGA	E	GAA	100				G	GGA	100			
8214	Nef	E	GAA	K	AAA	100				D	GAT	100			
8301	Nef	E	GAG	K	AAG	100				E	GAG	94.79	E	GAA	5.21
8313	Nef	M	ATG	I	ATA	100				M	ATG	100			
8382	Nef	D	GAT	N	AAT	100				D	GAT	96.59			
8394	Nef	E	GAA	K	AAA	100				E	GAA	100			
8397	Nef	D	GAT	N	AAT	100				D	GAT	100			
8400	Nef	D	GAT	N	AAT	100				D	GAT	67.51	N	AAT	32.49
8403	Nef	D	GAC	N	AAC	100				D	GAC	75.75	N	AAC	24.25

Supplemental Table 3. Codon frequencies at candidate adaptation loci.

Majority variant present at each candidate locus for all samples. Where no percentage is listed, that variant was present at greater than 98% frequency.

Supplemental Table 3 (Continued)

Sample			Candidate rhesus-specific adaptations					
Cohort	**Animal**	**Week p.i.**	**MA 128**	**CA 98**	**IN 256**	**Vif 74**	**Vif 199**	**Nef 78**
SIVmac239	**STOCK**		S	S	D	H	K	E
	MM001	**2**	S	S	D	H	K	E
		40	S (94%) P (6%)	S	D	H	K	E
	MM002	**2**	S	S	D	H	K	E
		40	S	S	D	H (65%) N (35%)	K	E
	MM003	**2**	S	S	D	H	K	E
		40	S	S	D	H	K	E
SIVmac251	**STOCK**		S	S	D	H	K	E
	MM004	**6**	S	S	D	H	K	E
		41	S	S	D	H	K	E (92%) K (5%) G (3%)
	MM005	**5**	S	S	D	H	K	E
		41	S	S	D	H	K	E
	MM006	**4**	S	S	D	H	K	E
		40	S	S	D	N (93%) H (7%)	E (69%) K (31%)	E (85%) K (15%)
	MM007	**5**	S	S	D	H	K	E
		42	S	S	D	N (94%) H (6%)	K	E (85%) K (15%)
SIVsmE543	**STOCK**		P	R	E	N	E	A
	MM008	**2**	P	R	E	N	E	A
		41	P	R	E	H	E	A
	MM009	**2**	P	R	E	N	E	A
		48	S (65%) P (30%) L (4%)	S	G	H	E	A
	MM010	**3**	P	R	E	N	E	A
		72	P	G	E	N	K	A
	MM011	**3**	P	R	E	N	E	A
		56	P (53%) S (47%)	T (51%) S (48%)	E	H	E	A
SIVsmE660	**STOCK**		P	R	E	N	E	A (66%) E (21%) T (12%)
	MM012	**2**	P	R	E	N	E	E
		44	P	S	D (57%) G (42%)	N	E	E
	MM013	**2**	P	R	E	N	E	E
		44	P (91%) L (5%) S (4%)	S	D	N	E (93%) K (7%)	A
	MM014	**2**	P	R	E	N	E	T
		44	P	S	G	N	E	T
	MM015	**2**	P	R	E	N	E	A (42%) T (36%) E (19%) K (3%)
		44	P	S	E	H	K	T

Overall legend for Supplemental Figures 5-13: dN/dS and dN hotspot analysis by gene

Each figure shows the dN/dS and dN hotspot results for each cohort, which are displayed in the following order: SIVsmE543, SIVsmE660, SIVmac239, and SIVmac251.

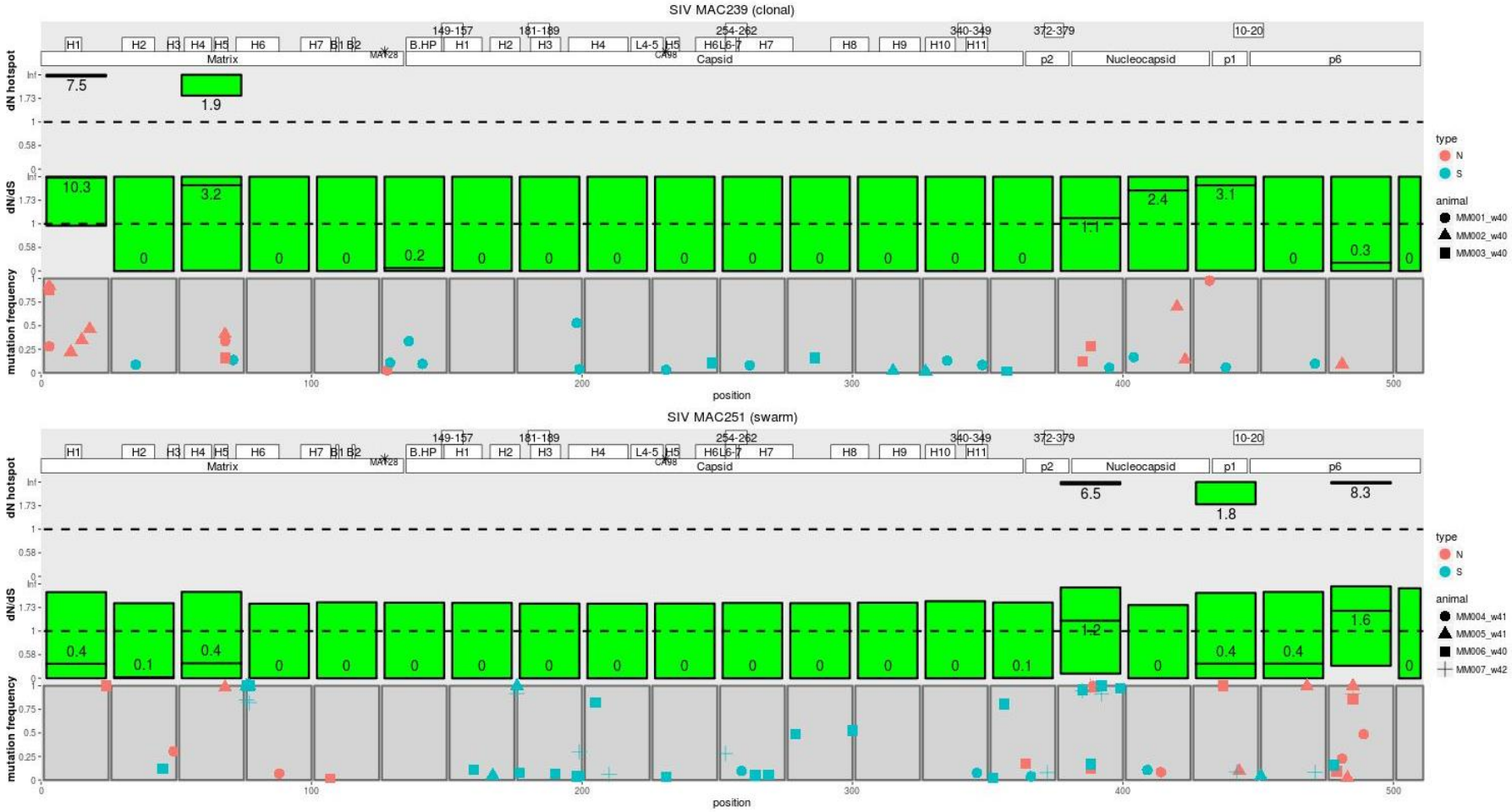
For each cohort, the bottom panel (mutation frequency) shows the frequency of non-synonymous (red) and synonymous (blue) mutations across the four animals in the cohort, with the x-axis representing the nucleotide position at the start of each codon. In the legend, the animal number is followed by the week post infection. Codons are grouped in windows according to the rectangles shown.

Second from the bottom, the dN/dS panel shows the 80% confidence interval (CI) for codons grouped by windows. Numbers shown represent the estimated dN/dS value.

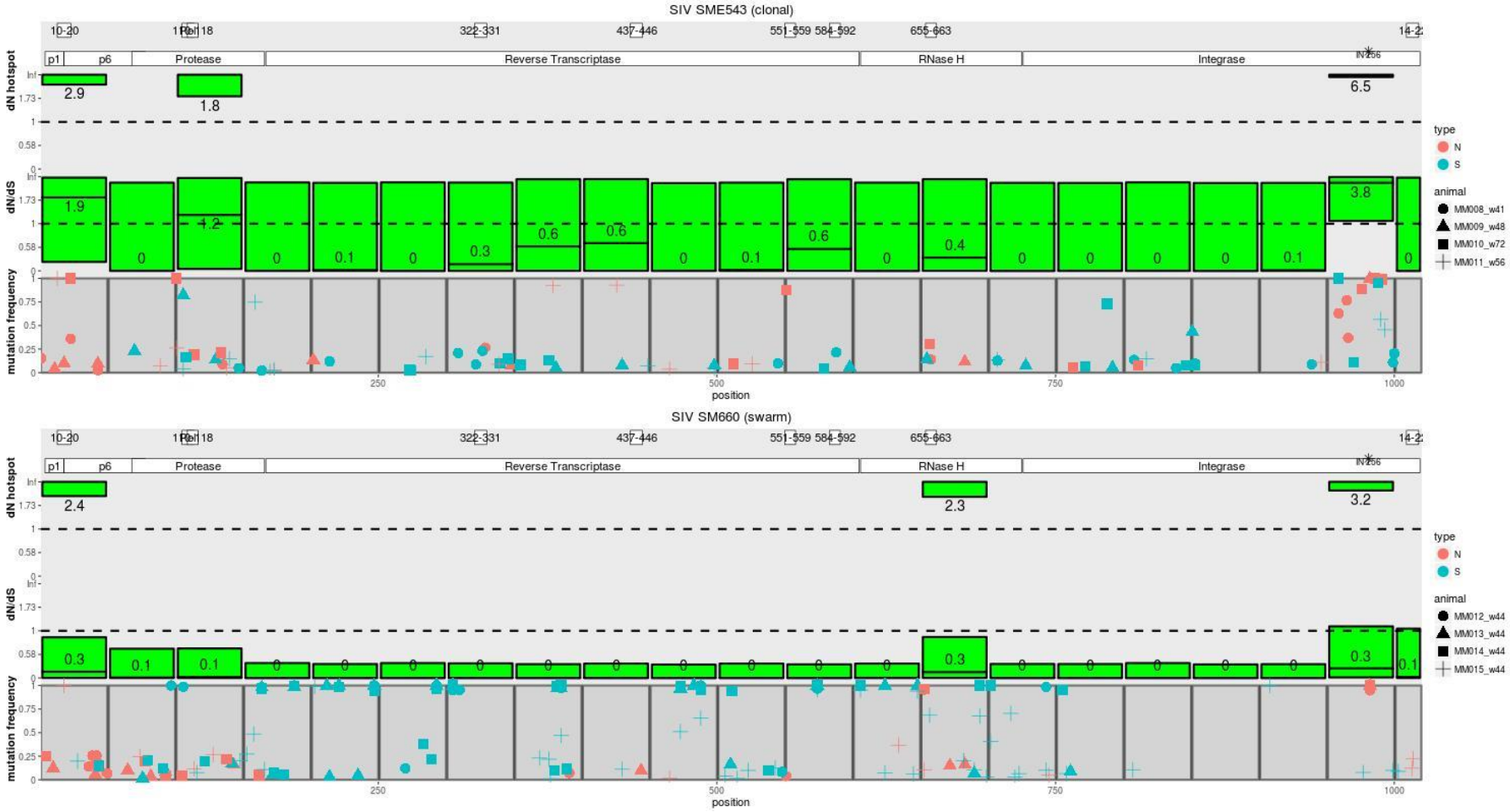
Third from the bottom, the dN hotspot panel shows CIs for dN hotspot estimates. Numbers shown represent the lower end of the CI for p, the dN hotspot estimate, rather than the p itself. Note the y-axis is not on a linear scale, for orientation the dashed black lines shows dN/dS=1.

Finally, the top panel provides protein annotation information, with the protein names at the bottom, structural features in the middle (for Gag figures only, H = helix, B = beta strand, B.HP = beta hairpin, L = loop), and known CTL epitopes at the top. Asterisks mark locations of candidate adaptations identified in section 2.5.4.1.

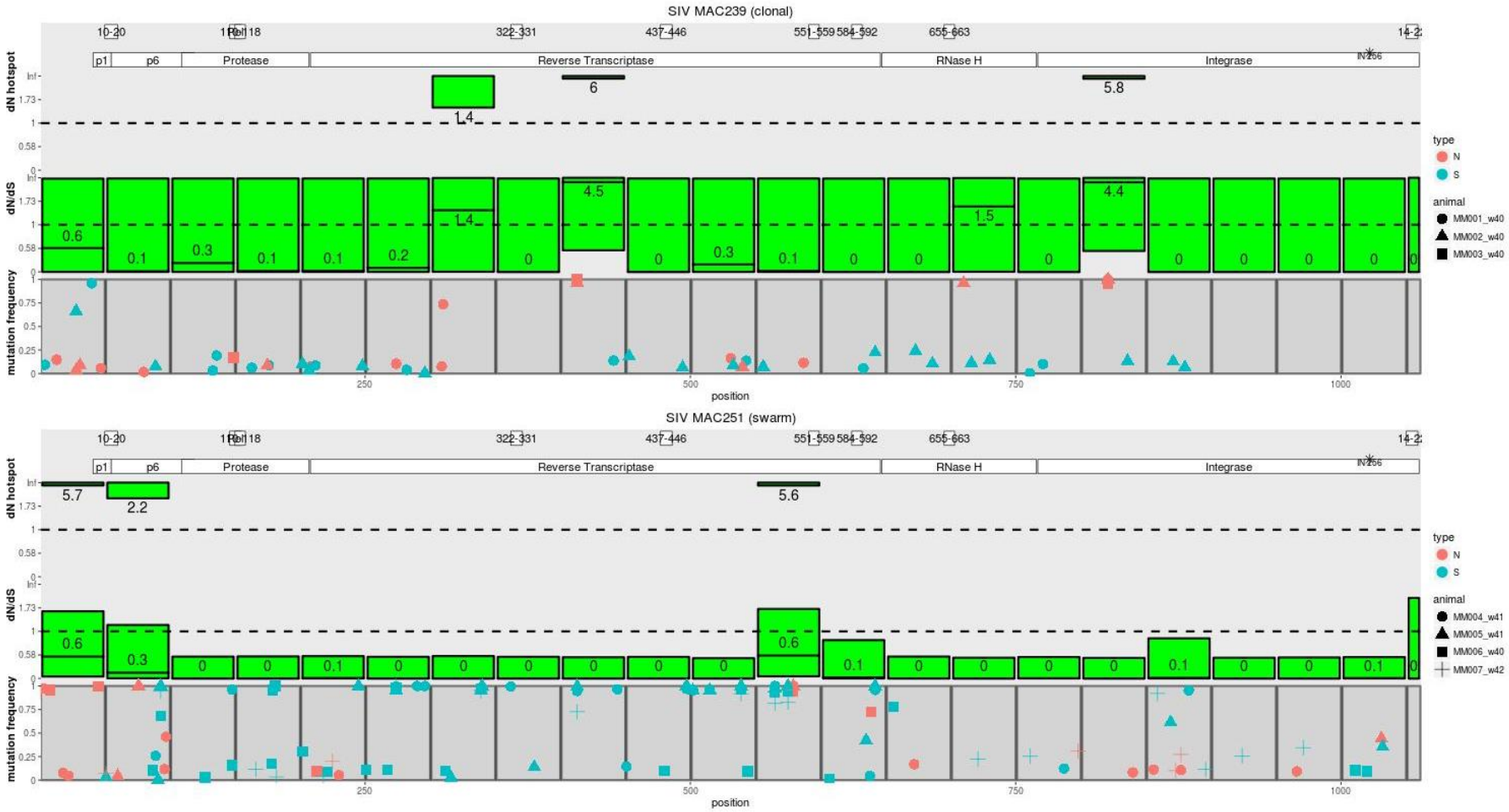
Supplemental Figure 5. dN/dS and dN hotspot analyses: Gag.



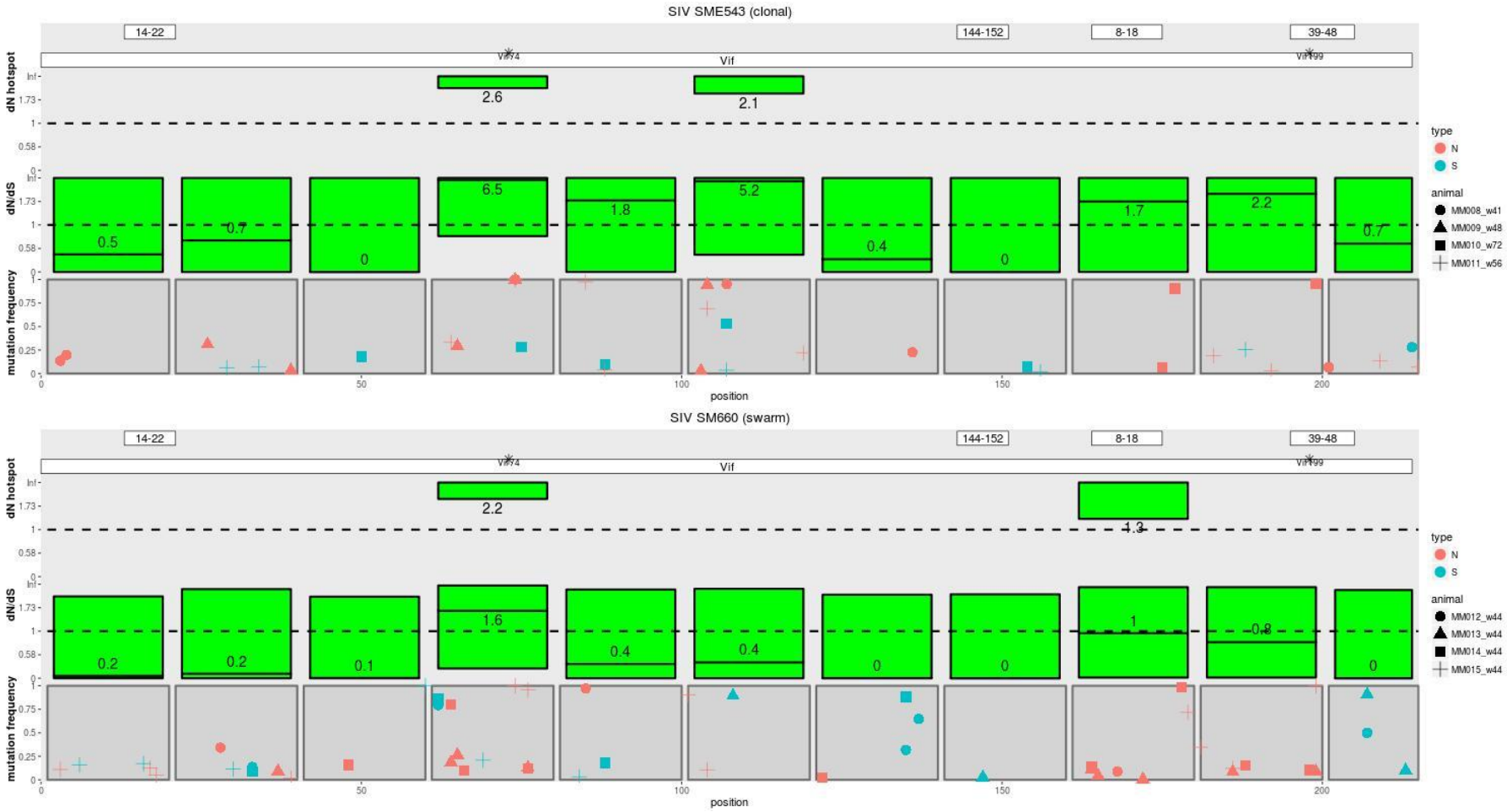
Supplemental Figure 5 (Continued).



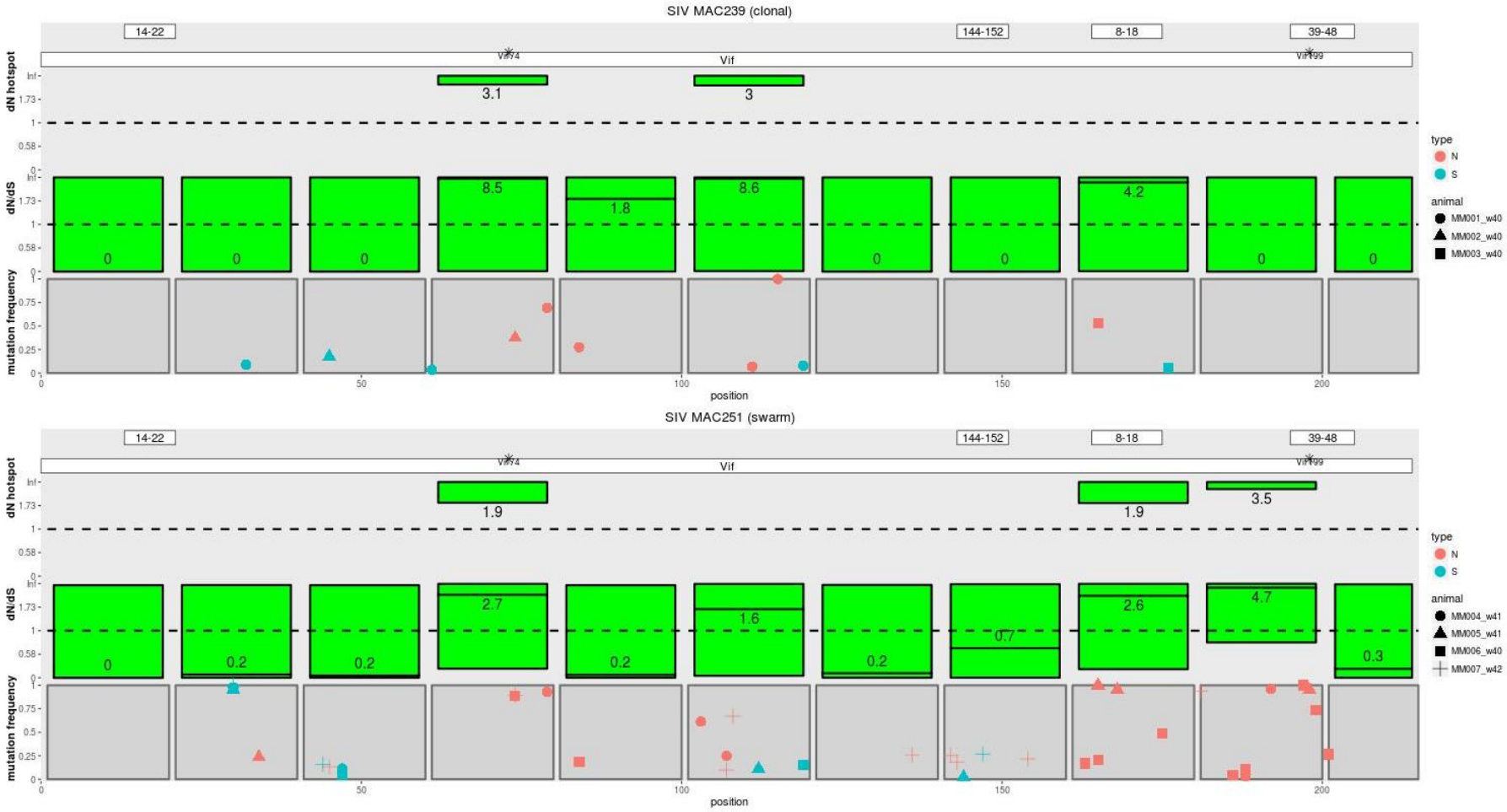
Supplemental Figure 6. dN/dS and dN hotspot analyses: Pol



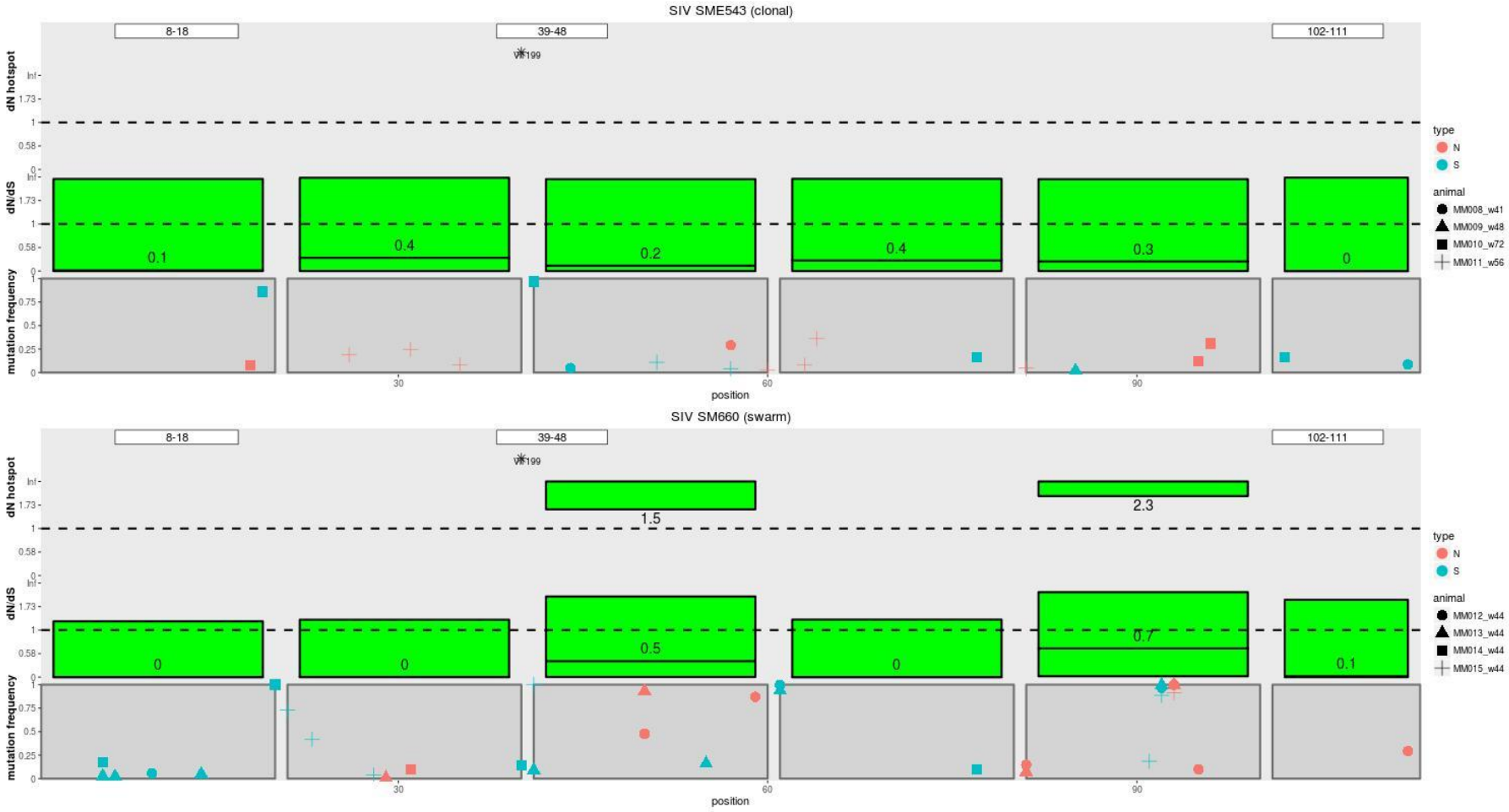
Supplemental Figure 6 (Continued).



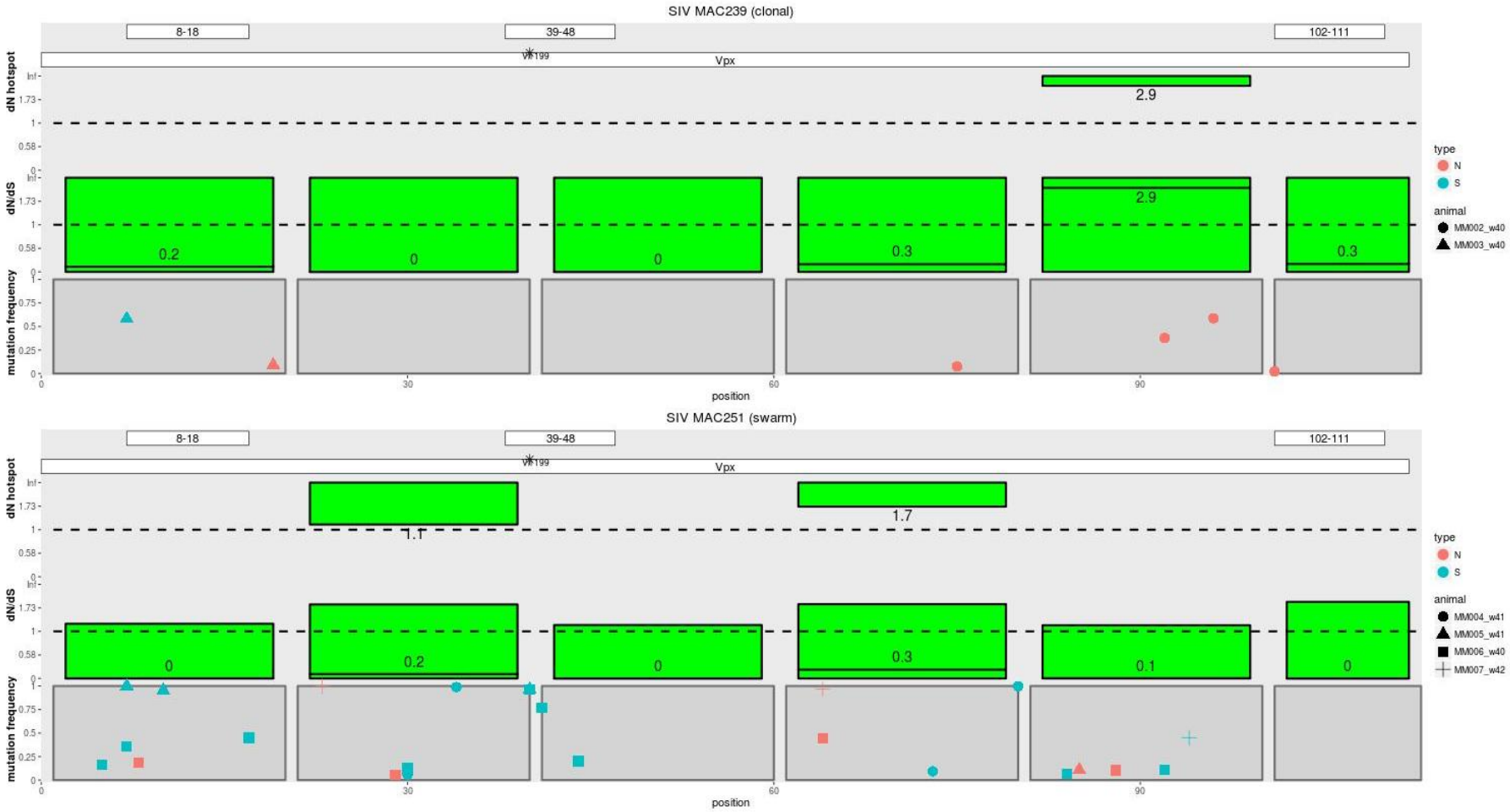
Supplemental Figure 7. dN/dS and dN hotspot analyses: Vif



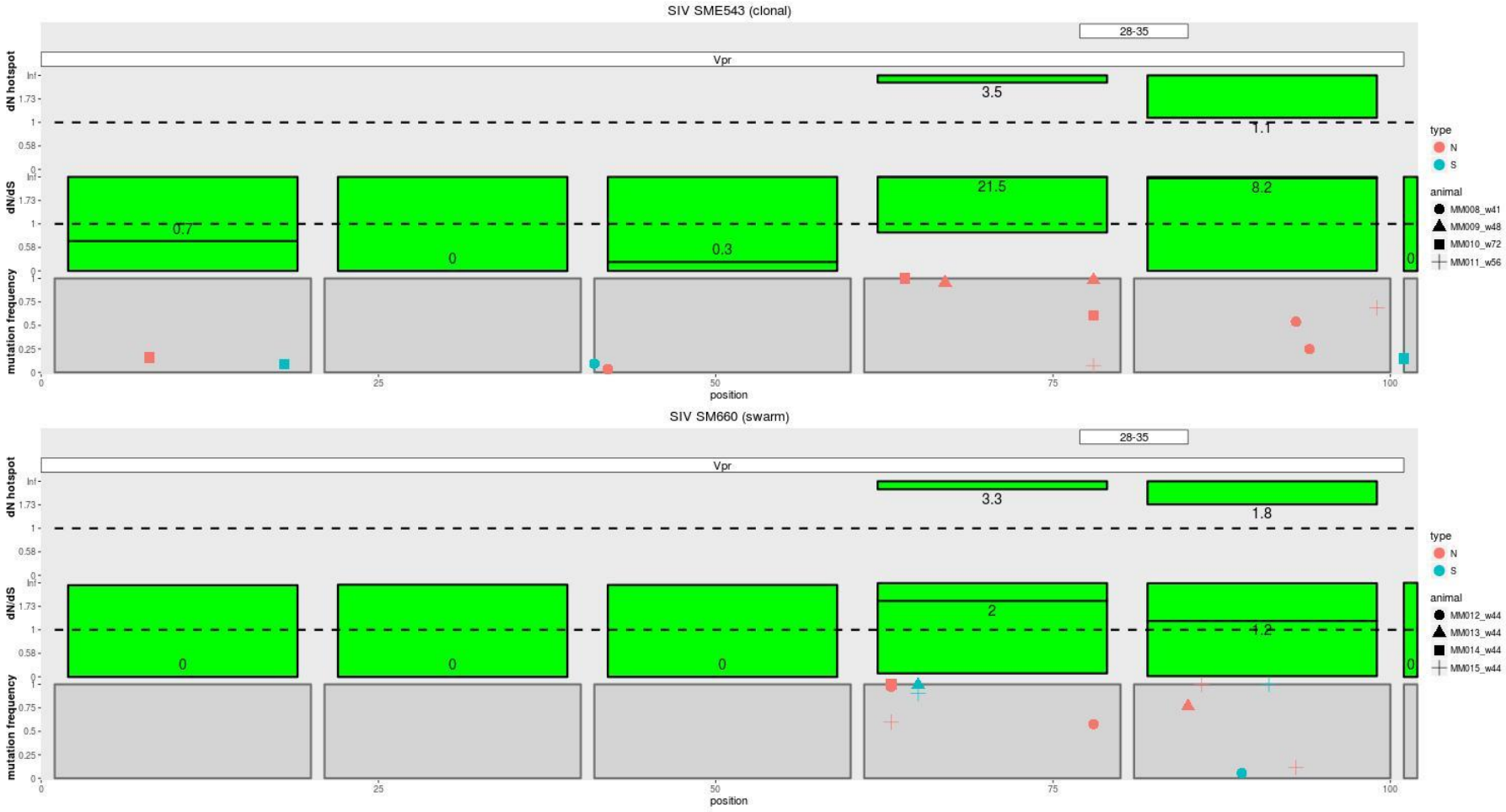
Supplemental Figure 7 (Continued)



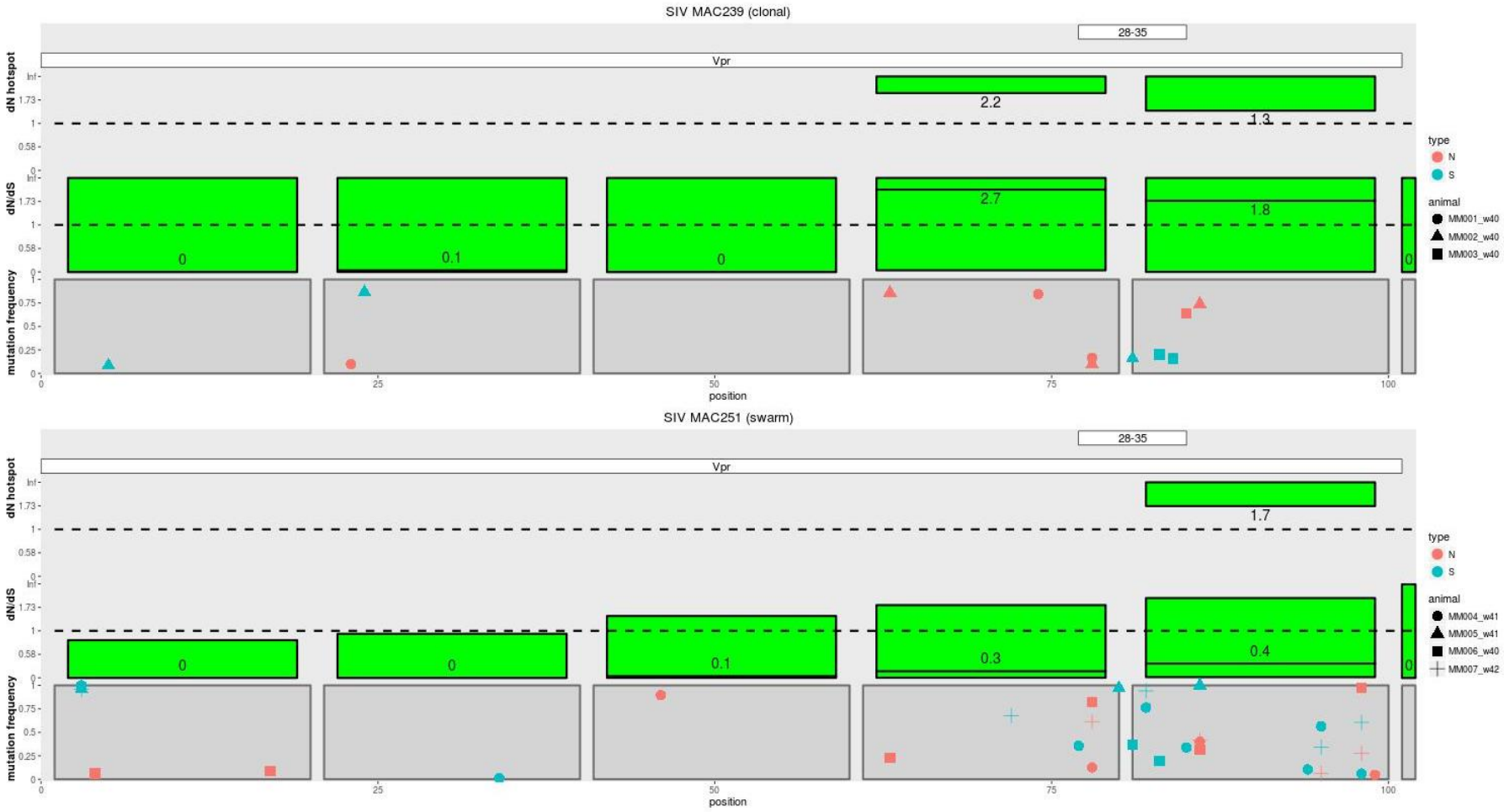
Supplemental Figure 8. dN/dS and dN hotspot analyses: Vpx



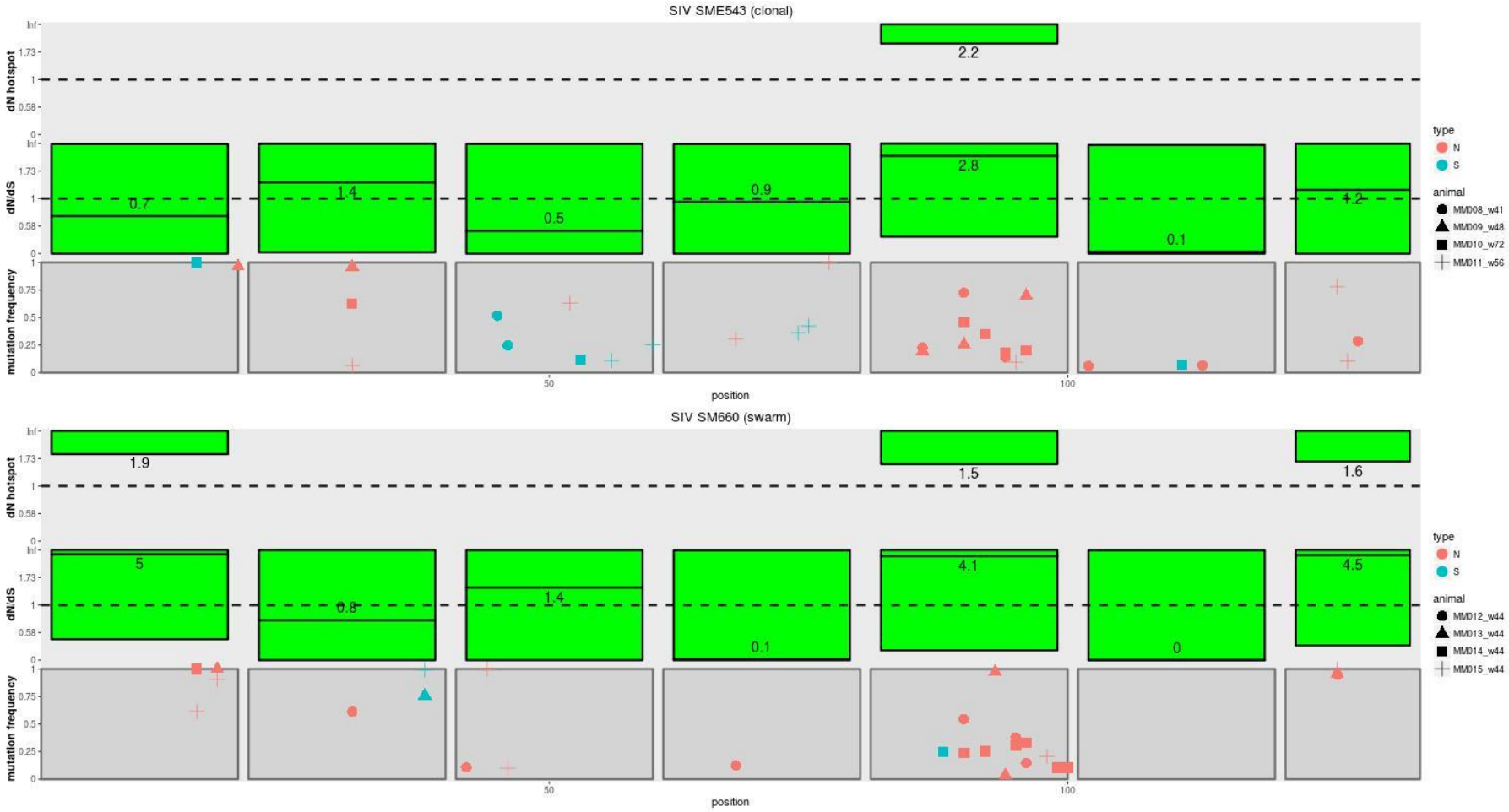
Supplemental Figure 8 (Continued)



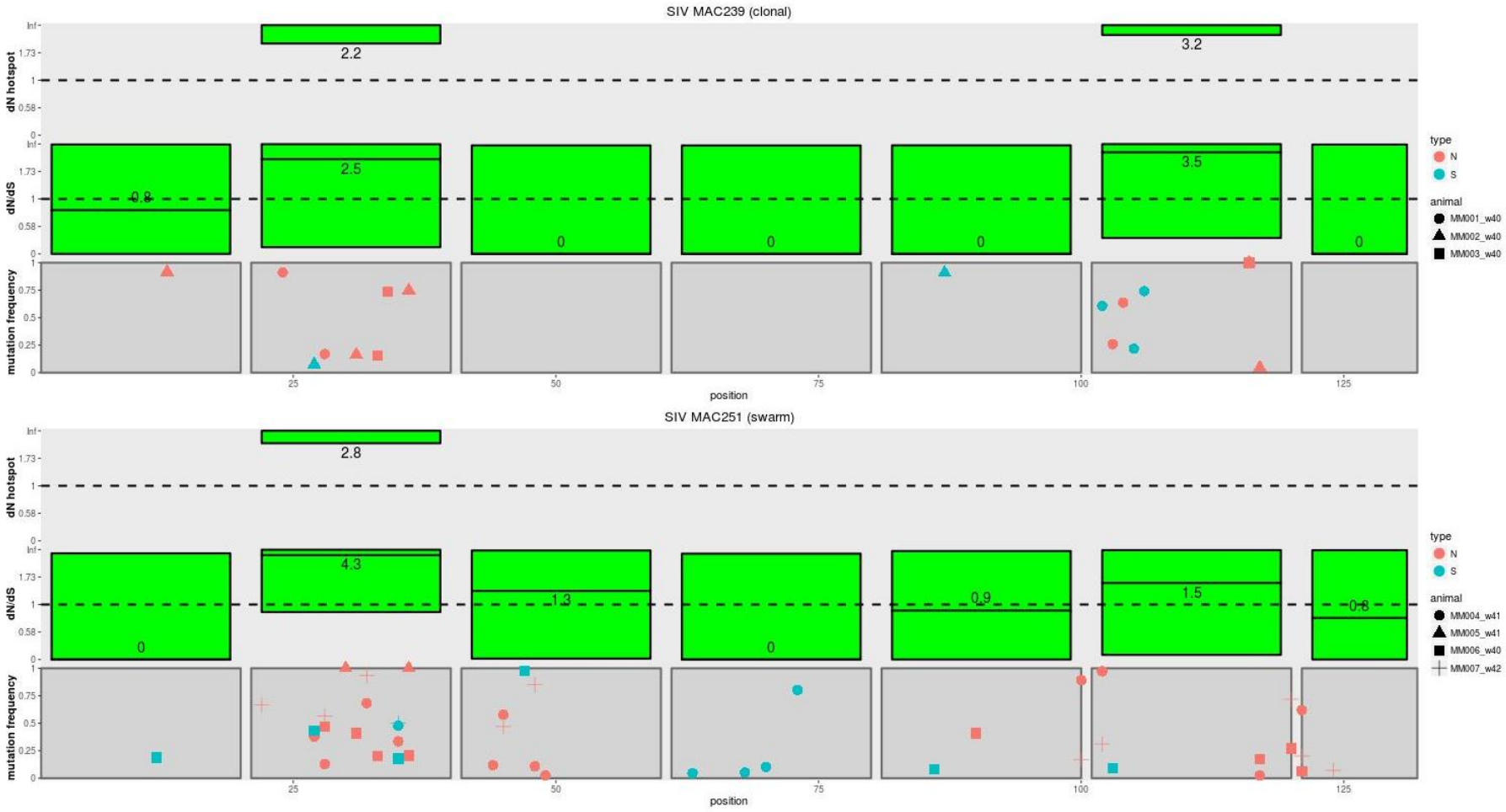
Supplemental Figure 9. dN/dS and dN hotspot analyses: Vpr



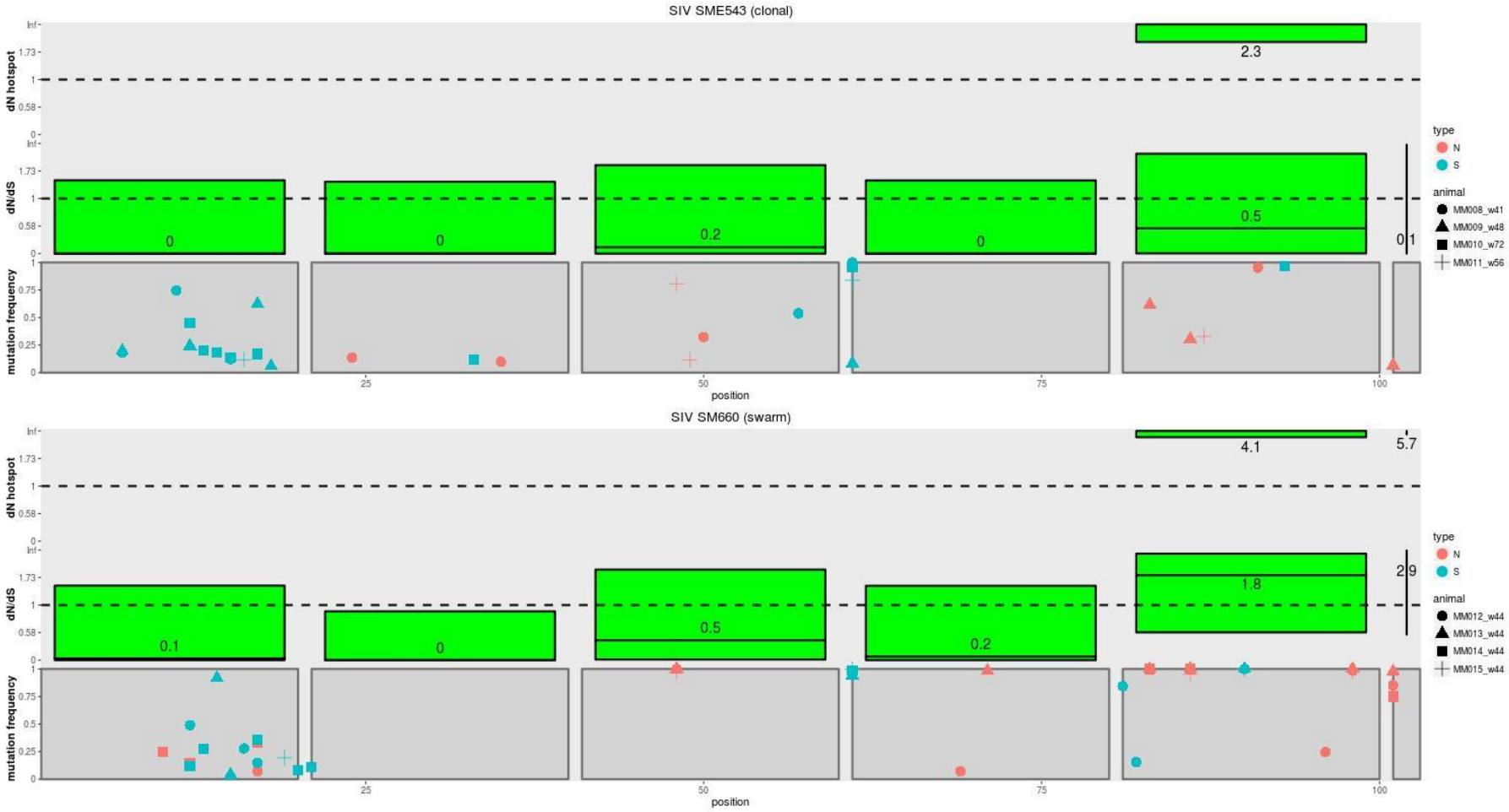
Supplemental Figure 9 (Continued)



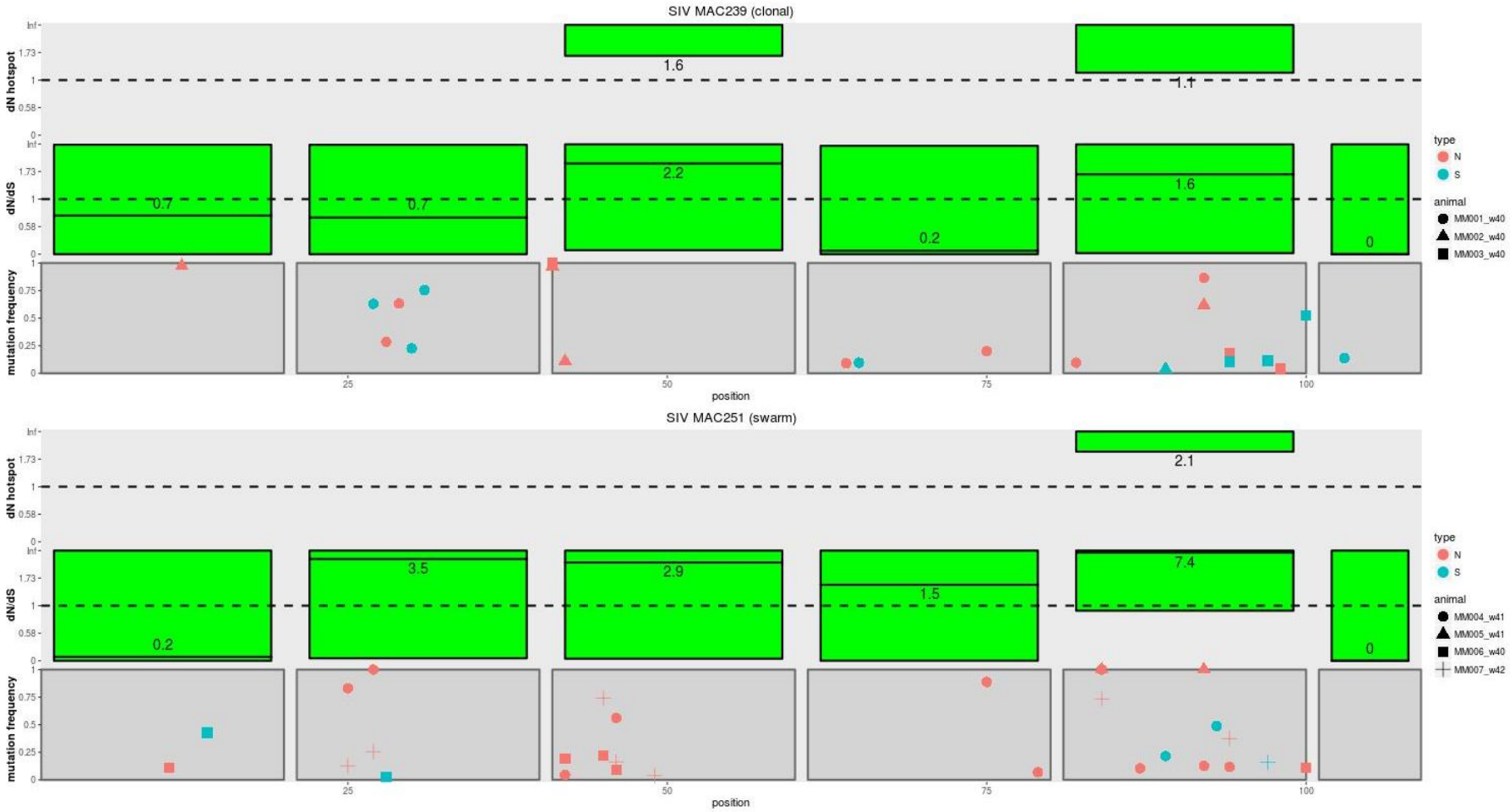
Supplemental Figure 10. dN/dS and dN hotspot analyses: Tat



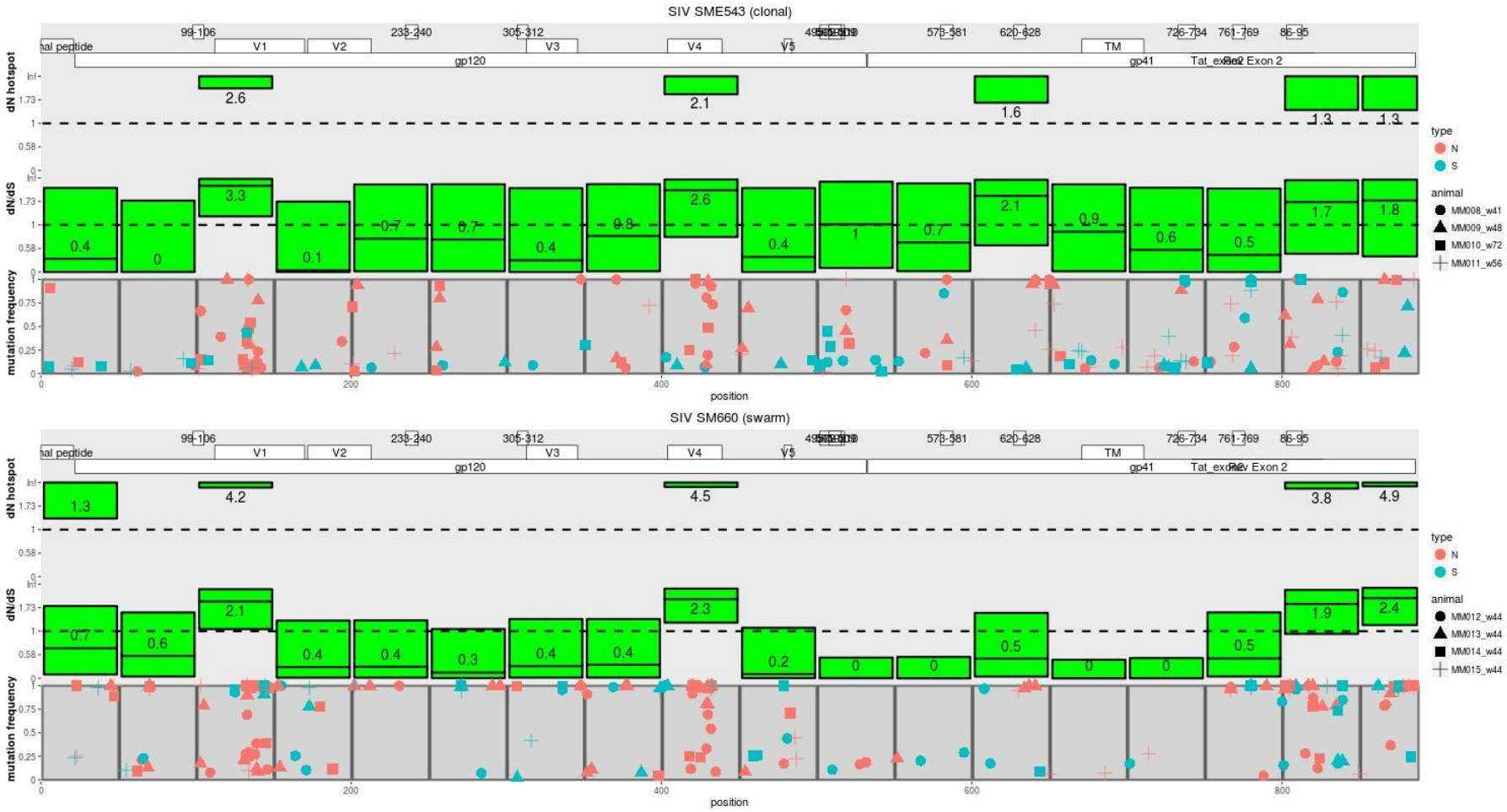
Supplemental Figure 10 (Continued)



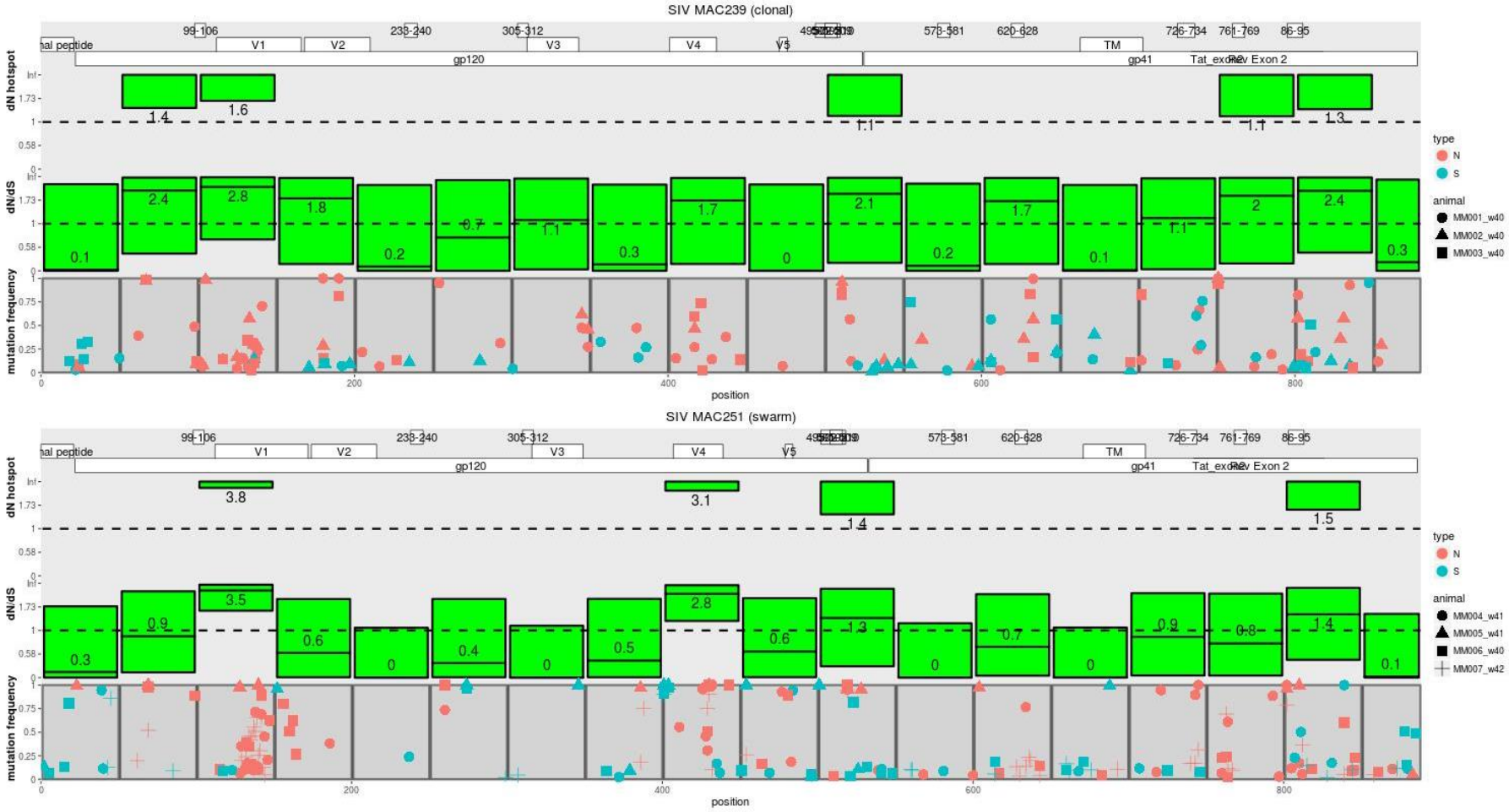
Supplemental Figure 11. dN/dS and dN hotspot analyses: Rev



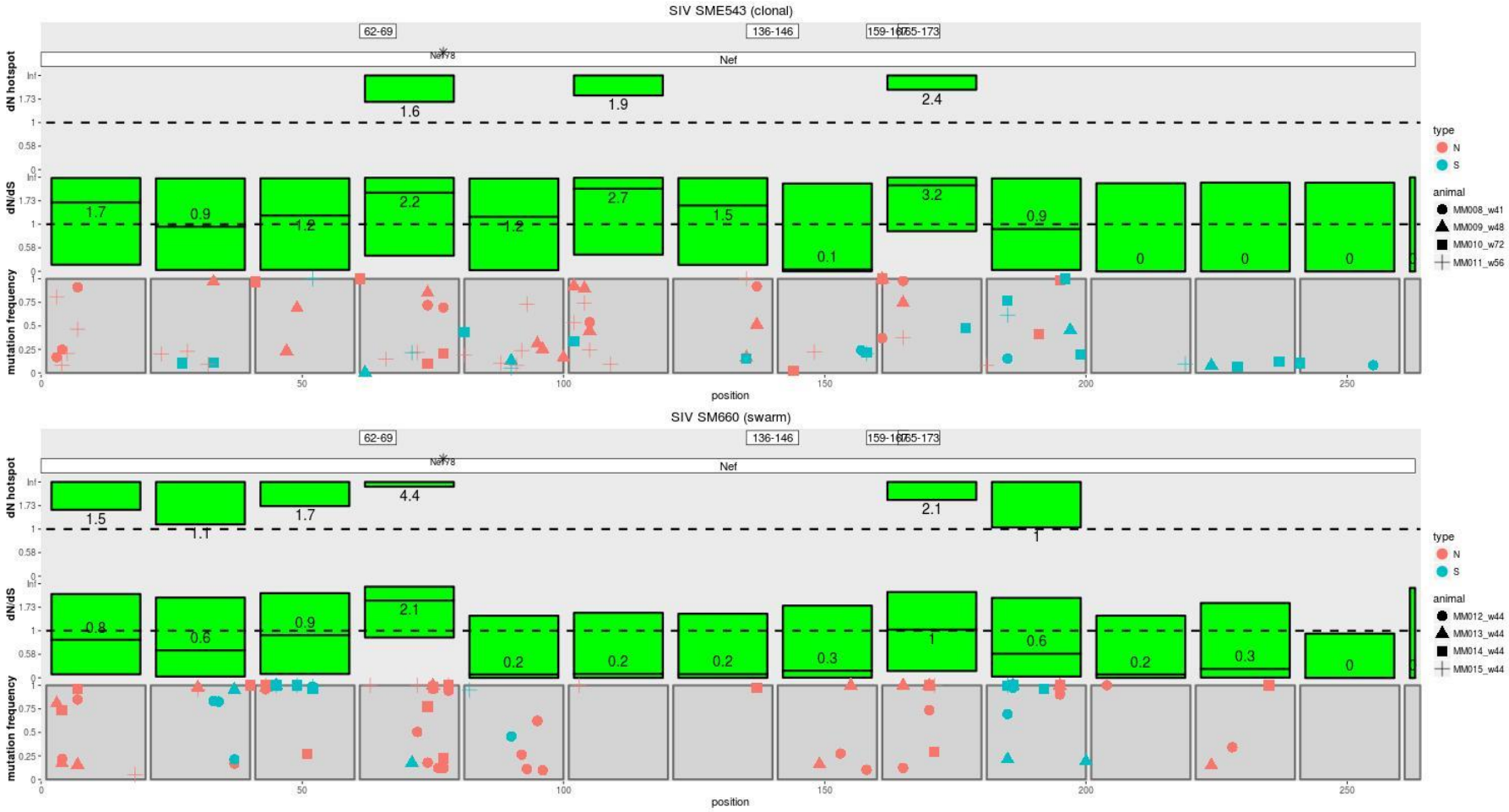
Supplemental Figure 11 (Continued)



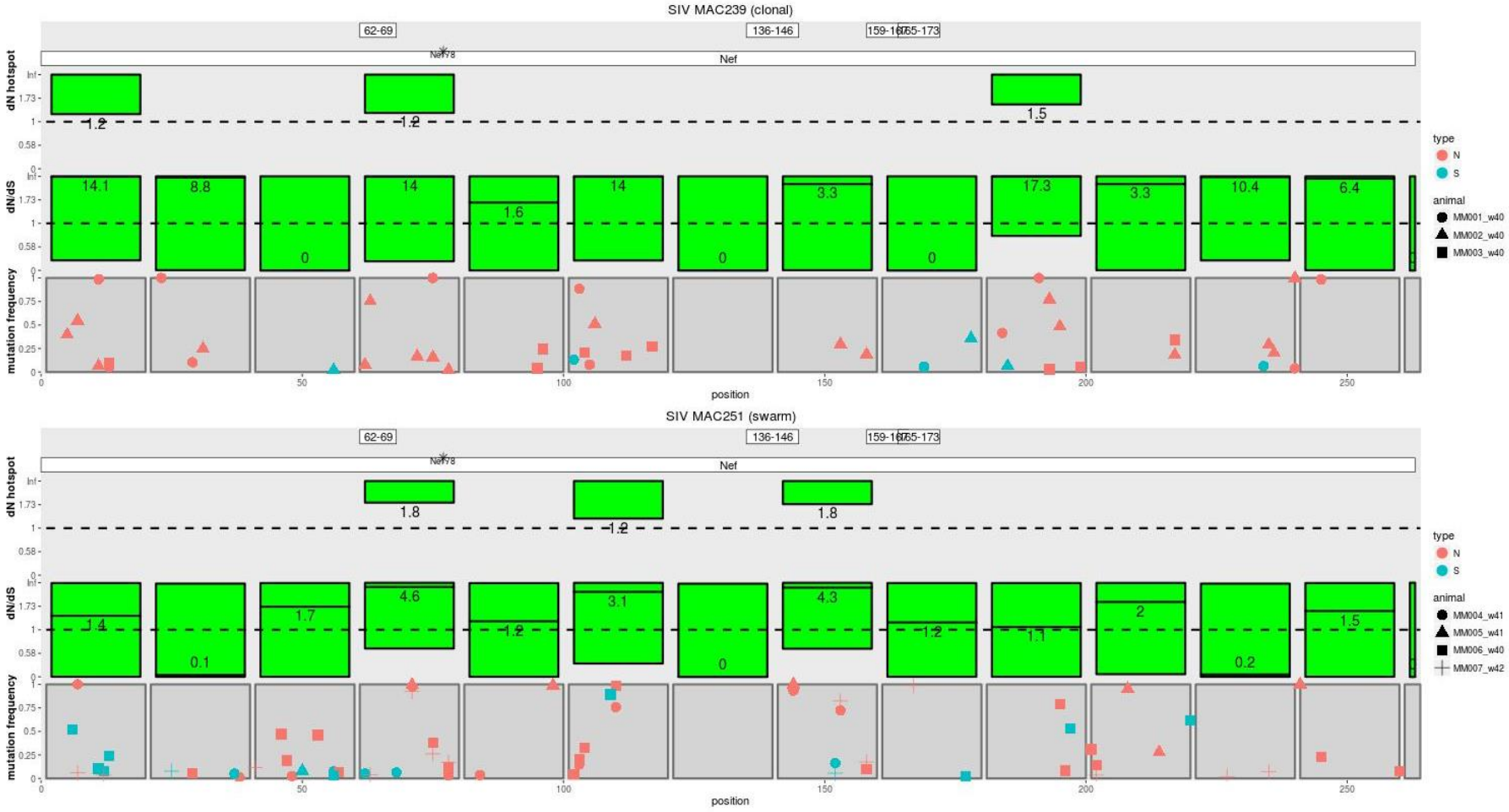
Supplemental Figure 12. dN/dS and dN hotspot analyses: Env



Supplemental Figure 12 (Continued)



Supplemental Figure 13. dN/dS and dN hotspot analyses: Nef



Supplemental Figure 13 (Continued)

Supplemental Table 4. Virus-specific primer sequences for FitSeq assay.
Primer sequences are oriented 5' to 3'. All primers were syntheized by IDT with the following 5' Illumina adapter sequences: forward, TCGTCGGCAGCGTCAGATGTGTATAAGAGACAG; reverse, GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAG.

Target site	Virus	Forward	Reverse
CA98	SIVmac239	GGAGACCATCAAGCGGCTATGCAGAT	GCCTACTGGTATGGGGTTCTGTTGTC
IN256	SIVmac239	GAAGGCAGAGATCAACTGTGGAAGGG	CGGTATCCTCCATGTGGGAACTGCTA
MA128	SIVmac239	GAGGAAGCAAAACAGATAGTGCAGAG	CCAGGCATTTAATGTTCTCGGGCTTAATGG
Vif74	SIVmac239	CATTTTAAGGTCGGATGGGCATGGTG	CGCTGTAAAGCAAGGGAAATAAGTGC
CA98	SIVsmE543	CTGCCGATTGGGATTTACAACACCCGC	GCCTACTGGTATGGGGTTTTGTTGCCTG
IN256	SIVsmE543	AAGGCAGAGACCAGCTGTGGA	CGGTATCCTCCAAGTGGGAACCACTA
MA128	SIVsmE543	GAGGAAGCAAAACAAATAGTGCAGAG	CCAAGCATTTAATGTTCTTGGACTTAAGGG
Vif74	SIVsmE543	CATCATAAAGTTGGATGGGCATGGTG	CTCTGAAAAGCAAGGGAAATAAGTGC