, YEAR, , 1–4
DOI added during production
Published: Date added during production
Paper

wavess 1.2: Presenting an HLA-aware within-host virus sequence simulation framework

Zena Lapp[1] **and Thomas Leitner**[1,*]

[1]Theoretical Biology and Biophysics, Los Alamos National Laboratory, New Mexico, United States of America
*Corresponding author. tkl@lanl.gov

Abstract

Motivation: Understanding how virus sequences are shaped by selection can inform vaccine design and transmission inference. Modeling within-host evolution to interrogate these questions requires a detailed mechanistic framework that accurately captures sequence diversification. The $CD8^+$ cytotoxic T-lymphocyte (CTL) response plays an important role in immune-mediated selection and can leave strong signatures in virus sequences; however, existing sequence-based within-host virus modeling frameworks do not explicitly include an HLA-aware CTL response.
Results: We extended our previously published within-host sequence evolution simulator, `wavess`, to include an explicit CTL response, and share a method for identifying HLA-specific CTL epitopes given a founder virus sequence. We also updated the model to permit a variable recombination rate, which allows for modeling recombination hotspots, non-adjacent genes, and segmented genomes. These extensions to `wavess` allow for more accurate simulation of viruses and virus genes, particularly in regions of the genome where the immune response is dominated by CTLs (rather than antibodies). It also provides the foundation for investigations of how these newly-added biological mechanisms influence within-host evolution.
Availability and implementation: The core of `wavess` is written in Python 3, with helper functions written in R. It is available at https://github.com/MolEvolEpid/wavess.

Keywords molecular evolution, immunology, recombination, simulation, virology

Introduction

Forward-in-time simulation of virus sequence evolution in a host can provide fundamental knowledge relevant to public health. It can help tease apart how different selective pressures influence virus survival, which may inform vaccine development. It can also improve our understanding of how to correctly infer transmission networks from sequences under strong selective pressures, which may inform inference tools used to suggest interventions. To investigate these questions, the model must accurately simulate mutational signatures in virus sequences.

An important selective pressure that shapes the virus fitness landscape, and therefore sequence signatures, is the $CD8^+$ cytotoxic T-lymphocyte (CTL) immune response (Carlson et al., 2015). CTLs recognize non-self peptides presented by human leukocyte antigen (HLA) class I molecules and mount an immune response against infected cells. There is strong pressure on the virus to escape this response, which it can do by mutating such that the peptide can no longer be presented by the host's HLA molecules. This escape signature can be identified in virus sequence data from infected individuals. However, sequence-based within-host simulators (Haller and Messer, 2023; Jariani et al., 2019) do not explicitly model a CTL immune response that can vary depending on the host's HLA molecules.

We previously developed a model and R package, `wavess` (Sambaturu et al., 2025), that simulates within-host virus evolution, with optional recombination and latency, under various selective pressures including conserved sites, replicative fitness, and a generic adaptive immune response with cross-reactivity. Here, we describe two updates to this model, i) an explicit CTL immune response, and ii) variable recombination rates to allow for modeling of recombination hostspots and non-adjacent or segmented sequences. These features are included in `wavess` version 1.2.

New developments to `wavess`

`wavess` is an individual-based forward-simulation model that explicitly simulates virus sequences including mutation, recombination, and selection (Sambaturu et al., 2025).

HLA-specific CTL immune response

`wavess` 1.0 modeled a generic adaptive immune response with cross-reactivity and no option for complete escape; however, the CTL response leads to distinctly observable escape mutations, particularly early in infection (Fischer et al., 2010). As immune-driven evolution in many virus genes is mostly due to the host CTL response rather than the antibody response, we modified `wavess` to separately model these two components of the immune system. The B-cell antibody immune response is modeled identically to the original `wavess` immune response.

Received: Date Month Year. **Revised received:** Date Month Year. **Accepted:** Date Month Year
© The Author(s) YEAR. Copyright and licence statements to be updated by the publisher during production.

In addition, we added functionality to model complete escape from a CTL response at user-defined amino acid positions. One amino acid is recognized by the immune system; mutation to any other amino acid confers complete and immediate escape from recognition. All T-cell epitopes have an identical user-defined maximum fitness cost c_{max} (default = 0.5), and each T-cell epitope i has a(n optionally) distinct time to reach maximum fitness cost t^i_{max}, so T-cell responses can mature at different rates. All responses mature linearly from 0 to t^i_{max} and once t^i_{max} is reached, that cost is maintained for the remainder of the infection. So the fitness of a recognized epitope i in generation t is defined as:

$$F^i_{I_T}(t) = 1 - c_{max} \times min(t/t^i_{max}, 1) \quad (1)$$

Epitopes that have escaped recognition have a fitness of 1. The T-cell immune fitness for a virus at generation t is the product of the individual fitness for each epitope:

$$F_{I_T}(t) = \prod_{i=1}^{m} F^i_{I_T}(t) \quad (2)$$

where m is the total number of CTL epitopes in the individual.

The overall fitness for a virus at time t is the product of the virus's conserved fitness F_C, replicative fitness F_R, B-cell (antibody) fitness F_{I_B}, and T-cell (CTL) fitness F_{I_T}:

$$F(t) = F_C \times F_R \times F_{I_B} \times F_{I_T} \quad (3)$$

Variable recombination rate

`wavess` 1.0 used an identical recombination rate (a combined rate of dual infection followed by recombination) across all potential breakpoints. To enable modeling of recombination hotspots, recombination between non-adjacent genes in a sequence, and reassortment of genes on separate segments of segmented viruses, we implemented the option to have a variable recombination rate across the simulated sequence.

Within `wavess`, we convert the recombination rate into a probability of recombination by assuming that the number of recombination events follows a Poisson process. In `wavess` 1.0, we assumed that the recombination rate was small enough that, per generation, only zero or one recombination events would occur at any given breakpoint; however, larger recombination rates break this assumption. Therefore, we now calculate the probability of an odd number of recombination events occurring as $(1 - e^{-2n\lambda})/2$, because if an even number occurs then it is not observable in the sequence. Rates over 3 recombinations per breakpoint per generation lead to a probability of about 0.5.

As identifying recombination events at each breakpoint individually substantially increases runtime, we implemented an algorithm that determines, based on the recombination probabilities, the optimal method to identify recombination events across the sequences. When the majority of positions ($> 95\%$) have the same baseline rate, binomial sampling is used to identify recombination events at those positions. For positions that differ from the baseline, or if no baseline exists, recombination events are identified at each position individually. We used GitHub Co-Pilot to help with optimization.

We envision that variable recombination rates can be used to model i) recombination hotspots, ii) non-adjacent genes, with a higher recombination rate between the genes, and iii) segmented genomes where the rate indicates the reassortment rate. For the second scenario, the rate can be calculated as $n\lambda$ where n is the number of breakpoints between the genes and λ is the baseline recombination rate for a single breakpoint.

Usage updates and availability

We have updated the `wavess` R package vignettes to describe, and provide examples of, the new options for a CTL immune response and a variable recombination rate. All updates can be found at https://github.com/MolEvolEpid/wavess.

HIV-1 example with an HLA-aware framework

Input data and simulation settings

To test our model, we simulated within-host HIV-1 evolution of *pol* and *gp120* (spliced together) using the `wavess` default values unless otherwise noted. *pol*, the polymerase gene, is under strong CTL immune pressure but not strong antibody pressure, while *gp120*, an envelope protein, is heavily targeted by the host antibody response. We used as the founder sequence *pol* and *gp120* from subtype B sequence DEMB11US006 (GenBank accession number KC473833) (Sanchez et al., 2014). Conserved sites were identified as recommended in `wavess` for *pol* and *gp120* and mapped to the founder sequence. The LANL HIV database consensus sequence (Linchangco et al., 2022) was used as the reference for both genes. B-cell epitope locations were identified for *gp120* only, using the recommended method from `wavess` with the LANL HIV database ENV sequence features. The same 10 B-cell epitopes were used for all simulations.

CTL epitopes

We identified founder T-cell epitopes for the IEDB panel of 27 HLA-A and HLA-B alleles (at the subtype, i.e. 4-digit, level) using the IEDB T-cell prediction tool with the NetMHCpan 4.1 EL prediction method for MHC-1 Binding/Elution (Reynisson et al., 2020; Vita et al., 2025). We also calculated class I pMHC immunogenicity with 1,2,C terminals masked. We calibrated the epitope filtering criteria to align with values in the literature, while allowing different sequences and HLAs to lead to different numbers of recognized epitopes. We wanted a filtering criteria that did not lead to zero CTL epitopes, but resulted in few enough to be realistic. It is estimated that between 9-18 escape mutations become fixed in the population within six months of infection (Carlson et al., 2015), with the epitope-specific time to maximum T-cell response ranging from less than two weeks to over two months (Turnbull et al., 2009). With this in mind, for any given set of 2 HLA-A and 2 HLA-B alleles, we identified all epitopes of length 9 amino acids under a NetMHCpan percentile of 0.1 and with an immunogenicity score of less than $1/90$. Across all HLA combinations, this resulted in a median of 12 epitopes across *pol* and *gp120* (range 2-25; Figure 1A). The number of days to reach maximum CTL immune cost was considered to be the inverse of the epitope immunogenicity score. Across all epitopes, the median time was 7 days (range 2-42; Figure 1B). For each epitope, the anchor positions (2 and

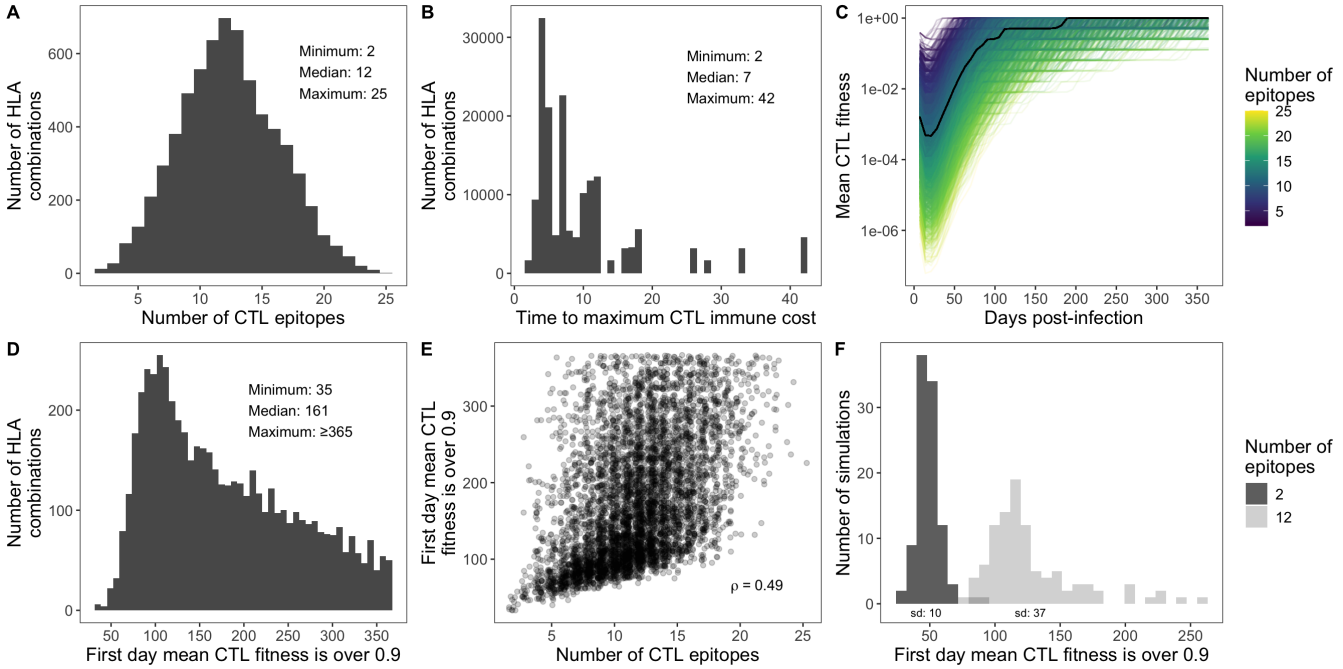

Figure 1 Escape from the CTL response. (A) Number of CTL epitopes across all combinations of two HLA-A and HLA-B molecules from the IEDB 27 CTL panel. (B) Distribution of time to maximum CTL immune response for all epitopes from panel A. (C) Mean CTL fitness output by `wavess` for each simulation. (D) First day the mean CTL fitness was > 0.9 for each simulation. (E) Correlation between the number of CTL epitopes and the first day the mean CTL fitness was > 0.9. ρ is Spearman's. (F) First day the mean virus CTL fitness was > 0.9 for each of 100 replicates of two HLA combinations with different numbers of epitopes. The standard deviation is indicated under each histogram.

9) were considered positions at which a mutation would lead to escape from the CTL response (Carlson et al., 2015).

Recombination rate

We set the baseline recombination rate within each gene to $r = 1.5 \times 10^{-5}$ per breakpoint per generation, and the rate between the two genes to $r \times n$, where $n = 1128$, the number of nucleotides between *pol* and *gp120* in the founder, yielding a rate of 0.17.

To investigate signals of recombination breakpoints in the simulated sequences, we used 3seq (Lam et al., 2018). To minimize compute time, we identified breakpoints using sequences sampled from the last 10 weeks of the simulation ($n = 200$ sequences) for 10 randomly selected simulations.

Simulations

For each combination of 2 HLA-A and 2 HLA-B epitopes ($n = 6600$), we simulated within-host evolution using `wavess` for one year. Weekly, we sampled 20 sequences and recorded the mean CTL fitness. For the HLA combination with the smallest number of epitopes (A*23:01, A*01:01, B*40:01, B*08:01) and one with a median number of epitopes (A*02:01, A*01:01, B*15:01, B*57:01) we also performed 100x replicates to investigate the stochasticity of CTL fitness over time. As complete immune escape may or may not occur at conserved sites (Li et al., 2007; Peyerl et al., 2004), we set the conserved cost and the maximum T-cell immune cost to 0.5.

Model output exploration and analysis

CTL response

Different HLA combinations led to different days at which the mean CTL fitness was > 0.9 (median 161, range 35 to ≥ 365; Figure 1C-D), which was correlated with the number of epitopes recognized by the CTL response (Spearman $\rho = 0.49$; Figure 1E). This also aligns with reports of escape occurring between a few weeks to several years post-infection (Goonetilleke et al., 2009). Furthermore, the time to escape for viruses with fewer recognized epitopes was less variable than the time to escape for viruses with more recognized epitopes (standard deviation of 10 days for 2 epitopes vs. 37 days for 12 epitopes; Figure 1F).

Recombination breakpoints

As expected, breakpoints were often identified at location where *pol* and *gp120* were spliced in the simulated sequence, although recombination events were detected throughout the sequence (Figure 2).

Conclusion

`wavess` can now model non-adjacent or segmented sequences, variable recombination, latency, conserved sites fitness, replicative fitness, antibody fitness, and CTL fitness. The new features we added here allow for more accurate within-host modeling of virus genes under CTL-mediated selective pressures. These updates may also prove useful for modeling non-virus pathogens.

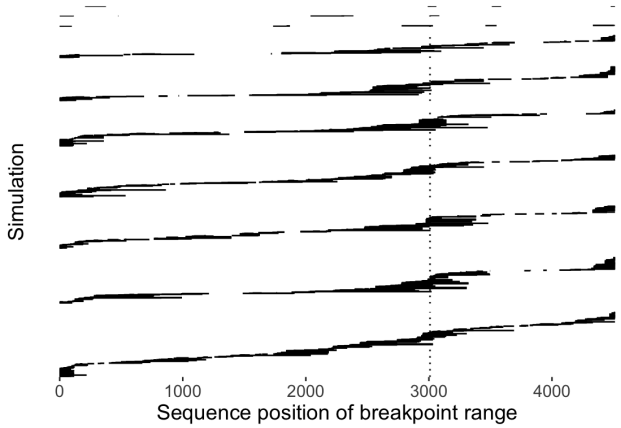

Figure 2 Example recombination breakpoints. Recombination breakpoints identified by 3seq for 10 simulations. 3seq returns a range for each recombination breakpoint; this is what is depicted in each row. The vertical dotted line indicates where *pol* and *gp120* were spliced; to the left is *pol* and to the right is *gp120*.

Conflicts of interest

The authors declare that they have no competing interests.

Funding

This work was supported by the National Institutes of Health [R01AI087520 to T.L.] and the Los Alamos National Laboratory [Laboratory Directed Research and Development program fellowship project no. 20230873PRD4 to Z.L.].

Data availability

The data and code in this article are available at `https://github.com/MolEvolEpid/wavess_ctl_manuscript`.

Author contributions statement

Z.L. and T.L. conceptualized the project, acquired funding, developed the methodology, and reviewed and edited the manuscript. T.L. supervised the project. Z.L. performed formal analysis and investigation, developed the software, visualized the results, and wrote the original draft of the manuscript.

Acknowledgments

We thank Narmada Sambaturu, Jennifer Mamrosh, Ruy Ribeiro, and Alan Perelson for discussion about the CTL response.

References

J. M. Carlson, A. Q. Le, A. Shahid, and Z. L. Brumme. Hiv-1 adaptation to hla: a window into virus–host immune interactions. *Trends in microbiology*, 23(4):212–224, 2015.

W. Fischer, V. V. Ganusov, E. E. Giorgi, P. T. Hraber, B. F. Keele, T. Leitner, C. S. Han, C. D. Gleasner, L. Green, C.-C. Lo, et al. Transmission of single hiv-1 genomes and dynamics of early immune escape revealed by ultra-deep sequencing. *PloS one*, 5(8):e12303, 2010.

N. Goonetilleke, M. K. Liu, J. F. Salazar-Gonzalez, G. Ferrari, E. Giorgi, V. V. Ganusov, B. F. Keele, G. H. Learn, E. L. Turnbull, M. G. Salazar, et al. The first t cell response to transmitted/founder virus contributes to the control of acute viremia in hiv-1 infection. *Journal of experimental medicine*, 206(6):1253–1272, 2009.

B. C. Haller and P. W. Messer. SLiM 4: Multispecies Eco-Evolutionary Modeling. *The American Naturalist*, 201(5): E127–E139, May 2023. ISSN 0003-0147. doi: 10.1086/723601. Publisher: The University of Chicago Press.

A. Jariani, C. Warth, K. Deforche, P. Libin, A. J. Drummond, A. Rambaut, F. A. Matsen IV, and K. Theys. SANTA-SIM: simulating viral sequence evolution dynamics under selection and recombination. *Virus Evolution*, 5(1):vez003, Jan. 2019. ISSN 2057-1577.

H. M. Lam, O. Ratmann, and M. F. Boni. Improved algorithmic complexity for the 3seq recombination detection algorithm. *Molecular biology and evolution*, 35(1):247–251, 2018.

B. Li, A. D. Gladden, M. Altfeld, J. M. Kaldor, D. A. Cooper, A. D. Kelleher, and T. M. Allen. Rapid reversion of sequence polymorphisms dominates early human immunodeficiency virus type 1 evolution. *Journal of virology*, 81(1):193–201, 2007.

G. V. Linchangco, B. Foley, and T. Leitner. Updated HIV-1 Consensus Sequences Change but Stay Within Similar Distance From Worldwide Samples. *Frontiers in Microbiology*, 12, Jan. 2022. ISSN 1664-302X. Publisher: Frontiers.

F. W. Peyerl, H. S. Bazick, M. H. Newberg, D. H. Barouch, J. Sodroski, and N. L. Letvin. Fitness costs limit viral escape from cytotoxic t lymphocytes at a structurally constrained epitope. *Journal of virology*, 78(24):13901–13910, 2004.

B. Reynisson, B. Alvarez, S. Paul, B. Peters, and M. Nielsen. Netmhcpan-4.1 and netmhciipan-4.0: improved predictions of mhc antigen presentation by concurrent motif deconvolution and integration of ms mhc eluted ligand data. *Nucleic acids research*, 48(W1):W449–W454, 2020.

N. Sambaturu, Z. Lapp, F. D. Tria, E. Romero-Severson, C. Molina-París, and T. Leitner. wavess: An r package for simulation of adaptive within-host virus sequence evolution. *PLOS Computational Biology*, 21(9):e1013437, 2025.

A. M. Sanchez, C. T. DeMarco, B. Hora, S. Keinonen, Y. Chen, C. Brinkley, M. Stone, L. Tobler, S. Keating, M. Schito, M. P. Busch, F. Gao, and T. N. Denny. Development of a contemporary globally diverse HIV viral panel by the EQAPOL program. *Journal of Immunological Methods*, 409:117–130, July 2014. ISSN 0022-1759.

E. L. Turnbull, M. Wong, S. Wang, X. Wei, N. A. Jones, K. E. Conrod, D. Aldam, J. Turner, P. Pellegrino, B. F. Keele, et al. Kinetics of expansion of epitope-specific t cell responses during primary hiv-1 infection. *The Journal of Immunology*, 182(11):7131–7145, 2009.

R. Vita, N. Blazeska, D. Marrama, I. C. T. Members, D. Shackelford, L. Zalman, G. Foos, L. Zarebski, K. Chan, B. Reardon, S. Fitzpatrick, M. Busse, S. Coleman, C. Sedwick, L. Edwards, C. MacFarlane, M. Ennis, S. Duesing, J. Bennett, J. Greenbaum, M. De Almeida Mendes, J. Mahita, D. K. Wheeler, et al. The immune epitope database (iedb): 2024 update. *Nucleic Acids Research*, 53(D1):D436–D443, 2025.