Adaptation to Different Human Populations by HIV-1 Revealed by Codon-Based Analyses

Sergei L. Kosakovsky Pond[1], Simon D. W. Frost[1], Zehava Grossman[2], Michael B. Gravenor[3], Douglas D. Richman[1,4], Andrew J. Leigh Brown[5*]

1 Department of Pathology, University of California San Diego, La Jolla, California, United States of America, **2** National HIV Reference Lab, Public Health Laboratory, Ministry of Health, Tel Hashomer, Israel, **3** School of Medicine, University of Swansea, Swansea, Wales, United Kingdom, **4** VA San Diego Health Care System, San Diego, California, United States of America, **5** Institute for Evolutionary Biology, School of Biological Sciences, University of Edinburgh, Edinburgh, Scotland, United Kingdom

Several codon-based methods are available for detecting adaptive evolution in protein-coding sequences, but to date none specifically identify sites that are selected differentially in two populations, although such comparisons between populations have been historically useful in identifying the action of natural selection. We have developed two fixed effects maximum likelihood methods: one for identifying codon positions showing selection patterns that persist in a population and another for detecting whether selection is operating differentially on individual codons of a gene sampled from two different populations. Applying these methods to two HIV populations infecting genetically distinct human hosts, we have found that few of the positively selected amino acid sites persist in the population; the other changes are detected only at the tips of the phylogenetic tree and appear deleterious in the long term. Additionally, we have identified seven amino acid sites in protease and reverse transcriptase that are selected differentially in the two samples, demonstrating specific population-level adaptation of HIV to human populations.

Citation: Kosakovsky Pond SL, Frost SDW, Grossman Z, Gravenor MB, Richman DD, et al. (2006) Adaptation to different human populations by HIV-1 revealed by codon-based analyses. PLoS Comput Biol 2(6): e62. DOI: 10.1371/journal.pcbi.0020062

Introduction

Differences in allele frequencies in different populations were used as evidence of natural selection in some classic studies [1–3]. Since the first identification of positive selection within protein sequences [4,5], however, estimation of the relative frequency of synonymous and non-synonymous nucleotide substitution has become a standard tool in molecular evolutionary studies. The power of these analyses to detect selection was substantially increased through the development of codon-based likelihood models that allow selection to vary across sites [6]. We have developed a method, within a maximum likelihood framework, that combines these two approaches to yield novel insights into adaptive evolutionary differences between populations. We have applied this method to investigate the hypothesis that HIV has adapted specifically to the distinct human host population.

The immune system is widely recognized as one of the factors that exert a selective effect on pathogen populations. Mutations that allow HIV-1 strains to escape MHC-restricted CTL killing arise in both acute and chronic infection [7–11], but in the absence of an immune response they can reduce fitness [12], and can be selected against on transmission [13,14]. Recently, from correlations of variability and CTL epitopes, it was suggested that adaptation to human CTL responses had led to genetic adaptations in the HIV genome [15]. Both individual- and population-based studies have found overwhelming evidence of positive selection in HIV [6,16–22] but have been unable to discriminate between possible mechanisms. Generally speaking, phylogenetic studies that rely on within-population sequence polymorphism to identify non-neutral evolution may not be able to detect selective sweeps, expressed by substitutions localized to a single branch of the phylogeny of serially sampled sequences; selected alleles driven to fixation in all sequences of the sample, resulting in within-population monomorphic sites; or the direction of selection, for instance acquisition of immune escape mutations versus reversion in the absence of selective forces [23].

We have developed a suite of fixed effects likelihood-based approaches [22,24] which we have used here to discriminate substitutions occurring on internal branches from those occurring at the tips of the tree, and to estimate whether selective pressure at a given site is different between two populations of HIV-infected individuals. This allows one to distinguish sites that are positively selected because they are associated with adaptation to the individual host from those associated with adaptation to the population.

Results

We analyzed the sequences of the protease (PR) and reverse transcriptase (RT) coding regions of HIV clade C from each of 74 individuals from KwaZulu Natal [25] and Zambia (ZA sample), and from 63 Ethiopian Falasha immigrants arriving in Israel between 1998 and 2003 [26] (ET sample). These sequences were obtained by population-based sequencing

Editor: Eddie Holmes, The Pennsylvania State University, United States of America

Received November 17, 2005; **Accepted** April 21, 2006; **Published** June 23, 2006

A previous version of this article appeared as an Early Online Release on April 21, 2006 (DOI: 10.1371/journal.pcbi.0020062.eor).

DOI: 10.1371/journal.pcbi.0020062

Abbreviations: IFEL, internal fixed effects likelihood; MHC, major histocompatability complex; MLE, maximum likelihood (parameter) estimate; PPV, positive predictive value; PR, protease transcriptase; RT, reverse transcriptase

* To whom correspondence should be addressed. E-mail: A.Leigh-Brown@ed.ac.uk

Synopsis

Despite the efforts devoted to surveying HIV genetic diversity and the development of an effective vaccine, there is still no consensus on the extent to which the former prejudices the latter. Experimental studies show that escape from cell-mediated immunity is selected for in HIV and SIV, and sometimes very quickly. Conversely, escape mutants may be selected against at transmission, so how much does this selection within individuals influence the genotype of the circulating HIV population overall? Kosakovsky Pond, Leigh Brown, and colleagues have developed a new statistical approach to address this question. Using sequences from the globally most abundant HIV subtype (subtype C), the authors directly compared virus of the same subtype infecting genetically different human populations. They show at least half of the amino acid sites selected within individuals are not selected at a population level, and they identify six amino acid sites in the RT gene that are selected differentially between populations. We can now partition molecular adaptation between individual and population components for whatever genes may be included in candidate vaccines, in the target populations themselves. The methods are also important beyond the HIV world, where analogous issues arise in the more general question of the evolution of virulence in pathogens.

from viral RNA PCR-amplified direct from the blood plasma of the infected individual: each sequence thus reflects the predominant nucleotides in the plasma viral population of that individual at that time. Within-host recombination will not affect a phylogeny based on these sequences. Our maximum likelihood calculations treated ambiguities as partially missing data. Amino acid positions 4 through 99 of PR and 41 through 240 of RT were analyzed. The absence of first 40 positions in RT and three positions in PR is an artifact of the sequencing procedure used for generating some sequences in the ET sample. The alignments used in this study did not contain any insertions or deletions, which is common for HIV *pol* alignments. We screened all sequences for mutations known to confer strong drug resistance and found that none of the sequences included in the study had such mutations [27–29]. Each sample formed a separate clade (Figure S1) in the maximum likelihood tree built from PR and RT jointly (reconstructed using the GTR+G+I model with the PhyML package [30]). This clustering was confirmed by Bayesian phylogenetic reconstruction, performed under the same model using the MrBayes software package [31]. 10^6 MCMC samples, thinned by a factor of 100, were generated, and the first 25% of the samples were discarded as burn-in. Sequence alignments and maximum likelihood trees for each sample can be downloaded (in NEXUS format) from http://www.hyphy.org/pubs/DS/sequences.tgz.

Phylogenetic methods can perform poorly if viral sequences have experienced sufficient recombination to result in discordant phylogenetic signal in different parts of the alignment [32,33]. We carried out a simple procedure, similar to the ideas in [34] and [35], to screen for evidence of recombination. Given an alignment, we split the data into two contiguous fragments, reconstructed neighbor joining trees [36] for each segment, fitted the trees using maximum likelihood, and investigated whether having two trees improves the small sample AIC score [37] over the model with one tree for the entire alignment. Additionally, we verified phylogenetic incongruence using the Shimodaira–Hasegawa [38] test. This test was carried out for all possible placements of the breakpoint. For the four alignments in this study, no two-tree model fitted the data significantly better than the single-tree model and yielded a significant SH test result. While this screening procedure does not rule out the presence of recombinant sequences in our alignments, it suggests that the impact of recombination is insufficient to cause detectably discordant phylogenies, and to heavily bias the inference procedure. As an additional check, we carried out the RDP test [39], which also failed to detect any recombination events in any of the four alignments.

Gene by gene maximum likelihood codon rate analyses revealed strong rate heterogeneity of both synonymous and nonsynonymous substitution rates in all four samples (Table 1). Consequently, it is imperative that site-to-site variation in synonymous rates be accounted for, lest the tests for selection suffer high rates of false positives [22,40]. Using three independent maximum likelihood methods for detecting selection at an individual site [22], we identified four sites in PR and seven sites in RT subject to significant positive selection, as detected by the consensus of the methods. Codons 12, 19, and 63 in PR were positively selected in both populations, as were codons 48, 166, 173, and 207 in RT. Other sites gave significant results using only one test, or were positively selected in only one population (Table 2).

Individual versus Population-Level Adaption in HIV

Sharp et al. [41] previously estimated *dN/dS* ratios for all branches in a phylogeny of HIV-1 and SIVcpz *env* sequences and correlated that quantity with the "depth" of the branch in the tree to deduce that *dN/dS* across the entire sequence was smaller for branches that were far from the tips of the

Table 1. Sequences Analyzed in This Study

Sample	Gene	Sequences	Codons	Diversity, %	*CV*(α)	Mean β	*CV*(β)
Ethiopian	PR	63	96	7.8	0.87 ($p < 0.001$)	0.22	3.0
Ethiopian	RT	63	200	7.2	0.72 ($p < 0.001$)	0.20	2.2
South African	PR	74	96	6.5	0.79 ($p < 0.001$)	0.23	2.17
South African	RT	74	200	6.8	0.58 ($p < 0.001$)	0.18	3.1

Diversity was computed by averaging pairwise phylogenetic distances between, based on branch lengths fitted using the Dual MG94×(012232) GDD 3 × 3 codon substitution model [40]. The same model was used to infer the mean and coefficients of variation (*CV*) for the distributions of synonymous and (α) and nonsynonymous substitution (β) rates. The *p*-values for the likelihood ratio test of *CV* (α) > 0 were computed using the likelihood ratio test as described in [40].

DOI: 10.1371/journal.pcbi.0020062.t001

Table 2. Sites found to be under positive and/or differential selection.

Codon	Ethiopian Composition	SLAC	FEL	IFEL	REL	South African Composition	SLAC	FEL	IFEL	REL	Different
PR12S	$T_{43}S_{20}$	**0.04**	0.07	**0.02**	**21**	$S_{57}T_{14}P_2A_1$	**0.004**	**0.02**	**0.03**	**360**	0.55
PR15V	$V_{50}I_{13}$	**0.01**	**0.02**	0.17	**>1000**	$V_{67}I_7$	0.29	0.55	0.73	0.05	0.32
PR19I	$L_{36}I_{18}T_4V_4A_1$	**0.02**	0.20	0.10	**>1000**	$I_{49}V_{15}L_4T_4A_1E_1$	**0.04**	**0.03**	0.14	**>1000**	0.40
*PR60*D	$D_{62}E_1$	0.93	0.20	0.16	0.5	$D_{66}E_8$	0.14	0.16	0.11	2.5	**0.03**
PR63L	$L_{34}P_{13}T_6S_3V_2H_2A_1C_1I_1$	**0.005**	0.12	0.11	**>1000**	$L_{43}P_{17}T_5V_4A_2S_2H_1$	**0.005**	**0.05**	**0.04**	**>1000**	1
RT48T	$S_{34}T_{28}Q_1$	**0.001**	**0.001**	**0.008**	**>1000**	$T_{64}S_6P_3E_1$	**0.01**	**0.007**	0.99	**42.5**	0.64
RT82K	$K_{61}R_2$	0.76	0.46	0.16	<0.01	$K_{73}R_1$	0.99	**.04**	0.12	<0.01	**0.04**
RT98A	$A_{44}S_{19}$	0.10	0.22	0.11	**278**	$A_{72}G_1S_1$	0.95	0.08	0.11	0.02	**0.02**
RT123G	$S_{35}D_{16}G_{11}N_1$	0.74	0.75	0.15	0.11	$D_{26}G_{25}S_{14}N_6A_1$	**0.01**	**0.001**	**0.012**	**>1000**	0.26
RT135I	$I_{46}T_{11}R_3V_2M_1$	0.45	0.63	0.23	**131**	$I_{52}V_{11}R_7T_3M_1$	0.24	0.08	**0.05**	**>1000**	0.27
RT165T	$T_{61}I_2$	0.89	0.25	0.18	<0.01	$T_{68}I_4A_1P_1$	0.14	0.09	0.13	**41**	**0.04**
RT166K	$K_{55}R_8$	0.09	**0.04**	1	**178**	$K_{52}R_{22}$	**0.03**	**0.03**	0.06	**480**	0.96
RT173A	$A_{52}T_7I_3V_1$	**0.02**	0.10	0.36	**>1000**	$A_{49}T_{19}K_3G_1I_1V_1$	**0.006**	**0.006**	**0.01**	**40**	1
RT174Q	$Q_{43}K_{18}R_2$	0.14	0.14	0.55	**347**	$Q_{47}K_{20}N_3E_1H_1R_1T_1$	**0.04**	**0.04**	**0.01**	**>1000**	0.53
RT177E	$E_{59}D_3Q_1$	0.63	0.96	0.30	0.11	$E_{64}D_6G_3N_1$	0.11	0.053	**0.04**	**24**	**0.04**
RT196G	$G_{45}K_{10}E_7R_1$	**0.01**	**0.001**	**0.05**	**>1000**	$G_{67}E_5R_1V_1$	0.93	0.31	0.51	0.03	**0.01**
RT202I	$I_{53}V_{10}$	0.52	0.30	0.10	0.17	$I_{71}V_3$	0.98	0.20	0.10	0.9	**0.02**
RT207E	$E_{52}G_3N_3A_2D_2Q_1$	**0.02**	**0.02**	0.34	**>1000**	$E_{44}G_7A_6D_5K_3N_2R_2S_2Q_1V_1$	**0.01**	0.06	0.29	**>1000**	0.45

Subtype C consensus (based of the Los Alamos HIV database) amino acid is listed for each reported position. Amino acid composition at a site was derived after resolving ambiguous codons to the appropriate most frequent character at that site.

SLAC, FEL, IFEL and Differential test significance levels are given as *p*-values, while empirical Bayes factors are given for REL.

Bold values are those deemed significant ($p \leq 0.05$, Bayes factors ≥ 20).

Underlined values are codon sites found to be evolving differentially between the samples.

DOI: 10.1371/journal.pcbi.0020062.t002

tree. Holmes [42] performed alignment-wide comparisons of dengue virus sequences sampled from individual hosts and from populations of infected hosts and found that average selective pressures were substantially more purifying in between-host samples. We have extended one of the methods used to detect positive selection (fixed effects likelihood) to permit the estimation of the ratio of nonsynonymous (β or dN) and synonymous substitution rates (α or dS) separately in internal and terminal branches of the tree connecting these sequences. This has revealed that many recent nonsynonymous substitutions, i.e., those in the terminal branches of the tree, were not represented on internal branches. For both the ZA and ET populations, there are more codons in both PR and RT with only recent nonsynonymous substitutions than there are codons with substitutions on internal branches (Figures 1 and 2). The difference was particularly striking in RT where the ratio was 35:17 in the ZA sample and 54:16 in the ET sample. This disparity alone may be statistically insignificant because the cumulative length of internal branches in the tree is smaller than that of terminal branches (Figure S1). However, at those codons where internal substitutions are seen, the strength of selection (measured by the dN/dS ratio) along terminal branches is in all cases higher (Figures 1 and 2): in all four comparisons this difference was significant based on the likelihood ratio test ($p < 0.05$ using parametric bootstrap to guard against the effect of small sample sizes).

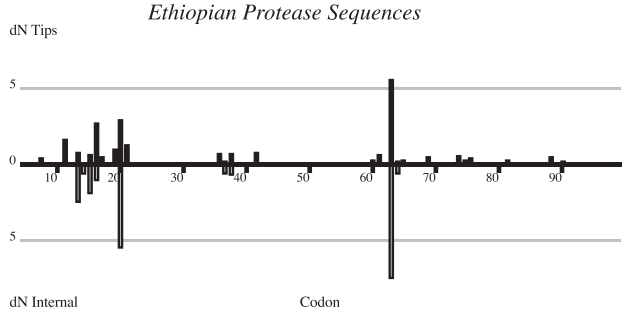
At the level of individual sites, three sites were positively selected ($p \leq 0.05$) along internal branches in the ET sample and seven in the ZA sample. Simulation results (see Materials and Methods) suggest that the test is conservative for model parameters chosen to resemble those likely to have generated our samples, and is capable of reliably detecting sites that are subject to strong selection. Positive predictive value (PPV) of the test was calculated at 98.8%, hence it is unlikely that detected sites are false positives. In particular, a high PPV estimate strongly suggests that site-wise testing procedures in this context do not require a correction for multiple testing.

Positively selected nonsynonymous substitutions on internal branches (persistent substitutions) must of necessity be adaptive at both individual host and population level. As we have analysed consensus sequences of the within-individual populations, the substitutions must have reached a high frequency in the infected individuals, but are transient at the population level, suggesting their removal by purifying selection. Based on the elevated rate of adaptation within individuals detected at codons subject to population-level selection, relative to the codons where only recent substitutions have been inferred, we conclude that recent substitutions are, on average, maladaptive at the level of the human population. We note that when longitudinal data are not available, comparative phylogenetic methods may be unable to detect directional selection if the population had undergone a selective sweep. Population level adaptation inferred for our samples could also be due to transient directional selection, or diversifying selection maintained by acquisition and transmission of escape mutants and reversion to wild type. However, because time scales of transmission and reversion processes are not known for this sample, a single mechanism cannot be distinguished.

Differential Adaptive Evolution in Different Populations

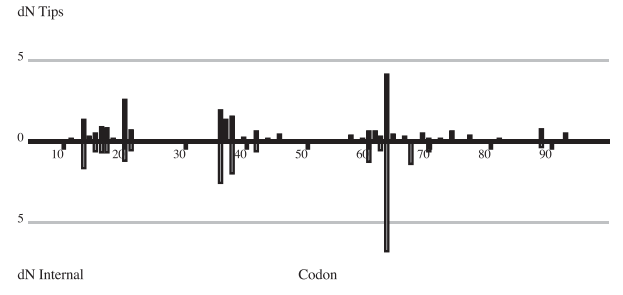
Human major histocompatibility complex (MHC) alleles are remarkably old, and some have been maintained since the human–chimpanzee divergence [43]. Thus, adaptation to human MHC alleles may not only reflect adaptation since the zoonotic transfer of HIV from chimpanzee to humans, but also include the prior history in the chimpanzee

Ethiopian Protease Sequences



Site Class	Codon Count	dN/dS along Terminal Branches	Is dN/dS elevated when dN internal > 0
Only dN Tips >0	17	0.49	LRT = 5.55
dN Internal >0	9	1.42	$p_A = 0.018$, **$p_B = 0.02$**

South African Protease Sequences



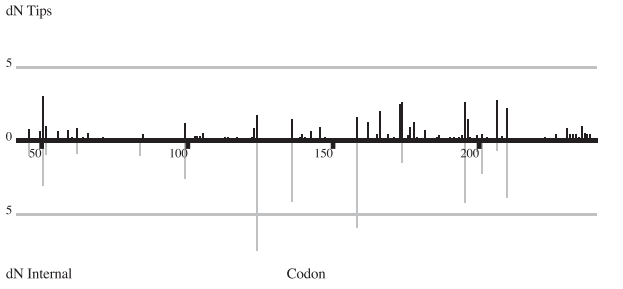
Site Class	Codon Count	dN/dS along Terminal Branches	Is dN/dS elevated when dN internal > 0
Only dN Tips >0	18	0.42	LRT = 10.46
dN Internal >0	15	1.19	$p_A = 0.001$, **$p_B = <0.01$**

Figure 1. Patterns of Population and Individual Level Nonsynonymous Evolution in the Two Protease Samples

MLEs of nonsynonymous substitution rates along internal (dN Internal) and terminal (dN Tips) branches were derived using the IFEL method. *dN/dS* estimates and *p*-values in the tables were obtained with the Population Level Adaptation test. p_A and p_B and denote, respectively, the asymptotic and parametric bootstrap significance levels for the likelihood ratio test.

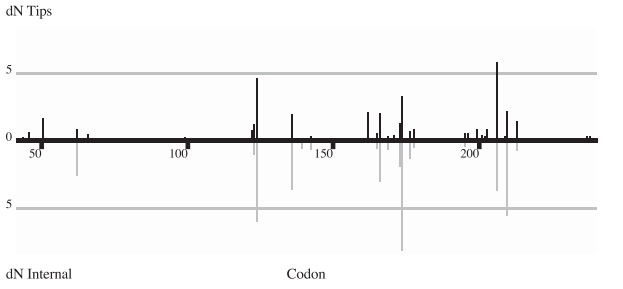
DOI: 10.1371/journal.pcbi.0020062.g001

Ethiopian Reverse Transcriptase Sequences



Site Class	Codon Count	dN/dS along Terminal Branches	Is dN/dS elevated when dN internal > 0?
Only dN Tips >0	54	0.36	LRT = 47.5
dN Internal >0	16	2.03	$p_A < 0.001$, **$p_B < 0.01$**

South African Reverse Transcriptase Sequences



Site Class	Codon Count	dN/dS along Terminal Branches	Is dN/dS elevated when dN internal > 0?
Only dN Tips >0	35	0.35	LRT = 43.6
dN Internal >0	17	1.55	$p_A < 0.001$, **$p_B < 0.01$**

Figure 2. Patterns of Population and Individual Level Nonsynonymous Evolution in the Two RT Samples

MLEs of nonsynonymous substitution rates along internal (*dN* Internal) and terminal (*dN* Tips) branches were derived using the IFEL method. *dN/dS* estimates and *p*-values in the tables were obtained with the Population Level Adaptation test. p_A and p_B denote, respectively, the asymptotic and parametric bootstrap significance levels for the likelihood ratio test.

DOI: 10.1371/journal.pcbi.0020062.g002

population. However, differential adaptation to different human populations could only be due to a species-specific process. Although many comparisons between HIV sequences from different host populations are confounded by a substantial phylogenetic difference between the viral populations, HIV-1 M group clade C has infected a number of ethnically distinct populations. We compared the sequence dataset from southern Africa with another subtype C dataset sampled from Ethiopian Falasha immigrants arriving in Israel between 1998 and 2003 [26]. An earlier study has shown that the Falasha (Amharic) share most genetic markers with other Ethiopian groups [44], and Ethiopian populations have quite distinct allele frequency spectra at HLA loci [45–47]. This comparison allows us to test the explicit hypothesis that passage through different human populations has led to adaptive divergence in the virus genome.

As transient substitutions would not contribute to inter-population adaptive divergence, only internal branches were tested for population-specific positive selection. A novel maximum likelihood test for differential selection (see Materials and Methods) permits the direct comparison of selection pressures on individual amino acid sites between populations. The test takes into account nucleotide substitution biases and weights over all possible ancestral codons, while avoiding assumptions regarding the distribution of dN and dS across sites. With this test we identified one codon in PR (60) and six codons in RT (82, 98, 165, 177, 196, and 202) as selected differentially in the two populations, at $p \leq 0.05$. Thus there is evidence for differential selection between

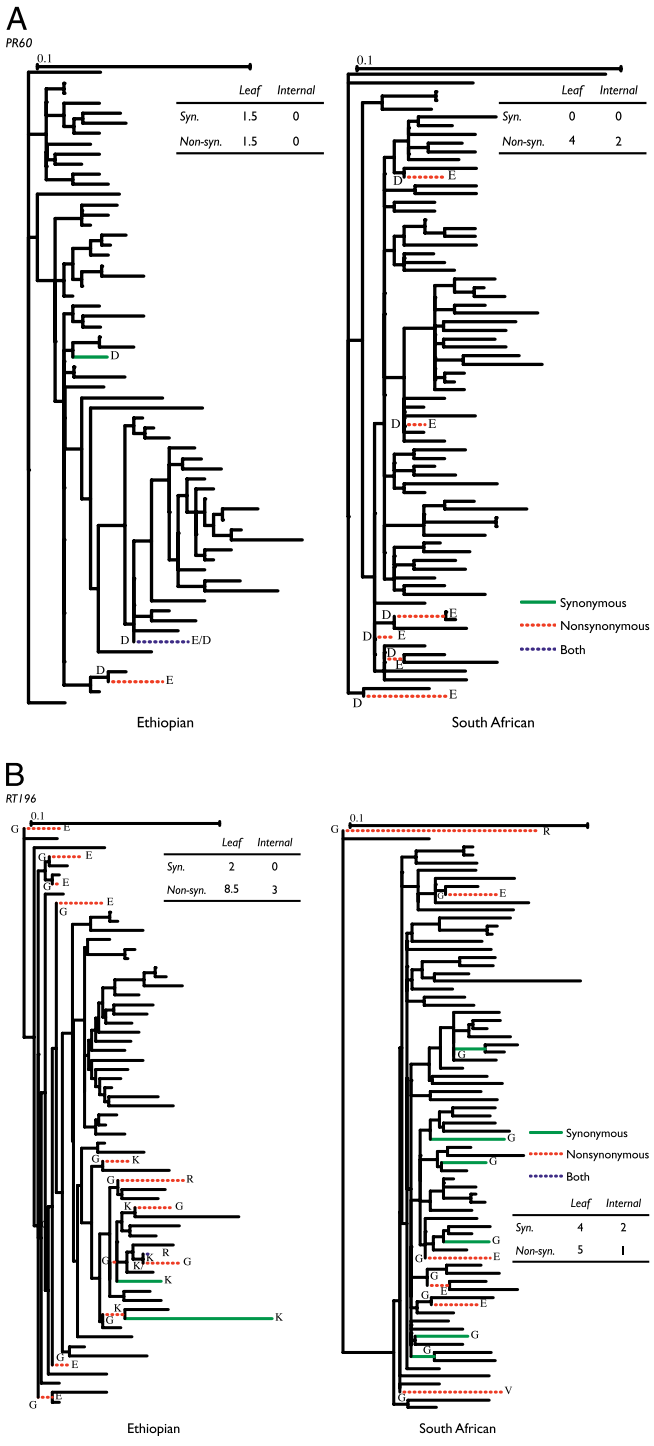


Figure 3. Inferred Evolutionary History at Two Codons in *pol*

(A) Codon 60 in protease. The trees were rooted using subtype B reference sequences from the Los Alamos HIV database, and ancestral states were inferred using maximum likelihood under the MG94×012232 Dual GDD 3 × 3 [40] model of codon evolution.

(B) Codon 196 in RT. The trees were rooted using subtype B reference sequences from the Los Alamos HIV database, and ancestral states were inferred using maximum likelihood under the MG94×012232 Dual GDD 3 × 3 [40] model of codon evolution.

DOI: 10.1371/journal.pcbi.0020062.g003

these two populations at seven codons out of 296 compared in RT and PR. Based on the high (98.2%) PPV of the test achieved on simulated data (see Materials and Methods) and the overall low power of the test for relatively small sample sizes, we conclude that these seven codons are unlikely to be false positive results, and that they probably constitute only a portion of codons which evolve differentially between the samples.

Figure 3A and 3B (and Figures S2–S6) show a codon-based maximum likelihood reconstruction of evolutionary history at these codon positions. We note that at many codons, evolution in both populations involves the same residues, but drastically different patterns of substitutions throughout the tree, with one population showing synonymous and non-synonymous evolution along terminal branches only, while the other displays nonsynonymous substitutions along internal branches as well as ongoing evolution at the tips (e.g., Figure 3B).

We also note that some sites (e.g., PR12 and RT48 in Table 2) show evidence of selection in both samples, but sequences appear to be driven towards different residues in each sample. Our method does not distinguish such sites as differentially selected, because they are subject to similar selective pressures in both samples, regardless of which residue appears to be selected for.

Discussion

The MHC-restricted host immune response represents a continuous selective force on pathogens whose effect is dependent on the pathogen genotype. In the case of HIV, viral escape mutations can arise soon after infection and can be transmitted onward, when their fate will depend on the MHC genotype of the new host [13]. In the absence of an active CTL response, due to MHC discordance, such escape mutations can be lost relatively quickly, implying that a second, antagonistic, selective force acts on the same genetic variant, possibly replication rate [48,49]. The extent to which the HIV and other viral genomes are shaped by the human immune response will therefore depend on the balance between these two effects. Only those mutations that either do not incur a significant cost in replicative efficiency, or have such a low probability of being recognized in the human population, would persist at the population level, and such population-level adaptation would be observed on internal branches of a phylogenetic tree of viral sequences.

Analyzing viral *pol* gene sequences from two populations infected with HIV subtype C, we have found many codons with amino acid substitutions only at the tips. At these sites variation is much lower than at those with both internal and tip substitutions, suggesting long-term purifying selection has removed many recent substitutions that may have arisen as adaptations to individual hosts. This suggests there are substantial long-term constraints on the extent to which the genome of HIV can be modified by human MHC-restricted immune responses. However, at seven codons there is evidence that substitutions on internal branches are selected differentially between the two human populations studied, confirming that these constraints are sequence context–specific. As the density of CTL epitopes in *pol* is low relative to that of other genes such as *gag*, *tat*, and *nef* [50,51], the level

of population adaptation in other genes could well be even higher.

Our approach looks for differences in evolutionary forces exerted deep in the phylogenetic trees, which are not always readily manifested in the amino-acid composition at a given site, or raw numbers of inferred synonymous or nonsynonymous substitutions. This approach can augment simpler but less sensitive methods [15,23], which rely on observed amino-acid composition of a site, or on detecting mutations toward (reversion) or away from (escape) a reference sequence (e.g., subtype consensus), thought to represent a variant with higher fitness in absence of selection.

Our methodology offers an alternative and more general approach to the "branch-site" class methods [52], which attempt to identify site-by-site positive selection along a single branch using a random effects approach and empirical Bayes inference. For example, Travers et al. [21] used such methods to locate sites under selection along predefined branches in a phylogeny of HIV-1 sequences from different clades. Our approaches are able to test selection operating along a set of tree branches, without assuming an a priori, perhaps unnecessarily restrictive, parametric form for all possible selection regimes. For instance, Bielawski and Yang [53] assumed that there are at most three modes of selection, with fixed selection strength at every site in a given mode. In contrast, by adopting a fixed effects phylogenetic likelihood framework [22] and inferring various selection regimes directly at every site, we can sidestep the problems inherent in model mis-specification in the context of branch-site models [54,55] and uncertainties associated with phylogenetic empirical Bayes inference in general.

In addition, we have developed a novel test to identify differential adaptation in different populations. This test is particularly suitable for exploring adaptation of parasites to genetically different host populations, and it allowed us to identify a subset of amino acid sites in PR and RT coding regions of HIV that were differentially selected in two human populations. Previous studies [56] have drawn upon observed correlations between location of sites subject to selection to hypothesize concordant or discordant selective pressures on gene regions among populations. While suggestive, correlational studies are unable to rigorously examine two populations for selective forces that differ at the level of an individual site, or a very short sequence region. Our test is capable of directly testing for such differences, including the case when we are only concerned with a subset of tree branches (e.g., internal or tips), and provides a rigorous significance level for such comparisons. Recent studies [40] provide strong evidence that site-to-site variation in synonymous substitution is pervasive in many genes, especially in HIV. Furthermore, it has been shown that failure to model such rate variation can result in uncontrollable rates of false positives and misidentify variable sites under relaxed selective constraints as those under strong positive selection pressure [22]. We have demonstrated that the new tests yield well-controlled false positive rates and high (>95%) PPV on data simulated with parameters realistic for HIV evolution. Additionally, the methods have been implemented as a part of parallelized software package HyPhy [57], can be run very quickly (≈10 min per 74 sequence sample) on a small computer cluster, and lend themselves to practical investigation of statistical properties of the method based on simulations, which can be tailored to the specific dataset being analyzed.

Adaptation to the host occurs at many levels in HIV: to the intracellular, intra-individual, and intrapopulation levels we have added an interpopulation level. Novel statistical methodology has allowed us to discriminate adaptation occurring at the last two levels and to answer questions raised by earlier correlation studies [15,50]. We have shown that within-host adaptation is often transient and that the codons at which persistent substitutions occur (which would include the immunological footprint) are subject to a substantially stronger ongoing selective force than those at which transient substitutions are seen. The ability to distinguish transient from persistent substitutions could be important for the development of an effective vaccine [58], as well as opening new routes to the analysis of selection in other settings.

Materials and Methods

Phylogeny reconstruction and substitution models. We used an iterative process [24] to reconstruct a phylogeny of the sample and to select an appropriate nucleotide substitution model, a special case of the general time-reversible Markov model [59]. Independent substitution bias parameters and branch lengths were fitted to each alignment, using the pruning algorithm [60], modified [61] for faster evaluation of phylogenetic likelihood functions, and numerical optimization routines, implemented in HyPhy [57], to obtain maximum likelihood parameter estimates (MLEs). The $MG94 \times REV$ codon model, which estimates synonymous α_s^b and nonsynonymous β_s^b rates independently at every site of the alignment (and possibly differing between branches), was then fitted, while holding branch length and nucleotide substitution bias parameters fixed at MLE values obtained with a nucleotide model using the entire alignment. The rate matrix for this model is a modification of the $MG94$ [62] model (see also [40]), to allow for variable rates across different branches in the tree and to correct for all possible nucleotide substitution biases, given by

$$Q_{x,y}^{MG94\times REV} = \begin{cases} \alpha_s^b \theta_{ij} \pi_{n_y}, & x \to y \text{ 1—step synonymous substitution of nucleotide } i \text{ with nucleotide } j, \\ \beta_s^b \theta_{ij} \pi_{n_y}, & x \to y \text{ 1—step nonsynonymous substitution of nucleotide } i \text{ with nucleotide } j, \\ 0, & \text{otherwise.} \end{cases} \quad (1)$$

To ensure time-reversibility we set $\theta_{ji} = \theta_{ij}$. Because only the products of rates and times are estimable, one of the parameters θ_{ij} cannot be identified, and we choose to set $\theta_{AG} = 1$; θ_{ij} estimates are obtained from the entire alignment and reflect the rate of substituting nucleotide i with nucleotide j relative to the A↔G substitution. π_{n_y} denotes the relative frequency of the nucleotide in position n (1,2,3) in codon y. For instance, the target nucleotide for the synonymous ACG to ACT substitution is T in the third codon position, and its corresponding rate is $\alpha_s^b \theta_{GT} \pi_T^3$. Under $MG94 \times REV$, the stationary frequency of codon y composed of nucleotides i, j, and k is the product of the constituent nucleotide frequencies, scaled to account for the stop codons. For sequences using the universal genetic code, these frequencies are given by

$$\pi_{ijk} = \frac{\pi_i^1 \pi_j^2 \pi_k^3}{1 - \pi_T^1 \pi_A^2 \pi_A^3 - \pi_T^1 \pi_G^2 \pi_A^3 - \pi_T^1 \pi_A^2 \pi_G^3}. \quad (2)$$

Other genetic codes can be easily accommodated by adjusting the list of stop codons. The (x,y) entry of the transition probability matrix $T(t) = \exp(Qt)$ defines the probability of substituting codon x with codon y in time $t \geq 0$. All data analyses were conducted on a 40-processor Linux cluster and all simulation studies were run on 64 processors of the Swansea Blue-C IBM cluster, using the message passing interface (MPI) distributed framework.

Testing for temporal differences in evolution within a sample. Every codon s can be endowed with a single synonymous rate α_s and two nonsynonymous rates: β_s^L (for terminal branches, or leaves) and β_s^I (for internal branches). If the latter two rates differ significantly,

we deduce that evolution along internal branches (historical, e.g., influenced primarily by selection for transmission in HIV) and along terminal branches (recent, e.g., influenced by within-patient evolution in HIV) are subject to differing selective constraints. Formally,

$$\begin{array}{ll} H_0 : \beta_s^L = \beta_s^I. & \text{No selective difference.} \\ H_A : \alpha_s, \beta_s^I, \beta_s^L \text{ are free to vary.} & \text{Temporally differential evolution.} \end{array} \quad (3)$$

A straightforward modification of the null hypothesis can be used to test for non-neutral evolution only along internal branches of the tree:

$$\begin{array}{ll} H_0 : \beta_s^I = \alpha_s. & \text{Neutral evolution.} \\ H_A : \alpha_s, \beta_s^I, \beta_s^L \text{ are free to vary.} & \text{Positive or negative selection.} \end{array} \quad (4)$$

We refer to the latter test as IFEL (internal fixed effects likelihood). Significance is assessed by the likelihood ratio test with one degree of freedom. Our simulations (see simulation strategy details below) have shown that the use of the χ_1^2 asymptotic distribution leads to a conservative test, and actual false positive rates (in our simulation scenario) are lower than the nominal significance level of the test (Figure S7). For a given sample size, the power of the test depends on the divergence level and the disparity between levels of selection between internal and terminal branches. For example, at $p = 0.05$, the overall power of the test to detect non-neutral evolution is only 25%. This rather low number can be partially explained by a large proportion of codon sites with low degree of polymorphism. Such sites are nearly impossible to classify within the current phylogenetic framework. However, if we narrow our focus to strongly selected sites (i.e., sites where $K = max(\beta_s^I/\alpha_s, \alpha_s/\beta_s^I) \geq 5$) with an above average level of divergence ($\alpha_s > 1$), the power increases to 41%. For very strongly selected sites ($K \geq 16$), the power is boosted to 68%. Overall, the PPV of the test is 98.8%.

Population level adaptation test. Given two partitions (Tips only $dN > 0$, or A for brevity, and Internal $dN > 0$, or B) of sites in an alignment, we performed a maximum likelihood fitting of the *MG*94 × *REV* model with each of the partitions having a single synonymous (α_A,α_B) and two nonsynonymous substitution rate parameters: the rate for internal branches (β_A^I, β_B^I), and that for terminal branches (β_A^L, β_B^L). Rate parameters are shared by all codons in the partition and estimated by maximum likelihood, whilst branch length parameters are held at values estimated from the entire alignment previously. We then tested whether the average selective pressure, measured by the ratio β^L/α, along terminal branches was different between partitions A and B. Formally,

$$\begin{array}{ll} H_0 : \beta_A^L = R\alpha_A, \beta_B^L = R\alpha_B. & \text{Same average selective pressure } (R). \\ H_A : \alpha_A, \beta_A^I, \beta_A^L, \alpha_B, \beta_B^I, \beta_B^L \text{ are free to vary.} & \text{Different average selective pressure.} \end{array} \quad (5)$$

Significance was assessed by the likelihood ratio test with one degree of freedom using the asymptoticχ_1^2 distribution of the LR statistic. Note that this test is an extension of the fixed sites approach of Yang and Swanson [63] to allow for variable selective pressures in different parts of the tree across partitions. When the sample size is small, the asymptotic distribution of the LR statistic may not be appropriate, hence we verified significance of the test using parametric bootstrap with 100 replicates.

Testing for differential evolution between populations. Having fitted $\alpha_s, \beta_s^L, \beta_s^I$ to each codon independently in sequences sampled from two different populations, we can test whether the selective pressure along internal branches was discordant between populations. We say that differential historical evolution has acted on codon s when $\omega_I = \beta_s^I/\alpha_s$ differs significantly between two populations. Formally,

$$\begin{array}{ll} H_0 : \beta_s^{1,I} = R\alpha_s^1, \beta_s^{2,I} = R\alpha_s^2. & \text{No differential evolution.} \\ H_A : \alpha_s^1, \beta_s^{1,I}, \beta_s^{1,L}, \alpha_s^2, \beta_s^{2,I}, \beta_s^{2,L} \text{ are free to vary.} & \text{Differential evolution.} \end{array} \quad (6)$$

Significance of the difference can be assessed assuming the χ_1^2 distribution for the likelihood ratio test statistic. Our simulations (see below) suggest that the χ_1^2-based determination of significance leads to a conservative test (Figure S8) and actual false positive rates (in our simulation scenario) are lower than the nominal significance level of the test. For a given sample size, the power of the test depends on the divergence level and the disparity between levels of internal branch selection between the samples. The proportion of sites correctly identified as evolving under temporally differential selection also depends on how different the ratio $D = \beta_s^{1,I}/\alpha_s^1 \div \beta_s^{2,I}/\alpha_s^2$ is from one. For example, at nominal $p = 0.05$, the overall power of the test to identify differential evolution along internal tree branches is merely 8%. Low overall power is attributable to the small extent of polymorphism at many codon positions and small sample sizes. However, for the sites with medium to high levels of divergence ($\min(\alpha_s^1, \alpha_s^2) \geq 1$) and where $\max(D, 1/D) \geq 8$, the power increases to 40%. If $\max(D, 1/D) \geq 32$, the power goes up to 64%. Overall, the PPV of the test is 98.2%.

Multiple test correction. Likelihood ratio tests for selection at an individual site have been applied in similar contexts in at least three studies [22,64,65]. If the main objective is to test whether or not there is evidence for selection *somewhere* in the sequence, based on the results of a series of site-by-site tests, then one would have to employ a multiple-test correction procedure—for example, the Bonferroni correction or a less conservative false discovery rates [66] approach. However, at the level of any given site, as argued in the three cited manuscripts, it is appropriate to use uncorrected *p*-values. Furthermore, our Type I error simulation studies (Figures S7 and S8) show that the size of the test at the level of an individual site is actually less than the nominal *p*-value.

Simulation strategy. Error rates and the power of the tests reported in the previous sections were derived using sequence data simulated under the following protocol. We have used trees, base frequencies, branch lengths (assuming neutral evolution), and nucleotide substitution biases fitted to the two ZA RT sample from our study, with 74 sequences each to simulate 100 (200 codons in each). A neighbor joining tree (using the Tamura–Nei distance metric [67]) was reconstructed from each data replicate and used for further inference, allowing us to investigate whether the power and error rates of the tests were unduly influenced by errors in phylogenetic reconstruction. Previous studies [22] and simulations results presented here (Figure S7) suggest that fixed effect likelihood methods are able to infer site-specific substitution rates accurately, on average, with moderate smoothing effects for larger rates (due to a fairly small sample size). With that in mind, we set out to generate sequences under a distribution of substitution rates that is similar to those which have influenced our real samples. Having fitted the IFEL model (and thus three rates: α_s, β_s^L, and β_s^I) to all four samples, we pooled each type of estimated rates into the following seven bins: [0,0.25), [0.25,0.5), [0.5,1), [1,1.5), [1.5,2.0), [2.0,4.0), and [4.0,∞), and represented each bin with its midpoint (except the final bin, which was represented by 8). For each codon, we drew α_s, β_s^L, and β_s^I from the appropriate estimated rate distribution (also shown in Figure S8). Sampling from distributions with identical supports ensured that a sufficient proportion of sites was generated under the null distribution (e.g., $\alpha_s = \beta_s^I$ for IFEL and $\beta_s^{1,I}/\alpha_1^s = \beta_s^{2,I}/\alpha_2^s$). For the evaluation of the differential selection test, we picked successive pairs of simulations (1–2, 2–3, 3–4, . . . , 99–100) for a total of 99 runs of the analysis.

Implementation. All the tests have been implemented as scripts in the HyPhy [57] batch language and are either a part of the standard distribution of the package, or can be obtained upon request from the authors.

Supporting Information

Figure S1. Joint PR and RT Phylogenies Inferred from Combined Sample Data

Ethiopian and South African sequence samples appear reciprocally monophyletic, both in the maximum likelihood and in the 50% consensus Bayesian trees.

Found at DOI: 10.1371/journal.pcbi.0020062.sg001 (112 KB EPS).

Figure S2. Inferred Evolutionary History at Codon 82 in RT

The trees were rooted using subtype B reference sequences from the Los Alamos HIV database, and ancestral states were inferred using maximum likelihood under the MG94×012232 Dual GDD 3 × 3 [40] model of codon evolution.

Found at DOI: 10.1371/journal.pcbi.0020062.sg002 (127 KB EPS).

Figure S3. Inferred Evolutionary History at Codon 98 in RT

The trees were rooted using subtype B reference sequences from the Los Alamos HIV database, and ancestral states were inferred using

maximum likelihood under the MG94×012232 Dual GDD 3 × 3 [40] model of codon evolution.

Found at DOI: 10.1371/journal.pcbi.0020062.sg003 (134 KB EPS).

Figure S4. Inferred Evolutionary History at Codon 165 in RT

The trees were rooted using subtype B reference sequences from the Los Alamos HIV database, and ancestral states were inferred using maximum likelihood under the MG94×012232 Dual GDD 3 × 3 [40] model of codon evolution.

Found at DOI: 10.1371/journal.pcbi.0020062.sg004 (128 KB EPS).

Figure S5. Inferred Evolutionary History at Codon 177 in RT

The trees were rooted using subtype B reference sequences from the Los Alamos HIV database, and ancestral states were inferred using maximum likelihood under the MG94×012232 Dual GDD 3 × 3 [40] model of codon evolution.

Found at DOI: 10.1371/journal.pcbi.0020062.sg005 (136 KB EPS).

Figure S6. Inferred Evolutionary History at Codon 202 in RT

The trees were rooted using subtype B reference sequences from the Los Alamos HIV database, and ancestral states were inferred using maximum likelihood under the MG94×012232 Dual GDD 3 × 3 [40] model of codon evolution.

Found at DOI: 10.1371/journal.pcbi.0020062.sg006 (128 KB EPS).

Figure S7. The Distribution of Rates Used to Simulate the Data, together with the Boxplots of Generating Rates and Corresponding Rates for Each of the Three Types of Rates at a Site

The bottom right panel depicts the rate of false positives (identifying sites with $\alpha_s = \beta_s^I$ as non-neutrally evolving along internal branches) versus the significance level of the IFEL selection test. Solid gray line shows the expected error rate. Because the actual rate of false positives (for this simulation scenario) is lower than predicted by the significance level of the test, we deduce that the IFEL test behaves conservatively. (More detail is available in the Materials and Methods section.)

Found at DOI: 10.1371/journal.pcbi.0020062.sg007 (487 KB EPS).

Figure S8. False Positive Rate for the Differential Selection Test

The bottom right panel depicts the rate of false positives (identifying sites with $\beta_s^{1,I}/\alpha_s^1 == \beta_s^{2,I}/\alpha_s^2$ as evolving differentially along internal branches) versus the significance level of the differential selection test.
The solid gray line shows the expected error rate. Because the actual rate of false positives (for this simulation scenario) is lower than predicted by the significance level of the test, we deduce that the test behaves conservatively.

Found at DOI: 10.1371/journal.pcbi.0020062.sg008 (124 KB EPS).

Acknowledgments

Author contributions. SLKP, SDWF, DDR, and AJLB conceived and designed the experiments. SLKP, SDWF, and AJLB analyzed the data. ZG and MBG contributed reagents/materials/analysis tools. SLKP, SDWF, ZG, and AJLB wrote the paper.

Funding. This research was supported by grants AI27670, AI38858, AI43638, UCSD Center for AIDS Research (AI36214), AI29164, AI47745, and AI57167 from the National Institutes of Health, and the Research Center for AIDS and HIV Infection of the San Diego Veterans Affairs Healthcare System, and by the University of California Universitywide AIDS Research Program (grant IS02-SD-701).

Competing interests. The authors have declared that no competing interests exist.

References

1. Dobzhansky T (1948) Genetics of natural populations. XVI. Altitudinal and seasonal changes produced by natural selection in certain populations of *Drosophila pseudoobscura* and *Drosophila persimilis*. Genetics 33: 158–176.
2. Johnson F, Schaffer H (1973) Isoenzyme variability in species of genus *Drosophila*. Genotype–environment relationships in populations of *D. melanogaster* from eastern United States. Biochem Genet 10: 149–163.
3. Oakeshott JG, Gibson J, Anderson PR, Knibb WR, Anderson D, et al. (1982) Alcohol dehydrogenase and glycerol-3-phosphate dehydrogenase clines in Drosophila melanogaster on different continents. Evolution 36: 86–96.
4. Hill RE, Hastie ND (1987) Accelerated evolution in the reactive centre regions of serine protease inhibitors. Nature 326: 96–99.
5. Hughes AL, Nei M (1988) Pattern of nucleotide substitution at major histocompatibility complex class I loci reveals overdominant selection. Nature 335: 167–170.
6. Nielsen R, Yang Z (1998) Likelihood models for detecting positively selected amino acid sites and applications to the HIV-1 envelope gene. Genetics 148: 929–936.
7. Borrow P, Lewicki H, Wei X, Horwitz M, Peffer N, et al. (1997) Antiviral pressure exerted by HIV-1-specific cytotoxic T lymphocytes (CTLs) during primary infection demonstrated by rapid selection of CTL escape virus. Nat Med 3: 205–211.
8. Goulder P, Phillips R, Colbert R, McAdam S, Ogg G, et al. (1997) Late escape from an immunodominant cytotoxic T-lymphocyte response associated with progression to AIDS. Nat Med 3: 212–217.
9. Kelleher AD, Long C, Holmes EC, Allen RL, Wilson J, et al. (2001) Clustered mutations in HIV-1 *gag* are consistently required for escape from HLA-B27-restricted cytotoxic T lymphocyte responses. J Exp Med 193: 375–386.
10. Goulder PJR, Brander C, Tang Y, Tremblay C, Colbert RA, et al. (2001) Evolution and transmission of stable CTL escape mutations in HIV infection. Nature 412: 334–338.
11. O'Connor DH, Allen TM, Vogel TU, Jing P, DeSouza IP, et al. (2002) Acute phase cytotoxic T lymphocyte escape is a hallmark of simian immunodeficiency virus infection. Nat Med 8: 493–499.
12. Yang OO, Sarkis PTN, Ali A, Harlow JD, Brander C, et al. (2003) Determinant of HIV-1 mutational escape from cytotoxic T lymphocytes. J Exp Med 197: 1365–1375.
13. Leslie AJ, Pfafferott KJ, Chetty P, Draenert R, Addo MM, et al. (2004) HIV evolution: CTL escape mutation and reversion after transmission. Nat Med 10: 282–289.
14. Friedrich TC, Dodds EJ, Yant LJ, Vojnov L, Rudersdorf R, et al. (2004) Reversion of CTL escape-variant immunodeficiency viruses in vivo. Nat Med 10: 275–281.
15. Moore CB, John M, James IR, Christiansen FT, Witt CS, et al. (2002) Evidence of HIV-1 adaptation to HLA-restricted immune responses at a population level. Science 296: 1439–1443.
16. Yang Z, Nielsen R, Goldman N, Pedersen AM (2000) Codon-substitution models for heterogeneous selection pressure at amino acid sites. Genetics 155: 431–449.
17. Zanotto PM, Kallas EG, de Souza RF, Holmes EC (1999) Genealogical evidence for positive selection in the *nef* gene of HIV-1. Genetics 153: 1077–1089.
18. Williamson S (2003) Adaptation in the *env* gene of HIV-1 and evolutionary theories of disease progression. Mol Biol Evol 20: 1318–1325.
19. Choisy M, Woelk CH, Guegan JF, Robertson DL (2004) Comparative study of adaptive molecular evolution in different human immunodeficiency virus groups and subtypes. J Virol 78: 1962–1970.
20. de Oliveira T, Salemi M, Gordon M, Vandamme AM, van Rensburg EJ, et al. (2004) Mapping sites of positive selection and amino acid diversification in the HIV genome: An alternative approach to vaccine design? Genetics 167: 1047–1058.
21. Travers SAA, O'Connell MJ, McCormack GP, McInerney JO (2005) Evidence for heterogeneous selective pressures in the evolution of the *env* gene in different human immunodeficiency virus type 1 subtypes. J Virol 79: 1836–1841.
22. Kosakovsky Pond SL, Frost SDW (2005) Not so different after all: A comparison of methods for detecting amino acid sites under selection. Mol Biol Evol 22: 1208–1222.
23. Ray SC, Fanning L, Wang XH, Netski DM, Kenny–Walsh E, et al. (2005) Divergent and convergent evolution after a common-source outbreak of hepatitis C virus. J Exp Med 201: 1753–1759.
24. Kosakovsky Pond SL, Frost SDW (2005) A simple hierarchical approach to modeling distributions of substitution rates. Mol Biol Evol 22: 223–234.
25. Gordon M, De Oliveira T, Bishop K, Coovadia HM, Madurai L, et al. (2003) Molecular characteristics of human immunodeficiency virus type 1 subtype C viruses from KwaZulu–Natal, South Africa: Implications for vaccine and antiretroviral control strategies. J Virol 77: 2587–2599.
26. Grossman Z, Vardinon N, Chemtob D, Alkan ML, Bentwich Z, et al. (2001) Genotypic variation of HIV-1 reverse transcriptase and protease: Comparative analysis of clade C and clade B. AIDS 15: 1453–1460.
27. Shafer R, Stevenson D, Chan B (1999) Human immunodeficiency virus reverse transcriptase and protease sequence database. Nucleic Acids Res 27: 348–352.
28. Kantor R, Machekano R, Gonzales MJ, Dupnik K, Schapiro JM, et al. (2001) Human immunodeficiency virus reverse transcriptase and protease sequence database: An expanded data model integrating natural language text and sequence analysis programs. Nucleic Acids Res 29: 296–299.
29. Johnson V, Brun–Vézinet F, Bonaventura C, Conway B, Kuritzkes D, et al. (2005) Update to the drug resistance mutations in HIV-1: Fall 2005. IAS Top HIV Med 13: 125–131.
30. Guindon S, Gascuel O (2003) A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. Syst Biol 52: 696–704.

31. Huelsenbeck JP, Ronquist F (2001) MRBAYES: Bayesian inference of phylogenetic trees. Bioinformatics 17: 754–755.
32. Anisimova M, Nielsen R, Yang Z (2003) Effect of recombination on the accuracy of the likelihood method for detecting positive selection at amino acid sites. Genetics 164: 1229–1236.
33. Shriner D, Nickle DC, Jensen MA, Mullins J (2003) Potential impact of recombination on sitewise approaches for detecting positive natural selection. Genet Res 81: 115–121.
34. Holmes EC, Worobey M, Rambaut A (1999) Phylogenetic evidence for recombination in dengue virus. Mol Biol Evol 16: 405–409.
35. Chare ER, Gould EA, Holmes EC (2003) Phylogenetic analysis reveals a low rate of homologous recombination in negative-sense RNA viruses. J Gen Virol 84: 2691–2703.
36. Saitou N, Nei M (1987) The neighbor-joining method: A new method for reconstructing phylogenetic trees. Mol Biol Evol 4: 406–425.
37. Sugiura N (1978) Further analysis of the data by Akaike's information criterion and the finite corrections. Comm Stat Th Meth A7:: 13–26.
38. Shimodaira H, Hasegawa M (1999) Multiple comparisons of log-likelihoods with applications to phylogenetic inference. Mol Biol Evol 16: 1114–1116.
39. Martin DP, Williamson C, Posada D (2005) RDP2: Recombination detection and analysis from sequence alignments. Bioinformatics 21: 260–262.
40. Kosakovsky Pond SL, Muse SV (2005) Site-to-site variation of synonymous substitution rates. Mol Biol Evol 22: 2375–2385.
41. Sharp PM, Bailes E, Chaudhuri RR, Rodenburg CM, Santiago MO, et al. (2001) The origins of acquired immune deficiency syndrome viruses: Where and when? Phil Trans R Soc Lond B Biol Sci 356: 867–876.
42. Holmes EC (2003) Patterns of intra- and interhost nonsynonymous variation reveal strong purifying selection in Dengue virus. J Virol 77: 11296–11298.
43. Lawlor DA, Ward FE, Ennis PD, Jackson AP, Parham P (1988) HLA-A and B polymorphisms predate the divergence of humans and chimpanzees. Nature 335: 268–271.
44. Lucotte G, Smets P (1999) Origins of Falasha Jews studied by haplotypes of the Y chromosome. Hum Biol 71: 989–993.
45. Fort M, de Stefano GF, Cambon–Thomsen A, Giraldo–Alvarez P, Dugoujon JM, et al. (1998) HLA class II allele and haplotype frequencies in Ethiopian Amhara and Oromo populations. Tissue Antigens 51: 327–336.
46. Williams F, Meenagh A, Darke C, Acosta A, Daar AS, et al. (2001) Analysis of the distribution of HLA-B alleles in populations from five continents. Hum Immunol 62: 645–650.
47. Middleton D, Williams F, Meenagh A, Daar AS, Gorodezky C, et al. (2000) Analysis of the distribution of HLA-A alleles in populations from five continents. Hum Immunol 61: 1048–1052.
48. Fernandez CS, Stratov I, De Rose R, Walsh K, Dale CJ, et al. (2005) Rapid viral escape at an immunodominant simian–human immunodeficiency virus cytotoxic T-lymphocyte epitope exacts a dramatic fitness cost. J Virol 79: 5721–5731.
49. Asquith B, Edwards C, Lipsitch M, McLean A (2004) Inefficient cytotoxic T lymphocyte–mediated killing of HIV-1–infected cells in vivo. PLoS Biol 4: e90. DOI: 10.1371/journal.pbio.004090
50. Yusim K, Kesmir C, Gaschen B, Addo MM, Altfeld M, et al. (2002) Clustering patterns of cytotoxic T-lymphocyte epitopes in human immunodeficiency virus type 1 (HIV-1) proteins reveal imprints of immune evasion on HIV-1 global variation. J Virol 76: 8757–8768.
51. Leitner T, Foley B, Hahn BH, Marx P, McCutchan F, et al. (2005) HIV sequence compendium 2005. Technical report LA-UR-06–0680. Los Alamos (New Mexico): Los Alamos National Laboratory.
52. Yang ZH, Nielsen R (2002) Codon-substitution models for detecting molecular adaptation at individual sites along specific lineages. Mol Biol Evol 19: 908–917.
53. Bielawski J, Yang Z (2004) A maximum likelihood method for detecting functional divergence at individual codon sites, with application to gene family evolution. J Mol Evol 59: 121–132.
54. Zhang JZ (2004) Frequent false detection of positive selection by the likelihood method with branch-site models. Mol Biol Evol 21: 1332–1339.
55. Zhang J, Nielsen R, Yang Z (2005) Evaluation of an improved branch-site likelihood method for detecting positive selection at the molecular level. Mol Biol Evol 22: 2472–2479.
56. Ross HA, Rodrigo AG (2002) Immune-mediated positive selection drives human immunodeficiency virus type 1 molecular variation and predicts disease duration. J Virol 76: 11715–11720.
57. Kosakovsky Pond SL, Frost SDW, Muse SV (2005) HyPhy: Hypothesis testing using phylogenies. Bioinformatics 21: 676–679.
58. McMichael A, Klenerman P (2002) HLA leaves its footprints on HIV. Science 296: 1410–1411.
59. Lanave C, Preparata G, Saccone C, Serio G (1984) A new method for calculating evolutionary substitution rates. J Mol Evol 20: 86–93.
60. Felsenstein J (1981) Evolutionary trees from DNA sequences: A maximum likelihood approach. J Mol Evol 17: 368–376.
61. Kosakovsky Pond SL, Muse SV (2004) Column sorting: Rapid calculation of the phylogenetic likelihood function. Syst Biol 53: 685–692.
62. Muse SV, Gaut BS (1994) A likelihood approach for comparing synonymous and nonsynonymous nucleotide substitution rates, with application to the chloroplast genome. Mol Biol Evol 11: 715–724.
63. Yang ZH, Swanson WJ (2002) Codon-substitution models to detect adaptive evolution that account for heterogeneous selective pressures among site classes. Mol Biol Evol 19: 49–57.
64. Suzuki Y (2004) New methods for detecting positive selection at single amino acid sites. J Mol Evol 59: 11–19.
65. Massingham T, Goldman N (2005) Detecting amino acid sites under positive selection and purifying selection. Genetics 169: 1753–1762.
66. Storey J (2002) A direct approach to false discovery rates. J Royal Stat Soc B 64: 479–479.
67. Tamura K, Nei M (1993) Estimation of the number of nucleotide substitutions in the control region of mitochondrial DNA in humans and chimpanzees. Mol Biol Evol 10: 512–526.