

Adaptation in the *env* Gene of HIV-1 and Evolutionary Theories of Disease Progression

Scott Williamson

Department of Ecology and Evolutionary Biology, University of Kansas

The exact mechanisms by which HIV overwhelms the immune system remain poorly understood. Among the several explanations of HIV disease progression, most include adaptation of the viral genome to the host environment as a causal factor. Therefore, quantifying the rate and pattern of adaptive evolution within infected patients is critical to understanding the development of AIDS. Using sequence data from infected individuals sampled at multiple time points, I estimate the within-host adaptation rate of the HIV-1 *env* gene for viral populations from 50 different patients. I find that, averaging across patients, one adaptive substitution occurs every 3.3 months. Also, one adaptive mutation is driven to a high frequency (>50% but <100%) every 2.5 months. Taken together, such adaptive events occur once every 25 viral generations, which is the fastest adaptation rate ever recorded for a single protein-coding gene. Within the entire *env* gene, I estimate that a majority (~55%) of both nonsynonymous substitutions and high-frequency polymorphisms are adaptive. Further, in the C2-V5 region of *env*, I find that patients with longer asymptomatic periods have virus populations with higher adaptation rates, corroborating the notion that a broad, strong immune response against epitopes in the *env* gene product leads to longer asymptomatic periods. I conclude by discussing the distribution of nonsynonymous changes over the *env* gene.

Introduction

Infection by the human immunodeficiency virus (HIV) follows a distinct progression of events (Clark et al. 1991; Pantaleo, Graziosi, and Fauci 1993). Flu-like symptoms and high viral load characterize the first 3 to 4 weeks after initial infection. After this initial period of viremia, the patient's adaptive immune response reduces viral load, and a long but variable (0 to 20 years) asymptomatic period follows. During this time CD4⁺ T-cell counts (the primary targets of HIV infection) slowly decrease, and viral load steadily increases. Finally, the immune system is no longer able to limit viral replication, viral load dramatically increases, and the infection progresses to acquired immune deficiency syndrome (AIDS). The variability of the asymptomatic period and the exact mechanisms by which HIV overwhelms the immune system remain poorly understood.

Several major hypotheses of disease progression posit that HIV pathogenicity is a direct result of virus adaptation to the host environment (Tersmette et al. 1989; Nowak et al. 1991; Wodarz, Klenerman, and Nowak 1998; Wolinsky and Learn 1999). Specifically, the evolution of new viral phenotypes that evade the host's immune responses is thought to play a central role in disease progression. Prompted by these hypotheses, a number of empirical studies have investigated the role of viral evolution in disease progression by following genetic divergence and diversity in vivo over the course of infection (Wolfs et al. 1990; Holmes et al. 1992; Strunnikova et al. 1995; Wolinsky et al. 1996; Ganeshan et al. 1997; Markham et al. 1998; Strunnikova et al. 1998; Shankarappa et al. 1999; Viscidi 1999). Unfortunately, the results of these studies have been somewhat contradictory. For example, some studies have found a positive relationship between the accumulation rate of genetic diversity and disease

progression rate (Strunnikova et al. 1995; Markham et al. 1998; Strunnikova et al. 1998), whereas others have found the opposite pattern (Wolinsky et al. 1996; Ganeshan et al. 1997). I suggest that there are two reasons for such inconsistencies. First, in analyzing these longitudinal studies of sequence evolution, investigators have not explicitly differentiated between adaptive and selectively neutral changes (with the notable exception of Zanotto et al. [1999] and Ross and Rodrigo [2002]). Characterizing the rate and pattern of adaptation is essential to determining the clinical significance of sequence evolution in vivo. Thus, it is necessary to filter out the "evolutionary noise" of neutral mutations. Also, the number of patients sampled in each of these longitudinal studies is simply too small (generally between five and 10) to make broad generalizations regarding disease progression. A combined analysis of all the available longitudinal sequence data is urgently needed to evaluate the role of viral evolution in disease progression. In this paper, I present such a combined analysis. In addition, I adapt a method that explicitly differentiates between adaptive and neutral changes (Smith and Eyre-Walker 2002).

The *env* Gene of HIV-1

The *env* gene codes for the envelope glycoprotein gp160, which is a precursor to two glycoproteins: gp41 and gp120. I will focus on the region of *env* that ultimately gives rise to gp120. This protein is embedded in and extends exterior to the viral lipid membrane and is primarily responsible for host cell receptor binding and host cell tropism. Additionally, due partly to its physical location in the virion, gp120 contains a number of recognition sites for various adaptive immune responses, including neutralizing antibodies (e.g., Goudsmit et al. 1988), helper T lymphocytes (Fenoglio et al. 2000), and cytotoxic T lymphocytes (Walker et al. 1986; Tsubota et al. 1989). Therefore, two potentially important positive selective forces acting on the *env* gene are changes in optimal host cell receptor affinity and evasion of host immune responses. The gp120 portion of *env* has been

Key words: HIV evolution, positive selection, adaptation rate.

E-mail: scottw@ku.edu.

Mol. Biol. Evol. 20(8):1318–1325. 2003

DOI: 10.1093/molbev/msg144

Molecular Biology and Evolution, Vol. 20, No. 8,

© Society for Molecular Biology and Evolution 2003; all rights reserved.

broadly categorized into five hypervariable regions (V1 to V5) with conserved regions interspersed (Modrow et al. 1987).

The action of natural selection on the *env* gene is evident from patterns of synonymous and nonsynonymous substitutions in *env* sequences. The rate of nonsynonymous substitution is greater than the rate of synonymous substitution in some regions of *env* (Bonhoeffer, Holmes, and Nowak 1995; Yamaguchi-Kabata and Gojobori 2000), which is a clear indication of positive selection. Estimates have been obtained for the relative frequency of adaptive mutation, the strength of positive selection, and the exact location of positively selected sites (Yamaguchi-Kabata and Gojobori 2000; Ross and Rodrigo 2002). Further, within infected individuals, the strength of selection and the relative frequency of adaptive mutation are positively associated with the time to disease progression (Ross and Rodrigo 2002). If the primary selective force acting on the *env* gene is evasion of immune responses, then this result implies that those patients who mount a broad and strong immune response to HIV are able to control the virus for a longer period of time.

Methods

Distinguishing Neutral and Adaptive Changes

To distinguish among neutral and adaptive changes, I adapt the estimator originally used by Smith and Eyre-Walker (2002), which is a simple extension of the McDonald-Kreitman test (McDonald and Kreitman 1991). The method contrasts patterns of nonsynonymous and synonymous divergence and polymorphism. I have modified the estimator in two ways. First, I apply a bias-correction term (see *Appendix*). Second, I divide polymorphisms into rare and common polymorphisms. Assuming that observed synonymous changes and rare, nonsynonymous polymorphisms are neutral, the relation $(D_n - a_d)/R_n = D_s/R_s$ should hold, where D_n is the total number of nonsynonymous substitutions (i.e., new mutations that have spread to complete fixation), a_d is the number of adaptive, nonsynonymous substitutions, R_n is the number of rare (where the frequency of the new mutation is less than 50%), nonsynonymous polymorphisms, D_s is the number of synonymous substitutions, and R_s is the number rare, synonymous polymorphisms. Given this relationship, the estimator for the number of adaptive substitutions is:

$$a_d = D_n - \frac{D_s R_n}{R_s} \left(1 + \frac{1}{R_s} \right)$$

where $1 + 1/R_s$ is a bias-correction term (see *Appendix*).

In the *env* gene of HIV, positive selection may contribute to high-frequency polymorphisms as well as substitutions. Frequency-dependent selection for rare mutations may be prevalent in genes that code for the targets of immune response (Nielsen 1999). Also, latently infected cells (e.g., CD4⁺ memory cells) may serve as a reservoir for the virus population—that is, archaic virus genomes may circulate at low frequencies among the contemporary virus population (Pierson, McArthur, and Siliciano 2000; Müller, Viguera-Gómez, and Bonhoeffer 2002; Kelly et al. in press). Therefore, even though a

mutation is (at least initially) positively selected, it may never reach complete fixation. To estimate the number of adaptive mutations that have reached high frequency (greater than 50% but less than 100%) since initial infection, I use the same approach that is used for adaptive substitutions:

$$a_p = C_n - \frac{C_s R_n}{R_s} \left(1 + \frac{1}{R_s} \right)$$

where C_n and C_s are the numbers of nonsynonymous and synonymous common polymorphisms, respectively. Further, if one defines an “adaptive event” as when an adaptive mutation attains a frequency greater than 50%, regardless of whether it ultimately becomes fixed, then the estimator for the total number of adaptive events is:

$$a_t = (D_n + C_n) - \frac{(D_s + C_s) R_n}{R_s} \left(1 + \frac{1}{R_s} \right)$$

For most evolutionary models, these three estimators will underestimate the actual numbers of adaptive substitutions, adaptive common polymorphisms, and adaptive events, respectively (see *Appendix*).

Analysis of Longitudinal Sequence Samples

I obtained *env* sequence data from several longitudinal studies of HIV-1 infection (Wolfs et al. 1990; Holmes et al. 1992; Strunnikova et al. 1995; Wolinsky et al. 1996; Ganeshan et al. 1997; Markham et al. 1998; Strunnikova et al. 1998; Shankarappa et al. 1999). Data sets are available from three nonoverlapping regions in the *env* gene that are approximately the same length: V1–V2 (seven patients, ~290 bp, average 22 clones screened/time point [Strunnikova et al. 1998]), C2–V3 (43 patients, ~300 bp, average 10 sequences/time point [Wolfs et al. 1990; Holmes et al. 1992; Strunnikova et al. 1995; Wolinsky et al. 1996; Ganeshan et al. 1997; Markham et al. 1998; Shankarappa et al. 1999]), and V4–V5 (20 patients, ~340 bp, average 12 sequences/time point [Wolinsky et al. 1996; Ganeshan et al. 1997; Shankarappa et al. 1999]).

It should be noted that the sequences from the V1–V2 region were originally sampled by using a heteroduplex mobility assay to identify similarity groups, and then one clone from each similarity group was sequenced (Strunnikova et al. 1998). Also, the frequencies from each similarity group are no longer available (R. Viscidi, personal communication), so I treated each sequence as equally frequent. Therefore the samples were not truly random. Based on the sample sizes reported in the original study (Strunnikova et al. 1998), I repeated the adaptation rate analysis assuming a completely skewed frequency distribution—that is, one common similarity group in each sample, with the remaining groups represented only once. These results from analyses differed little from the analysis based on equal frequencies (data not shown).

For each infected individual, data sets were selected on the basis that the first sample was taken less than 3 years after seroconversion, and the last sample was taken at least 1 year after the first sample. The DNA sequences used were isolated from either plasma RNA or peripheral

Table 1
The Average Adaptive Substitution Rate, Accumulation Rate for Adaptive Polymorphisms, and Rate of Adaptive Events in Different Regions of the HIV-1 *env* Gene

Region	Patient	<i>n</i>	a_d /Month (95% CI)	α_d	a_p /Month (95% CI)	α_p	a_e /Month (95% CI)	α_e
V1–V2	Child	7	0.188 (0, 0.393)	0.537	0.242 (0.057, 0.428)	0.485	0.408 (0.074, 0.741)	0.479
C2–V3	Child	11	0.095 (0.010, 0.180)	0.547	0.268 (0, 0.643)	0.559	0.355 (0, 0.754)	0.627
	Adult	32	0.058 (0.031, 0.084)	0.578	0.054 (0.023, 0.085)	0.525	0.106 (0.066, 0.146)	0.595
	All	43	0.067 (0.038, 0.097)	0.570	0.109 (0.008, 0.209)	0.534	0.170 (0.061, 0.279)	0.603
V4–V5	Child	6	0.026 (0, 0.055)	0.396	0.063 (0, 0.163)	0.199	0.070 (0, 0.179)	0.198
	Adult	15	0.052 (0.027, 0.076)	0.614	0.042 (0.017, 0.067)	0.450	0.087 (0.047, 0.127)	0.558
	All	21	0.044 (0.025, 0.064)	0.552	0.048 (0.015, 0.080)	0.379	0.082 (0.041, 0.123)	0.455
Total	Child		0.309	0.506	0.573	0.447	0.833	0.477
	All		0.299	0.561	0.399	0.483	0.660	0.547

NOTE.— a_d , a_p , and a_e are the numbers of adaptive substitutions, adaptive polymorphisms, and adaptive events, respectively. Confidence intervals (CI) of 95% for the means were obtained by jackknifing across rate estimates from each patient. The α values are the proportion of changes that are adaptive for each class (see text). The value n is the number of patients analyzed.

blood mononuclear cells. These data were combined. Shankarappa et al. (1999) isolated sequences from both sources at the same times in the same patients. They found that patterns of polymorphism and divergence were virtually identical for the two types of data.

Because sequence data are available over time, and because the *env* gene is genetically homogeneous early in infection, divergence (D_n and D_s) was measured relative to a “founding” ancestral sequence. This ancestral sequence was reconstructed as the consensus sequence of the first time point sampled in each patient. Also, at polymorphic sites, the ancestral sequence was used to determine which nucleotides are derived and which are ancestral. Samples from the last time point available in each patient were aligned with the ancestral sequence using the default parameters of ClustalX (Thompson et al. 1997), and then hand corrected. Nonsynonymous and synonymous changes were classified and counted using the SITES program (Hey and Wakeley 1997). Substitutions were not corrected for saturation. Using samples from intermediate time points, I screened for multiple substitutions in a few of the faster-evolving virus populations and found little evidence for saturation.

For each patient in the basic analyses, adaptation rates were estimated as the numbers of adaptive substitutions (a_d), adaptive common polymorphisms (a_p), or adaptive events (a_e) in the last time point, divided by the time between the first and last samples. For each patient in the detailed analyses, a_d , a_p , and a_e were first estimated at each sampled time point except the first one, and then regressed on time since the first sample to estimate adaptation rates.

Results and Discussion

Adaptation Rates in Different Regions of the *env* Gene

Estimates of the average rate of adaptive substitution (a_d /month), the average accumulation rate for adaptive polymorphisms (a_p /month), and the average rate of adaptive events (a_e /month) are given in table 1. Also shown is the average proportion of nonsynonymous substitutions that are adaptive, α_d , (defined as a_d/D_n), and analogous proportions for common polymorphisms ($\alpha_p = a_p/C_n$) and adaptive events ($\alpha_e = a_e/[C_n + D_n]$). Longitudinal data sets from infected adults and perinatally infected children were analyzed separately, and a combined analysis was also

done. The V1–V2 region has the highest adaptation rates, averaging one adaptive event every 2.5 months. This result is surprising, given that most studies have focused on the V3 region as the primary location for adaptive evolution. However, because all data sets in the V1–V2 region are from perinatally infected children, the adaptation rates may not be representative of the infected population as a whole. In the combined analysis of adults and children in the C2–V3 region, one adaptive event occurs, on average, every 5.9 months, and in the V4–V5 region, one adaptive event occurs every 12.2 months. Adding the adaptation rates from different regions, I find that one adaptive event occurs every 45 days. Considering that this estimate is conservative, and considering that adaptive evolution is probably common in other parts of the HIV genome, such as the *nef* gene (Zanotto et al. 1999), such a high adaptation rate helps explain how HIV populations can maintain very high levels of replication despite strong immune responses (Ho et al. 1995). If the virus population generates mutant phenotypes that evade the current immune response (“escape mutants”) every few weeks, then a substantial portion of replicating viruses at any one time could be inaccessible to the current immune response.

Estimates of the generation time of HIV-1 range from 1.2 to 2.6 days (Perelson et al. 1996; Rodrigo et al. 1999; Fu 2001; Seo et al. 2002), so the results presented here translate to approximately one adaptive event every 25 generations. To my knowledge, this is the fastest adaptation rate ever recorded for a single protein-coding gene. For comparison, using the same method, Smith and Eyre-Walker (2002) estimated that one adaptive substitution occurs every 450 generations in the entire genomes of *Drosophila simulans* and *Drosophila yakuba*, which translates to one adaptive substitution every 6,000,000 generations in the average protein-coding gene in these two species.

Adaptation Rates and Disease Progression

I investigated the relationship between disease progression rates and adaptation rates through a detailed analysis of a subset of the available longitudinal sequence data (nine patients [Shankarappa et al. 1999]). The longitudinal data sets from these nine patients were selected because samples were taken very frequently over the entire course of infection, sample sizes were relatively large

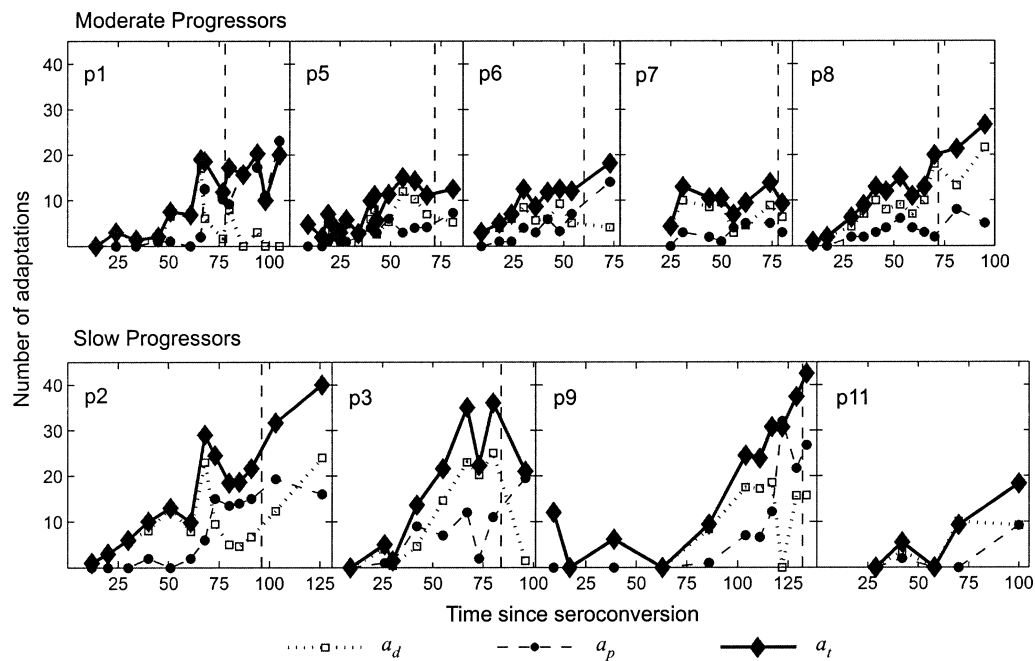


FIG. 1.—The accumulation of adaptive changes as a function of time in HIV populations from five patients with moderate disease progression rates and four patients with slow disease progression rates. Slow progressors have significantly faster adaptation rates than moderate progressors (Mann-Whitney U -test, $P < 0.05$). The vertical dashed lines represent progression times, measured as the time between seroconversion and the moment the patient's CD4⁺ T-cell count dropped below 200 cells/ μ l.

(typically 10 to 20 sequences), and a large portion of the *env* gene was sequenced, including both the C2–V3 and the V4–V5 regions (i.e., the C2–V5 region). Progression time was measured as the time between initial infection and the onset of clinical AIDS (CD4⁺ T-cell count below 200 cells/ μ l). For comparisons, patients were categorized into two groups based on their progression times: moderate progressors (progression time between 5 and 7 years; patients 1, 5, 6, 7, and 8) and slow progressors (progression time greater than 7 years; patients 2, 3, 9, and 11).

The numbers of adaptive substitutions (a_d), common polymorphisms (a_p), and adaptive events (a_t) as a function of time in each patient are shown in figure 1. The most striking pattern evident from this analysis is that slow progressors tend to have higher adaptation rates than the moderate progressors. This holds true for adaptive substitutions, common polymorphisms, and overall adaptive events. To investigate this pattern statistically, I estimated the different adaptation rates as the slope of a best-fit line for each class of adaptation in each patient. A comparison of the adaptation rates between slow and moderate progressors is shown in table 2. The rate of adaptive events is significantly higher in the slow progressors (Mann-Whitney U test, $P < 0.05$). The rate of adaptive substitutions and the accumulation rate of adaptive polymorphisms were also higher in slow progressors, but these differences were not statistically significant.

These results complement the recent results of Ross and Rodrigo (2002), who, in analyzing the same sequence data from the same nine patients, found that positive selection is more prevalent in slow progressors than in the moderate progressors. Their analysis used a different method (Nielsen and Yang 1998) for quantifying adapta-

tion and rested on very different assumptions (e.g., they assume complete linkage, whereas this study assumes free recombination). Taken together, the results of Ross and Rodrigo (2002) and the results presented here establish that viral adaptation in the C2–V5 region is related to disease duration. If the primary positive selective force acting on the C2–V5 region is the host's immune response, as seems likely, then these results further strengthen the hypothesis that a broad, strong immune response to gp120 epitopes leads to a slower rate of disease progression.

The Distribution of Adaptations over the *env* Gene

Because a large proportion of nonsynonymous substitutions and common polymorphisms are adaptive, we can make some generalizations about the spatial distribution of adaptations over the *env* gene. Of particular interest is whether most nonsynonymous changes are limited to the five hypervariable regions and whether common polymorphisms share the same distribution as substitutions. To determine the distributions of nonsynonymous substitutions and common polymorphisms, each sample was aligned with the NL4-3 genome (GenBank accession number M19921) as a reference, and the locations of the relevant nonsynonymous changes were recorded.

The distributions of nonsynonymous substitutions and common polymorphisms are shown in figure 2. The most striking pattern evident from these distributions is the remarkable lack of selective constraint in the V3 loop flanking regions and in the V3 loop itself. In the 100 codons of the C2–V3 region, there are apparently no highly conserved regions. This result is especially noteworthy, considering that this region is largely responsible for initi-

Table 2
Disease Progression Time and Adaptation Rates in the C2–V5 Region of the *env* Gene Among Moderate and Slow Progressors

Patient	Progression			
	Time	a_d /Month	a_p /Month	a_i /Month
Moderate progressors				
1	78	0.240	0.000	0.221
5	72	0.090	0.078	0.168
6	60	0.202	0.018	0.220
7	78	0.054	0.000	0.045
8	72	0.069	0.217	0.286
Average	72	0.131	0.061	0.188
Slow progressors				
2	96	0.201	0.123	0.324
3	84	0.185	0.203	0.388
9	132	0.201	0.081	0.280
11	>144	0.111	0.137	0.246
Average	>114	0.175	0.136	0.310

NOTE.—Progression time is measured as the number of months between seroconversion and the moment the patient's CD4⁺ T-cell count drops below 200 cells/μl. The patient numbers are from Shankarappa et al. (1999). Adaptation rate estimates were obtained as the slope of a best-fit line for the estimated number of adaptations as a function of time.

ating viral entry into the host cell. Further, the abundance of fixations and common polymorphisms in the regions flanking the V3 corroborates earlier results that suggested that adaptive changes may not be limited to the hyper-variable regions (Nielsen and Yang 1998; Yamaguchi-Kabata and Gojobori 2000; Ross and Rodrigo 2002). Another interesting pattern is that, in the V1–V2 region, the distribution of nonsynonymous substitutions is significantly different from the distribution of nonsynonymous common polymorphisms (G -test for goodness-of-fit, $G = 10.96$, $df = 4$, $P < 0.05$; observed changes were binned into five groups, each 50 base pairs long), with more substitutions occurring in the V1 loop and more common polymorphisms occurring in the V2 loop and flanking regions. The reasons for this difference are unclear. One possibility is that, assuming that immune-mediated, frequency-dependent selection is operating, the relationship between frequency and fitness may be variable across the V1–V2 region.

Conclusions

I have shown that adaptive evolution of the *env* gene is both extremely fast and widespread in HIV populations within infected individuals. If the primary selective pressure is imposed by host immune responses, then such high adaptation rates help explain some unique aspects of HIV life history, such as persistently high levels of viral replication (Ho et al. 1995) in the face of strong adaptive immune responses. I have also presented evidence that adaptation in the C2–V5 region of *env* is related to the time to disease progression, suggesting that those patients that mount a stronger immune response can control the infection for a longer time and that adaptation in the *env* gene could have a causal role in disease progression.

One of the major unsolved problems in the study of HIV and related viruses is determining the relative importance of neutralizing antibodies, helper T lympho-

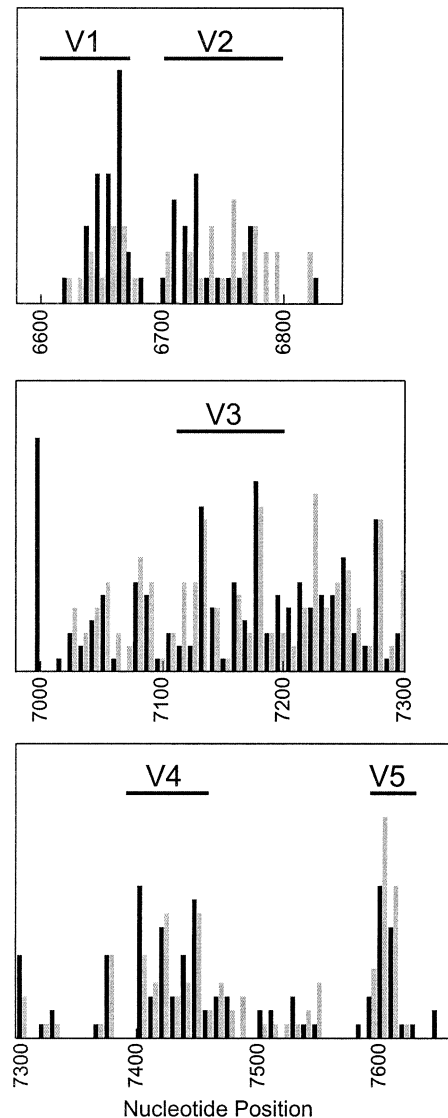


FIG. 2.—The spatial distribution of nonsynonymous substitutions (black bars) and common polymorphisms (gray bars) in the gp120 region of *env*. Physical location is plotted relative to a reference genome, and changes are binned into 10-bp regions for plotting.

cytes, and cytotoxic T lymphocytes in controlling viral replication. The methods presented here, in combination with detailed epitope analyses of longitudinally sampled patients, could help elucidate the roles played by each of these three types of immune response. This would provide new insight into the mechanisms by which HIV populations overwhelm the immune system and lead to AIDS.

Appendix

I use a method of moments estimator to find the number of adaptive substitutions (a_d), adaptive common polymorphisms (a_p), and adaptive events (a_i). In the case of adaptive substitutions, the evolutionary model assumes that: (1) both adaptive and neutral evolution contribute to divergence, (2) adaptive substitutions occur so rapidly that they are rarely observed as rare polymorphisms, (3)

deleterious changes are not observed as polymorphisms, and (4) synonymous changes are neutral. Under these assumptions, the expected relationship between nonsynonymous and synonymous divergence and rare polymorphism is $(D_n - a_d)/R_n = D_s/R_s$. Solving for the number of adaptations, $a_d = D_n - (D_s R_n / R_s)$.

To explore the bias inherent to this method, I evaluated the expectation of a_d , a_p , and a_i using a propagation of error analysis (Rice 1987). Let $\mu_{dn} = E(D_n)$, $\mu_{ds} = E(D_s)$, etc. The expectation of a_d is

$$E(a_d) = E\left(D_n - \frac{D_s R_n}{R_s}\right) = \mu_{dn} - E\left(\frac{D_s R_n}{R_s}\right)$$

By a Taylor expansion, the far right-hand term is approximately

$$E\left(\frac{D_s R_n}{R_s}\right) \approx \frac{E(D_s R_n)}{\mu_{rs}} + \frac{Var(R_s)E(D_s R_n)}{\mu_{rs}^3} - \frac{Cov(R_s, D_s R_n)}{\mu_{rs}^2}$$

where

$$E(D_s R_n) = \mu_{ds} \mu_{rn} + Cov(D_s, R_n)$$

and

$$Cov(R_s, D_s R_n) = E[(R_s - \mu_{rs})(D_s - \mu_{ds})(R_n - \mu_{rn})] \\ + \mu_{ds} Cov(R_s, R_n) + \mu_{rn} Cov(R_s, D_s)$$

The Poisson random field (PRF) model (Sawyer and Hartl 1992; Bustamante et al. 2001) is the only evolutionary model for which the covariances and variances in the above expressions are known. The PRF model assumes free recombination. Therefore, sites evolve independently, and all of the above covariances are equal to 0. Further, $Var(R_s) = \mu_{rs}$ because R_s is Poisson distributed. So, by substitution, the expression for $E(a_d)$ simplifies substantially:

$$E(a_d) = \mu_{dn} - \left(\frac{\mu_{ds} \mu_{rn}}{\mu_{rs}} + \frac{\mu_{ds} \mu_{rn}}{\mu_{rs}^2}\right) \\ = \mu_{dn} - \frac{\mu_{ds} \mu_{rn}}{\mu_{rs}} \left(1 + \frac{1}{\mu_{rs}}\right).$$

By setting observed values to their expectations, we arrive at the bias-corrected estimator of a_d found in the text. The same method can be used to arrive at the corrected estimators for a_p and a_i .

Violations of the assumptions of the evolutionary model should make the estimates more conservative. I assumed that all nonsynonymous, rare polymorphisms are neutral. If in fact some of these polymorphisms are either beneficial or slightly deleterious, then R_n would overestimate the number of neutral, nonsynonymous, rare polymorphisms, and a_d , a_p , and a_i would be underestimates. Furthermore, I did not correct for saturation, so D_n and C_n will underestimate the actual numbers of nonsynonymous substitutions and common polymorphisms, respectively. This would also cause a_d , a_p , and a_i to be downwardly-biased. Considering these two factors, a_d , a_p , and a_i are probably conservative estimators.

The above analyses also assume free recombination.

A violation of this assumption should not strongly bias the estimates, but if it does, it may make the estimates more conservative. First, note that the estimates will not be strongly affected by genetic hitchhiking—that is, the fixation of neutral variants tightly linked to and in association with beneficial mutations (Maynard Smith and Haigh 1974). This is because positive selection does not affect the expected substitution rate at linked neutral loci; it only affects the variance (Kelly 1994; Gillespie 2000). Hitchhiking will reduce the expected levels of polymorphism at linked neutral sites, but the effect should be proportional for nonsynonymous and synonymous sites. Thus, the effects on R_n and R_s should roughly cancel out. In summary, hitchhiking does not affect $E(D_n)$, $E(D_s)$, or the ratio $E(R_n)/E(R_s)$, so it should not affect the adaptation rate estimates.

Furthermore, considering the dynamics of polymorphism and divergence in the case of limited recombination, the covariances in the above propagation of error analysis may cause a_d (and a_p and a_i) to be even more conservative. In the expression for $E((D_s R_n)/R_s)$, the most prominent covariance is $Cov(D_s, R_n)$ because it is only divided by μ_{rs} , rather than μ_{rs}^2 or μ_{rs}^3 . This covariance should be negative, which would cause the free-recombination estimate of a_d to be more conservative. To illustrate why this covariance should be negative, consider two independent realizations of the evolutionary process, starting with initial infection. In one realization, due to chance, the time back to the most recent common ancestor (MRCA) of all the sequences in the population is short. Therefore, levels of polymorphism will be relatively low. However, the time between initial infection and the MRCA will be long, so the number of substitutions will be relatively high. In the other realization, also due to chance, the time back to the MRCA might be longer, which would lead to higher levels of polymorphism but a shorter time between initial infection and the MRCA and correspondingly lower numbers of substitutions. Therefore, covariances between polymorphism (i.e., R_n) and divergence (i.e., D_s) statistics should be negative.

Acknowledgments

This manuscript was greatly improved by many helpful comments from J. Kelly, M. Orive, M. Smith, and A. Eyre-Walker. I would also like to thank R. Shankarappa, S. Ray, and R. Viscidi for help in obtaining sequence information. This work was supported by an NIH grant (1 R01 GM 60792-01A1) to J. Kelly and M. Orive.

Literature Cited

- Bonhoeffer, S., E. C. Holmes, and M. A. Nowak. 1995. Causes of HIV diversity. *Nature* **376**:125.
- Bustamante, C. D., J. Wakeley, S. A. Sawyer, and D. L. Hartl. 2001. Directional selection and the site-frequency spectrum. *Genetics* **159**:1779–1788.
- Clark, S. J., M. S. Saag, W. D. Decker, S. Cambell-Hill, J. L. Roberson, P. J. Veldkamp, J. C. Kappes, B. H. Hahn, and G. M. Shaw. 1991. High titers of cytopathic virus in plasma of patients with symptomatic primary HIV-1 infection. *N. Engl. J. Med.* **324**:954–960.

- Fenoglio, D., G. Li Para, L. Lozzi et al. (11 co-authors). 2000. Natural analogue peptides of an HIV-1 gp120 T-helper epitope antagonize response of gp120-specific human CD4 T-cell clones. *J. Acquir. Immune Defic. Syndr.* **23**:1–7.
- Fu, Y.-X. 2001. Estimating mutation rate and generation time from longitudinal samples of DNA sequences. *Mol. Biol. Evol.* **18**:620–626.
- Ganeshan, S., R. E. Dickover, B. T. M. Korber, Y. J. Bryson, and S. M. Wolinsky. 1997. Human immunodeficiency virus type 1 genetic evolution in children with different rates of development of disease. *J. Virol.* **71**:663–677.
- Gillespie, J. H. 2000. Genetic drift in an infinite population: the pseudohitchhiking model. *Genetics* **155**:909–919.
- Goudsmit, J., C. Debrouck, R. H. Meloen, L. Smit, M. Bakker, D. M. Asher, A. Wolff, C. J. Gibbs, and D. C. Gajdusek. 1988. Human immunodeficiency virus type 1 neutralization epitope with conserved architecture elicits early type-specific antibodies in experimentally infected chimpanzees. *Proc. Natl. Acad. Sci. USA* **85**:4478–4482.
- Hey, J., and J. Wakeley. 1997. A coalescent estimator of the population recombination rate. *Genetics* **145**:833–846.
- Ho, D. D., A. U. Neumann, A. S. Perelson, W. Chen, J. M. Leonard, and M. Markowitz. 1995. Rapid turnover of plasma virions and CD4 lymphocytes in HIV-1 infection. *Nature* **373**:123–126.
- Holmes, E. C., L. Q. Zhang, P. Simmonds, C. A. Ludlam, and A. J. Brown. 1992. Convergent and divergent sequence evolution in the surface envelope glycoprotein of human immunodeficiency virus type 1 within a single infected patient. *Proc. Natl. Acad. Sci. USA* **89**:4835–4839.
- Kelly, J. K. 1994. An application of population genetic theory to synonymous gene sequence evolution in the human immunodeficiency virus (HIV). *Genet. Res.* **64**:1–9.
- Kelly, J. K., S. Williamson, M. E. Orive, M. S. Smith, and R. D. Holt. 2003. Linking ecological and genetic models of intra-host viral dynamics. I. Infection of multiple cell types. *Am. Nat.* (in press).
- Markham, R. B., W.-C. Wang, A. E. Weisstein et al. (11 co-authors). 1998. Patterns of HIV-1 evolution in individuals with differing rates of CD4 T cell decline. *Proc. Natl. Acad. Sci. USA* **95**:12568–12573.
- Maynard Smith, J., and J. Haigh. 1974. The hitch-hiking effect of a favourable gene. *Genet. Res.* **23**:23–35.
- McDonald, J. H., and M. Kreitman. 1991. Adaptive protein evolution at the Adh locus in *Drosophila*. *Nature* **351**:652–654.
- Modrow, S., B. H. Hahn, G. M. Shaw, R. C. Gallo, F. Wong-Staal, and H. Wolf. 1987. Computer-assisted analysis of envelope protein sequences of seven human immunodeficiency virus isolates: prediction of antigenic epitopes in conserved and variable regions. *J. Virol.* **61**:570–578.
- Müller, V., J. F. Viguera-Gómez, and S. Bonhoeffer. 2002. Decelerating decay of latently infected cells during prolonged therapy for human immunodeficiency virus type 1 infection. *J. Virol.* **76**:963–965.
- Nielsen, R. 1999. Changes in d_s/d_n in the HIV-1 env gene. *Mol. Biol. Evol.* **16**:711–714.
- Nielsen, R., and Z. Yang. 1998. Likelihood models for detecting positively selected amino acid sites and applications to the HIV-1 envelope gene. *Genetics* **148**:929–936.
- Nowak, M. A., R. M. Anderson, A. R. McLean, T. F. Wolfs, J. Goudsmit, and R. M. May. 1991. Antigenic diversity thresholds and the development of AIDS. *Science* **254**:963–969.
- Pantaleo, G., C. Graziosi, and A. S. Fauci. 1993. New concepts in the immunopathogenesis of human immunodeficiency virus infection. *N. Engl. J. Med.* **328**:327–335.
- Perelson, A. S., A. U. Neumann, M. Markowitz, J. M. Leonard, and D. D. Ho. 1996. HIV-1 dynamics in vivo: virion clearance rate, infected cell life-span, and viral generation time. *Science* **271**:1582–1586.
- Pierson, T., J. McArthur, and R. F. Siliciano. 2000. Reservoirs for HIV-1: mechanisms for viral persistence in the presence of antiviral immune responses and antiretroviral therapy. *Annu. Rev. Immunol.* **18**:665–708.
- Rice, J. A. 1987. *Mathematical statistics and data analysis*. Wadsworth, Pacific Grove, Calif.
- Rodrigo, A. G., E. G. Shaper, E. L. Delwart, A. K. Iversen, M. V. Gallo, J. Brojatsch, M. S. Hirsch, B. D. Walker, and J. I. Mullins. 1999. Coalescent estimates of HIV-1 generation time in vivo. *Proc. Natl. Acad. Sci. USA* **96**:2187–2191.
- Ross, H. A., and A. G. Rodrigo. 2002. Immune-mediated positive selection drives human immunodeficiency virus type 1 molecular variation and predicts disease duration. *J. Virol.* **76**:11715–11720.
- Sawyer, S. A., and D. L. Hartl. 1992. Population genetics of polymorphism and divergence. *Genetics* **132**:1161–1176.
- Seo, T.-K., J. L. Thorne, M. Hasegawa, and H. Kishino. 2002. Estimation of effective population size of HIV-1 within a host: a pseudomaximum-likelihood approach. *Genetics* **160**:1283–1293.
- Shankarappa, R., J. B. Margolick, S. J. Gange et al. (12 co-authors). 1999. Consistent viral evolutionary changes associated with the progression of human immunodeficiency virus type 1 infection. *J. Virol.* **73**:10489–10502.
- Smith, N. G. C., and A. Eyre-Walker. 2002. Adaptive protein evolution in *Drosophila*. *Nature* **415**:1022–1024.
- Strunnikova, N., S. C. Ray, C. Lancioni, M. Nguyen, and R. P. Viscidi. 1998. Evolution of human immunodeficiency virus type 1 in relation to disease progression in children. *J. Hum. Virol.* **1**:224–239.
- Strunnikova, N., S. C. Ray, R. A. Livingston, E. Rubalcaba, and R. P. Viscidi. 1995. Convergent evolution within the V3 loop domain of human immunodeficiency virus type 1 in association with disease progression. *J. Virol.* **69**:7548–7558.
- Tersmette, M., R. A. Gruters, F. de Wolf, R. E. de Goede, J. M. Lange, P. T. Schellekens, J. Goudsmit, H. G. Huisman, and F. Miedema. 1989. Evidence for a role of virulent human immunodeficiency virus (HIV) variants in the pathogenesis of acquired immunodeficiency syndrome: studies on sequential HIV isolates. *J. Virol.* **63**:2118–2125.
- Thompson, J. D., T. J. Gibson, F. Plewniak, F. Jeanmougin, and D. G. Higgins. 1997. The ClustalX windows interface: flexible strategies for multiple sequence alignment aided by quality analysis tools. *Nucleic Acids Res.* **24**:4876–4882.
- Tsubota, H., C. I. Lord, D. I. Watkins, C. Morimoto, and N. L. Letvin. 1989. A cytotoxic T lymphocyte inhibits acquired immunodeficiency syndrome virus replication in peripheral blood lymphocytes. *J. Exp. Med.* **169**:1421–1434.
- Viscidi, R. P. 1999. HIV evolution and disease progression via longitudinal studies. Pp. 346–389 in K. A. Crandall, ed. *The evolution of HIV*. John Hopkins, Baltimore.
- Walker, C. M., D. J. Moody, D. P. Stites, and J. A. Levy. 1986. CD8+ lymphocytes can control HIV infection in vitro by suppressing virus replication. *Science* **234**:1563–1566.
- Wodarz, D., P. Klenerman, and M. A. Nowak. 1998. Dynamics of cytotoxic T-lymphocyte exhaustion. *Proc. R. Soc. Lond. B Biol. Sci.* **265**:191–203.
- Wolfs, T. F., J. J. de Jong, H. Van den Berg, J. M. Tijnagel, W. J. Krone, and J. Goudsmit. 1990. Evolution of sequences encoding the principal neutralization epitope of human immunodeficiency virus 1 is host-dependent, rapid, and continuous. *Proc. Natl. Acad. Sci. USA* **87**:9938–9942.
- Wolinsky, S. M., B. T. M. Korber, A. U. Neumann et al. (11 co-authors). 1996. Adaptive evolution of human immunodeficiency virus type 1. *J. Virol.* **70**:1048–1058.

- ciency virus-type 1 during the natural course of infection. *Science* **272**:537–542.
- Wolinsky, S. M., and G. H. Learn. 1999. Levels of diversity within and among host individuals. Pp. 275–314 *in* K. A. Crandall, ed. *The evolution of HIV*. John Hopkins, Baltimore.
- Yamaguchi-Kabata, Y., and T. Gojobori 2000. Reevaluation of amino acid variability of the human immunodeficiency virus type 1 gp120 envelope glycoprotein and prediction of new discontinuous epitopes. *J. Virol.* **74**:4335–4350.
- Zanotto, P. M. A., E. G. Kallas, R. F. de Souza, and E. C. Holmes. 1999. Genealogical evidence for positive selection in the *nef* gene of HIV-1. *Genetics* **153**:1077–1089.

Edward Holmes, Associate Editor

Accepted April 7, 2003