Impact of CCR5delta32 host genetic background and disease progression on HIV-1 intra-host evolutionary processes: efficient hypothesis testing through hierarchical phylogenetic models

Diana Edo-Matas, Philippe Lemey, Jennifer Tom, Cèlia Serna-Bolea, Agnes E. van den Blink, Angélique B. van 'T Wout, Hanneke Schuitemaker, Marc Suchard

Impact of CCR5delta32 host genetic background and disease progression on HIV-1 intra-host evolutionary processes: efficient hypothesis testing through hierarchical phylogenetic models

Research article

Diana Edo-Matas[1], Philippe Lemey[2,†], Jennifer A. Tom[3], Cèlia Serna-Bolea[1,a], Agnes E. van den Blink[1], Angélique B. van 't Wout[1,b], Hanneke Schuitemaker[1,b], Marc A. Suchard[3,4]

[1] Department of Experimental Immunology, Sanquin Research, Landsteiner Laboratory, Center for Infection and Immunity Amsterdam (*CINIMA*) at the Academic Medical Center of the University of Amsterdam, Amsterdam, The Netherlands

[2] Rega Institute for Medical Research, Leuven, Belgium

[3] Department of Biostatistics, School of Public Health; University of California, Los Angeles, CA 90095, USA

[4] Departments of Biomathematics and Human Genetics, David Geffen School of Medicine, University of California, Los Angeles, CA, 90095, USA

[a] Present address: Barcelona Centre for International Health Research, Hospital Clinic, Institut d'Investigacions Biomediques August Pi i Sunyer, Universtitat de Barcelona, Barcelona, Spain.

[b] Present address: Crucell Holland BV, Leiden, The Netherlands.

[†]Correspondence should be addressed to:

Philippe Lemey

Rega Institute, Minderbroedersstraat 10, 3000 Leuven, Belgium

Tel: +32 (0)16 332160

Fax: +32 (0)16 332131

Email: philippe.lemey@uz.kuleuven.be

Key words:

- CCR5
- Envelope
- HIV-1
- Hierarchical Phylogenetic models
- Disease progression
- Bayesian inference

Running head: HIV-1 Hierarchical Phylogenetic Hypothesis Testing

ABSTRACT

The interplay between CCR5 host genetic background, disease progression and intra-host HIV-1 evolutionary dynamics remains unclear because differences in viral evolution between hosts limit the ability to draw conclusions across hosts stratified into clinically relevant populations. Similar inference problems are proliferating across many measurably evolving pathogens for which intra-host sequence samples are readily available. To this end, we propose novel hierarchical phylogenetic models (HPMs) that incorporate fixed-effects to test for differences in dynamics across host-populations in a formal statistical framework employing stochastic search variable selection and model averaging. To clarify the role of CCR5 host genetic background and disease progression on viral evolutionary patterns, we obtain *gp*120 envelope sequences from clonal HIV-1 variants isolated at multiple time points in the course of infection from populations of HIV-1 infected individuals who only harbored CCR5-using HIV-1 variants at all time points. Presence or absence of a CCR5 wt/Δ32 genotype and progressive or long-term non-progressive course of infection stratify the clinical populations in a two-way design. As compared to the standard approach of analyzing sequences from each patient independently, the HPM provides more efficient estimation of evolutionary parameters such as nucleotide substitution rates and d_N/d_S rate ratios, as shown by significant shrinkage of the estimator variance. The fixed-effects also corrects for non-independence of data between populations and results in even further shrinkage of individual patient estimates. Model selection suggests an association between nucleotide substitution rate and disease progression, but a role for CCR5 genotype remains elusive. Given the absence of clear d_N/d_S differences between patient groups, delayed onset of AIDS symptoms appears to be solely associated with lower viral replication rates rather than with differences in selection on amino acid fixation.

INTRODUCTION

The high mutation rate and rapid viral turnover that characterize HIV-1 infection (Ho et al. 1995; Wei et al. 1995) generate a highly diverse genetic viral population within an HIV-1 infected individual (Shankarappa et al. 1999). Continuous emergence of new HIV-1 variants facilitates rapid viral adaptation to humoral and cellular immune responses of the host (Borrow et al. 1997; Goulder et al. 1997; Wei et al. 2003; Jones et al. 2004), escape from antiretroviral drugs (Coffin 1995) and the selection for optimal biological properties such as replication capacity and use of the entry complex (Koning et al. 2003; Kwa et al. 2003) (Sterjovski et al. 2007; Repits et al. 2008).

Following primary infection, an asymptomatic phase with a gradual loss of $CD4^+$ T cells and T-cell function characterizes the clinical course of HIV-1 infection (Lane et al. 1985; Polk et al. 1987; Miedema et al. 1988), resulting eventually in the development of AIDS. The duration of this asymptomatic phase in the absence of antiretroviral therapy varies among patients, from several months to more than two decades, and determines their rate of disease progression (Veugelers et al. 1994; Munoz, Sabin, and Phillips 1997). Many selective forces may play a role in intra-host viral evolution and disease progression such as neutralizing antibodies (nAbs) and cytotoxic T cell (CTL) response, immune activation, target cell availability, co-receptor expression levels and emergence of CXCR4-using viruses among others. The severity of HIV infection may be further complicated by co-infections and heritable viral genetic factors (Hollingsworth et al. 2010). Largely stimulated by a comprehensive longitudinal analysis demonstrating common patterns of sequence divergence, diversity and emergence of CXCR4-using variants in chronic HIV-1 infections (Shankarappa et al. 1999), phylogenetic analyses have been widely used as a means of elucidating how host factors impact HIV within-host dynamics. More specific evolutionary parameters such as evolutionary rate (Lemey et al. 2007; Lee et al.

2008), adaptation rates (Williamson 2003), positively selected sites (Ross and Rodrigo 2002), compartmentalization (Kemal et al. 2003) and recombination (Carvajal-Rodriguez et al. 2008) have been scrutinized, but consistent associations with disease progression have rarely been revealed.

Here, we focus on a polymorphism in the CCR5 gene, which is a host factor known to influence disease progression. The CCR5 gene encodes one of the main coreceptors required for HIV-1 entry, and a heterozygous genotype for a 32 base pair deletion (CCR5 wt/Δ32) associates with a lower viral load set point, defined as the viral load between 18 and 24 months after seroconversion which is stable in most HIV-1 infected individuals and predictive for clinical course of infection (Mellors et al. 1996; de Wolf et al. 1997), and a slower HIV-1 disease progression (de Roda Husman et al. 1997; Ioannidis et al. 2001). Given the reported lower percentages of CCR5 expressing target cells and higher levels of RANTES production in HIV-1 infected individuals with a CCR5 wt/Δ32 genotype (de Roda Husman et al. 1999a; Blaak et al. 2000), it is likely that target cell and CCR5 availability influence HIV-1 intra-patient evolution and contributes to the progression to AIDS.

To investigate these influences, we compared the evolution of CCR5-using HIV-1 variants (R5) in individuals with either a CCR5 wt/wt or CCR5 wt/Δ32 genotype who only harbored CCR5-using HIV-1 variants in their progressive or long-term non-progressive course of infection. Such comparisons require asking questions across multiple populations of individuals about the evolutionary histories that occur within each individual. Traditional modelling of evolutionary histories across individuals generally assumes that within-individual processes vary independently and are fit separately from individual to individual (Shankarappa et al. 1999; Ross and Rodrigo 2002; Potter et al. 2006; Lemey et al. 2007; Carvajal-Rodriguez et al. 2008). Often, this approach results in poor estimates of the underlying evolutionary parameters, as the

111 informative content within a single intra-host dataset is sparse. Not surprisingly, Carvajal-
112 Rodriguez et al. (2008) arrived at the conclusion that the statistical characterization of HIV
113 within-host evolutionary processes in relationship to disease progression is a difficult task and
114 suffers from a lack of power. To overcome the data sparsity, one may enforce strict equality
115 between within-individual evolutionary parameters (Rodrigo et al. 2003). In both cases, however,
116 the ability to formally assess similarities or differences between populations of individuals is lost.
117 Hierarchical modelling (Laird and Ware 1982)(Gelman et al. 1995), and in particular hierarchical
118 phylogenetic models [HPMs] (Suchard et al. 2003), furnish an advantageous statistical
119 framework in which to consider drawing conclusions across populations of individuals about the
120 evolutionary processes within individuals. In general, the Bayesian hierarchical framework
121 allows different evolutionary histories of the intra-host variants and pressures driving their
122 evolution from individual to individual, while providing overall or across-individual summaries
123 of important evolutionary measures, such as the DNA sequence mutation rate or
124 synonymous/non-synonymous rate ratio (d_N/d_S) identifying positive selection. Critically, the
125 HPM allows the within-individual-level parameters to vary about, for example, an unknown
126 common mean for each population. This occurs through the employment of a hierarchical prior
127 distribution on the parameters that are in turn characterized by unknown estimable
128 hyperparameters. Then conveniently, hypothesis testing reduces to asking if these common mean
129 parameters differ between populations. Fortuitously, the hierarchical prior embedded in the HPM
130 also affords a borrowing of strength of information from one individual by another, providing
131 more precise within-individual-level estimates (Suchard et al. 2003; Kitchen et al. 2004; Kitchen
132 et al. 2006; Kitchen et al. 2009).
133 In this study, we extend the HPM across multiple populations of individuals through the
134 introduction of population-specific, fixed effects. These effects allow the expected evolutionary

parameter estimated within a population to potentially vary across populations. We then exploit ideas from Bayesian model averaging (Hoeting et al. 1999) and selection (Suchard, Weiss, and Sinsheimer 2001) to formally ask if these effects statistically differ between populations. We use this approach to estimate viral evolutionary rates and selective pressures within hosts and to evaluate whether these quantities differ with respect to CCR5 wt/Δ32 host genetic background and disease progression.

MATERIALS AND METHODS

Study subjects

18 men who have sex with men (MSM) participants in the Amsterdam Cohort Studies on HIV and AIDS, 11 with a CCR5 wt/wt genotype (patients P1 to P11) and 7 with a CCR5 wt/Δ32 genotype (patients P12 to P18), who at all times tested during follow-up harbored only R5 HIV-1 variants were selected. All patients were either seropositive at entry in the cohort studies (seroprevalent cases with an average imputed seroconversion (SC) date of 18 months before entry in the cohort (Geskus 2000)) or seroconverted during active follow-up in the cohort studies. Nine individuals were classified as long-term non-progressors (LTNP) (defined as HIV-1 infected patients that at the end of follow-up (April 1997) had an asymptomatic seropositive follow-up of at least 11 years with relatively stable $CD4^+$ T cell counts that were still above 400 cells/ml in the ninth year of follow-up in the absence of antiretroviral therapy). The remaining nine individuals progressed to AIDS during the study period (median time to AIDS = 8.2 (2.7-10.8) years after SC or imputed SC date) and were classified as Progressors (P). Individuals included in this study did not receive effective antiretroviral therapy during the study period. Clinical parameters and time points of virus isolation are shown per patient in Figure 1.

The Amsterdam Cohort Studies are conducted in accordance with the ethical principles set out in the declaration of Helsinki and written informed consent was obtained prior to data collection. The study was approved by the Academic Medical Center institutional medical ethics committee.

Isolation of clonal HIV-1 variants

Clonal HIV-1 variants were isolated by co-cultivation of serial dilutions of patient Peripheral Blood Mononuclear Cells (PBMC) from two to eight time points in the course of their infection

and expanded to viral stocks for further study as described previously (Schuitemaker et al. 1992; van 't Wout, Schuitemaker, and Kootstra 2008). For each patient, time points of virus isolation and number of clonal HIV-1 variants per time point are summarized in Supplementary Table S1. The R5 phenotype of all clonal HIV-1 variants that were isolated was confirmed by inability to replicate in the MT2 cell-line, in PHA-PBMC from a donor with a CCR5Δ32 homozygous genotype and in astroglioma cells transfected with CD4 and CCR3 or CXCR4 (de Roda Husman et al. 1999b) and predicted co-receptor use based on the V3 amino acid sequence using the position specific scoring matrix (PSSM) NSI/SI (http://indra.mullins.microbiol.washington.edu/pssm/)(Jensen et al. 2003).

DNA isolation, PCR and sequencing

Total DNA was isolated from PBMCs infected with clonal HIV-1 variants using a modification of the L6 isolation method (Kootstra and Schuitemaker 1999). Precipitated DNA was dissolved in 100µl of distilled water and 5µl were used for PCR amplification of the gp120 (C1-C4) region corresponding to HXB2 nucleotide positions 6444 to 7595. Amplification was performed by PCR with primers TB3 forward (5'-GGCCTTATTAGGACACATAGTTAGCC-3') and OFM19 reverse (5'-GCACTCAAGGCAAGCTTTATTGAGGCTTA-3') using the expand high fidelity Taq polymerase kit (Roche) and the following amplification cycles: 2 min 30s 94˚C, 9 cycles of 15s 94˚C, 45s 50˚C, 6 min 68˚C, 30 cycles of 15s 94˚C, 45s 53˚C, 6 min 68˚C, followed by a 10 min extension at 68˚C and subsequent cooling to 4˚C. Nested PCR was performed with two different inner PCR primer combinations: Seq1 forward (5'-TACATAATGTTTGGGCCACACATGCC -3'), Seq4 reverse (5'-CTTGTATTGTTGTTGGGTCTTGTAC -3'), Seq5 forward (5'-GTCAACTCAACTGCTGTTAAATGGC -3') and Seq2 reverse (5'-

188 TCCTTCATATCTCCTCCTCCAGGTC -3'). Nested PCRs were performed using Promega Taq
189 polymerase in the presence of 2mM $MgCl_2$ using the following amplification cycles: 5 min 94˚C,
190 40 cycles of 15s 95˚C, 30s 59˚C, 2 min 72˚C, followed by a 10 min extension at 72˚C and
191 subsequent cooling to 4˚C.
192 PCR products were purified using ExoSAP-IT (USB, Cleveland, Ohio, USA) according to
193 manufacturer's protocol. Sequencing conditions consisted of 5' at 94˚C, 30 cycles of 15'' at 94˚C,
194 10'' at 50˚C, 2' at 60˚C and a 10' extension at 60˚C. Sequencing was performed using BigDye
195 Terminator v1.1 Cycle Sequencing kit (ABI Prism, Applied Biosystems, Warrington, UK)
196 according to the manufacturer's protocol using the nested PCR primers. Sequences were analyzed
197 on the Applied Biosystems 3130 xl Genetic Analyzer. The nucleotide sequences are available
198 from Genbank under the accession numbers EU743973.1-EU44009.1, EU744014.1-EU744046.1,
199 EU744055.1-EU744093.1, EU744097.1-EU744129.1, EU744146.1-EU744175.1, GU455514-
200 GU455525 and HQ644787-HQ645012.
201
202 **Bayesian inference of within-host HIV evolutionary rates and selection pressures**
203 Nucleotide sequences for all clonal HIV-1 gp120 (C1-C4) variants isolated from the individual
204 patients were aligned using ClustalW (Thompson, Higgins, and Gibson 1994) and manually
205 edited. Cross-contamination was excluded using phylogenetic analysis.
206 *(a) Independent estimates of within-host evolutionary rates.* Nucleotide substitution rates were
207 estimated for each patient using strict and relaxed (uncorrelated lognormal) molecular clock
208 models implemented in BEAST v.1.4.8 (Drummond et al. 2006; Drummond and Rambaut 2007).
209 We used a general time-reversible (GTR) model of nucleotide substitution with discrete gamma-
210 distributed rate variation among sites. Posterior distributions were obtained using Bayesian

211 Markov chain Monte Carlo (MCMC) analysis. MCMC chains were run sufficiently long to
212 ensure stationarity and adequate effective sample sizes (ESS > 100) as diagnosed using Tracer
213 (http://tree.bio.ed.ac.uk/software/tracer/). The uncertainty of continuous parameter estimates is
214 expressed as 95% highest posterior density (HPD) intervals.
215 *(b) Hierarchical estimates of evolutionary parameters*. To draw inference about different
216 evolutionary patterns across populations of patients, we implement a novel HPM in BEAST
217 (Suchard et al. 2003). HPMs analyze viral sequence data from multiple patients simultaneously
218 and have found extensive use in uncovering common patterns of intra-host HIV evolution
219 (Kitchen et al. 2004; Kitchen et al. 2006; Kitchen et al. 2009). At the heart of the HPM lies a
220 Bayesian mixed effects model that pools information across patients. Pooling information
221 through random effects affords more precise individual-patient parameter estimates when the data
222 are sparse for a patient. Further, unique to the work here, the introduction of fixed effects (see
223 below) offers a formal hypothesis testing framework from which to identify differences in
224 evolutionary process between patient population groups.
225 Let θ_i for $i = 1, \ldots, N$ patients represent the evolutionary process parameter of interest; this could
226 be, for example, the overall rate of nucleotide substitution or the nonsynonymous/synonymous
227 substitution rate ratio (d_N/d_S) in a codon substitution process across the unknown genealogy
228 relating the sequences from within patient i. In the analysis of four different patient groups:
229 Progressors, Long-term non-progressors (LTNP), CCR5 wt/wt (WT) and CCR5 wt/Δ32 (Δ32),
230 we assume that either log θ_i or θ_i is drawn from an underlying normal distribution where the
231 mean and variance of this underlying prior distribution are also unknown and simultaneously
232 estimated along with all sequence data. The choice of a log transform is convenient for modeling
233 strictly positive parameters. Importantly, fixing this mean and variance to known values does not

234 return a hierarchical model, but rather results in complete independence across individuals. On
235 the other hand, estimating the mean or variance imparts both an approach to make comparisons
236 across populations and the borrowing of strength for poorly informed within-individual model
237 parameters.
238 For nucleotide analyses, we apply this hierarchical setup to the strict clock evolutionary rate (on
239 the log-scale), the mean evolutionary rate parameter of the lognormal relaxed clock (log), the
240 constant population size (log) of the demographic prior, the GTR substitution parameters (log)
241 and the shape parameter (log) of the discrete gamma distribution modeling rate variation among
242 sites. For codon model analyses, a hierarchical transition/transversion rate parameter and a
243 hierarchical d_N/d_S rate ratio (Goldman and Yang 1994) replace the GTR model parameters.
244 *(c) Hierarchical estimation with population-specific, fixed effects.* For hypothesis testing
245 purposes, we extend the HPM to include across-population fixed effects. Each patient belongs to
246 one of four fixed population groups that we can designate using two indicator factors: $LTNP_i = 0$
247 (1) for short (long) term progressors and $\Delta 32_i = 0$ (1) for deletion 32 absent (present) patients.
248 Our HPM assumes
249
250 $$\log \theta_i = \beta_0 + \delta_{LTNP}\beta_{LTNP}LTNP_i + \delta_{\Delta 32}\beta_{\Delta 32}\Delta 32_i + \varepsilon_i,$$
251
252 where β_0 is an unknown grand-mean, δ_{LTNP} and $\delta_{\Delta 32}$ are binary indicator variables, β_{LTNP} and $\beta_{\Delta 32}$
253 are conditional effective sizes and ε_i are independent and normally distributed random variables
254 with mean 0 and an estimable variance. The inclusion of the indicator variables follows from a
255 Bayesian stochastic search variable selection approach (Kuo and Mallick 1998; Chipman, George,
256 and McCulloch 2001) that simultaneously estimates the posterior probabilities of all possible

257 linear models that may or may not include LTNP or Δ32 status effects. When an indicator equals
258 1, this effect is included in the model, demonstrating that the evolutionary process parameter
259 differs with high probability between patient population groups. Lemey et al. (2009) discuss
260 Bayesian stochastic search variable selection in further detail (Lemey et al. 2009).
261 We complete this HPM model with variable selection through assigning independent Bernoulli
262 prior probability distributions on δ_{LTNP} and $\delta_{\Delta 32}$. These distributions place equal probability on
263 each factor's inclusion and exclusion. We further assume diffuse priors on the unknown grand-
264 mean and error variance and specify that *a priori* β_{LTNP} and $\beta_{\Delta 32}$ are normally distributed with
265 mean 0 and a variance of 1/2. We choose 1/2 as, before seeing the data, we believe that, if a
266 factor does result in different evolutionary parameters across population groups, process
267 parameters should differ by at most an order of magnitude on their original scale. The
268 introduction of HPMs into BEAST necessitates the development of MCMC transition kernels to
269 efficiently explore that space of the grand-mean and effect-size, model indicator, and random-
270 effects variance parameters. Given our judicious prior choices, the full conditional distributions
271 of these parameters are in standard-form: multivariate-normal, binomial and inverse-gamma,
272 respectively. This enables us to build highly effective Gibbs samplers (Casella and George 1992;
273 Suchard et al. 2003) over the joint space of these parameters. Suchard et al. (2003) provide
274 detailed derivations of the full condition distributions and their Gibbs samplers (Suchard et al.
275 2003). We implement these Gibbs samplers as regular BEAST "operators" that are now
276 accessible to interested readers through BEAST's XML model specification language.
277 Supplementary material to this paper reports the transition kernels' XML syntax and gives
278 examples on their use to implement HPMs.

To assign statistical significance to differences between population groups, we employ Bayes factors (Jeffreys 1998; Suchard, Weiss, and Sinsheimer 2001) that report how much the data change our prior opinion (here, 1:1 odds) about the inclusion of each factor. These Bayes factors are straightforward to estimate through the variable selection procedure, as the Bayes factor equals the posterior odds that a factor indicator equals 1 divided by the corresponding prior odds. The posterior odds follow immediately from the marginal posterior probability that a factor indicator equals 1 that we estimate through the posterior expectation of the factor indicator. In cases where an estimate of this expectation approaches very closely to 0 or 1, an estimator based on a Rao-Blackwellization procedure is available (Casella and Robert 1996).

RESULTS

Independent versus hierarchical estimation of evolutionary parameters

We first explored the nucleotide substitution rate as a hierarchical parameter estimated across patients in four separate patient groups: Progressors, LTNP, CCR5 wt/wt and CCR5 wt/Δ32. Using a strict clock model, a higher mean evolutionary rate was estimated in the Progressors group (mean = 7.65×10^{-4}, 95% HPD = [6.45×10^{-4},8.84×10^{-4}]) compared to the LTNP group (5.87×10^{-4} [$4.30 \cdot 10^{-4}$-7.55×10^{-4}]) (Figure 2A). While these estimates demonstrate overlapping marginal posterior credible intervals (CIs), immediately concluding that their difference is not significant ignores the correlation between the rates; we return to a formal test later. A less pronounced difference in evolutionary rate was estimated between the CCR5 wt/wt (7.27×10^{-4} [5.74×10^{-4}-8.75×10^{-4}]) and CCR5 wt/Δ32 (6.00×10^{-4} [4.21×10^{-4}-7.89×10^{-4}]) groups. Similar rate differences, with somewhat less overlapping CIs between Progressors (7.57×10^{-4} [6.49×10^{-4}-8.67×10^{-4}]) and LTNPs (5.63×10^{-4} [4.19×10^{-4}-7.06×10^{-4}]), were observed using a relaxed clock model (Figure 2B), in which the log of the mean evolutionary rate across all branches in a patient genealogy is drawn from an underlying normal distribution. For both strict and relaxed evolutionary rate estimates (Figure 3 A and B), as well as other substitution model and population genetic parameters (data not shown), we observed significant shrinkage in uncertainty under the standard hierarchical fit, which clearly demonstrates the HPM improvement. Moreover, separate fit of parameter-rich models such as the uncorrelated relaxed clock required informative priors to achieve efficient sampling. To demonstrate the impact of such priors on our posterior rate estimates obtained by separate model fitting, and compare these with the hierarchical estimates that did not require such priors, we plot the marginal posterior rate estimates for the three least informative (lowest number of time points and/or sequences per time point) and three

most informative patients within the LTNP group (P10, P16 and P17 versus P9, P11 and P13 respectively) as violin plots in Figure 4. Violin plots are box plots overlaid with (rotated) kernel density estimates in order to show to the probability density at different parameter values. The patients for which only two or three time points were available resulted in rate estimates that only weakly diverged from their respective prior (uniform[0,0.004] or lognormal(-7.5,1); Figure 4A and C respectively), whereas many time points provide sufficient information to dominate over these priors (Figure 4 B and D). Under the hierarchical model, even weakly informative patient-specific data sets with extremely diffuse priors on the rate yield relative precise posteriors (Figure 4E), and the individual patient estimates are only marginally higher than for the three most informative patients (Figure 4F). This demonstrates that comparing the mean rates for individual estimates would is inappropriate to assess differences among patient groups. Weakly informative patients result in relatively high mean rates, but their high variances ensure that the contribution to the population rate (LTNP group) in the hierarchical model remains low.

While the application of relaxed clock models to individual data sets with few time points or sequences may be questionable, analysis under a HPM, in which information is pooled between patients, enables us to side-step this limitation. Marginal likelihood estimates for the both strict and relaxed clock analyses of the different patient groups (Supplementary Table S2) indicate a better fit of the relaxed clock model, with log Bayes factors (BFs) of 7.8, 6.1, 4.4 and 4.2 in favor of the relaxed clock for Progressors, LTNP, CCR5 wt/wt and CCR5 wt/Δ32 respectively. The fact that a strict clock could often not be rejected for individual patient analysis also indicates the HPM draws on increased statistical power of HPMs to reject simpler models. Because of the increased model fit, we employ relaxed clocks in further codon model analyses and hypothesis tests incorporating fixed effects.

Analyses using a codon model revealed comparable codon substitution rate differences between Progressors/LTNP and between CCR5 wt/wt and CCR5 wt/Δ32 compared to the nucleotide analyses (Supplementary Figure 1A vs. Figure 2B). Hierarchical d_N/d_S estimates, however, were comparable for the four patient groups (Supplementary Figure 1B).

Hypothesis testing using HPMs incorporating across-population fixed effects

The four different groups considered previously are not comprised of independent patient sets; some patients fall in more than one group. Hence, direct comparison of the marginal parameter estimates fit to each group independently does not generate independent estimates. For more appropriate hypothesis testing of difference, the HPM for the evolutionary rate was extended to accommodate fixed effects (see methods), enabling estimation of hierarchical parameters across all patients. Successfully, hierarchical estimation with fixed effects across all patients resulted in even further shrinkage of individual patient estimates compared to hierarchal models applied to separate groups (Figure 3B). Bayes factor comparison of the fixed-effects HPM model to a model that assumes either completely linked or unlinked parameters (log BF of 51.7 and 57.2 respectively) provides strong evidence that the shrinkage is accompanied by improved goodness-of-fit. The main results of the fixed effect HPM analyses are listed in Table 1. For the nucleotide analysis, the LTNP versus Progressor and CCR5 wt/wt versus CCR5 wt/Δ32 effects were employed to model the evolutionary rates. Through examining the posterior distribution of the rate indicators (δ_{effect}), we estimate the posterior probability for including the LTNP versus Progressor effect at 0.72 resulting in a moderate Bayes factor support of 2.6 in agreement with the group-by-group hierarchical rate estimates obtained above. Importantly, the rate decrease attributable to this fixed-effect returns a credible interval that does not include 0. This approach

appropriately controls for the non-independence missed in the group-by-group analyses and rejects the null hypothesis of no difference between LTNP and Progressor patients.

There was no support in favor of a CCR5 wt/Δ32 effect. Even after conditioning on the effect-indicator equaling 1 to estimate the potential effect-size, the posterior CCR5 wt/Δ32 effect-size parameter distribution remained centered close to 0 with symmetric CIs. In the codon analysis, the same effects were tested on both the substitution rate and d_N/d_S. A very similar LTNP effect was observed for the codon substitution rate, although the CIs now included 0. Interestingly, the conditional effect size of LTNP versus Progressor on codon substitution rate remains very similar to the effect size on nucleotide substitution rate. Further, there was more support against than in favor of a CCR5 wt/Δ32 effect. Finally, no support for a LTNP effect or CCR5 wt/Δ32 effect was observed on the hierarchical d_N/d_S estimates.

DISCUSSION

In this study, we adopted a HPM approach to estimate within-host HIV evolutionary parameters and test evolutionary hypotheses regarding host susceptibility and disease progression. We sought to investigate whether the CCR5 wt/Δ32 genotype, which is associated with a lower viral load set point and a slower HIV-1 disease progression (de Roda Husman et al. 1997; Ioannidis et al. 2001), also impacts the evolutionary rate of the virus by limiting target cell or CCR5 availability. Furthermore, we wanted to evaluate the contribution of CCR5 availability and CCR5 use on the selection pressure directed against the viral envelope protein by estimating d_N/d_S.

HPMs have been used for HIV evolutionary enquiry before, but this is the first study that develops HPMs to estimate evolutionary rate, d_N/d_S and demographic parameters. In a HPM framework, we assume that the patient-specific HIV-1 evolutionary parameters can be drawn from a population distribution. Estimations of the evolutionary process based on a limited sample from each patient are riddled with noise and the improvement of a HPM follows from the reduced uncertainty on individual patient estimates. Bayes factor comparison further confirms a considerable improvement in goodness-of-fit of the HPM with respect to a completely linked and unlinked model. This can be explained by the fact that the completely linked model inappropriately ignores any difference among patients on the one hand, and a completely linked model suffers from an unnecessarily high effective number of parameters (Spiegelhalter et al. 2002) arising from the independent prior specifications on the other hand. The HPM sits in between these two extremes and reduces the effective number of parameters without sacrificing fit to the data. Furthermore, we demonstrate that the HPM is more powerful in rejecting simpler evolutionary models, like the constant rate assumption, which is frequently violated for HIV.

The hierarchical estimates for the Progressors, LTNP, CCR5 wt/wt and CCR5 wt/Δ32 groups indicated a pronounced strict and relaxed clock rate difference between the Progressors and

LTNP, whereas differences between CCR5 wt/wt and CCR5 wt/Δ32 rates were less pronounced. The same patterns were observed for relaxed codon substitution rates, but no real differences were noted in terms of d_N/d_S. These comparisons are based on non-independent data because patients will be part of two different groups. For more appropriate hypothesis testing, we incorporated fixed effects and employed Bayesian stochastic search variable selection to estimate the posterior probability that different patient group characteristics influence within-host evolutionary parameters. The advantage of a Bayesian model averaging approach that simultaneously explores the space of models and regression coefficients is the opportunity to distinguish between the relative size of an effect and its importance, which can be formalized in terms of standard Bayes factor support. The latter effectively becomes independent of the scale of the predictors, which otherwise may confound drawing conclusions on the effect sizes only. Because both predictors we considered only achieve 0 or 1, controlling for scale is not an issue in the current study, but it does contribute to a more general framework for evolutionary hypothesis testing. While the statistical support is not decisive, the fixed-effects HPM approach produces substantially more efficient parameter estimates and conditional effect sizes confirm rate differences among LTNP and Progressors. Despite the elevated power, more elaborate sampling in terms of numbers of patients, within-host time points or maybe even larger genome regions would be desirable.

The HPM estimates suggest an association between evolutionary rate and disease progression, but the CCR5 genotype does not account for the rate differences. Given the absence of clear d_N/d_S differences – if anything, they are slightly higher in LTNP – we cannot attribute the rate nuances to differences in selection on amino acid fixation. Therefore, we conclude that these differences are due to variations in the product of mutation rate and generation time. In particular, lower replication rates may be associated with delayed onset of AIDS symptoms. In agreement

with this, a codon-model extension of the Bayesian relaxed-clock analysis of more extensively sampled patients has shown that absolute synonymous substitutions are correlated with disease progression (Lemey et al. 2007). These authors argued that synonymous substitutions were a marker of replication rate and most probably reflect the action of immune activation, which in itself is a marker of disease progression. In the current study, we employed standard codon model implementation in the Bayesian framework, rather than evaluating genealogies under nucleotide models as a proxy. This approach comes at a computational expense, and further extensions - such as codon models to estimate absolute rates of synonymous and nonsynonymous substitutions (Seo, Kishino, and Thorne 2004) - may prove even more computationally intensive. Fortunately, recent advances in GPU computation provide significant increases in computation speed for high state space models (Suchard and Rambaut 2009). These advances promise to stimulate further development of various codon models in the Bayesian framework, the parameters of which could be efficiently estimated in hierarchical models.

CCR5 genotype has a measurable impact on disease progression (de Roda Husman et al. 1997; Ioannidis et al. 2001) but there appears to be no absolute relationship (not all CCR5 wt/Δ32 infected individuals are LTNP). This implies a more complex scenario, in which the combination of CCR5 availability with other host genetic factors, in particular cellular and humoral immune pressures, and immune activation, will determine the viral replication rate and progression of the disease in a patient. While lower CCR5 availability does not appear to exert selection pressure on the viral envelope during the chronic phase of infection, it cannot be excluded that in HIV-1 infected individuals with CCR5 WT/Δ32 genotype, in whom $CCR5^{+}$ target cells and CCR5 expression are already limiting in the acute phase, selection for viruses with optimal CCR5 use occurs in a very early stage. Moreover, we performed analyses on sequences in which ambiguously aligned hypervariable regions were deleted, which may play an important role in

440 both humoral immune responses (Cao et al. 1997; Chackerian, Rudensey, and Overbaugh 1997;
441 Stamatatos and Cheng-Mayer 1998; Pinter et al. 2004; Sagar et al. 2006; Gray et al. 2007) and
442 selection for optimal CCR5 use (Hubert and Arabie 1985; Stamatatos, Wiskerchen, and Cheng-
443 Mayer 1998; Wang et al. 1999; Sagar et al. 2006; Repits et al. 2008).
444 Studying evolutionary dynamics within hosts has become an integral part of HIV research, but
445 one that still faces the challenge of fully unraveling the relationship between evolutionary
446 parameters and clinical outcome. There may be several reasons for the difficulty in establishing
447 the role of evolutionary processes in disease progression. Within-host dynamics appear to be
448 highly complex, with many host-specific and environmental (co-infections) factors interacting
449 with various evolutionary processes such as hypermutation, diversifying and directional selection,
450 recombination and compartmentalization. Untangling this complex interplay requires accurate
451 measurement of all host factors involved and evolutionary models that explicitly accommodate
452 the relevant evolutionary forces. Without the latter, many simplifying assumptions are at risk of
453 being violated when considering HIV evolution. Parameter-rich models may be limited by
454 current sampling as they require highly informative data. To our knowledge, the most elaborate
455 sampling dates back to over a decade ago (Shankarappa et al. 1999), which, differently from this
456 study, included patients with HIV populations harboring CXCR4-using variants. Next generation
457 sequencing may offer new opportunities for within host HIV genetic analyses, but produces data
458 with particular challenges for comparative analyses (Vrancken et al. 2010). Here, we have
459 adopted a modeling approach that efficiently pools the information from multiple individuals and
460 we demonstrate how this can be employed for rigorous testing across patient populations. We
461 hope that this stimulates further model-based inference of evolutionary processes, which
462 ultimately may lead to more profound insights into persistent viral infections.
463

FUNDING

This work was supported by Netherlands AIDS fund (grant nr 6006) and The European Community's Seventh Framework Programme NGIN (FP7/2007-2013) under grant agreement n° 201433. PL was supported by a postdoctoral fellowship from the Fund for Scientific Research (FWO) Flanders. MAS and JAT are supported by the National Institutes of Health R01 grant GM86887. The research leading to these results has received funding from the European Research Council under the European Community's Seventh Framework Programme (FP7/2007-2013) / ERC Grant agreement n° 260864.

ACKNOWLEDGEMENTS

The Amsterdam Cohort Studies on HIV infection and AIDS, a collaboration between the Public Health Service of Amsterdam, the Academic Medical Center of the University of Amsterdam, the Sanquin Blood Supply Foundation, the University Medical Center Utrecht, and the Jan van Goyen Medical Center, are part of the Netherlands HIV Monitoring Foundation and financially supported by the Center for Infectious Disease Control of the Netherlands National Institute for Public Health and the Environment.

TABLES

Table 1. Estimates of the long-term non-progressor (LTNP) and Δ32 effects on nucleotide substitution rates, codon substitution rates and d_N/d_S.

Evolutionary parameter	Effect support/size	LTNP effect	Δ32 effect
Nucleotide substitution rate	Posterior probability $\delta_{effect} = 1$	0.72	0.27
	Bayes factor$_{effect}$	2.6	0.4
	$\beta_{effect}\|\delta_{effect} = 1$*	-0.275 (-0.524,-0.016)	-0.007 (-0.940,0.920)
Codon substitution rate:	Posterior probability $\delta_{effect} = 1$	0.726	0.324
	Bayes factor$_{effect}$	2.6	0.5
	$\beta_{effect}\|\ \delta_{effect} = 1$*	-0.265 (-0.523,0.019)	-0.012 (-0.700,0.692)
d_N/d_S	Posterior probability $\delta_{effect} = 1$	0.502	0.393
	Bayes factor$_{effect}$	1.0	0.6
	$\beta_{effect}\|\ \delta_{effect} = 1$*	0.083 (-0.101,0.25)	-0.005 (-0.228,0.242)

*these are effective sizes conditional on the effect being included (the binary effect indicator δ_{effect} being 1). For the rates these effective sizes are in log space.

FIGURES

Figure 1. $CD4^+$ T cell numbers, viral loads, and antiretroviral treatments of 18 participants from the Amsterdam Cohort Studies who were selected for this study. Time points of clinical AIDS diagnosis are indicated with open downward triangles. Arrows indicate time points of clonal virus isolation. The length and type of antiretroviral therapy are indicated in the top part of the panels.

Figure 2. Evolutionary rate estimates using a hierarchical phylogenetic model applied separately to four patient groups (Progressors, LTNP, CCR5 wt/wt and CCR5 wt/Δ32). Evolutionary rate estimated under strict clock model (A). Mean evolutionary rate estimated under relaxed clock model (B). CCR5 wt/wt (WT); CCR5 wt/Δ32 (Δ32).

Figure 3. Improved statistical efficiency (shrinkage effect) of the hierarchical phylogenetic model. Strict clock (A). Relaxed clock (B). Posterior variance of estimated evolutionary rate from the independent analyses of each patient (white); evolutionary rate variance from the hierarchical analysis of LTNPs and Progressors (black); evolutionary rate variance from the hierarchical analysis of LTNPs and Progressors incorporating fixed effects (grey).

Figure 4. Marginal posterior rate distributions for LTNP patients with different numbers of sampling time points. Least informative patients (lowest number of time points or sequences per time point): P10, P16 and P17. Most informative patients: P9, P11 and P13. A & B: Assuming a uniform[0,0.004] rate prior. C & D: lognormal(-7.5,1) rate prior. E & F: hierarchical phylogenetic model with unknown mean and variance and diffuse priors.

REFERENCES

Blaak, H., L. J. Ran, R. Rientsma, and H. Schuitemaker. 2000. Susceptibility of in vitro stimulated PBMC to infection with NSI HIV-1 is associated with levels of CCR5 expression and beta-chemokine production. Virology **267**:237-246.

Borrow, P., H. Lewicki, X. Wei, M. S. Horwitz, N. Peffer, H. Meyers, J. A. Nelson, J. E. Gairin, B. H. Hahn, M. B. Oldstone, and G. M. Shaw. 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.

Cao, J., N. Sullivan, E. Desjardin, C. Parolin, J. Robinson, R. Wyatt, and J. Sodroski. 1997. Replication and neutralization of human immunodeficiency virus type 1 lacking the V1 and V2 variable loops of the gp120 envelope glycoprotein. J Virol **71**:9808-9812.

Carvajal-Rodriguez, A., D. Posada, M. Perez-Losada, E. Keller, E. J. Abrams, R. P. Viscidi, and K. A. Crandall. 2008. Disease progression and evolution of the HIV-1 env gene in 24 infected infants. Infect Genet Evol **8**:110-120.

Casella, G., and E. I. George. 1992. Explaining the Gibbs sampler. The American Statistician **46**:167-174.

Casella, G., and C. Robert. 1996. Rao-Blackwellisation of sampling schemes. Biometrika **83**:81-94.

Chackerian, B., L. M. Rudensey, and J. Overbaugh. 1997. Specific N-linked and O-linked glycosylation modifications in the envelope V1 domain of simian immunodeficiency virus variants that evolve in the host alter recognition by neutralizing antibodies. J Virol **71**:7719-7727.

Chipman, H., E. George, and R. McCulloch. 2001. The practical implementation of Bayesian model selection. IMS Lecture Notes - Monograph Series **38**:67-134.

Coffin, J. M. 1995. HIV population dynamics in vivo: implications for genetic variation, pathogenesis, and therapy. Science **267**:483-489.

de Roda Husman, A. M., H. Blaak, M. Brouwer, and H. Schuitemaker. 1999a. CC chemokine receptor 5 cell-surface expression in relation to CC chemokine receptor 5 genotype and the clinical course of HIV-1 infection. J Immunol **163**:4597-4603.

de Roda Husman, A. M., M. Koot, M. Cornelissen, I. P. Keet, M. Brouwer, S. M. Broersen, M. Bakker, M. T. Roos, M. Prins, F. de Wolf, R. A. Coutinho, F. Miedema, J. Goudsmit, and H. Schuitemaker. 1997. Association between CCR5 genotype and the clinical course of HIV-1 infection. Ann Intern Med **127**:882-890.

de Roda Husman, A. M., R. P. van Rij, H. Blaak, S. Broersen, and H. Schuitemaker. 1999b. Adaptation to promiscuous usage of chemokine receptors is not a prerequisite for human immunodeficiency virus type 1 disease progression. J Infect Dis **180**:1106-1115.

de Wolf, F., I. Spijkerman, P. T. Schellekens, M. Langendam, C. Kuiken, M. Bakker, M. Roos, R. Coutinho, F. Miedema, and J. Goudsmit. 1997. AIDS prognosis based on HIV-1 RNA, CD4+ T-cell count and function: markers with reciprocal predictive value over time after seroconversion. Aids **11**:1799-1806.

Drummond, A. J., S. Y. W. Ho, M. J. Phillips, and A. Rambaut. 2006. Relaxed phylogenetics and dating with confidence. PLoS Biol **4**.

Drummond, A. J., and A. Rambaut. 2007. BEAST: Bayesian evolutionary analysis by sampling trees. BMC Evol Biol **7**:214.

558 Gelman, A., J. B. Carlin, H. S. Stern, and D. B. Rubin. 1995. Bayesian Data Analysis. Chapman
559 & Hall/CRC, New York.
560 Geskus, R. B. 2000. On the inclusion of prevalent cases in HIV/AIDS natural history studies
561 through a marker-based estimate of time since seroconversion. Stat Med **19**:1753-1769.
562 Goldman, N., and Z. Yang. 1994. A codon-based model of nucleotide substitution for protein-
563 coding DNA sequences. Mol. Biol. Evol. **11**:725-736.
564 Goulder, P. J., R. E. Phillips, R. A. Colbert, S. McAdam, G. Ogg, M. A. Nowak, P. Giangrande,
565 G. Luzzi, B. Morgan, A. Edwards, A. J. McMichael, and S. Rowland-Jones. 1997. Late
566 escape from an immunodominant cytotoxic T-lymphocyte response associated with
567 progression to AIDS. Nat Med **3**:212-217.
568 Gray, E. S., P. L. Moore, I. A. Choge, J. M. Decker, F. Bibollet-Ruche, H. Li, N. Leseka, F.
569 Treurnicht, K. Mlisana, G. M. Shaw, S. S. Karim, C. Williamson, and L. Morris. 2007.
570 Neutralizing antibody responses in acute human immunodeficiency virus type 1 subtype
571 C infection. J Virol **81**:6187-6196.
572 Ho, D. D., A. U. Neumann, A. S. Perelson, W. Chen, J. M. Leonard, and M. Markowitz. 1995.
573 Rapid turnover of plasma virions and CD4 lymphocytes in HIV-1 infection. Nature
574 **373**:123-126.
575 Hoeting, J., D. Madigan, A. Raftery, and C. Volinsky. 1999. Bayesian Model Averaging.
576 Statistical Science **14**:382-401.
577 Hollingsworth, T. D., O. Laeyendecker, G. Shirreff, C. A. Donnelly, D. Serwadda, M. J. Wawer,
578 N. Kiwanuka, F. Nalugoda, A. Collinson-Streng, V. Ssempijja, W. P. Hanage, T. C.
579 Quinn, R. H. Gray, and C. Fraser. 2010. HIV-1 transmitting couples have similar viral
580 load set-points in Rakai, Uganda. PLoS Pathog **6**:e1000876.
581 Hubert, L., and P. Arabie. 1985. Comparing Partitions. Journal of Classification **2**:193.
582 Ioannidis, J. P., P. S. Rosenberg, J. J. Goedert, L. J. Ashton, T. L. Benfield, S. P. Buchbinder, R.
583 A. Coutinho, J. Eugen-Olsen, T. Gallart, T. L. Katzenstein, L. G. Kostrikis, H. Kuipers, L.
584 G. Louie, S. A. Mallal, J. B. Margolick, O. P. Martinez, L. Meyer, N. L. Michael, E.
585 Operskalski, G. Pantaleo, G. P. Rizzardi, H. Schuitemaker, H. W. Sheppard, G. J. Stewart,
586 I. D. Theodorou, H. Ullum, E. Vicenzi, D. Vlahov, D. Wilkinson, C. Workman, J. F.
587 Zagury, and T. R. O'Brien. 2001. Effects of CCR5-Delta32, CCR2-64I, and SDF-1 3'A
588 alleles on HIV-1 disease progression: An international meta-analysis of individual-patient
589 data. Ann Intern Med **135**:782-795.
590 Jeffreys. 1998. Theory of Probability. Oxford University Press, New York.
591 Jensen, M. A., F. S. Li, A. B. van 't Wout, D. C. Nickle, D. Shriner, H. X. He, S. McLaughlin, R.
592 Shankarappa, J. B. Margolick, and J. I. Mullins. 2003. Improved coreceptor usage
593 prediction and genotypic monitoring of R5-to-X4 transition by motif analysis of human
594 immunodeficiency virus type 1 env V3 loop sequences. J. Virol. **77**:13376-13388.
595 Jones, N. A., X. Wei, D. R. Flower, M. Wong, F. Michor, M. S. Saag, B. H. Hahn, M. A. Nowak,
596 G. M. Shaw, and P. Borrow. 2004. Determinants of human immunodeficiency virus type
597 1 escape from the primary CD8+ cytotoxic T lymphocyte response. J Exp Med **200**:1243-
598 1256.
599 Kemal, K. S., B. Foley, H. Burger, K. Anastos, H. Minkoff, C. Kitchen, S. M. Philpott, W. Gao,
600 E. Robison, S. Holman, C. Dehner, S. Beck, W. A. Meyer, 3rd, A. Landay, A. Kovacs, J.
601 Bremer, and B. Weiser. 2003. HIV-1 in genital tract and plasma of women:
602 compartmentalization of viral sequences, coreceptor usage, and glycosylation. Proc Natl
603 Acad Sci U S A **100**:12972-12977.

Kitchen, C. M., J. Lu, M. A. Suchard, R. Hoh, J. N. Martin, D. R. Kuritzkes, and S. G. Deeks. 2006. Continued evolution in gp41 after interruption of enfuvirtide in subjects with advanced HIV type 1 disease. AIDS Res Hum Retroviruses **22**:1260-1266.

Kitchen, C. M., S. Philpott, H. Burger, B. Weiser, K. Anastos, and M. A. Suchard. 2004. Evolution of human immunodeficiency virus type 1 coreceptor usage during antiretroviral Therapy: a Bayesian approach. J Virol **78**:11296-11302.

Kitchen, C. M. R., V. Marconi, D. R. Kuritzkes, E. W. Bloomquist, S. G. Deeks, and M. A. Suchard. 2009. Two-way Bayesian hierarchical phylogenetic models: an application to the co-evolution of gp120 and gp41 during partial treatment interruptions of enfuvirtide. Computational Statistics and Data Analysis **53**:766-775.

Koning, F. A., D. Kwa, B. Boeser-Nunnink, J. Dekker, J. Vingerhoed, H. Hiemstra, and H. Schuitemaker. 2003. Decreasing sensitivity to RANTES (regulated on activation, normally T cell-expressed and -secreted) neutralization of CC chemokine receptor 5-using, non-syncytium-inducing virus variants in the course of human immunodeficiency virus type 1 infection. J Infect Dis **188**:864-872.

Kootstra, N. A., and H. Schuitemaker. 1999. Phenotype of HIV-1 lacking a functional nuclear localization signal in matrix protein of gag and Vpr is comparable to wild-type HIV-1 in primary macrophages. Virology **253**:170-180.

Kuo, L., and B. Mallick. 1998. Variable selection for regression models. Sankhya B **60**:65-81.

Kwa, D., J. Vingerhoed, B. Boeser, and H. Schuitemaker. 2003. Increased in vitro cytopathicity of CC chemokine receptor 5-restricted human immunodeficiency virus type 1 primary isolates correlates with a progressive clinical course of infection. J Infect Dis **187**:1397-1403.

Laird, N. M., and J. H. Ware. 1982. Random-effects models for longitudinal data. Biometrics **38**:963-974.

Lane, H. C., J. M. Depper, W. C. Greene, G. Whalen, T. A. Waldmann, and A. S. Fauci. 1985. Qualitative analysis of immune function in patients with the acquired immunodeficiency syndrome. Evidence for a selective defect in soluble antigen recognition. N Engl J Med **313**:79-84.

Lee, H. Y., A. S. Perelson, S. C. Park, and T. Leitner. 2008. Dynamic correlation between intrahost HIV-1 quasispecies evolution and disease progression. PLoS Comput Biol **4**:e1000240.

Lemey, P., S. L. Kosakovsky Pond, A. J. Drummond, O. G. Pybus, B. Shapiro, H. Barroso, N. Taveira, and A. Rambaut. 2007. Synonymous substitution rates predict HIV disease progression as a result of underlying replication dynamics. PLoS Comput Biol **3**:e29.

Lemey, P., A. Rambaut, A. J. Drummond, and M. A. Suchard. 2009. Bayesian phylogeography finds its roots. PLoS Comput Biol **5**:e1000520.

Mellors, J. W., C. R. Rinaldo, Jr., P. Gupta, R. M. White, J. A. Todd, and L. A. Kingsley. 1996. Prognosis in HIV-1 infection predicted by the quantity of virus in plasma. Science **272**:1167-1170.

Miedema, F., A. J. Petit, F. G. Terpstra, J. K. Schattenkerk, F. de Wolf, B. J. Al, M. Roos, J. M. Lange, S. A. Danner, J. Goudsmit, and et al. 1988. Immunological abnormalities in human immunodeficiency virus (HIV)-infected asymptomatic homosexual men. HIV affects the immune system before CD4+ T helper cell depletion occurs. J Clin Invest **82**:1908-1914.

Munoz, A., C. A. Sabin, and A. N. Phillips. 1997. The incubation period of AIDS. AIDS **11 Suppl A**:S69-76.

651 Pinter, A., W. J. Honnen, Y. He, M. K. Gorny, S. Zolla-Pazner, and S. C. Kayman. 2004. The
652 V1/V2 domain of gp120 is a global regulator of the sensitivity of primary human
653 immunodeficiency virus type 1 isolates to neutralization by antibodies commonly induced
654 upon infection. J Virol **78**:5205-5215.
655 Polk, B. F., R. Fox, R. Brookmeyer, S. Kanchanaraksa, R. Kaslow, B. Visscher, C. Rinaldo, and J.
656 Phair. 1987. Predictors of the acquired immunodeficiency syndrome developing in a
657 cohort of seropositive homosexual men. N Engl J Med **316**:61-66.
658 Potter, S. J., P. Lemey, W. B. Dyer, J. S. Sullivan, C. B. Chew, A. M. Vandamme, D. E. Dwyer,
659 and N. K. Saksena. 2006. Genetic analyses reveal structured HIV-1 populations in serially
660 sampled T lymphocytes of patients receiving HAART. Virology **348**:35-46.
661 Repits, J., J. Sterjovski, D. Badia-Martinez, M. Mild, L. Gray, M. J. Churchill, D. F. Purcell, A.
662 Karlsson, J. Albert, E. M. Fenyo, A. Achour, P. R. Gorry, and M. Jansson. 2008. Primary
663 HIV-1 R5 isolates from end-stage disease display enhanced viral fitness in parallel with
664 increased gp120 net charge. Virology **379**:125-134.
665 Rodrigo, A. G., M. Goode, R. Forsberg, H. A. Ross, and A. Drummond. 2003. Inferring
666 evolutionary rates using serially sampled sequences from several populations. Mol Biol
667 Evol **20**:2010-2018.
668 Ross, H. A., and A. G. Rodrigo. 2002. Immune-mediated positive selection drives human
669 immunodeficiency virus type 1 molecular variation and predicts disease duration. J Virol
670 **76**:11715-11720.
671 Sagar, M., X. Wu, S. Lee, and J. Overbaugh. 2006. Human immunodeficiency virus type 1 V1-
672 V2 envelope loop sequences expand and add glycosylation sites over the course of
673 infection, and these modifications affect antibody neutralization sensitivity. J Virol
674 **80**:9586-9598.
675 Schuitemaker, H., M. Koot, N. A. Kootstra, M. W. Dercksen, R. E. de Goede, R. P. van
676 Steenwijk, J. M. Lange, J. K. Schattenkerk, F. Miedema, and M. Tersmette. 1992.
677 Biological phenotype of human immunodeficiency virus type 1 clones at different stages
678 of infection: progression of disease is associated with a shift from monocytotropic to T-
679 cell-tropic virus population. J Virol **66**:1354-1360.
680 Seo, T. K., H. Kishino, and J. L. Thorne. 2004. Estimating absolute rates of synonymous and
681 nonsynonymous nucleotide substitution in order to characterize natural selection and date
682 species divergences. Mol Biol Evol **21**:1201-1213.
683 Shankarappa, R., J. B. Margolick, S. J. Gange, A. G. Rodrigo, D. Upchurch, H. Farzadegan, P.
684 Gupta, C. R. Rinaldo, G. H. Learn, X. He, X. L. Huang, and J. I. Mullins. 1999.
685 Consistent viral evolutionary changes associated with the progression of human
686 immunodeficiency virus type 1 infection. J. Virol. **73**:10489-10502.
687 Spiegelhalter, D. J., N. G. Best, B. P. Carlin, and A. Van Der Linde. 2002. Bayesian measures of
688 model complexity and fit. Journal of the Royal Statistical Society: Series B (Statistical
689 Methodology) **64**:583–639.
690 Stamatatos, L., and C. Cheng-Mayer. 1998. An envelope modification that renders a primary,
691 neutralization-resistant clade B human immunodeficiency virus type 1 isolate highly
692 susceptible to neutralization by sera from other clades. J Virol **72**:7840-7845.
693 Stamatatos, L., M. Wiskerchen, and C. Cheng-Mayer. 1998. Effect of major deletions in the V1
694 and V2 loops of a macrophage-tropic HIV type 1 isolate on viral envelope structure, cell
695 entry, and replication. AIDS Res Hum Retroviruses **14**:1129-1139.
696 Sterjovski, J., M. J. Churchill, A. Ellett, L. R. Gray, M. J. Roche, R. L. Dunfee, D. F. Purcell, N.
697 Saksena, B. Wang, S. Sonza, S. L. Wesselingh, I. Karlsson, E. M. Fenyo, D. Gabuzda, A.

L. Cunningham, and P. R. Gorry. 2007. Asn 362 in gp120 contributes to enhanced fusogenicity by CCR5-restricted HIV-1 envelope glycoprotein variants from patients with AIDS. Retrovirology **4**:89.

Suchard, M. A., C. M. Kitchen, J. S. Sinsheimer, and R. E. Weiss. 2003. Hierarchical phylogenetic models for analyzing multipartite sequence data. Syst. Biol. **52**:649-664.

Suchard, M. A., and A. Rambaut. 2009. Many-core algorithms for statistical phylogenetics. Bioinformatics **25**:1370-1376.

Suchard, M. A., R. E. Weiss, and J. S. Sinsheimer. 2001. Bayesian selection of continuous-time Markov chain evolutionary models. Mol Biol Evol **18**:1001-1013.

Thompson, J. D., D. G. Higgins, and T. J. Gibson. 1994. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Res. **22**:4673-4680.

van 't Wout, A. B., H. Schuitemaker, and N. A. Kootstra. 2008. Isolation and propagation of HIV-1 on peripheral blood mononuclear cells. Nat Protoc **3**:363-370.

Veugelers, P. J., K. A. Page, B. Tindall, M. T. Schechter, A. R. Moss, W. W. Winkelstein, Jr., D. A. Cooper, K. J. Craib, E. Charlebois, R. A. Coutinho, and et al. 1994. Determinants of HIV disease progression among homosexual men registered in the Tricontinental Seroconverter Study. Am J Epidemiol **140**:747-758.

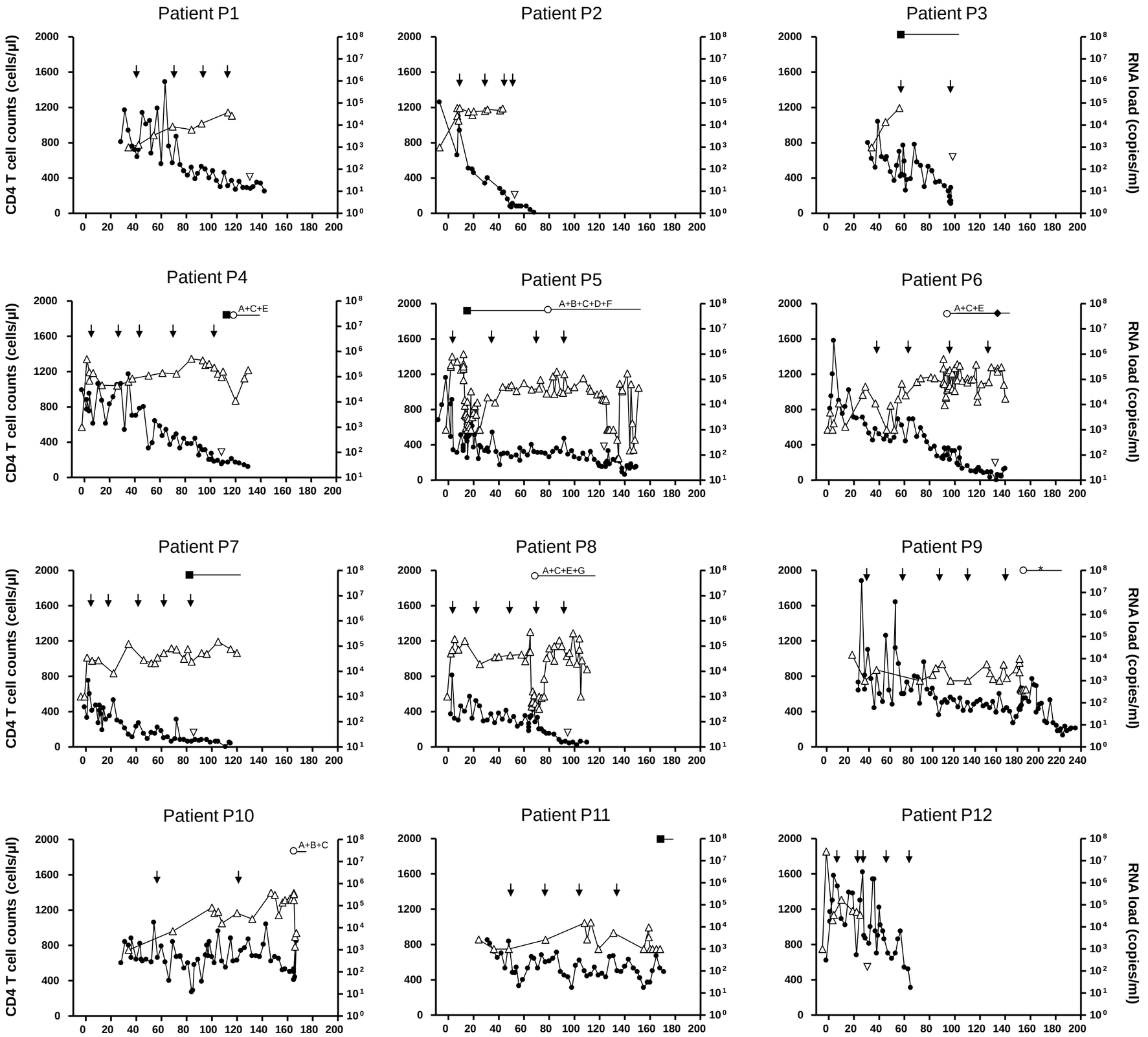
Vrancken, B., S. Lequime, K. Theys, and P. Lemey. 2010. Covering all bases in HIV research: unveiling a hidden world of viral evolution. AIDS Rev **12**:89-102.

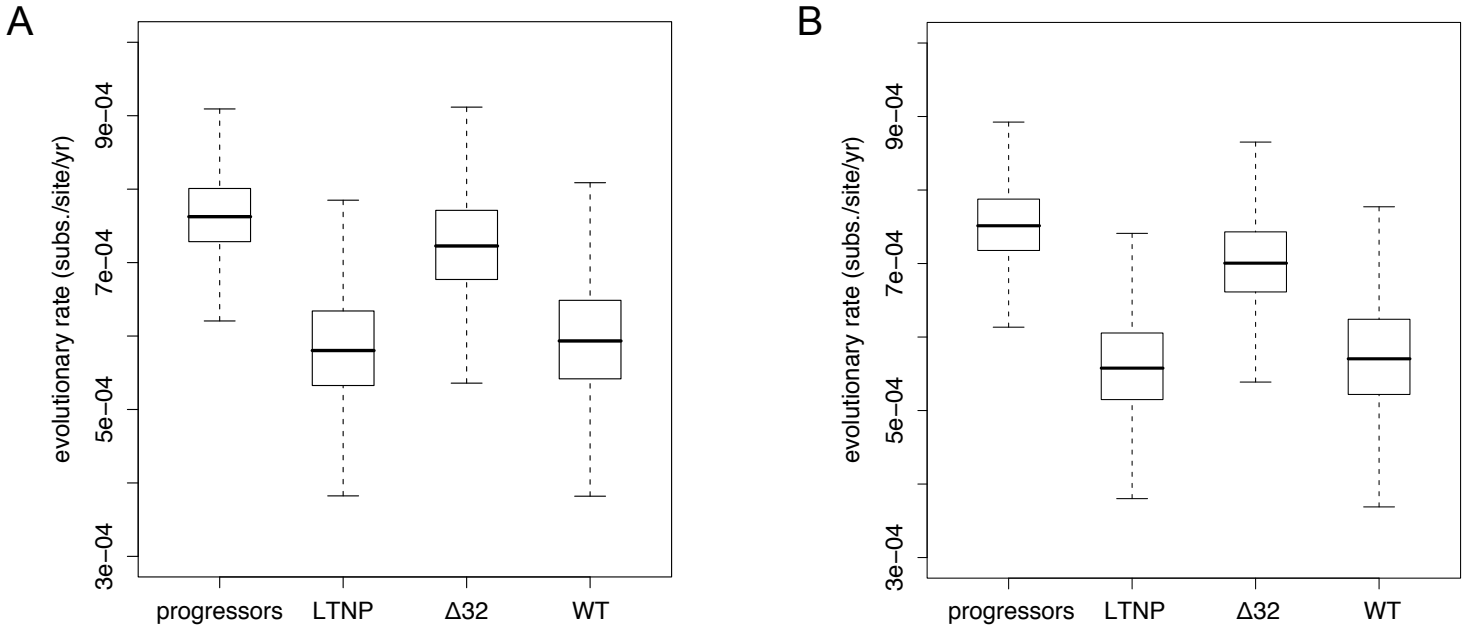
Wang, W. K., T. Dudek, M. Essex, and T. H. Lee. 1999. Hypervariable region 3 residues of HIV type 1 gp120 involved in CCR5 coreceptor utilization: therapeutic and prophylactic implications. Proc Natl Acad Sci U S A **96**:4558-4562.

Wei, X., J. M. Decker, S. Wang, H. Hui, J. C. Kappes, X. Wu, J. F. Salazar-Gonzalez, M. G. Salazar, J. M. Kilby, M. S. Saag, N. L. Komarova, M. A. Nowak, B. H. Hahn, P. D. Kwong, and G. M. Shaw. 2003. Antibody neutralization and escape by HIV-1. Nature **422**:307-312.

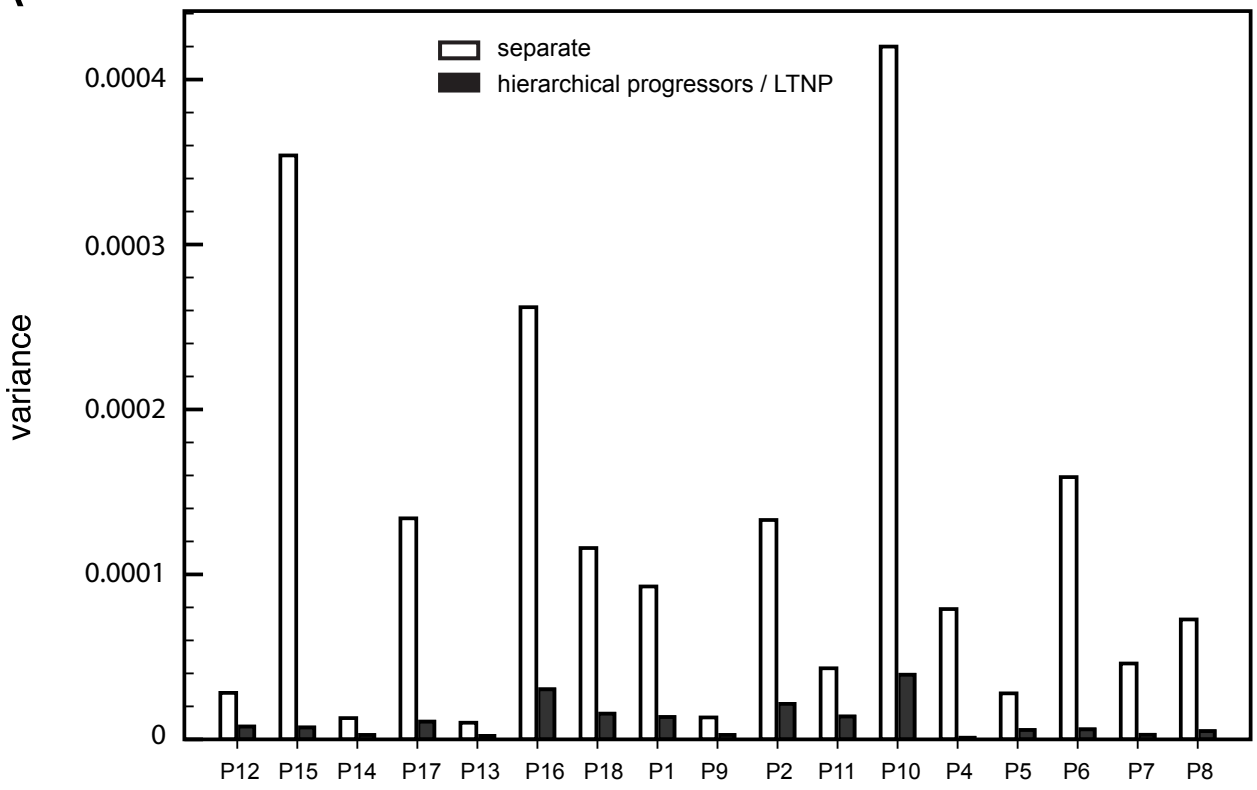
Wei, X., S. K. Ghosh, M. E. Taylor, V. A. Johnson, E. A. Emini, P. Deutsch, J. D. Lifson, S. Bonhoeffer, M. A. Nowak, B. H. Hahn, M. S. Saag, and G. M. Shaw. 1995. Viral dynamics in human immunodeficiency virus type 1 infection. Nature **373**:117-122.

Williamson, S. 2003. Adaptation in the env gene of HIV-1 and evolutionary theories of disease progression. Mol Biol Evol **20**:1318-1325.

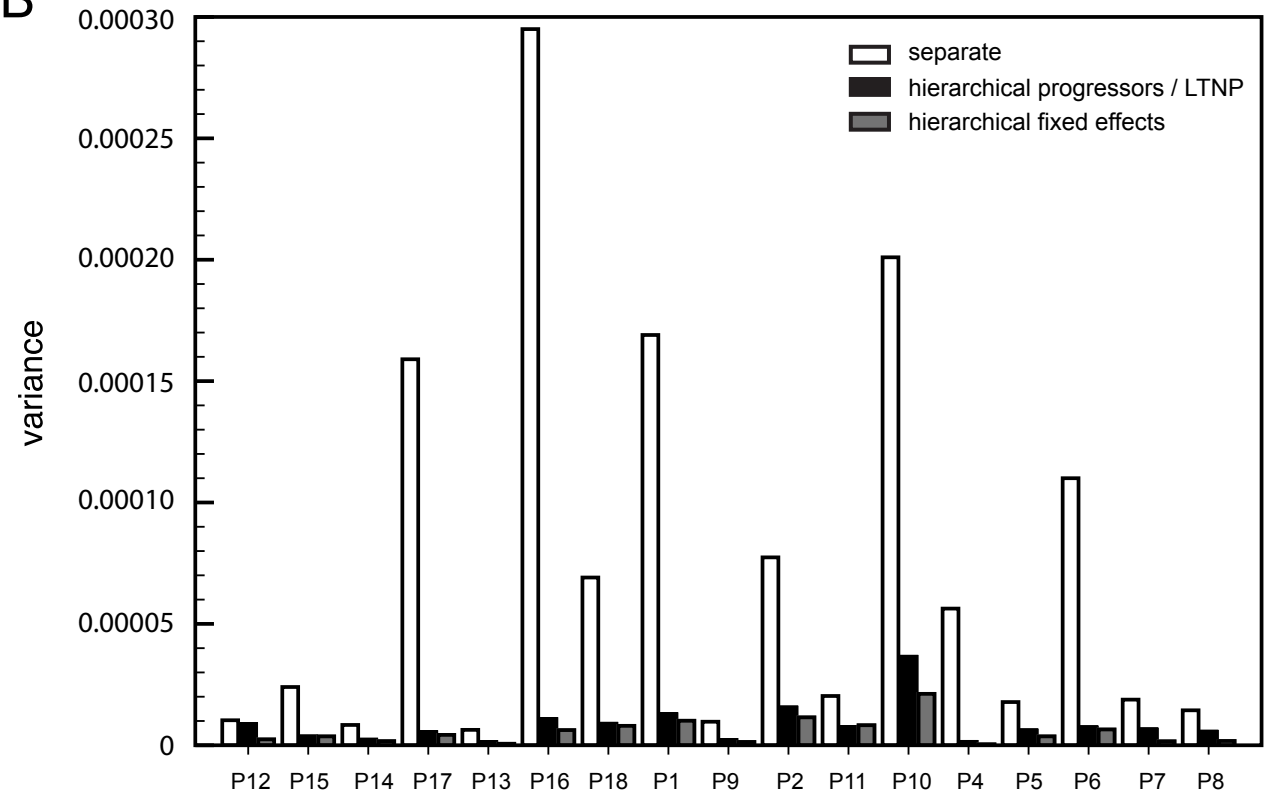

Patient P1
Patient P2
Patient P3
Patient P4
Patient P5
Patient P6
Patient P7
Patient P8
Patient P9
Patient P10
Patient P11
Patient P12
Patient P13
Patient P14
Patient P15
Patient P16
Patient P17
Patient P18
CD4 T cell counts (cells/µl)
RNA load (copies/ml)
Time to SC (months)
A+C+E
A+B+C+D+F
A+C+E+G
A+B+C
*
CD4+ T cells/µl serum
RNA load (copies/ml)
Time point of clonal virus isolation
AIDS diagnosis
Zidovudine (AZT)
Lamivudine (3TC)
cART (Combined Antiretroviral Therapy)
A: Zidovudine (AZT)
B: Stavudine (d4T)
C: Lamivudine (3TC)
D: Didanosine (ddI)
E: Salcitabine (ddC)
F: Indinavir
G: Sanquinavir
* unknown

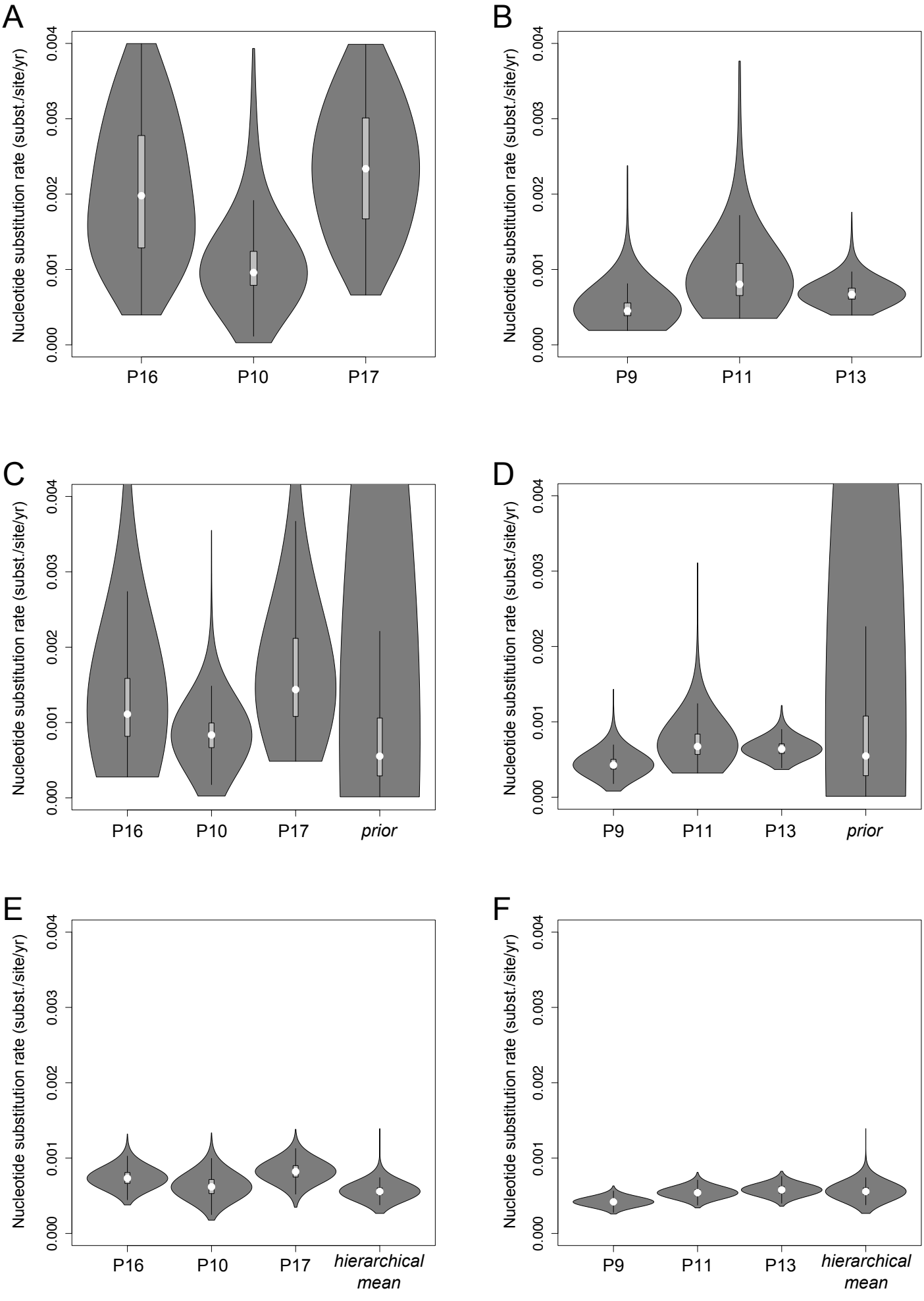

A
B
evolutionary rate (subs./site/yr)
3e−04
5e−04
7e−04
9e−04
progressors
LTNP
Δ32
WT

A


separate
hierarchical progressors / LTNP
variance
0.0004
0.0003
0.0002
0.0001
0
P12 P15 P14 P17 P13 P16 P18 P1 P9 P2 P11 P10 P4 P5 P6 P7 P8

B


separate
hierarchical progressors / LTNP
hierarchical fixed effects
variance
0.00030
0.00025
0.00020
0.00015
0.00010
0.00005
0
P12 P15 P14 P17 P13 P16 P18 P1 P9 P2 P11 P10 P4 P5 P6 P7 P8


A
B
C
D
E
F
Nucleotide substitution rate (subst./site/yr)
0.000
0.001
0.002
0.003
0.004
P16
P10
P17
P9
P11
P13
prior
hierarchical mean

SUPPLEMENTARY INFORMATION

SUPPLEMENTARY TABLES

Table S1. Patients, time points and number of sequences analyzed.

Subject	Patient number	CCR5 genotype	Disease progression	Sampling time after SC (months)	gp120 env (nr of clones)
P1	19858	WT/WT	P	42*	8
				69*	5
				92*	4
				113*	6
P2	19576	WT/WT	P	7	2
				29	4
				43	5
				51	5
P3	19947	WT/WT	P	56*	3
				98*	3
P4	19999	WT/WT	P	4	14
				26	16
				42	5
				74	7
				107	20
P5	19768	WT/WT	P	2	21
				36	15
				67	17
				93	12
P6	19659	WT/WT	P	2	1
				30	7
				62	22
				95	20
				128	5
P7	19542	WT/WT	P	2	4
				20	5
				43	7
				63	17
				86	15
P8	18969	WT/WT	P	2	25
				22	21
				47	10
				68	7
				91	15
P9	19559	WT/WT	LTNP	39*	3
				71*	5
				106*	1
				133*	5
				170*	3
P10	19932	WT/WT	LTNP	54*	3
				120*	5
P11	19417	WT/WT	LTNP	48*	3
				77*	6
				101*	5
				131*	5
P12	19828	Δ32/WT	P	4	5
				22	2
				25	4

				47	11
				63	4
P13	19383	Δ32/WT	LTNP	39*	2
				50*	2
				62*	4
				71*	4
				95*	3
				107*	6
				133*	7
				148*	2
P14	19922	Δ32/WT	LTNP	39*	5
				82*	6
				111*	5
				135*	5
P15	19663	Δ32/WT	LTNP	47*	5
				91*	6
				111*	6
				140*	5
P16	19984	Δ32/WT	LTNP	19	6
				109	4
P17	19566	Δ32/WT	LTNP	13	2
				19	7
				101	3
				116	4
P18	19956	Δ32/WT	LTNP	28*	5
				51*	3
				78*	2
				123*	1
				146*	1

P: Progressor; LTNP: Long-term non-progressor; SC: seroconversion; *Sampling time after imputed SC date.

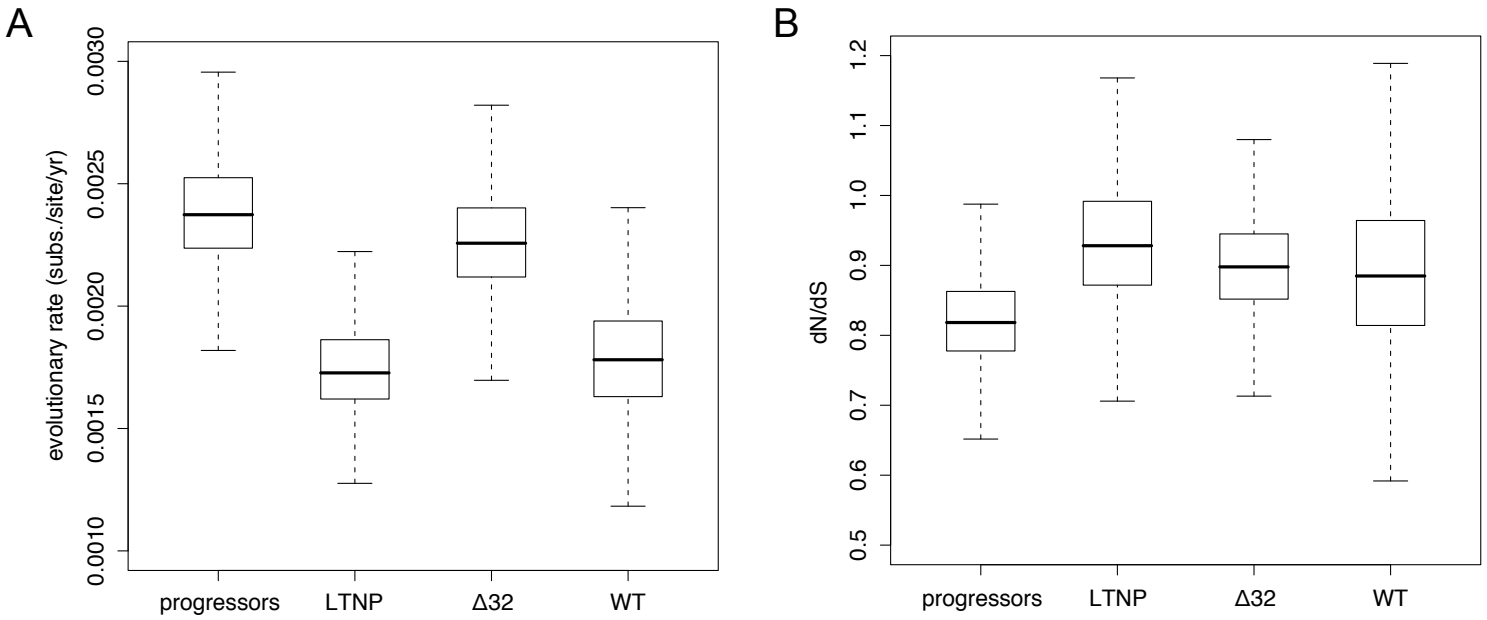
Table S2. Log marginal likelihood estimates for strict and relaxed clock analyses of four patient groups.

	Progressors	LTNP	WT	Δ32
Strict clock	-28534.3	-26762.8	-33501.0	-21793.5
Relaxed clock	-28526.5	-26756.7	-33496.6	-21789.3

WT: CCR5 wt/wt; Δ32: CCR5 wt/Δ32.

SUPPLEMENTARY FIGURES

Supplementary Figure 1. Codon model estimates. Mean evolutionary rate estimated under relaxed clock model with codon model for four patient groups: Progressors, LTNP, CCR5 wt/wt and CCR5 wt/Δ32 (A)**.** *dN/dS* rate ratios estimated for the same for patient groups (B). CCR5 wt/wt (WT); CCR5 wt/Δ32 (Δ32).


A
evolutionary rate (subs./site/yr)
0.0010
0.0015
0.0020
0.0025
0.0030
progressors
LTNP
Δ32
WT
B
dN/dS
0.5
0.6
0.7
0.8
0.9
1.0
1.1
1.2
progressors
LTNP
Δ32
WT