Evolution during primary HIV infection does not require adaptive immune selection

David A. Swan[a], Morgane Rolland[b,c], Joshua T. Herbeck[d], Joshua T. Schiffer[a,e,f], and Daniel B. Reeves[a,1]

[a]Vaccine and Infectious Disease Division, Fred Hutchinson Cancer Research Center, Seattle, WA 98109; [b]US Military HIV Research Program, Walter Reed Army Institute of Research, Silver Spring, MD 20910; [c]Henry M. Jackson Foundation for the Advancement of Military Medicine, Bethesda, MD 20817; [d]International Clinical Research Center, Department of Global Health, University of Washington, Seattle, WA 98195; [e]Department of Medicine, University of Washington, Seattle, WA 98195; and [f]Clinical Research Division, Fred Hutchinson Cancer Research Center, Seattle, WA 98109

Edited by Malcolm Martin, Laboratory of Molecular Microbiology, National Institute of Allergy and Infectious Diseases, Bethesda, MD; received May 17, 2021; accepted December 16, 2021

Modern HIV research depends crucially on both viral sequencing and population measurements. To directly link mechanistic biological processes and evolutionary dynamics during HIV infection, we developed multiple within-host phylodynamic models of HIV primary infection for comparative validation against viral load and evolutionary dynamics data. The optimal model of primary infection required no positive selection, suggesting that the host adaptive immune system reduces viral load but surprisingly does not drive observed viral evolution. Rather, the fitness (infectivity) of mutant variants is drawn from an exponential distribution in which most variants are slightly less infectious than their parents (nearly neutral evolution). This distribution was not largely different from either in vivo fitness distributions recorded beyond primary infection or in vitro distributions that are observed without adaptive immunity, suggesting the intrinsic viral fitness distribution may drive evolution. Simulated phylogenetic trees also agree with independent data and illuminate how phylogenetic inference must consider viral and immune-cell population dynamics to gain accurate mechanistic insights.

HIV primary infection | viral dynamics modeling | phylogenetics | phylodynamics | viral evolution

Longitudinal sequencing of HIV over time in an infected person provides invaluable insights into the pathogenesis of disease. Phylogenetic tools (1) help illuminate evolutionary relationships between viral sequences and phylodynamics leverages such relationships to further infer underlying processes governing evolution (2, 3). However, phylodynamic tools often do not encompass the details of within-host HIV infection, which include massive exponential expansions and contractions of viral populations, mounting immune responses, target cell limitation, and existence of short- and long-lived cell populations. Mechanistic mathematical models of HIV explicitly include these predator/prey interactions and are amenable to generalized nonlinear processes including therapeutic interventions (4). Therefore, unifying intrahost mechanistic modeling with phylodynamics is a potentially powerful approach to reveal the mechanisms underlying intrahost viral evolution that hinder HIV prevention and/or cure.

There is an extensive history of modeling unified phylodynamics for viruses within and between hosts (5–7). Within-host HIV models have modeled nucleotide sequences and assumed multistrain phenotypes with fitness distributions (6–12). Particular aspects of HIV biology including recombination (13, 14), drug resistance (15), antibody evolution (16, 17), adaptive immunity (18, 19), and latency (20) have been considered. Several general forward simulation packages are available (21–23). Our work grows from these and other models. Here we sought to identify mechanistic drivers of HIV evolution by identifying and validating a within-host phylodynamic (WiPhy) model against a range of experimentally collected HIV data. By building in capabilities of the model to simulate data that can be exported and analyzed by existing phylogenetic inference software, we corroborate the body of work showing that without accurate models for population dynamics, phylogenetic trees can inaccurately infer phylodynamics.

The WiPhy model is then used to address the ongoing question of how and how much the host immune system influences within-host HIV evolution. Previously, immune control over viremia has been elegantly demonstrated in nonhuman systems through rapid increases in viral loads following CD8+ T cell depletion (24–27). Yet, the extent of immune pressure on viral evolution is harder to observe directly. Selective pressure and coevolution of CD8+ T cells has been inferred from matching HIV mutations to circulating CD8 epitopes (28–31), computing the ratio of synonymous to nonsynonymous mutations in HIV sequences (32), linking the prevalence of escape mutations to host-genetic predispositions such as HLA type (33) (human leukocyte antigen, genes that regulate immune function), and modeling (19, 34). However, explicit escape from cellular immunity is not always obvious. Using data from 125 adults in the SPARTAC trial (33) Roberts et al. found most individuals had no detectable escape mutants within 2 y of infection. Early mutations (<6 mo after seroconversion) were mostly transmitted, rather than arising from rapid de novo escape in the new host. Even in an HLA-matched host who mounted a measurable and HIV-specific CD8 response, the average time before the targeted epitope evolved an escape mutation was longer than 2 y. Lee et al. also found four clade-C-infected individuals had little indication of cytotoxic T cell-driven immune selections

Significance

HIV evolution within infected individuals creates large barriers to successful vaccination and therapy. Here, we used a model that matches viral loads and mutation rates to characterize the driving forces behind HIV evolution early during infection. Surprisingly, the best model of the data did not require explicit pressure from the host immune system. Instead, the model predicts most new viral variants are intrinsically worse at infecting new cells relative to their parents. Thus, most variants do not persist and only by occasional chance does a new fit variant come to dominate. These findings also highlight the tight connection between viral population dynamics and evolution, warranting more modeling to disentangle these processes in the future.

Author contributions: J.T.S. and D.B.R. designed research; D.A.S. and D.B.R. performed research; M.R. and D.B.R. contributed new reagents/analytic tools; D.A.S. and D.B.R. analyzed data; and D.A.S., M.R., J.T.H., J.T.S., and D.B.R. wrote the paper.

The authors declare no competing interest.

This article is a PNAS Direct Submission.

[1]To whom correspondence may be addressed. Email: dreeves@fredhutch.org.

This article contains supporting information online at http://www.pnas.org/lookup/suppl/doi:10.1073/pnas.2109172119/-/DCSupplemental.

Published February 10, 2022.

in the first year (35). Neutralizing and broadly neutralizing antibodies (bNAbs) also are thought to interact and coevolve with founder viruses over the course of HIV infection (36, 37). Yet, bNAbs can arise quickly without many mutations (38) and in infants (39). Recently, Strauli et al. analyzed data of unprecedented detail on both HIV and antibody repertoire sequences, ultimately finding that HIV/Ab coevolution is at minimum hard to detect, if not rare entirely (40).

In that context, we find that the most parsimonious model of HIV primary infection requires adaptive immunity to control viral load but does not require the adaptive immune system to directly select for certain variants. Sequence evolution can instead be controlled by a distribution of intrinsic viral fitness where most variants are less infectious than their parents. We show the model-predicted viral fitness distributions agree with that of HIV deep mutational scanning (DMS) which quantifies fitness in vitro, necessarily in the absence of immune pressure. By building a model that includes host and virus population dynamics as well as mutation, our work highlights crucial questions about HIV evolution relevant to vaccines and therapeutics.

Results

Viral Dynamics and Phylogenetics during Primary HIV Infection. We sought to identify an optimal model for HIV primary infection phylodynamics. Therefore, we first collected four datasets relevant to early HIV infection: 1) nonlinear viral dynamics during early HIV infection, 2) longitudinal divergence and diversity, 3) post antiretroviral therapy HIV reservoir size and composition in terms of defective and intact sequences, and 4) tree-balance measures (see *SI Appendix*, Table S1 for details and references).

Next, to quantitatively score models against these data, we developed 10 phylodynamic metrics: viral kinetic measures (peak, set point, and set point variability), evolutionary measures (HIV envelope, or env, divergence and diversity on days 20 and 40 after infection), and ratio of intact to all proviral sequences in the HIV reservoir (see *Materials and Methods* and all definitions and values in *SI Appendix*, Table S2). Because no individual dataset had sufficiently granular data for all types, we opted to fit to population data across types. This decision implies the model describes a typical HIV infection. Parameter values are therefore less specific and have higher variance than those that might be estimated by fitting to individuals. Alternatively, tighter individual estimates could be more biased by features of the specific cohort and potentially less reflective of the entire range of HIV infections.

WiPhy Model for HIV Primary Infection. We began with a general stochastic WiPhy model that extends the canonical viral dynamics model (41) by adding latency and adaptive immunity and includes viral variants that mutate (Fig. 1*A*). Each variant corresponds to a unique genotype (signified by an integer g).

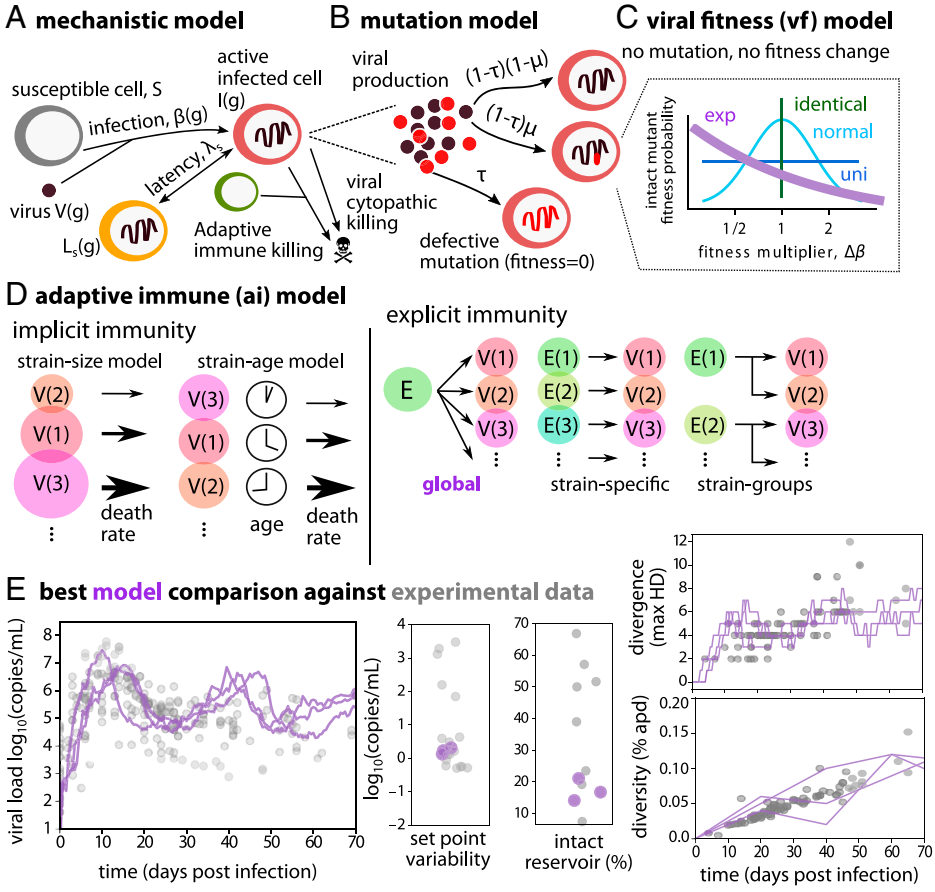

Fig. 1. Mechanistic WiPhy models and the optimal fit to experimental data. (*A*) Mechanistic model schematic. Susceptible cells are infected by viral variant with a genotype $V(g)$, generating new infected cells (latent and active), producing more virus, and engendering immune responses. (*B*) Mutation model governs new variants that are defective (probability τ) or intact. (*C*) If intact, point mutations (probability μ) can occur that change variant fitness (infectivity) based the viral fitness (vf) model—exponential model was optimal. (*D*) Adaptive immune (ai) models were also varied—global was optimal. (*E*) Three stochastic replicate simulations of the best model (exponential-global, purple) against the five types of experimental data (gray, see *SI Appendix*, Table S1 for details on data and cohorts).

Evolution is then tracked by recording the complete genealogy (or transmission record) of these genotypes ($g \rightarrow g' \rightarrow g'' \ldots$). We record attributes for each genotype that allow reconstruction of phylogenetic trees and calculation of evolutionary summary statistics: parent genotype, infectivity, Hamming distance (HD) to the founder sequence, number of each nucleic acids (ACTG), and age.

Upon cell infection, the virus can mutate (probability μ). Most mutants are terminally defective but some variants remain intact (with probability τ) (Fig. 1*B*). New intact variants are given a new genotype (g') and a new infectivity ($\beta_{g'}$) drawn from a distribution (Fig. 1*C*). Throughout, we characterize variant fitness using infectivity. Changes cannot be ascribed to mutation of a certain genomic locus, that is, there is no genotype/phenotype link in the model. However, nucleotide sequences can be reconstructed from the genealogy to enable alignment and phylogenetic tree reconstruction (see Fig. 4). Together, this formulation allowed us to simulate large population sizes (10^9 viruses in 10 mL of blood) with good temporal resolution ($\Delta t = 0.01$ d), compute all phylodynamic metrics, and connect simulated data to phylogenetic trees. All code is freely available at https://github.com/FredHutch/WiPhy_HIV.

Mathematical Model Selection against Phylodynamic Metrics from Primary HIV Infection. To determine the mechanisms required to accurately match experimental data, we attempted to fit 24 distinct mechanistic models. For viral fitness (vf, Fig. 1*C*), we tested four models: all new variants have the same fitness (vf-identical), all new variants have a randomly assigned fitness (vf-random), and two models where variant fitness was inherited, either based upon an exponentially (vf-exp) or normally distributed (vf-normal) change $\Delta\beta$ from its parent sequence. For adaptive immunity (ai, Fig. 1*D*), we tested six models: one with no adaptive immunity; two where immune pressure was implicit, either based upon the size of a certain sequence population (ai-size) or the length of time that a certain sequence existed (ai-age); and three where immune pressure was explicit—meaning a compartment of the model $E(g)$ was added—either based upon a scenario where adaptive immune cells can kill any HIV sequence (ai-global) or adaptive immune cells have a specific cognate genotype which they can only kill (ai-specific) or adaptive immune cells can kill a range of genotypes (ai-groups). In the third category, a typical dynamical model for adaptive immune cells was included that allows adaptive cells to grow and shrink in number based on the commensal infected cell count. Immune cell generation is not limitless, and creation saturates if a certain value is reached (see Eq. **3**).

For each model, 100 parameter sets per model parameter were tested and 20 stochastic replicates were attempted, stopping if 10 were reached. This amounted to 104,497 simulations (or roughly 72 simulation days at 1 min per run). We determined successful models as those with a balanced fit to all phylodynamic metrics: an approximate Bayesian computation (ABC) approach (see *Materials and Methods*) (42). *SI Appendix*, Fig. S1 shows the total scores of the best single stochastic runs, as well as the best parameter sets that fit well across stochastic runs.

The best single run and the best average score across stochastic replicates was achieved by the "exponential-global" model that is governed by an exponential intrinsic fitness distribution for viral offspring and a single adaptive immune compartment that kills globally, i.e., removes all viral strains equally. This model had a normalized residual sum of squares (RSS) two or more points lower than all other models. Regardless of the adaptive immune model, most of the top models had the exponential fitness distribution. The normal viral fitness model with no adaptive immunity at all came in third for the stochastic runs, suggesting it can do well on a given stochastic run, but did not fall into the top five models when averaged across stochastic runs. Individual model traces are compared to data metrics in *SI Appendix*, Fig. S2, which shows that visually several of the top five models appear to fit reasonably well. To go further, in *SI Appendix*, Fig. S3 we illustrate that viral load set point and diversity later than 40 d appear to put the strongest filter on models, with exponential-global outperforming all other models in these categories.

We were particularly interested in why a model in which individual adaptive immune compartments kill specific viral strains—the most literal version of host-on-pathogen selective force—was unsuccessful. *SI Appendix*, Fig. S4 shows that the exponential-strain model was not capable of achieving a low enough viral load set point, suggesting there is a balance between maintaining diversity and set-point level that is hard to achieve through asymmetric immune pressure to certain strains.

In summary, although it appeared that single stochastic runs of several models could perform reasonably well, averaged across stochastic replicates the exponential fitness distribution was optimal for all adaptive immune models. Furthermore, by this quantitative scoring system, global immune pressure was optimal, with total RSS greater than four points lower than the nearest competing models.

Implications of an Optimal Model with Nonspecific Immunity and Inherited Exponential Viral Fitness. Three stochastic replicate simulations of the best model and best parameter set are compared to data in Fig. 1*E*. Individual traces are imperfect but all metrics (viral load peak, nadir and set point, set point variability, sequence divergence, sequence diversity, and intact and defective latent reservoir size) were captured within our tolerance.

Several mechanistic results are implied by the optimal WiPhy model. Requiring inherited fitness indicates that lineages persist by chance beyond single-cell lifetimes. Quantitatively, our model predicts that advantageous mutation occurs in 11% of intact mutations, and intact are only ~5% of all mutations. The average intact mutant has roughly half (0.47) the infectivity of its parent [e.g., a nearly neutral process (43)] such that most lineages die out and leave room for new variants. Importantly, strain-specific adaptive immune pressure was not necessary to capture the major features of within-host HIV phylodynamics in early infection. Neither a model where viral strains implicitly lost fitness over time nor models with explicit strain-specific immunity matched the data as well as one with a broad immune response that killed all variants. Rather than immune-mediated selection sweeps, our model favors constant mutations and fluctuations in intrinsic viral fitness as the primary mechanistic driver of observed HIV evolution during primary infection.

Self-Consistency of Model Selection. To check that the model selection process was robust, we performed a self-consistency exercise. Data simulated by a given model and parameter set were used as the experimental data and the model selection process was repeated. For most models (and most parameter sets), it was possible to correctly select the model that generated the data (*SI Appendix*, Fig. S5). We also assessed the effective dimensionality of model output, finding that a substantial amount (~80%) of model variation is encompassed by two principal components (pc1 and pc2, *SI Appendix*, Fig. S6) and that within these components many models overlap. The relatively low effective dimensionality helps to explain why one (or a few) metrics can provide unique signatures that differentiate between models. The overlap shows how many models can fit reasonably well (though not optimally). Together, these checks emphasize the uniqueness of model output and strengthen confidence in the selection process.

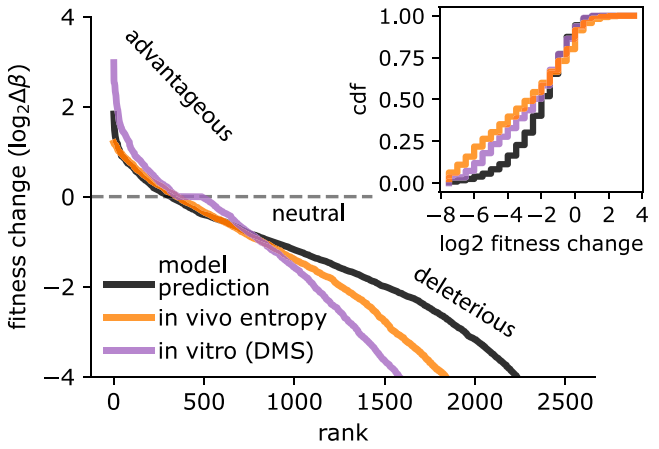

Fig. 2. Model-predicted fitness distribution resembles in vivo data not restricted to primary infection as well as in vitro data without the influence of adaptive immune pressure. The exponential distribution predicted by the model was compared to available in vivo sequence entropy and in vitro DMS data that quantified the relative fitness of all amino acid changes within env. Ranked fitness has similar fractions advantageous. Distributions were not significantly different by paired Kolmogorov–Smirnov tests. (*Inset*) Cumulative distribution functions (cdf).

Model-Estimated Fitness Distribution Resembles In Vivo Fitness Distributions Not Restricted to Primary Infection. Next, our model's prediction that HIV evolution can be explained in the absence of specific adaptive immunity was tested against a different dataset: in vivo entropy distributions from sequences not restricted to primary infection. A reranked Shannon's entropy was employed as a quantitative estimate of the relative effect of reducing fitness after a point mutation. After putting experimental and model-predicted distributions on the same scale (because there is no absolute scale for entropy), there the shape of the two distributions was not significantly different (Fig. 2). Disagreement arose only in ranges of larger fitness costs ($\Delta\beta < 0.25$). Fortunately, this range is less biologically relevant because variants experiencing large fitness costs, regardless of precise value, are subdominant and do not substantially influence data or simulations. Importantly, although our original model validation used primary infection data, the model-derived intrinsic viral fitness distribution also approximated fitness distributions from chronic infection data, suggesting that our conclusions about viral evolution might pertain in other stages of infection.

Model-Estimated Fitness Distribution Resembles In Vitro Fitness Distributions. It is difficult to know whether the viral fitness distribution could be conflating viral fitness and adaptive immune selection. Thus, we next compared the model-estimated distribution with another dataset: DMS of HIV env (44, 45). DMS quantifies the relative fitness of in vitro-generated variants (perturbing nearly all amino acids in env). The distribution resembles the model-estimated fitness distribution and both distributions contain a similar proportion (~10 to 15%) of advantageous mutations (Fig. 2) and distributions were not significantly different (Fig. 2, *Inset*). The largest fitness enhancements in vivo were generally less than those in vitro, such that we cannot rule out that intrinsically fit variants are reduced by immunity. This does not conflict with the model in which adaptive immunity reduces all variants. It also is unclear whether the same variants that are extremely fit in vitro would necessarily succeed as well in vivo for reasons aside from immunity. These data reinforce that it is sufficient to describe HIV phylodynamics during primary infection without including positive selection by adaptive immunity.

Global sensitivity analysis shows adaptive immune response has a strong impact on viral load but limited impact on viral evolution. To study the impact of the adaptive immune compartment further, we performed a global sensitivity analysis by simultaneously varying all parameters of the best model and calculating the Spearman correlation coefficient between all parameters and all summary statistics (*SI Appendix*, Fig. S7*A*). Importantly, of all parameters, the adaptive immune killing rate κ had the strongest impact on the drop to nadir and setpoint, illustrating the importance of the immune system on controlling viral dynamics, especially after a high peak viral load. Parameters regulating the maximal intensity of the immune response (saturation terms for both killing h_g and recruitment h_E) had minimal impact on all metrics compared with other parameters. There was minimal correlation among adaptive immune parameters and phylodynamic measures (diversity and divergence at days 20 and 40) and average infectivity was the strongest determinant of phylodynamic metrics—fitting scores for phylodynamic measures show this pattern even more strongly (*SI Appendix*, Fig. S7*B*). Together, these observations suggest that adaptive immunity's effect on evolution is indirect through viral load modulation. *SI Appendix*, Fig. S7*C* shows predicted connections between population dynamic and phylodynamic measures that cannot be calculated from these data (because metrics are from different individuals). In general, there were strong correlations within population dynamic measures and phylodynamic measures but little correlation between these two broad categories. There was a notable lack of correlation between peak viral load and phylodynamics at or after day 20. While relatively weak, nadir and set-point viral load were correlated with phylodynamics, emphasizing the secondary impact of adaptive immune pressure on evolution through reduced viral load.

Single-Variant Viral Dynamics during Early HIV-1 Infection. The best-fit model was employed to investigate individual variant viral dynamics. First, we tracked variants' viral loads by genotype (Fig. 3*A*). During the first 3 wk of infection there were only 1,000 variants, whereas by day 60 >300,000 productively infectious viral genotypes had been produced, meaning many more defective variants had been created. At approximately day 40, population sweeps appeared (a new variant achieving top abundance) and abundances of concurrent sequences became increasingly even. As explained above, the population sweeps are not caused by strain-specific targeting by the immune system but by continual mutations following the exponential intrinsic viral fitness distribution. Next, realigning variants to their time of emergence (setting $t = 0$ when the variant entered the top 10; Fig. 3*B*) identified two dominant kinetic profiles. The first were variants from before and during peak viremia, which have a large spike of $>10^5$ viral copies (red/yellow); the second were mostly generated after day 60 (when global adaptive immunity was appreciable), which peak at $\sim10^4$ viral copies and slowly decay (blue).

Coloring the variants instead by their HD from the founder virus (Fig. 3*C*) revealed a starlike phylogeny that dominates for approximately the first 40 d—meaning that while many distinct variants have emerged, they are all only one or two mutations away from the founder virus, and that the founder virus remains the mutual common ancestor. A shift from a starlike phylogeny arrives as sequential mutations occur; variants emerge with three or four base pair mutations from the founder around day 50. The predominance of the founder virus for the first 40 d is also evident by examining proportional abundance (calculated as the ratio of each variant viral load to the total; Fig. 3*D*). When the founder loses dominance the other variants are similarly competitive, and thus a more even balance of several variants becomes

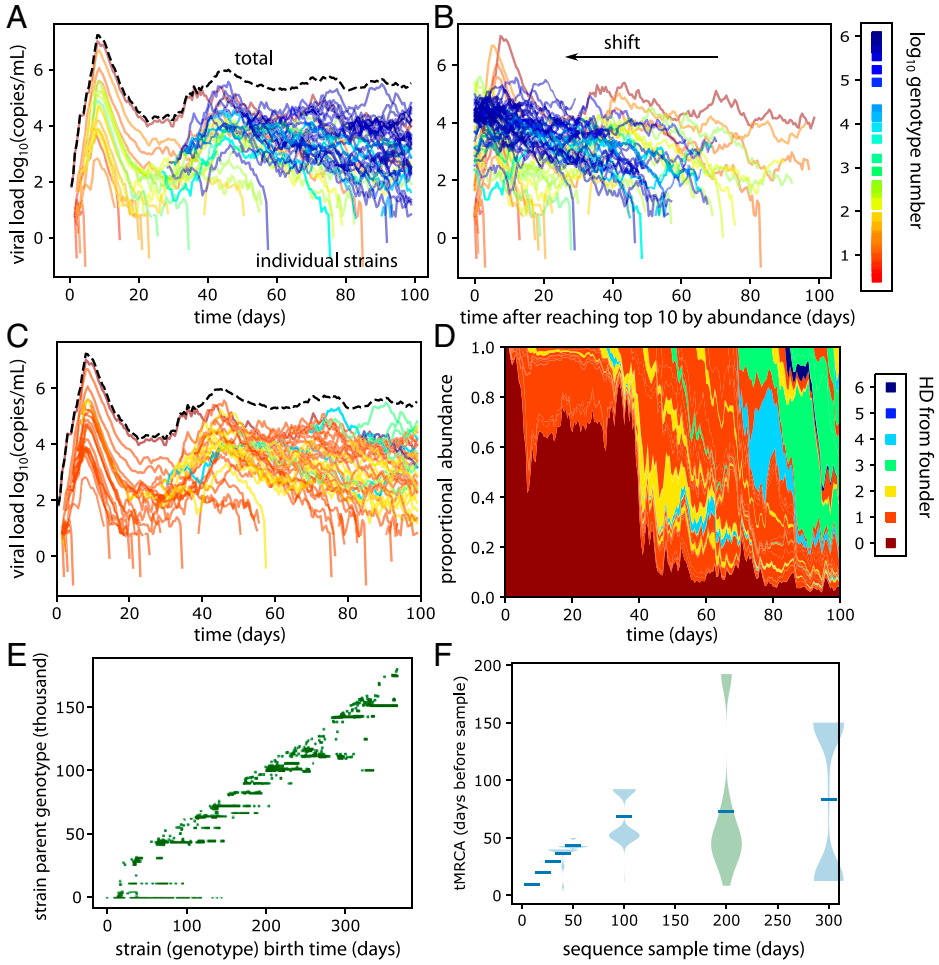

Fig. 3. Visualizing evolutionary dynamics in the optimal model. Example simulation of the best model (variants ever in top 10 and total viral loads). (*A*) Coloring by genotype number illustrates population sweeps and $>10^6$ intact variants; many more defective variants have been created. (*B*) Variant trajectories shifted to the time they entered the top 10 by abundance; variants emerging later in infection have different kinetic profiles than those from early infection (compare red and blue). (*C*) Coloring by HD to founder sequence illustrates most early (red) variants have approximately one point mutation from the founder sequence, whereas later sequential evolution has occurred, with variants emerging with more than two mutations from the founder sequence. (*D*) Proportional abundance colored by HD illustrates the stark shift from founder predominance to more evenness after viral load nadir. (*E*) The complete transmission record, or genealogy illustrates the "true tree"—the parental genotype of each variant created on each day. Certain lineages persist for more than a hundred days, meaning that offspring are generated from a parental sequence that was created months prior. (*F*) The tMRCA of 50 randomly sampled sequences on a given day is bimodal: Variants are created by both more ancestral and more recent parents.

apparent. The timing of these results agrees with independent data showing shifts from demographic to selective effects around day 50 (46).

Highly Granular Simulated Phylogenetic Trees. The ability to access the complete transmission record from these simulations allows examination of evolutionary relationships with high granularity. Fig. 3*E* demonstrates how long certain lineages persist by plotting the parental genotype of each variant sampled on a given day. For example, the founder variant ($g = 0$) and other variants (e.g., $g = 100$) are prolific, producing new direct descendant variants that can be found for months. This timescale far outlasts the lifespan of any single infected cell (~1 d) and this mechanistic model has no adaptive immune selection. Therefore, lineage persistence is a probabilistic balance between viral production and deleterious mutation of offspring. Additionally, some variants do persist at subdominant levels; gaps on the x axis indicate times between which a parental variant was not dominant to the point where its progeny were guaranteed to be sampled.

Calculating the times to most common ancestor (tMRCA) throughout infection for all pairs of subsampled sequences ($n =$ 50 at each time point) revealed a bimodal distribution with cocirculating lineages (Fig. 3*F*). For example, most sampled sequences on day 200 (green) coalesced to common ancestors ~10 to 40 d prior to the sampling date. This represents a time-localized quasispecies that is generated actively by a dominant circulating variant. However, a minority coalesced to more ancestral sequence, representing the continuing impact of prolific early variants. Note bimodality is not driven by latency and reactivation; similar results were found using a model without latent compartments.

Model Validation with Estimated Phylogenetic Trees in the First Year of Infection. Phylogenetic trees are commonly used to illustrate patterns in HIV evolution. We therefore tested whether the selected exponential-global model could show reasonable agreement with another independent dataset, a phylogenetic tree from a highly sampled individual in the first year of infection (47) (p1362, Fig. 4*A*). To accommodate all sources of variability in comparing to the experimental data, three simulations with the best model, three sequence samplings from each simulation, and three tree estimation replicates [in BEAST (48)] were performed on each sample set. This process admits 27 phylogenetic trees

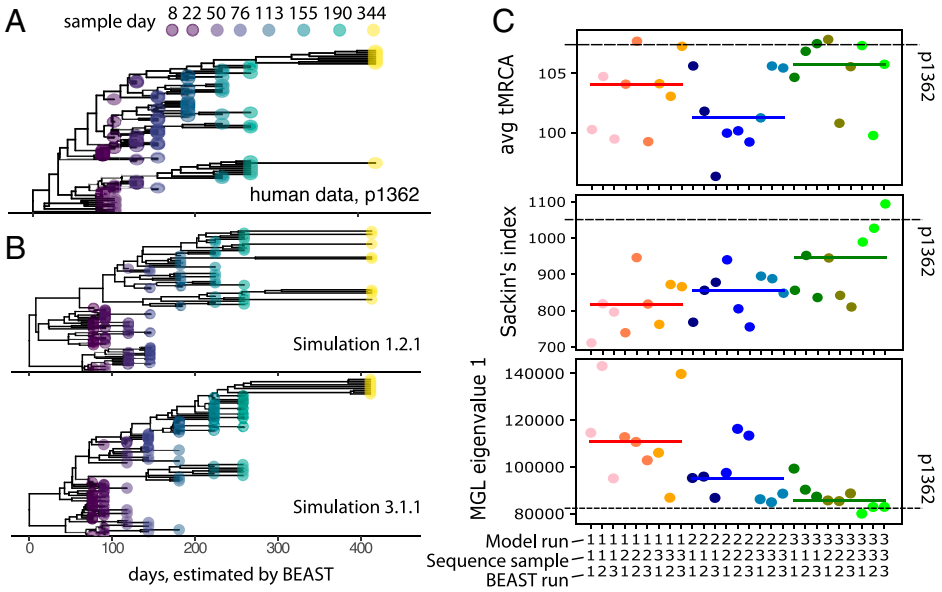

Fig. 4. Comparative analysis of experimental and model tree estimation. (*A*) Experimental tree (C1V2 env, p1362). All sampling schemes are based on this individual. (*B*) Running the best model three times (i), sampling sequences with identical timing and sample size three times (j), and with three tree estimate replicates (k) resulted in 27 trees enumerated i.j.k. Two example simulated trees visually match the experimental tree. (C) Quantitative comparison of trees using phylogenetic summary statistics show some simulations (dots) agree with data (dashed line) and that model run introduces the most variability (solid colored lines are medians across sequence sampling and BEAST run).

(1.1.1 → 3.3.3); two examples are illustrated in Fig. 4*B*, which appeared visually similar to the individual p1362.

To quantitatively compare trees, phylogenetic summary statistics from each simulated tree and the experimental tree were examined (Fig. 4*C*): average tMRCA, Sackin's index (a tree balance statistic calculated as the sum over the number of internal nodes between root and tip for all tips in the tree) (49), and the dominant eigenvalue of the tree's modified graph Laplacian spectrum (MGL). MGL is a robust measurement of tree shape that quantifies deep/shallow branching events and importantly was shown to be a surrogate for synonymous to nonsynonymous (dN/dS) ratio, a metric often used to quantify selection (50). Model run 3 (greens) matched experimental statistics well. Although sampling and tree estimation stochasticity affected summary statistic values, the most significant variability was introduced by rerunning the model—particularly average tMRCA (Fig. 4*C*; horizontal lines show median across sampling and tree estimation).

This process also highlights the potential for misclassifications in the absence of detailed population dynamics. This exercise employed the simplest assumption of constant population size in BEAST. Because viral loads peak early in infection, the tree inference substantially overestimated the distance between the root and the founder sequence: The purple samples observed at day 8 of infection were placed in the maximum clade credibility tree at ∼100 d. Such artifacts might be overcome with more complicated population dynamic models in BEAST. Yet, recent work on birth–death models with time-varying rates showed different scenarios generate the same trees such that scenarios are not distinguishable even with infinite data (51). Another immediate challenge arises from building bifurcating rather than polytomic trees on data from the simulation. The present model allows for a single ancestor to produce many different offspring variants without intermediates (polytomy)—thus additional internal nodes inferred by a bifurcating tree may be artifacts—a point warranting further investigation.

Discussion

By modeling human HIV data including viral population sizes and evolutionary dynamics, we uncovered several important characteristics of HIV pathogenesis. The most parsimonious model was governed by 1) an inherited distribution of viral infectivity drawn from an exponential distribution such that 2) most mutants are less fit than parental sequences. This distribution in turn implies a nearly neutral evolutionary process driven by intrinsic fluctuations in viral fitness. The optimal model also carried an adaptive immune system that was equally potent against all variants, suggesting that 3) although adaptive immunity is needed to control viremia, within-host pressure against specific strains was not needed to accurately model viral evolution.

Together these findings paint a picture of what is sufficient to describe early HIV infection: a viral quasi-species in which a fit variant can dominate or cocirculate with other dominant strains. However, any mutant progeny of currently dominant variants are probabilistically likely to be less fit such that new variants emerge and take over [similar to nearly neutral evolution (43)]. Such population sweeps are sufficiently modeled without any additional pressure from the immune system against specific variants. The imprint of the founder virus is also long-lasting (Fig. 3), which leads to a bimodal distribution of circulating variant sequence age (i.e., there is creation of infected cells by recent and ancestral strains) This finding might be relevant to understanding the discordance of within- and between-host evolutionary rates, but more work is warranted.

Next, using sequence entropy from individuals not necessarily sampled within primary infection, we found our model-estimated distribution was similar, suggesting that although we fit our model to primary infection, this distribution may hold during other stages of infection, and that evolution might be driven by intrinsic fitness in those stages too. Moreover, it might be questioned whether exponentially distributed fitness in the model is effectively modeling adaptive immune selection. Thus, we showed that in vitro DMS data (which guarantees no influence from adaptive immunity) were also relatively similar to the model-estimated distribution (Fig. 2). These results corroborated our hypothesis that much of HIV evolution is controlled at the viral rather than adaptive immune level.

Although our results imply that adaptive immunity to HIV may be broader and less directly influential on evolution than previously imagined, it remains a key component of viral control. This agrees with past experimental work: The timing of CD8+ T cell expansion correlated with reductions in viral loads (52) and depleting CD8+ T cells in SIV infected macaques led to viral expansion (53) [we note other studies show inefficient infected cell killing by CD8+ T cells, suggesting a more nuanced interpretation (54)]. Additionally, because our model effector cell killing rate κ was a strong determinant of viral load setpoint (*SI Appendix*, Fig. S7), we hypothesize that the overall HLA–antigen match (which depends on the specific host and the specific virus) determines disease severity. This agrees with the finding that certain host HLA genotypes are associated with delayed progression to AIDS (55) but that the founder sequence is correlated to pathogenesis (56).

There is strong evidence pointing to selection by adaptive immunity during chronic HIV infection. Observations range from fixed mutations that can be linked to detectable CD8+ T cell responses (31), a dose–response relationship (in one individual) between immune pressure and escape rate (57), an increase over time of escape mutations in HLA-matched hosts relative to HLA-mismatched hosts (33), and the emergence of bNAbs (37). Our results do not invalidate these findings. Instead, in the context of the "red-queen phenomenon" (58), a constant escape and chase, it may be that [as others have observed (33, 35)] sequential viral/host coevolution is not particularly relevant for early HIV pathogenesis. A surprisingly similar message arose from a deep analysis of HIV and antibody repertoires sequenced from the same individuals (40).

Our modeling has several limitations. The magnitude of our modeled adaptive immune response cannot be directly compared to existing values from the various studies because $E(g)$ is not precisely representing any specific cell type (e.g., CD8+ T cells, anti-HIV antibodies, or natural killer cells) and likely only captures the HIV-specific arm of the immune system. The landscape of HIV fitness costs has been modeled in more detail previously (59). Susceptible cells are also not clearly defined phenotypes. In ∼15% of cell infections viral progeny share genetic material from two parental sequences that infected the same cell (60, 61); we do not explicitly simulate such genetic recombination. While explicit modeling of recombination could be added as described previously (13, 62), the present approach effectively allows for some recombination signatures. For example, since all mutational distances are small during early infection, by allowing for many point mutations in a single infection event this could be seen as a recombination. We do not attempt to incorporate compartmental anatomy, instead relying on past studies that show HIV dynamics are reasonably consistent across tissues (63–65). We do not directly model nonsynonymous to synonymous ratio (dN/dS), which has been used to demonstrate selective pressure. However, dN/dS can be tricky to interpret for nonequilibrium scenarios (66) and dN/dS > 1, which implies mutations that meaningfully change proteins (nonsynonymous mutations) are more likely to survive, is not obvious in the first years of HIV infection (32). Additionally, the MGL summary statistic [a surrogate for dN/dS (50)] agreed between our model-derived phylogenetic trees and human trees sampled in the first year of infection.

In building simulated trees (Fig. 4), we also highlighted several challenges of tree estimation from real data in which depth and granularity of sampling is limited. Others have argued that positive selection can be obscured by or conflated with demography (67), have shown misclassification of phylodynamic parameters (51, 68–70), and demonstrated that nonequilibrium population dynamic "jackpot" events can resemble selection (71). Bearing these complexities in mind, we advocate for inclusion of population dynamic data whenever possible and continual enhancement of phylodynamic methods such as ours to disentangle these exquisitely coupled processes in practice.

Future applications of WiPhy models abound from optimizing sampling depth for phylogenetic inference using simulated data, estimating infection timing, and modeling therapies. Our results have important ramifications for vaccine design and therapeutic application of bNAbs to supplement the adaptive immune system. As within-host viral genetic data continue to be collected in treatment and prevention trials, phylodynamic models will be crucial for precise and comprehensive interpretation.

Materials and Methods

Mathematical Description of the Model. The model (Fig. 1) contains cells susceptible to HIV infection S, which are created with rate α_S and die with rate δ_S. HIV infection begins with the introduction of a founder HIV sequence with genotype g as an intact actively infected cell A_g^* (superscript $*$ denotes intactness). Infected cells produce virions and intact virions V_g^* infect new cells with rate $\beta_g S V_g^*$. Unproductive virions $V_g^{()}$ are also produced (hence empty superscript parentheses) from defective active infected cells (see third equation in Eq. **1**) but cannot go on to infect other cells.

When a new cell is infected, mutations occur with rate μ. Given mutation, the proviral sequence is intact with probability τ. A small proportion of infected cells enter one of two latent states $L_{s,g}^{(*)}$ with the small probability λ_s, where subscript s further subdivides latent classes to satisfy observed multiphasic decay patterns (72). Thus, we have three possible infected cell states, which can each be intact or defective—for brevity we express as $\boldsymbol{I}_g^{(*)} = \{A_g^{(*)}, L_{1,g}^{(*)}, L_{2,g}^{(*)}\}$. The rules of the mechanistic model can be approximately expressed as set of differential equations (∂_t denotes time derivative) that grows as genotypes are added. After each time step, new sequences are added by mutation such that $\{g\} \to \{g, g'\}$.

$$\begin{aligned} \partial_t S &= \alpha_S - \delta_S S - \Sigma_g \beta_g V_g^* S \\ \partial_t \boldsymbol{I}_g^{(*)} &= \boldsymbol{b}_s(g, \tau, \mu, \lambda_s) V_g^* S - \boldsymbol{d}_s\left(\boldsymbol{I}_g^{(*)}\right) \boldsymbol{I}_g^{(*)} \\ \partial_t V_g^{(*)} &= \pi A_g^{(*)} - \gamma V_g^{(*)} \end{aligned} \quad [1]$$

The generic creation and removal rates of each type of infected cell is governed by the birth and death vectors $\boldsymbol{b}_s$ and $\boldsymbol{d}_s$ such that, for example,

$$\boldsymbol{b}_{s=A} = \beta_g [\lambda_A \tau \mu,\ \lambda_A (1-\tau)\mu,\ \lambda_A \tau (1-\mu),\ \lambda_A (1-\tau)(1-\mu)] \quad [2]$$

represents the rate of creation of active cells of four types: defective mutated, intact mutated, defective nonmutated, and intact nonmutated. There are copies of these birth and death vectors for each overall infected cell state $s \in \{A, L_1, L_2\}$.

The removal rate of each type of infected cells depends on their state, intactness, and genotype $\boldsymbol{d}_s\left(\boldsymbol{I}_g^{(*)}\right)$, and the rate itself can also be different functions of the number of cells of that state. These rules vary in each of the adaptive immune (ai) models described below. Additionally, latently infected cells proliferate (added to the birth vector with rate α_s for all intactness/genotypes) and die (added to the death vector with rate δ_s for all intactness/genotypes) and reactivate to an active state (added to the death vector with rate ξ_s for intact genotypes).

To model mutational changes, we modify the HDs of mutated sequences by drawing a Poisson distributed number of nucleotide changes $\wp(\Delta g)$. For intact mutants we use an average of one nucleotide substitution $H_{g'} = H_g + \wp(\Delta g; 1)$ and for defectives—typically generated through APOBEC hypermutation or large insertions/deletions—we use $H_{g'} = H_g + \wp(\Delta g; 40)$, where 40 is the average number of base pair changes for DNA (73).

Viral Fitness (vf) Models. We created four models for the viral fitness (vf) of mutated intact sequences (Fig. 1*C*). The first is a trivial model where each viral strain has the same fitness. Thus, $p(\Delta\beta) = d(\Delta\beta)$, where d is the Dirac delta function equal to zero unless $\Delta\beta = 1$. The second assumes no inheritance from parental strains such that fitness changes are uniformly distributed up to a maximum value $p(\Delta\beta) = U[0, \beta_{max}]$. The third and fourth models assume heritability of viral fitness, either with an exponential distribution $p(\Delta\beta) = \exp(-\Lambda\Delta\beta)/\Lambda$ with rate Λ or a Gaussian distribution $p(\Delta\beta) = \mathcal{N}(1, \sigma_\beta)$. However, in both cases we also enforce the constraint that $\beta_{g'} \in [0, \beta_{max}]$. From a biological point of view, these models encompass a broad range of possibilities for phenotypic variation, ranging from simplest (constant), to most complicated (normal and exponential) that have symmetric or asymmetric fitness (74). These choices are also justified by maximum entropy distributions (75).

Adaptive Immune (ai) Models. The genotype-dependent death rate of actively infected cells $d_A(A_g)$ is used to incorporate six models of the ai response. The first model has no adaptive immunity such that $d_A(A_g) = \delta_A$.

The next two models have "implicit immunity," meaning that there are no additional compartments to represent immune cells. In the strain-size model, $d_A(A_g) = \phi_A \frac{A_g}{A_g + h_A}$. We interpret this to mean that the number of actively infected cells with viral genotype g attracts immune cells relative to that genotype's abundance. More abundant genotypes are removed faster. Rate ϕ_A is the maximum and h_A parameterizes maximal rate saturation. In the strain-age model infected cell death depends on genotype age, $d_A(A_g) = \delta_A \exp[\kappa_a(t - a_g)]$. This can be interpreted to mean older sequences have had more time to accrue adaptive immune pressure and thus are eliminated more rapidly—a mechanism which enforces strain replacement based on magnitude of rate constant κ_a.

The remaining versions explicitly model immunity. We add a state variable compartment representing effector cells $E_g(t)$ governed by

$$\partial_t E_g = \omega \frac{A_g}{A_g + h_g} E_g - \delta_E E_g - \phi_E \frac{\sum_g E_g}{\sum_g E_g + h_E}. \quad [3]$$

This part of the biology is the least understood and our mechanistic implementations may effectively capture several types of cells or molecules (CD8+ T cells, NK cells, and antibodies). We draw inspiration for construction from our prior work and published models of immune systems in viral dynamics (76–79).

In strainwise immune models, effector cells have their own genotype g which matches a viral genotype. Then, immune cells grow based on the prevalence of their cognate antigenic genotype (term with nonlinear growth rate ω and saturation constant h_g), die naturally with rate δ_E, and have another death term such that the total adaptive immune response (sum over genotypes) is constrained in size (term with saturation constant h_E). In the global model, $d_A(A_g) = [\delta_A + \kappa \sum_g E_g]$. We interpret this to imply that there is a single adaptive immune compartment that can kill any strain. In the strain-specific model, $d_A(A_g) = [\delta_A + \kappa_g E_g]$. We interpret this to imply that for each viral strain there is an adaptive immune compartment that can kill only that strain. The killing ability of each strain-specific adaptive immune compartment is κ_g. In the strain-group model, $d_A(A_g) = [\delta_A + \kappa_G \sum_{g \in G_i} E_g]$. We interpret this to mean there is some cross-immunity such that sequences with similar sequence numbers (within a group G_i where each group has the same size G) can be killed by a single immune compartment with killing rate κ_G which is assumed the same for all groups. The death rate of latently infected cells is simpler, and rates are not dependent on values, instead $\boldsymbol{d}_{L_s} = [\delta_s, \delta_s + \xi_s, \delta_s, \delta_s + \xi_s]$ such that intact cells die slightly faster in accordance with observed values.

Parameters from the Literature. By using previous estimates, we constrained the parameters that must be estimated. All information on fixed parameters, initial conditions, and fit parameter ranges is contained in *SI Appendix*, Table S3. Three parameters are estimated in all models, and different models have further parameters that must be estimated such that results range from three to nine estimated parameters.

Simulation Implementation. The model is implemented in C++ and is freely available (https://github.com/FredHutch/WiPhy_HIV). We use a discrete stochastic τ-leap simulation scheme in 10 mL of plasma and a simulation time interval of $\Delta t = 0.01$ d. The state variables $\boldsymbol{X} = \left\{S, \boldsymbol{I}_g^{(*)}, V_g^{(*)}, E \ldots\right\}$ represent the numbers of susceptible, active/latent infected cells and virions (for each genotype) and adaptive immunity if explicitly in the model. Thus, in each time interval, a Poisson-distributed number of events of each mechanistic transition is chosen by the reaction propensities $p_{\boldsymbol{X}}$ (80) such that $\boldsymbol{e} = \wp(p_{\boldsymbol{X}} \Delta t)$. Then, the state variables are updated using the event transition matrix $\boldsymbol{T}$ as $\Delta \boldsymbol{X} = \boldsymbol{eT}$. For example, in an interval we might observe the creation of a new latently infected cell of a new genotype by viral infection. For this example, $T = \left[S - 1, \ldots, L^*_{1,g'} + 1, \ldots, V^*_g - 1\right]$, meaning removal of a susceptible cell, removal of an intact virion of genotype g, and the creation of an intact first phase latently infected cell with genotype g'. In this same interval many other events could occur simultaneously.

Tracking the Complete Transmission Record (Genealogy). To capture evolution, each viral strain is given a genotype number (an integer g). This number specifies an fitness/infectivity (β_g), the number of base-pair mutations for this genotype relative to the founder virus (the HD H_g), its age (a_g the date of its emergence in time since the start of infection), and the number of each nucleic acids (n_x where $x \in \{A, C, T, G\}$ and the initial number is taken from the reference HXB2 sequence). The number of each state variable (e.g., infected cells) associated to that genotype is recorded at each time step. A substantial computational enhancement was achieved by tracking population size but not attributes and transmission records for defective variants. This choice is valuable because hypermutants and/or large deletions are typically removed before analysis of experimental data, but modeling the number of defective sequences was crucial to accurately populate the latent reservoir, which is well known to be predominantly defective (81).

Model Fitting Procedure. The best parameterization of each model was achieved by testing $k \times 100$ values of each parameter, where the number of model parameters is k. This approach means that for a model with eight parameters, a total of $800 \times 8 = 6{,}400$ parameterizations were tested, i.e., more complex models had more opportunities to find an optimum. Values were drawn from a grid search evenly spaced between a lower and upper bound for each parameter (often several orders of magnitude) based on previously determined HIV model rates. Because the model is stochastic, we attempted 20 replicate simulations for each parameter set, stopping if 10 replicates were successful. Each replicate was scored by computing the model value of each metric (m_i) and computing a variance-normalized residual sum of squares (RSS_i, also called the χ^2 statistic) against the metric from the data: $RSS_i = (m_i - \bar{M}_i)^2 / var(M_i)$, where M_i is the experimentally determine metric; the overbar denotes the mean and *var* denotes the variance or squared SD. We also calculated the total RSS as the sum $RSS = \sum_i RSS_i$.

Model Selection. Because we fit to correlated metrics and combined data sources, we ruled out typical model selection procedures based on likelihoods and information criteria (e.g., Akaike information criterion). Instead, we applied an ABC approach. The normalized RSS was calculated for each metric and runs with all individual metrics fitting reasonably well ($RSS_i < 5 \; \forall \; i$) were included. *SI Appendix*, Fig. S1 shows the RSS summed over individual metrics. We determined the best single run for the best parameter set, for each model. Then, to pick models that consistently work well, we ultimately accepted model parameterizations for which model output averaged across stochastic runs was within our RSS tolerance.

Global Sensitivity Analysis. Using the best model, we quantified the influence of model parameters on all metrics, and on the RSS error of all metrics, using global sensitivity analysis by calculating the Spearman correlation coefficient (*SI Appendix*, Fig. S7) (82).

Modeling Comparison to Entropy Distributions. We obtained filtered HIV-1 env alignments (type M without recombinants) from the LANL database (https://www.hiv.lanl.gov/content/index) and removed all but subtype B sequences resulting in 2,339 sequences. Entropy was calculated with default options on the LANL website and gaps are removed to resolve the consensus sequence and its entropy values. The relative abundance of each base b at each position ψ in the env is expressed such that a perfectly even distribution at some position is written $p_\psi(b) = [0.25, 0.25, 0.25, 0.25]$, whereas a perfectly uneven distribution at some position is written $p_\psi(b) = [0, 1, 0, 0]$, which represents that a single base (e.g., *T*) is found at that position for all individuals in the database. We calculated Shannon's entropy $\mathcal{S}_\psi = -\sum_b p_\psi(b) \log p_\psi(b)$ for each position. Next, because our model is agnostic to nucleotide-position-specific biology, we reranked entropy from most to least variable. We then use the distribution of entropy as a quantitative estimate of the relative effect of reducing fitness after a point mutation to positions in env. We then identified the factor y that would scale entropy (assumed constant over position) such that $y\mathcal{S}$ most closely resembled our best-fit viral-fitness distribution $p(\Delta\beta)$. We minimized the RSS between data and model distributions to find $y \sim 1.5$.

Sequence Sampling. To simulate sampling, we randomly select virions (recapitulating experimental sampling of viral RNA) from the simulation. At the time intervals matching the experimental data, cells are computationally sampled from the present virus until a given number of sequences are represented or until all active sequences are represented if the actual number is less.

Calculating tMRCA. Time to most recent common ancestor (Fig. 3) was calculated by examining all sequence pairs and determining their parental sequence. If the parent is identical, the procedure halts and the birth date of the parent is recorded as tMRCA. If the parent is different, we track back to parents of the parental sequences and repeat.

Calculation of Divergence and Diversity. Because we do not track nucleotide sequences, to calculate the number of base-pair differences between a pair of sequences we compute the sum of their HDs and subtract off the HD of their

parental sequence $\Delta(i,j) = H_i + H_j - H_{P(i,j)}$. The divergence is calculated as the $\max\Delta(i,j)$ where $i = 0$ is the founder sequence. Diversity is calculated as the average pairwise distance (83):

$$\nabla = \frac{2}{l}\Sigma_{i=2}^{N}\Sigma_{j=2}^{N} f_i f_j \Delta(i,j), \quad [4]$$

where the frequency of each sampled variant is $f_i = \frac{N_i}{N}$, where N is the total sample size and the number of nucleotides l is the length of HIV env.

Integration with BEAST. To harmonize simulation output with phylogenetic inference tools, we used genealogies and HDs to output a list of sampled nucleotide sequences. We applied an HKY nucleotide substitution model beginning with the HXB2 reference and keeping a fixed length genome to export a time labeled FASTA file which can be input to BEAST. Note that this coarse post facto site model could be expanded to include insertions and deletions but not recombination currently. Our model allowed for HKY variable frequency of transitions and transversions such that forward simulation and backward inference was congruent. We chose a strict molecular clock and a fixed population size and ensured convergence by testing different burn-in sizes. An example XML file in which BEAST settings can be found is provided within our GitHub repository.

Data Availability. All code is freely available at GitHub (https://github.com/FredHutch/WiPhy_HIV). Previously published data were used for this work (all citations for all data are given in the text and *SI Appendix*, Table S1). All other study data are included in the article and/or *SI Appendix*.

ACKNOWLEDGMENTS. This work was funded by a Washington Research Foundation postdoctoral fellowship and a National Institute of Allergy and Infectious Diseases K25 (AI155224) to D.B.R. D.B.R. is grateful to numerous colleagues for conversations including F. Boshier, A. Dingens, T. Bedford, B. Dearlove, E. Lewitus, A. Feder, P. Roychoudury, P. Edlefsen, and J. Mullins.

1. J. Felsenstein, Evolutionary trees from DNA sequences: A maximum likelihood approach. *J. Mol. Evol.* **17**, 368–376 (1981).
2. E. M. Volz, S. L. Kosakovsky Pond, M. J. Ward, A. J. Leigh Brown, S. D. W. Frost, Phylodynamics of infectious disease epidemics. *Genetics* **183**, 1421–1430 (2009).
3. T. Stadler, O. G. Pybus, M. P. H. Stumpf, Phylodynamics for cell biologists. *Science* **371**, eaah6266 (2021).
4. S. M. Ciupe, J. M. Heffernan, In-host modeling. *Infect. Dis. Model.* **2**, 188–202 (2017).
5. B. T. Grenfell *et al.* Unifying the epidemiological and evolutionary dynamics of pathogens. *Science* **303**, 327–332 (2004).
6. S. Alizon, C. Magnus, Modelling the course of an HIV infection: Insights from ecology and evolution. *Viruses* **4**, 1984–2013 (2012).
7. A. J. Kucharski, V. Andreasen, J. R. Gog, Capturing the dynamics of pathogens with many strains. *J. Math. Biol.* **72**, 1–24 (2016).
8. M. A. Nowak *et al.*, Antigenic diversity thresholds and the development of AIDS. *Science* **254**, 963–969 (1991).
9. L. M. Wahl, M. A. Nowak, Adherence and drug resistance: Predictions for therapy outcome. *Proc. Biol. Sci.* **267**, 835–843 (2000).
10. D. C. Nickle *et al.*, Evolutionary indicators of human immunodeficiency virus type 1 reservoirs and compartments. *J. Virol.* **77**, 5540–5546 (2003).
11. H. Y. Lee *et al.*, Modeling sequence evolution in acute HIV-1 infection. *J. Theor. Biol.* **261**, 341–360 (2009).
12. C. L. Ball, M. A. Gilchrist, D. Coombs, Modeling within-host evolution of HIV: Mutation, competition and strain replacement. *Bull. Math. Biol.* **69**, 2361–2385 (2007).
13. E. E. Giorgi, B. T. Korber, A. S. Perelson, T. Bhattacharya, Modeling sequence evolution in HIV-1 infection with recombination. *J. Theor. Biol.* **329**, 82–93 (2013).
14. T. T. Immonen, J. M. Conway, E. O. Romero-Severson, A. S. Perelson, T. Leitner, Recombination enhances HIV-1 envelope diversity by facilitating the survival of latent genomic fragments in the plasma virus population. *PLoS Comput. Biol.* **11**, e1004625 (2015).
15. S. Moreno-Gamez *et al.*, Imperfect drug penetration leads to spatial monotherapy and rapid evolution of multidrug resistance. *Proc. Natl. Acad. Sci. U.S.A.* **112**, E2874–E2883 (2015).
16. S. Luo, A. S. Perelson, The challenges of modelling antibody repertoire dynamics in HIV infection. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* **370**, 20140245–20140247 (2015).
17. N. M. Dixit, A. S. Perelson, HIV dynamics with multiple infections of target cells. *Proc. Natl. Acad. Sci. U.S.A.* **102**, 8198–8203 (2005).
18. C. H. van Dorp, M. van Boven, R. J. de Boer, Modeling the immunological preadaptation of HIV-1. BioRxiv [Preprint] (2020). https://www.biorxiv.org/content/10.1101/2020.01.08.897983v1 (Accessed 22 November 2021).
19. V. V. Ganusov *et al.*, Fitness costs and diversity of the cytotoxic T lymphocyte (CTL) response determine the rate of CTL escape during acute and chronic phases of HIV infection. *J. Virol.* **85**, 10518–10528 (2011).
20. H. M. Doekes, C. Fraser, K. A. Lythgoe, Effect of the latent reservoir on the evolution of HIV at the within- and between-host levels. *PLOS Comput. Biol.* **13**, e1005228-27 (2017).
21. F. Zanini, R. A. Neher, FFPopSim: An efficient forward simulation package for the evolution of large populations. *Bioinformatics* **28**, 3332–3333 (2012).
22. B. C. Haller, P. W. Messer, SLiM 3: Forward genetic simulations beyond the Wright-Fisher model. *Mol. Biol. Evol.* **36**, 632–637 (2019).
23. A. Jariani *et al.*, SANTA-SIM: Simulating viral sequence evolution dynamics under selection and recombination. *Virus Evol.* **5**, vez003 (2019).
24. J. E. Schmitz *et al.*, Control of viremia in simian immunodeficiency virus infection by CD8+ lymphocytes. *Science (80-)* **283**, 857–860 (1999).
25. B. Asquith, A. R. McLean, In vivo CD8+ T cell control of immunodeficiency virus infection in humans and macaques. *Proc. Natl. Acad. Sci. U.S.A.* **104**, 6365–6370 (2007).
26. E. K. Cartwright *et al.*, CD8(+) Lymphocytes are required for maintaining viral suppression in SIV-infected macaques treated with short-term antiretroviral therapy. *Immunity* **45**, 656–668 (2016).
27. Y. Cao, E. K. Cartwright, G. Silvestri, A. S. Perelson, CD8+ lymphocyte control of SIV infection during antiretroviral therapy. *PLoS Pathog.* **14**, e1007350 (2018).
28. P. Borrow *et al.*, 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 (1997).
29. D. A. Price *et al.*, Positive selection of HIV-1 cytotoxic T lymphocyte escape variants during primary infection. *Proc. Natl. Acad. Sci. U.S.A.* **94**, 1890–1895 (1997).
30. N. Goonetilleke *et al.*, CHAVI Clinical Core B, The first T cell response to transmitted/founder virus contributes to the control of acute viremia in HIV-1 infection. *J. Exp. Med.* **206**, 1253–1272 (2009).
31. T. M. Allen *et al.*, Selective escape from CD8+ T-cell responses represents a major driving force of human immunodeficiency virus type 1 (HIV-1) sequence diversity and reveals constraints on HIV-1 evolution. *J. Virol.* **79**, 13239–13249 (2005).
32. F. Zanini *et al.*, Population genomics of intrapatient HIV-1 evolution. *eLife* **4**, e11282 (2015).
33. H. E. Roberts *et al.*, SPARTAC trial investigators, Structured observations reveal slow HIV-1 CTL escape. *PLoS Genet.* **11**, e1004914 (2015).
34. H. R. Fryer *et al.*, SPARTAC Trial Investigators, Modelling the evolution and spread of HIV immune escape mutants. *PLoS Pathog.* **6**, e1001196 (2010).
35. G. Q. Lee *et al.*, HIV-1 DNA sequence diversity and evolution during acute subtype C infection. *Nat. Commun.* **10**, 2737 (2019).
36. P. L. Moore *et al.*, Evolution of an HIV glycan-dependent broadly neutralizing antibody epitope through immune escape. *Nat. Med.* **18**, 1688–1692 (2012).
37. H. X. Liao *et al.*, NISC Comparative Sequencing Program, Co-evolution of a broadly neutralizing HIV-1 antibody and founder virus. *Nature* **496**, 469–476 (2013).
38. N. A. Doria-Rose *et al.*, NISC Comparative Sequencing Program, Developmental pathway for potent V1V2-directed HIV-neutralizing antibodies. *Nature* **509**, 55–62 (2014).
39. L. Goo, V. Chohan, R. Nduati, J. Overbaugh, Early development of broadly neutralizing antibodies in HIV-1-infected infants. *Nat. Med.* **20**, 655–658 (2014).
40. N. Strauli *et al.*, The genetic interaction between HIV and the antibody repertoire. BioRxiv [Preprint] (2019). https://www.biorxiv.org/content/10.1101/646968v2 (Accessed 22 November 2021).
41. A. S. Perelson, A. U. Neumann, M. Markowitz, J. M. Leonard, D. D. Ho, HIV-1 dynamics in vivo: Virion clearance rate, infected cell life-span, and viral generation time. *Science (80-)* **271**, 1582–1586 (1996).
42. T. Toni, D. Welch, N. Strelkowa, A. Ipsen, M. P. H. Stumpf, Approximate Bayesian computation scheme for parameter inference and model selection in dynamical systems. *J. R. Soc. Interface* **6**, 187–202 (2009).
43. T. Ohta, The nearly neutral theory of molecular evolution. *Annu. Rev. Ecol. Syst.* **23**, 263–286 (1992).
44. H. K. Haddox, A. S. Dingens, J. D. Bloom, Experimental estimation of the effects of all amino-acid mutations to HIV's envelope protein on viral replication in cell culture. *PLoS Pathog.* **12**, e1006114 (2016).
45. H. K. Haddox, A. S. Dingens, S. K. Hilton, J. Overbaugh, J. D. Bloom, Mapping mutational effects along the evolutionary landscape of HIV envelope. *eLife* **7**, 1–29 (2018).
46. J. T. Herbeck *et al.*, Demographic processes affect HIV-1 evolution in primary infection before the onset of selective processes. *J. Virol.* **85**, 7523–7534 (2011).
47. Y. Liu *et al.*, Selection on the human immunodeficiency virus type 1 proteome following primary infection. *J. Virol.* **80**, 9519–9529 (2006).
48. A. J. Drummond, A. Rambaut, B. Shapiro, O. G. Pybus, Bayesian coalescent inference of past population dynamics from molecular sequences. *Mol. Biol. Evol.* **22**, 1185–1192 (2005).
49. S. D. Frost, E. M. Volz, Modelling tree shape and structure in viral phylodynamics. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* **368**, 20120208 (2013).
50. E. Lewitus, M. Rolland, A non-parametric analytic framework for within-host viral phylogenies and a test for HIV-1 founder multiplicity *Virus Evol.* **5**, 1–13 (2019).
51. S. Louca, M. W. Pennell, Extant timetrees are consistent with a myriad of diversification histories. *Nature* **580**, 502–505 (2020).
52. K. R. Demers *et al.*, Temporal dynamics of CD8+ T cell effector responses during primary HIV infection. *PLoS Pathog.* **12**, e1005805 (2016).
53. J. E. Schmitz *et al.*, Control of viremia in simian immunodeficiency virus infection by CD8+ lymphocytes. *Science* **283**, 857–860 (1999).

54. B. Asquith, C. T. T. Edwards, M. Lipsitch, A. R. McLean, Inefficient cytotoxic T lymphocyte-mediated killing of HIV-1-infected cells in vivo. *PLoS Biol.* **4**, e90 (2006).
55. F. Pereyra *et al.*, The major genetic determinants of HIV-1 control affect HLA class I peptide presentation. *Science* **330**, 1551–1557 (2010).
56. C. Fraser *et al.*, Virulence and pathogenesis of HIV-1 infection: An evolutionary perspective. *Science* **343**, 1243727 (2014).
57. M. R. Henn *et al.*, Whole genome deep sequencing of HIV-1 reveals the impact of early minor variants upon immune recognition during acute infection. *PLoS Pathog.* **8**, e1002529 (2012).
58. A. Nourmohammad, J. Otwinowski, J. B. Plotkin, Host-pathogen coevolution and the emergence of broadly neutralizing antibodies in chronic infections. *PLoS Genet.* **12**, e1006171-23 (2016).
59. A. L. Ferguson *et al.*, Translating HIV sequences into quantitative fitness landscapes predicts viral vulnerabilities for rational immunogen design. *Immunity* **38**, 606–617 (2013).
60. R. A. Neher, T. Leitner, Recombination rate and selection strength in HIV intra-patient evolution. *PLoS Comput. Biol.* **6**, e1000660 (2010).
61. T. E. Schlub *et al.*, Fifteen to twenty percent of HIV substitution mutations are associated with recombination. *J. Virol.* **88**, 3837–3849 (2014).
62. H. Song *et al.*, Tracking HIV-1 recombination to resolve its contribution to HIV-1 evolution in natural infection. *Nat. Commun.* **9**, 1928 (2018).
63. R. D. Hockett *et al.*, Constant mean viral copy number per infected cell in tissues regardless of high, low, or undetectable plasma HIV RNA. *J. Exp. Med.* **189**, 1545–1554 (1999).
64. S. von Stockenstrom *et al.*, Longitudinal genetic characterization reveals that cell proliferation maintains a persistent HIV type 1 DNA pool during effective HIV therapy. *J. Infect. Dis.* **212**, 596–607 (2015).
65. R. Rose *et al.*, HIV maintains an evolving and dispersed population in multiple tissues during suppressive combined antiretroviral therapy in individuals with cancer. *J. Virol.* **90**, 8984–8993 (2016).
66. S. Kryazhimskiy, J. B. Plotkin, The population genetics of dN/dS. *PLoS Genet.* **4**, e1000304 (2008).
67. C. Bank, G. B. Ewing, A. Ferrer-Admettla, M. Foll, J. D. Jensen, Thinking too positive? Revisiting current methods of population genetic selection inference. *Trends Genet.* **30**, 540–546 (2014).
68. D. Kühnert, T. Stadler, T. G. Vaughan, A. J. Drummond, Simultaneous reconstruction of evolutionary history and epidemiological dynamics from viral sequences with the birth-death SIR model. *J. R. Soc. Interface* **11**, 20131106 (2014).
69. T. Stadler *et al.*, Swiss HIV Cohort Study, Estimating the basic reproductive number from viral sequence data. *Mol. Biol. Evol.* **29**, 347–357 (2012).
70. E. Saulnier, O. Gascuel, S. Alizon, Inferring epidemiological parameters from phylogenies using regression-ABC: A comparative study. *PLos Comput. Biol.* **13**, e1005416 (2017).
71. O. Hallatschek, Selection-like biases emerge in population models with recurrent jackpot events. *Genetics* **210**, 1053–1073 (2018).
72. D. B. Reeves *et al.*, A majority of HIV persistence during antiretroviral therapy is due to infected cell proliferation. *Nat. Commun.* **9**, 4811 (2018).
73. J. M. Cuevas, R. Geller, R. Garijo, J. López-Aldeguer, R. Sanjuán, Extremely high mutation rate of HIV-1 in vivo. *PLoS Biol.* **13**, e1002251 (2015).
74. E. Bons, F. Bertels, R. R. Regoes, Estimating the mutational fitness effects distribution during early HIV infection. *Virus Evol.* **4**, vey029 (2018).
75. E. T. Jaynes, Information theory and statistical mechanics. *Phys. Rev.* **106**, 620–630 (1957).
76. J. M. Conway, A. S. Perelson, Post-treatment control of HIV infection. *Proc. Natl. Acad. Sci. U.S.A.* **112**, 5467–5472 (2015).
77. D. B. Reeves, C. W. Peterson, H. P. Kiem, J. T. Schiffer, Autologous stem cell transplantation disrupts adaptive immune responses during rebound simian/human immunodeficiency virus viremia. *J. Virol.* **91**, e00095-17 (2017).
78. R. J. De Boer, A. S. Perelson, Target cell limited and immune control models of HIV infection: A comparison. *J. Theor. Biol.* **190**, 201–214 (1998).
79. D. B. Reeves *et al.*, Timing HIV infection with a simple and accurate population viral dynamics model. *J. R. Soc. Interface* **18**, 20210314 (2021).
80. M. Voliotis, P. Thomas, R. Grima, C. G. Bowsher, Stochastic simulation of biomolecular networks in dynamic environments. *PLoS Comput. Biol.* **12**, e1004923 (2016).
81. K. M. Bruner *et al.*, Defective proviruses rapidly accumulate during acute HIV-1 infection. *Nat. Med.* **22**, 1043–1049 (2016).
82. S. M. Blower, H. Dowlatabadi, Sensitivity and uncertainty analysis of complex models of disease transmission: An HIV model, as an example. *Int. Stat. Rev./Rev. Int. Stat.* **62**, 229 (1994).
83. M. Nei, W. H. Li, Mathematical model for studying genetic variation in terms of restriction endonucleases. *Proc. Natl. Acad. Sci. U.S.A.* **76**, 5269–5273 (1979).