PLOS COMPUTATIONAL BIOLOGY

RESEARCH ARTICLE

Optimal sequence-based design for multi-antigen HIV-1 vaccines using minimally distant antigens

Eric Lewitus[1,2], **Jennifer Hoang**[1,2], **Yifan Li**[1,2], **Hongjun Bai**[1,2], **Morgane Rolland**[1,2]*

1 U.S. Military HIV Research Program, Walter Reed Army Institute of Research, Silver Spring, Maryland, United States of America, **2** Henry M. Jackson Foundation for the Advancement of Military Medicine, Inc., Bethesda, Maryland, Uniiited States of America

* mrolland@hivresearch.org

OPEN ACCESS

Citation: Lewitus E, Hoang J, Li Y, Bai H, Rolland M (2022) Optimal sequence-based design for multi-antigen HIV-1 vaccines using minimally distant antigens. PLoS Comput Biol 18(10): e1010624. https://doi.org/10.1371/journal.pcbi.1010624

Editor: Rob J. De Boer, Utrecht University, NETHERLANDS

Received: January 11, 2022

Accepted: October 3, 2022

Published: October 31, 2022

Peer Review History: PLOS recognizes the benefits of transparency in the peer review process; therefore, we enable the publication of all of the content of peer review and author responses alongside final, published articles. The editorial history of this article is available here: https://doi.org/10.1371/journal.pcbi.1010624

Data Availability Statement: Sequences analyzed in this study are available in GenBank under accession numbers: MN791130—MN792579, ON959609 - ON959788. The code and data

Abstract

The immense global diversity of HIV-1 is a significant obstacle to developing a safe and effective vaccine. We recently showed that infections established with multiple founder variants are associated with the development of neutralization breadth years later. We propose a novel vaccine design strategy that integrates the variability observed in acute HIV-1 infections with multiple founder variants. We developed a probabilistic model to simulate this variability, yielding a set of sequences that present the minimal diversity seen in an infection with multiple founders. We applied this model to a subtype C consensus sequence for the Envelope (Env) (used as input) and showed that the simulated Env sequences mimic the mutational landscape of an infection with multiple founder variants, including diversity at antibody epitopes. The derived set of multi-founder-variant-like, minimally distant antigens is designed to be used as a vaccine cocktail specific to a HIV-1 subtype or circulating recombinant form and is expected to promote the development of broadly neutralizing antibodies.

Author summary

Diverse HIV-1 populations are generally thought to promote neutralizing responses. Current leading HIV-1 vaccine design strategies maximize the distance between antigens to attempt to cover global HIV-1 diversity or serialize immunizations to recapitulate the temporal evolution of HIV-1 during infection. To date, no vaccine has elicited broadly neutralizing antibodies. As we recently demonstrated that infection with multiple HIV-1 founder variants is predictive of neutralization breadth, we propose a novel strategy that endeavors to promote the development of broadly neutralizing antibodies by replicating the diversity of multi-founder variant acute infections. By training an HIV-1 Env consensus sequence on the diversity from acute infections with multiple founders, we derived in silico a set of minimally distant antigens that is representative of the diversity seen in a multi-founder acute infection. As the model is particular to the input sequence, it can produce antigens specific to any HIV-1 subtype or circulating recombinant form (CRF). We

generated during this study are available at https://www.hivresearch.org/publication-supplements.

Funding: This work was supported by a cooperative agreement between The Henry M. Jackson Foundation for the Advancement of Military Medicine, Inc., and the U.S. Department of the Army [W81XWH-18-2-0040]. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing interests: I have read the journal's policy and the authors of this manuscript have the following competing interests: A patent application on invention disclosed in this publication is filed. MR and EL are the co-inventors.

applied this to HIV-1 subtype C and obtained a set of minimally distant antigens that can be used as a vaccine cocktail.

Introduction

The diversification of HIV-1 is mediated by low replication fidelity, large population size, high recombination rates [1–3] and escape from immune pressures, including that exerted by neutralizing antibodies (nAbs) [4,5]. In the first months of infection, nAbs direct their response to recognize Envelope (Env) targets in autologous viruses [6] before heterologous targets are recognized typically a couple of years later [7,8]. While most individuals living with HIV-1 develop nAbs, only 10–25% of individuals elicit nAbs that can neutralize >70% of a diverse virus panel [9,10]. The maturation time needed to induce broadly nAbs (bnAbs) indicate that it is a complex, multi-factorial process. So far, this process has not been reproduced by vaccine candidates. Several viral factors have been associated with the development of bnAbs, including viral load, viral subtype, CD4+ T cell count, infection duration and viral diversity [11–20].

Theoretical and empirical data show that increased Env diversity in acute or early infection may prime the immune system to develop bnAbs. A modeling study by Luo and Perelson showed that broadly neutralizing responses could emerge earlier in infections founded by multiple strains as bNAbs were less likely to develop after infection with single variants due to competitive exclusion by the autologous antibody response [21]. Relatedly, the large increase in diversity that occurs following a super-infection has been linked to the subsequent development of bnAbs in these individuals [22–24]. We recently analyzed 3,482 HIV-1 Env sequences sampled from 70 people living with HIV-1 (PLWH) who were diagnosed in acute infection [25–27] and compared the sequence data to neutralization breadth measurements performed on samples collected between six months and four years after infection [28]. Participants had been enrolled in the prospective HIV-1 acute infection cohort RV217 [29]. In this cohort, more than 3,000 seronegative individuals from four countries (Kenya, Tanzania, Thailand and Uganda) were tested twice-weekly for HIV-1 RNA and 155 acute infections were identified. RV217 PLWH were followed for up to five years after viremia was detected. We showed that individuals with infections established with multiple HIV-1 founder viruses were more likely to develop neutralization breadth than those with infections established with single founder viruses [27]. This finding was reproduced in the cohort of placebo recipients who were infected during the RV144 vaccine efficacy trial [27,30,31]. We interpret recent data from small cohorts of infants living with HIV-1 as also supporting this relationship between multivariant infections and the development of breadth. Infections with more diverse viral populations were associated with the development of neutralization breadth [32] and the proportion of infections with multiple founders was high among infants (7/12) in a different small cohort [33].

Infections with multiple HIV-1 founder variants account for approximately 25% of infections [25,34–37]. These multi-founder infections occur when multiple sequences are transmitted from a single transmitter who was likely in the chronic phase of their infection. Hence, all the sequences in the recipient are closely related, phylogenetically linked, and show around 1% intra-participant diversity (i.e., minimally distant). We previously showed that infections with multiple founder variants have clinically relevant features as viral load set point was 0.3 log10 higher in infections with multiple founders when compared to single founder infections [36]. We also recently showed that higher engagement of B cells in the first months of infection was associated with the development of bnAbs years later [28]. This finding together with the fact

that infections with multiple HIV-1 founder variants was predictive of the development of neutralization breadth [27] emphasize that the early events of HIV-1 infection play a critical role in the ontogeny of bnAbs. This led us to propose that a vaccine which would be constituted by multi-founder like, minimally distant antigens (differing by ~1% in Env) corresponding to the variability we observed in infections with multiple founder variants could initiate the induction of bnAbs. We hypothesize that a set of antigens that show minimal differences and differ primarily at surface sites, including sites that correspond to bnAb epitopes, would promote the maturation of neutralizing responses through toggle responses between variant epitopes.

An HIV-1 vaccine that elicits bnAbs is vital to prevent infection from circulating viruses. Serial administration of the bnAb VRC01 prevented HIV-1 acquisition, albeit blocking only the small fraction of circulating viruses that were highly sensitive to the bnAb (IC80 <1 μg/mL) (corresponding to 28 of 162 infections) [38]. These results emphasize the enormous challenge associated with the development of a protective HIV-1 vaccine and highlight the need for new strategies to develop vaccine candidates that could elicit neutralization breadth. Here we present a probabilistic model for simulating Env alignments that resemble a set of minimally distant antigens representative of the variability seen in multi-founder acute infections. We showed that the model recapitulated features of multi-founder variant infections, including divergence or pairwise distances across sequences and diversity at key Env antibody epitopes. Using this probabilistic model, we derived minimally distant antigens (differing by ~1%) for subtype C Env sequences and selected five sequences that can be used as a vaccine cocktail.

Results

A probabilistic model designed to simulate sequence alignments

The model was based on two training alignments, F_1 and F_2. F_1 corresponded to all the sequences descended from the major founder variant, i.e. the major founder lineage in a participant infected with multiple founder variants. F_2 grouped all sequences descended from minor founder variants in that participant (**Fig 1A**). Infections with multiple founder variants can be established with two founders that are clearly delineated; however, rare lineages and recombinant forms (based on the extant or unsampled founder variants) are often identified—these are sometimes found as singletons. While the model can be adapted to n founder lineages, for simplicity, here we considered the major founder lineage F_1 and grouped the other sequences in that participant (all closely phylogenetically related) as representing F_2. Hence, a group is not necessarily comprised of sequences descending from a unique founder variant but represents a genetically differentiable cluster with a divergent evolutionary pattern; identifying and distinguishing such clusters allow us to recapitulate the diversity of acute infections with multiple founders without biasing towards one founder variant population. The percentage of non-consensus residues at each site (π_j) was calculated for F_1 and F_2 separately, giving a mutational landscape for each founder lineage (major and minor viral population) (**Fig 1B**). Next, a transition probability matrix, Θ, was computed according to the procedure described by Le & Gascuel (LG matrix) [39] for the overall rate to define the probability of each amino acid (or gap) transitioning into any other amino acid (or gap) (**Fig 1C**); the empirical transition probabilities were computed on an alignment of 172 subtype C Env sequences sampled since 2011 from the LANL HIV Database. Finally, a single sequence, aligned to the training alignment reference frame, was required to seed the model (**Fig 1D**). The single seeding sequence corresponds to the design target and can be any sequence for which a set of multi-founder like antigens is to be derived, e.g. the consensus (or most recent common ancestor) from a set of sequences corresponding to a particular participant or to a specific HIV-1 subtype

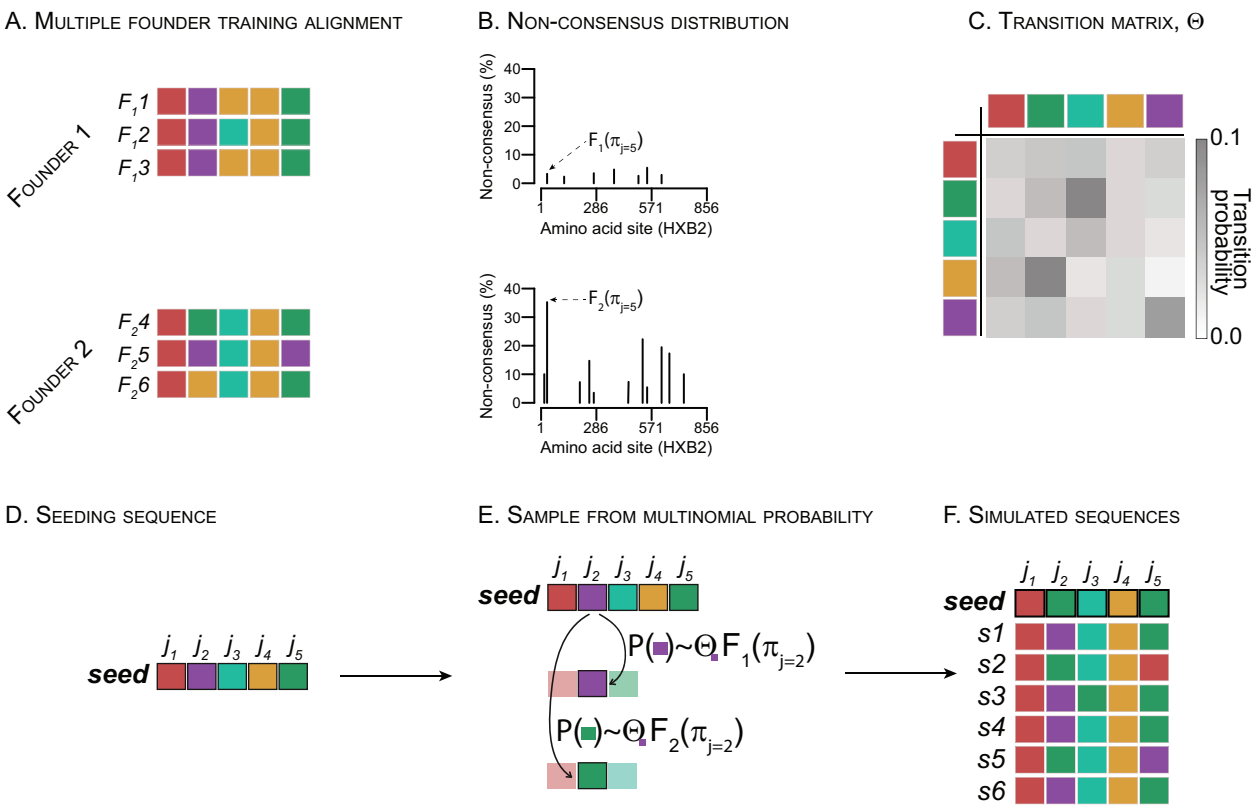

Fig 1. Protocol for simulating alignments with multiple founder lineages. (A) Alignments of the major (F_1) and minor (F_2) founder lineages from a participant infected with multiple founder variants are used to train the algorithm; this can also be done by pooling a set of participants in each training alignment. (B) The percentage of non-consensus amino acids is calculated at each site separately for the training alignments. (C) A transition probability matrix is calculated based on a set of empirical HIV-1 sequences. (D) A seeding sequence is input to seed the sequence simulation. (E) For each site, a residue is simulated from a multinomial probability distribution defined by the transition probability matrix and percentage of non-consensus amino acids at that site based on one of the founder variants, with an equal probability of sampling from each variant model. (F) The simulated sequences are output.

https://doi.org/10.1371/journal.pcbi.1010624.g001

or Circulating Recombinant Form (CRF). At each iteration, n, sequence simulation was trained on either F_1 or F_2, such that site j of the seeding sequence had a probability of mutating proportional to π_j to a new residue defined by Θ_j, where π was drawn from either F_1 or F_2 (Fig 1E). The model then outputs simulated sequences in an alignment (Fig 1F). Here we showed equal numbers of sequences derived from each founder lineage model, however, the proportion of simulated sequences trained on each founder lineages alignment can be specified. Instead of considering the variability in a given participant, the model can also rely on a pooled dataset of major vs. minor founder lineages from multiple participants. In that case, at each site, the mutational pattern can be that of any individual in the pooled set (to retain within-host diversity characteristics and forbid inter-host transitions).

Simulated sequences reproduced the variability of acute infections with multiple founder variants

We simulated sequences trained on alignments corresponding to the major (F_1) and minor (F_2) founder lineages sampled from six RV217 participants who had infections established with multiple founder variants and whose plasma, years later, neutralized >70% of viruses on a 34-virus panel [27,28] (Table 1). Sequences were obtained between 4 and 34 days post-

Table 1. RV217 infections with multiple founder variants used as training alignments. The RV217 participant ID, Env subtype/CRF, number of sequences sampled from each founder lineage, the days post-diagnosis that sequences were sampled, and the peak neutralization breadth reached by the participant within three years of diagnosis are reported.

ID	Subtype	Founder lineage 1	Founder lineage 2	Days post-diagnosis	Peak neutralization breadth (%)
10220	A1	7	13	15,31	74
30124	A1	4	14	4,32	82
20337	C	9	9	7,34	79
40123	CRF01_AE	8	13	7,29	82
40363	CRF01_AE	10	10	7,28	88
40436	CRF01_AE	5	14	4,28	77

https://doi.org/10.1371/journal.pcbi.1010624.t001

diagnosis. Highlighter plots (S1 Fig) and tree topologies (S2 Fig) indicated that each infection was established with multiple founder variants (Fig 2A and 2B). For the major founder lineage F_1, the mean percentage of non-consensus residues per site across the six participants was 0.06% (max = 10%) and median pairwise amino acid diversity was 0.001 (min = 0, max = 0.005). For the minor founder lineage F_2, mean percentage of non-consensus residues per site (1.11%, max = 50%) and median diversity (0.019, min = 0.006, max = 0.034) were both significantly higher than for F_1 (Mann-Whitney U test, $P<0.015$). An empirical transition probability matrix, Θ, was computed on an alignment of 172 subtype C Env sequences sampled since 2011 and simulations were seeded with the consensus derived from sequences from an independent RV217 participant (id = 10066). One thousand sequences were simulated and trained in equal proportions on major and minor founder lineages alignments for each participant separately, where the percentage of non-consensus residues at each site, π_j, was defined by either the major founder lineages, $F_1(\pi_j)$, or minor founder lineages, $F_2(\pi_j)$. The percentage of non-consensus residues in simulated sequences at each site was highly correlated (median $R^2 = 0.99$, min = 0.98, max = 100) with the percentage in the training alignment for each founder lineage (Fig 2C and 2D). One thousand sequences were also simulated while trained on a pool of all founder lineage alignments across participants (Fig 2E), where $F_1(\pi_j)$ was the maximum percentage of non-consensus residues found at site j in a given individual across all F_1 alignments and $F_2(\pi_j)$ was the maximum percentage of non-consensus residues found at site j in a given individual across all F_2 alignments. The 95% confidence intervals of median pairwise diversity in simulated sequences included median pairwise diversity of training alignments for sequences simulated on each participant and for the pooled set of participants (Fig 2F). Similarly, the 95% confidence intervals of the median number of polymorphic sites per simulated sequence as well as the number of polymorphisms per sequence at CD4bs, V1-V2, V3 and MPER contact sites included the median number for training alignments for sequences simulated on each participant (Fig 2G).

The model was designed to simulate sequences that are specific to the seeding sequence rather than mimic the composition of the training alignment. Therefore, the simulated alignment should be genetically closer to the seeding sequence than to the training alignment. Indeed, the percentage of mismatched ungapped sites between the consensus of simulated sequences and seeding sequence (0%) was significantly lower (Mann-Whitney U test, $P = 0.001$) than between the consensus of simulated and consensus of the training sequences (8.27–16.12%) for sequences simulated with each training alignment (S3 Fig).

Simulated sequences replicated the diversity found in infections with multiple founder variants

We compared diversity and divergence estimates for sequences simulated under the pooled set of major and minor founder lineage alignments to Env sequences sampled during acute

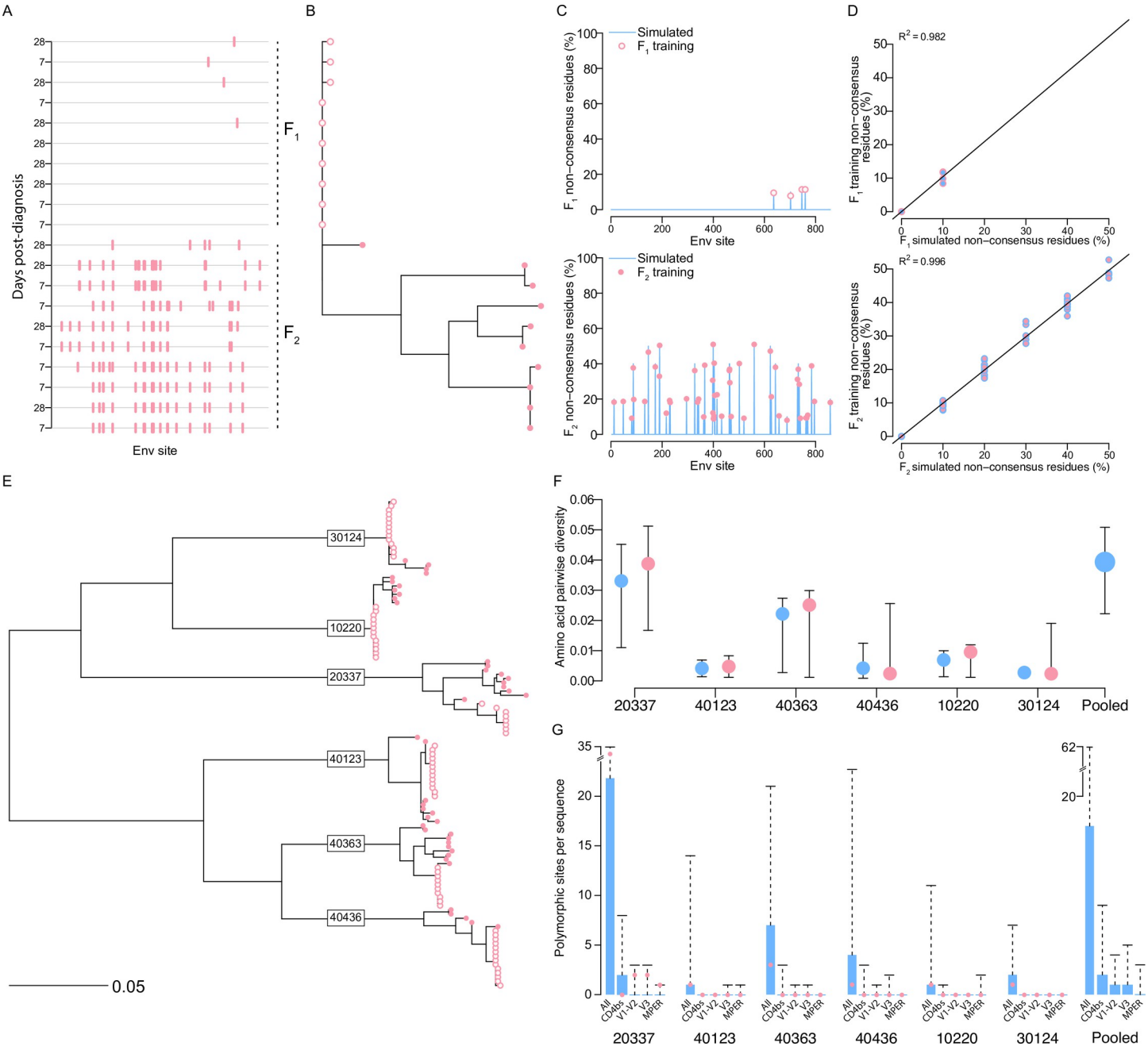

Fig 2. Simulated sequences reproduced the variability of acute HIV-1 infections with multiple founder variants. (A) A highlighter plot and (B) phylogeny for Env sequences sampled at 7 and 28 days post-diagnosis from a RV217 participant (id = 40363) with an infection with multiple founder variants. (C) The percentage of non-consensus residues at each Env site in sequences simulated from the major founder variant or major lineage 1 (top, blue line) and the minor founder variants grouped here as lineage 2 (bottom, blue line) after seeding with the consensus sequence from an independent acutely-infected RV217 participant (id = 10066); values for sequences belonging to the major and minor founder lineages in 40363 are shown in open and filled pink circles, respectively. (D) Regression plots of the percentage of non-consensus residues in the training alignment as a function of non-consensus residues in the simulated alignment for (top, blue fill and pink border) founder lineage 1 and (bottom, blue border and pink fill) founder lineage 2. (E) Phylogeny of sequences sampled at 4–34 days post-diagnosis from 6 RV217 participants with infections with multiple founder variants. Tips are colored to represent the population corresponding to the major (open circles) and minor (closed circles) founder populations for each participant (for simplicity, multiple founder variants or singleton sequences are grouped under the minor lineage). (F) For sequences simulated under each training alignment (see panel E), the pairwise diversity of the training alignment (pink) and of the sequences simulated under that training alignment (blue); and the pairwise diversity of sequences simulated under the pooled alignment (blue). Solid lines represent 25% and 75% interquartile ranges. (G) Barplots of the number of polymorphic sites per sequence in sequences simulated under each training alignment and the pooled-participants training alignment (blue) at all sites, CD4bs, V1-V2 contact sites, V3 contact sites, and MPER sites. Dashed whiskers indicate maximum values. Pink dots represent median values for training alignments.

https://doi.org/10.1371/journal.pcbi.1010624.g002

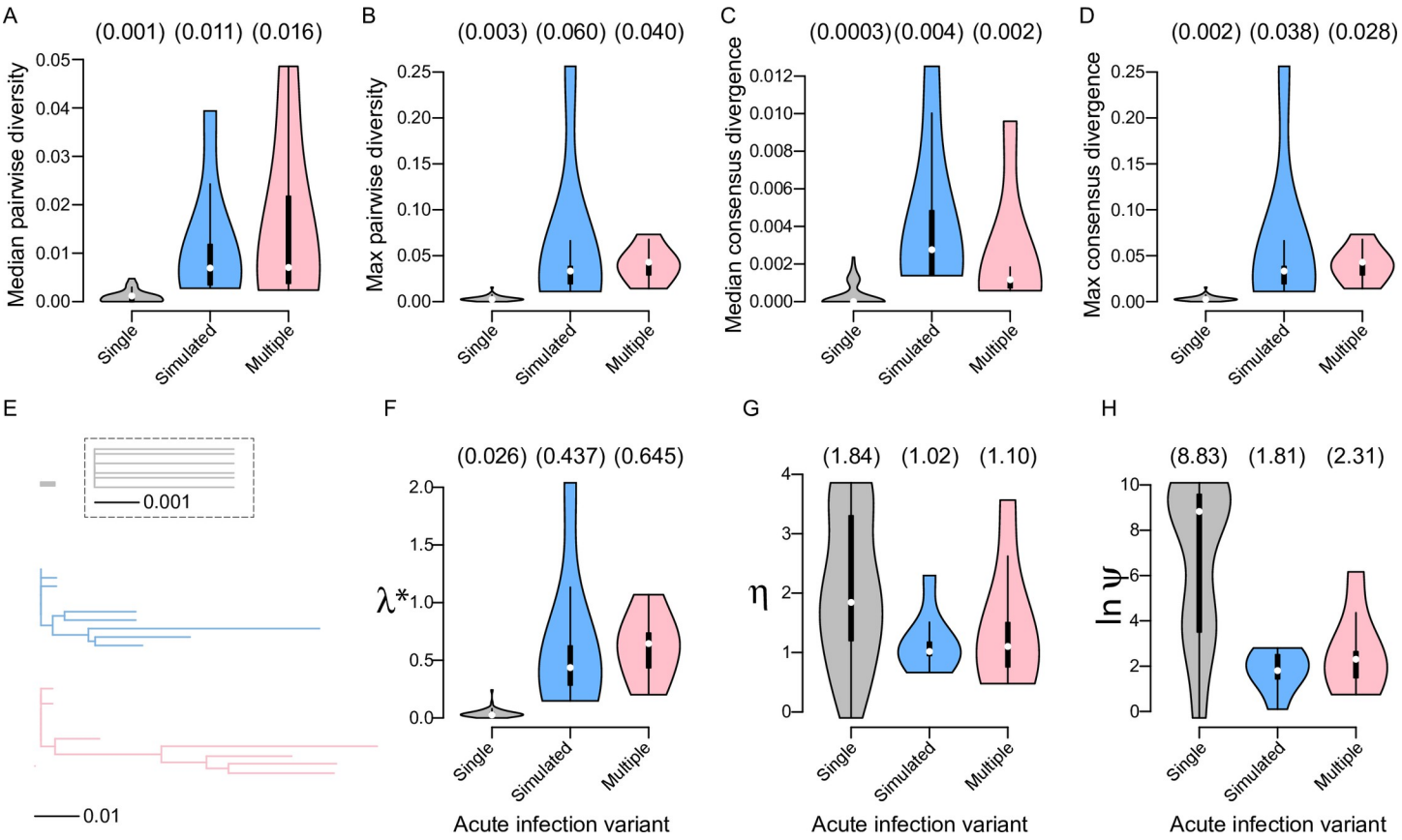

Fig 3. Simulated sequences replicated the diversity found in infections with multiple founder variants. Violin plots of (A) median pairwise diversity, (B) maximum pairwise diversity, (C) median divergence from the consensus, and (D) maximum divergence from the consensus for sequences sampled from RV217 participants during acute infection with a single founder variant (n = 53, grey), sequences simulated under the pooled-participants probabilistic model (n = 5000, blue), and sequences sampled from RV217 participants during acute infection with multiple founder variants (n = 6, pink). (E) Phylogenies constructed with sequences from a participant with a single founder variant sampled at 1 and 29 days post-diagnosis (id = 10066), a sample of simulated sequences from the pooled-participants model (five from major founder lineage 1 and five from minor founder lineage 2, blue), and sequences from a participant with multiple founder variants sampled at 7 and 28 days post-diagnosis (id = 40363) (pink). Phylogenies are shown at the same scale; the dashed box shows the top phylogeny at ten-times magnification. Violin plots of spectral density profile summary statistics λ^*(F), η (G) and ln-transformed ψ (H) for the same groups. Median values are shown above each plot.

https://doi.org/10.1371/journal.pcbi.1010624.g003

infection from RV217 participants infected with either a single founder variant (and who developed <35% neutralization breadth) (n = 12) or multiple founder variants (and who developed >70% neutralization breadth) (n = 6). The 95% confidence intervals of diversity and divergence estimates for simulated sequences included the median values for Env alignments of multi-founder variant acute infections and was higher than that of infections with single founders (Fig 3A–3D).

Similarly, phylogenies of simulated sequences were more similar to phylogenies of infections with multiple founder variants than to those with single founder variants (Fig 3E). Spectral density profiles of the modified graph Laplacian [40] were computed for Env alignments from infections with single and multiple founder variants in RV217 and for down-sampled alignments from sequences simulated under the pooled set of major and minor founder lineages alignment. Spectral density profile summary statistics each capture a unique aspect of phylogenetic topology: λ^* is proportional to non-synonymous/synonymous rates, η is inversely proportional to transition-transversion rate ratio, and ψ is proportional to rate heterogeneity [41]. The simulated alignment was downsampled 100 times for 5 sequences simulated under each of the founder lineage training alignments, F_1 and F_2. The 95% confidence

intervals of λ^*, η, and ln-transformed ψ for simulated sequences included the median values for Env alignments of acute infections with multiple founder variants but none included the median value for infections with a single founder variant (Fig 3F–3H).

In silico-derived antigenic sequences for HIV-1 Env subtype C

A consensus sequence was generated from 172 subtype C Env sequences sampled after 2010 (Fig 4A). One thousand sequences were simulated under the pooled founders training alignment and seeded with the subtype C Env consensus sequence derived from the subtype C alignment. One-hundred samples of ten sequences (five simulated under F_1 and five under F_2) were analyzed. The sampled sequences produced phylogenies with bimodal or multimodal topologies (Fig 4B). The median pairwise distance of subtype C Env sequences sampled after 2010 to the consensus was 0.160 [IQR = 0.149–0.172], while the mean of median pairwise distance to the subtype C consensus across samples of simulated sequences was 0.015 (IQR = 0.013–0.016) (Fig 4C). Across the subtype C alignment, a mean of 15.8% of the residues at a site were non-consensus residues and 79.1% of sites (681/861) were polymorphic,

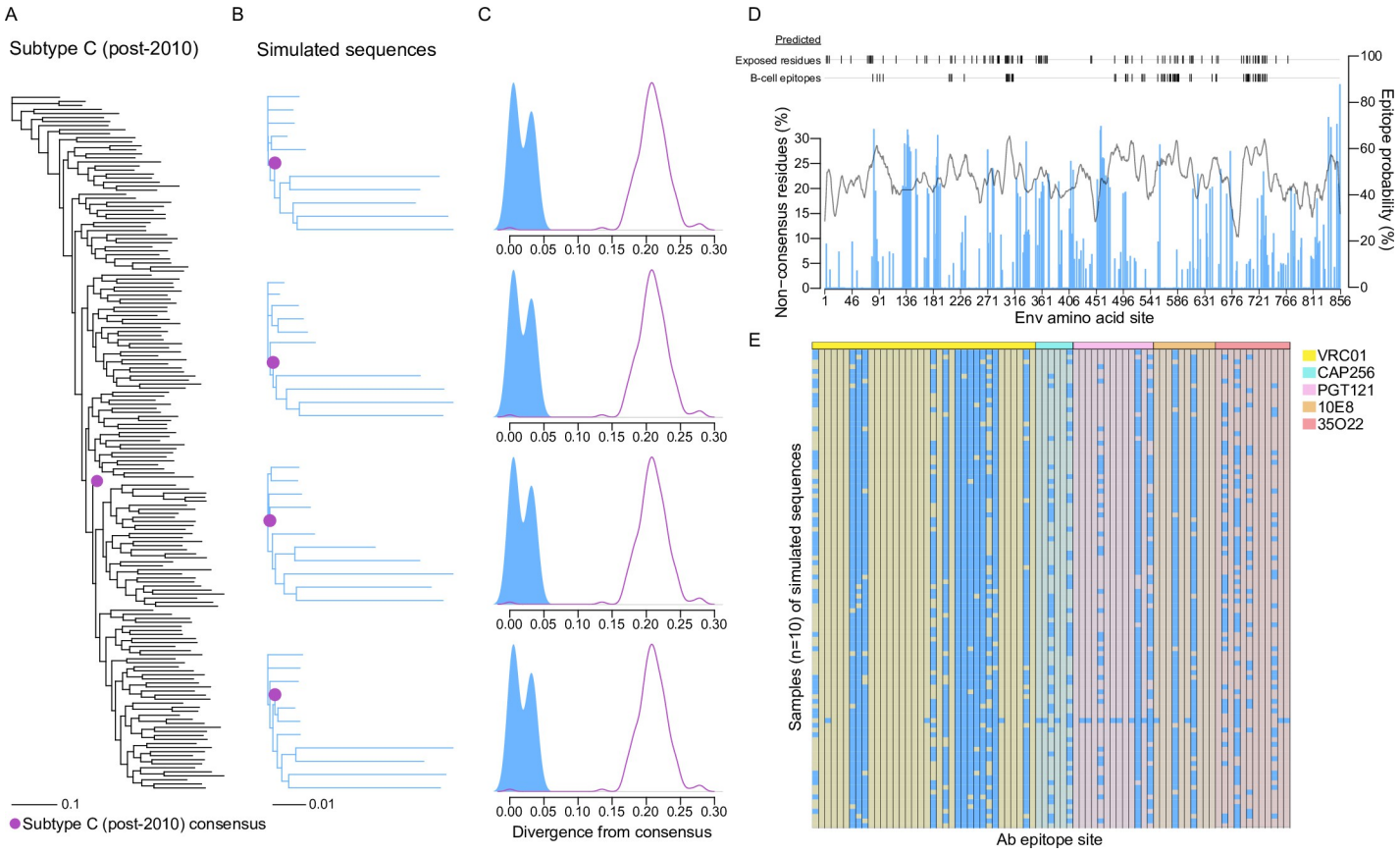

Fig 4. Sequences simulated from the HIV-1 subtype C Env consensus. (A) Phylogeny of subtype C Env sequences sampled after 2010. The consensus is marked in purple. (B) Four phylogenies constructed from sequences simulated under the pooled-participants model seeded with the subtype C consensus; each phylogeny is comprised of the subtype C consensus (purple) and five sequences randomly selected from simulations under major founder lineage 1 and five under minor founder lineage 2. (C) Density plots of divergence from the consensus for the alignment corresponding to each intra-host phylogeny of simulated sequences (blue) and for the inter-host subtype C phylogeny (purple). (D) Barplot of the percentage of non-consensus residues at each site in the simulated alignment. Epitope prediction probability for the subtype C consensus across sites (solid grey line). Dashes indicate the subtype C consensus predicted exposed residues and predicted B-cell epitopes that are polymorphic in the simulated alignment. (E) Polymorphisms (blue squares) at contact sites for five representative antibodies in 100 downsampled simulated alignments comprised of five sequences simulated under each founder variant lineage.

https://doi.org/10.1371/journal.pcbi.1010624.g004

whereas across samples of simulated sequences an average (mean of median) of 3.08% [IQR = 2.93–3.16%] of residues at a site were non-consensus and 17.3% [IQR = 15.2–16%] (148/856) of sites were polymorphic (Fig 4D). B-cell epitopes and exposed sites were predicted for the subtype C consensus. Across samples of simulated sequences, an average of 12.3% [IQR = 11.7–13.2%] of sites (26.73/213) with >50% epitope probability were polymorphic and 16.2% [IQR = 15.0–16.9%] of sites (42.65/260) at predicted exposed sites were polymorphic (Fig 4D). Finally, all samples of simulated sequences were polymorphic at one or more Ab epitope sites. Five Abs corresponding to critical Env targets for neutralization were considered: VRC01:CD4bs[42], CAP256-VRC26.25:V2 apex[43], PGT121:V3[44], 10E8:MPER[45], 35O22:interface between gp120 and gp41[46]. At VRC01 epitope sites (n = 36), a median of 12 sites were polymorphic per sample of simulated sequences and 17 sites were polymorphic in at least one sample; at CAP256-VRC26.25 epitope sites (n = 6), 1 site was polymorphic per sample and 4 were polymorphic in at least one sample; at PGT121 sites (n = 13), 2 and 8; at 10E8 sites (n = 10), 2 and 2; and at 35O22 sites (n = 12), 2 and 6 (Fig 4E).

Finally, an alignment of simulated sequences was constructed for five Ab epitopes representative of key Env targets (VRC01, CAP256-VRC26.25, PGT121, 10E8, 35O22). There were 166 sequences with non-consensus residues in at least three of the five Ab epitopes (Fig 5A–5F). From these, two candidate sequences generated by F_1 and two by F_2 that were maximally divergent (within the framework of minimal diversity) were selected as candidate antigens (Fig 5G and S1 File). Together with the subtype C consensus, the candidate sequences had 20 polymorphic sites with 2–5 different residues per polymorphic site (Fig 5H). We predicted the structure of the simulated sequences using AlphaFold2 [47]. Predicted local distance difference tests (pLDDTs) showed generally good confidence across all domains, with median pLDDTs between 86.25–88.44, which was comparable to the median pLDDT of the subtype C Env consensus (87.77) that was used to seed the simulations. The sequences generated by F_1 were more similar to the seed sequence than those generated by F_2 (Fig 5I), as expected, and the structure protein prediction indicated that all simulated sequences should fold to a structured protein.

Discussion

While bnAb infusions in humans can prevent HIV-1 infection [38], no vaccine candidate has shown the induction of such bnAbs in a vaccine efficacy trial [48–54], emphasizing the need for novel vaccine strategies. Here we developed a new vaccine design approach that emulates the diversity observed in HIV-1 infections with multiple founder variants. This multi-founder like vaccine design derives from the finding that individuals with infections established with multiple founder variants were more likely to develop bnAbs than individuals with infections established with single founder variants [27]. We showed that our design strategy reproduced the diversity seen in infections with multiple founder variants and we applied this approach to design a subtype C-specific vaccine candidate constituted of a set of five minimally distant Env sequences centered on an updated subtype C consensus.

First, we developed a probabilistic method to design antigens that reflect the diversity seen in acute infections with multiple founder variants. This method was trained on sequences descended from major and minor founder lineages sampled during acute infection from one of six individuals who developed broad neutralization breadth against HIV-1 or on a pooled alignment of sequences from all six individuals. When seeded with an independent acute infection sequence, our method generated a set of antigens that recapitulated the variability of infections with multiple founder variants, including polymorphisms at critical Env target epitopes (CD4bs, V1-V2, V3 and MPER). Importantly, the derived sequences remained close to the seeding sequence, suggesting that the simulated sequences would preserve the structural

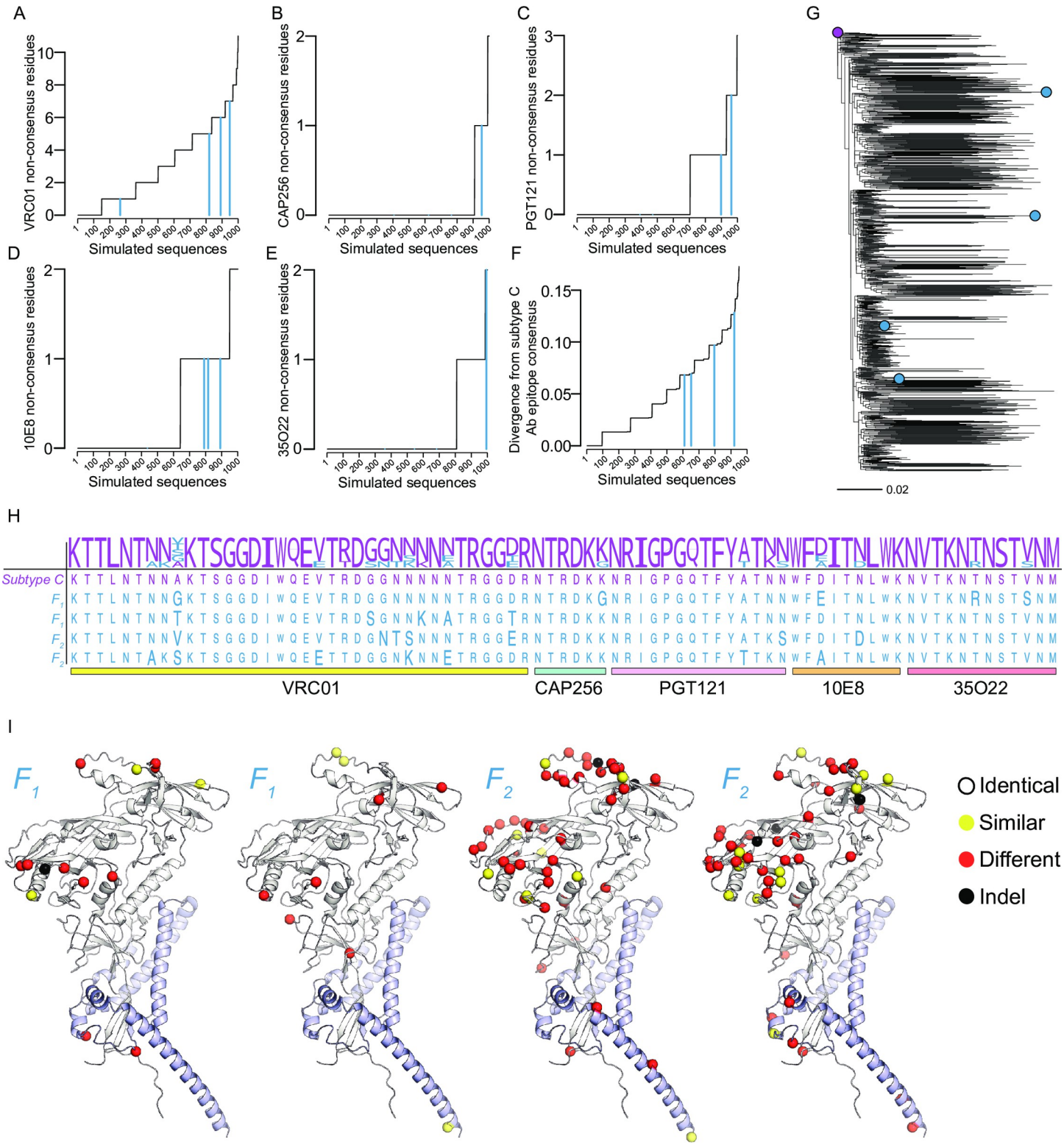

Fig 5. In silico-derived sequences for an HIV-1 subtype C vaccine cocktail. Line plots of the number of non-consensus residues per simulated sequence are represented for F_1 (dashed blue line) and F_2 (blue line) at epitope sites for five representative antibodies: (A) VRC01, (B) CAP256-VRC26.25, (C) PGT121, (D) 10E8, and (E) 35O22. (F) Divergence of simulated sequences from the subtype C consensus based on the five Ab epitope sites (n = 77). In A-F, sequences are sorted along the x-axis by their y-axis value; the black line traces the value across sequences. (G) Phylogeny of simulated subtype C Env sequences with tips corresponding to the subtype C consensus (purple) and candidate antigen sequences generated by F_1 (blue) and F_2 (blue). (H) Multiple sequence alignment of antibody epitope sites for the subtype C consensus and

candidate antigen sequences; non-consensus residues are shown in larger font and regions corresponding to the epitopes of VRC01, CAP256-VRC26.25, PGT121, 10E8, and 35O22 sites are highlighted below the alignment; a logo plot above the 5 candidate sequences represents the diversity found in circulating subtype C sequences at the 5 representative antibody epitopes. (I) Sequence differences from the consensus to candidate antigen sequences mapped on the predicted structure of the consensus C sequence. The C_α atoms of similar, different, and insertion/deletion (indel) sites are shown as yellow, red and black spheres, respectively.

https://doi.org/10.1371/journal.pcbi.1010624.g005

integrity of the chosen central sequence. By training on alignments sampled from different infections with multiple founder variants, we showed that the variability patterns of simulated sequences differed according to the training alignment. However, for all training alignments, our method generated sequences that reflected a median genetic distance of ~1%, conforming to the infections with multiple founder variants that we want to emulate in order to elicit bNAbs.

Second, we applied this method to design a subtype C specific vaccine candidate. Seeded with the Env subtype C consensus (derived from 172 independent subtype C Env sequences sampled since 2011), we generated a set of minimally distant antigens that preserved the composition of subtype C sequences while recapitulating the variability of infections with multiple founder variants. The divergence from the consensus of simulated sequences was a magnitude smaller than that of independent subtype C sequences and so met the criteria of representing diversity around the seeding sequence while reducing the total inter-host genetic space. We selected four candidate antigens that exemplified diversity at epitope sites for representative bNAbs. However, alternative selection criteria could be used to design antigens with features that could promote other types of immune responses, such as Fc effector functions [55,56]. An important limitation of this method is its reliance on alignments from only six individuals to simulate multi-founder-like sequences. While each individual developed bnAbs, precise mechanisms behind the association between multi-founder variant diversity and the development of neutralization breadth are unknown. We tried to overcome this limitation by pooling founder lineage alignments to capture the scope of diversity found in acute infection in these broad neutralizers rather than relying on one individual to define a prototypical infection with multiple founder variants. While we chose to use as seeding sequence a subtype C consensus in order to design a subtype C vaccine candidate (because over half of the people living with HIV-1 live with subtype C), only one of the six infections we used in our pooled founder lineage alignment corresponded to subtype C (A1 (n = 2) and CRF01_AE (n = 3)). It has not been reported that patterns of variability in acute infections with multiple founders differ by subtype, nonetheless, it is possible that subtype C specific patterns exist and would not necessarily be captured by our approach. Additionally, we separated sequences into two major and minor founder lineages—this does not account for the complexity seen in some infections with multiple founder variants which can include recombinants between extant and/or unsampled sequences and rare or even unique sequences. Hence, our model simplified the landscape of acute viral diversity in these individuals. Nonetheless, if datasets of increased depths are available, the model can be used to simulate the multitude of distinct sequences that can be identified in a set of hundreds of Env sequences from acute HIV-1 infection.

In silico methods for vaccine design have gained a foothold over the last decade. Computational approaches for updating vaccines against Influenza have proposed models for predicting antigenic diversity over time, including multivariate regression on physicochemical properties of circulating variants [57], phylogenetic weighting of antigenic evolution [58–60], and dynamic fitness models of antigenic alleles [61]. In silico vaccine design methods have been used for HIV-1 for over two decades to overcome the obstacle posed by HIV-1's extreme diversity [62–64]. There is more diversity within each HIV-1 subtype or CRF than what can be seen across a viral species [65]. Gaschen and colleagues showed that a centralized sequence,

consensus or ancestral, would better represent HIV-1 than any sequence derived from a PLWH [63,66]. To cope with HIV-1 diversity, some designs, such as the mosaic approach, seek to integrate a fraction of the diversity seen in HIV-1 sequences [67–69]. These variability-inclusive strategies are reminiscent of the diversity seen in super-infections, which have previously been associated with the development of neutralization breadth. As such, mosaic antigens are designed to be maximally distant to cover a large fraction of circulating viruses. The rationale is that immunizations with these diverse mosaic inserts, for example corresponding to consensus sequences for group M, subtype B and subtype C, could lead to the development of antibody responses against these distant viruses thereby potentiating broadly cross-reactive responses. Two vaccine efficacy trials are testing the Mosaic design (one reached futility criteria in 2021: https://www.jnj.com/johnson-johnson-and-global-partners-announce-results-from-phase-2b-imbokodo-hiv-vaccine-clinical-trial-in-young-women-in-sub-saharan-africa). An opposite strategy to Mosaic designs was to focus on only the most conserved elements of HIV-1 [70–73]. This stemmed from the realization that variable segments of HIV-1 functioned as decoys eliciting immune responses that were not optimal and that only a small fraction of HIV-1 diversity could be integrated in a vaccine candidate of practical size. Another currently leading strategy, the germline targeting approach, seeks to improve the longitudinal process seen in individuals infected who later developed breadth [74] by reproducing the directional process that leads to breadth in a minority of individuals through using antigens that correspond to stepwise stages of the co-evolution between the virus and the neutralizing response. Our approach is also based on a process seen in natural infections, whereby infections with multiple founder variants were linked to the subsequent development of neutralization breadth. This multi-founder-like design of minimally distant antigens is also akin to the conserved elements vaccine design rationale. We consider that the 'noisification' of a central consensus sequence at target sites for key antibodies will trigger responses to these antibody epitopes and that the toggle between these minimally distant epitopes will promote a desirable affinity maturation process leading to the development of bnAbs. The fact that our vaccine design was derived from the variability seen in multi-founder acute infections suggests that this strategy with a cocktail of minimally distant antigens may be best suited as a priming immunization. Whether subsequent immunizations should consist of the same set of antigens or a subset of them, or one or more distinct antigens, will need to be evaluated with experimental assays.

In summary, our in silico method generates a set of antigens that bear distinct epitopes, but maintain a minimal global distance across Env, constituting a projected formula for increasing the probability of eliciting bNAbs. We hypothesize that this generic approach can serve to design vaccine candidates with enhanced bnAb-eliciting properties for any given sequence. As such, this approach can be used to design cocktail vaccine candidates adapted to any HIV-1 subtypes and circulating recombinant forms. While this model can also be used to design an HIV-1 group M vaccine cocktail, the idea of a successful universal HIV-1 vaccine is far-fetched when considering lessons from the past forty years of HIV-1 vaccine research.

Materials and Methods

RV217 participant sequences

We used env sequences that we previously generated via single genome amplification of HIV-1 on plasma samples collected in the first five weeks after HIV-1 diagnosis in acute infection in participants from the RV217 cohort [25–27,29]. All participants were antiretroviral treatment naïve. We included Env sequences from twelve participants with infections with single founders (median = 10, min = 10, max = 28) who developed <35% neutralization breadth and

from six participants with infections with multiple founders who developed >70% neutralization breadth (median = 11, min = 10, max = 13) (**Table 1**) (another participant with multiple founder variants and >70% neutralization breadth was excluded because the development of neutralization breadth occurred following superinfection). Infections with multiple founder variants are illustrated with highlighter plots [34] (**S1 Fig**); we previously reported that these individuals neutralized 74–88% of a 34-virus panel at 435–2115 days post-diagnosis [28]. Sequences belonging to each founder lineage in multi-founder acute infections were used in the training dataset.

Independent subtype C sequences

Subtype C Env sequences sampled since 2011 were downloaded from the Los Alamos National Laboratory HIV Sequence Database (https://www.hiv.lanl.gov/components/sequence/HIV/search/search.html). Sequences were excluded if the individuals had been vaccinated, if the sequence did not have a complete open reading frame or did not have a sampling year. One sequence was downloaded per individual and sequences were removed if they were non-independent or an outlier. Hypermutated sequences identified with Hypermut 2.0 [75] (using https://github.com/philliplab/hypermutR) were removed with a Fisher's exact test $P<0.1$. Sequences were de-duplicated at 95% identity.

Probabilistic model for sequence simulation

A probabilistic model was developed to simulate Env sequences that replicated the variability of infections with multiple founder variants. The model is trained on two Env alignments, corresponding to the major (F_1) and minor (F_2) founder lineages identified in a given participant (the model can also be trained on a pooled set of individual lineages). A first-order Markov transition probability matrix, Θ, is estimated as described by Le and Gascuel (LG matrix) [39]; in brief, transition rates are directly computed between pairs of amino acids, including transition rates from/to gaps based on an alignment of empirical sequences. We suggest using a large dataset of independent sequences that would correspond to the subtype of the desired target vaccine candidate (i.e., subtype C sequences if the goal is a subtype C vaccine which implies that the seed sequence corresponds to subtype C). For each alignment,

1. The percentage of non-consensus residues was calculated for each alignment site.
2. For each site, j, percentage of non-consensus residues, π_j, an amino acid in the seeding sequence, k, and a transition probability, Θ_k, the probability of transitioning from k to k' was written as

$$P\left(k\prime|k,\pi_j,\Theta_k\right)=\frac{\Theta_k[k]\pi_j}{\Sigma(\Theta_k[-k])}$$

where $P_j(k')$ is calculated for each site j. For each simulated sequence, π_j is trained on either F_1 or F_2 and at each site in the sequence, j, the new residue is then drawn from a multinomial probability distribution based on $P_j(k')$. The denominator, $\Sigma(\Theta_k[-k])$, is included to force the probability distribution to sum to 1, such that $P_j(k'\neq k) = 1\text{-}P_j(k' = k)$. For each seeding sequence (e.g., a subtype C consensus sequence), whichever model is initially drawn to simulate the initial amino acid in a sequence is consistently applied to generate the following amino acids in that sequence. For the pooled sets, the less diverse major founder lineages were pooled together and the more diverse minor founder lineages constituted a second set. For each

pooled founder lineage alignment, at each site j, π_j was estimated as the maximum percentage of non-consensus residues in any individual alignment in that pool (i.e., this retained within-host diversity levels); however, this could be alternatively modeled such that, at each site j, π_j was randomly drawn from an individual alignment in the pooled alignment.

Sequence simulation

One thousand sequences were simulated using a probabilistic model with an empirical transition probability matrix, Θ, computed on an alignment of 172 subtype C Env sequences sampled since 2011 and seeded with a consensus sequence corresponding to sequences from an independent acutely-infected RV217 participant (id = 10066) or a subtype C consensus. Seven sets of sequences were simulated: one trained on each of the six individual training alignments separately and one on a pooled alignment of all of the training sequences. For the pooled model, the $F_1(\pi_j)$ and $F_2(\pi_j)$ were the maximum percentage of non-consensus residues found at site j in a given individual across all F_1 and F_2 alignments, respectively. The probability of drawing from each variant model was recorded, so the outputs could be analyzed separately.

Sequence analysis

Consensus sequences were computed with a majority rule. Sequences were aligned to the HXB2 reference in MAFFT v7.419 [76]. For alignments of sequences sampled from participants in RV217, the percentage of non-consensus (as well as non-gap, non-ambiguous) residues at each site was calculated as the number of residues at each site different from the majority consensus residue for that alignment divided by the total number of sequences. Polymorphic sites were defined as sites with at least one amino acid different from the consensus. For simulated sequences, the percentage of non-consensus residues and polymorphic sites were defined against the seeding sequence. Contact sites for known HIV-1 antibodies (n = 116) were previously reported in studies of natural HIV-1 infection (https://www.hiv.lanl.gov/components/sequence/HIV/featuredb/search/env_ab_search_pub.comp).

A maximum-likelihood model of pairwise sequence distance that corrects for sequence length was computed using the *dist.ml* function [77]. Sequence divergence was calculated against the seeding sequence for each alignment. Phylogenies of aligned sequences were constructed with IQ-TREE 2 [78] based on the model with the lowest BIC identified with ModelFinder [79]. The modified graph Laplacian (MGL) is computed for the distance matrix of the reconstructed phylogeny of sequences; eigenvalues calculated from the MGL define the connectivity of the phylogeny in terms of substitutions. Spectral density profile summary statistics represent different aspects of the topology of the phylogeny, such as the longest path through the phylogeny, λ^* which is a correlate of non-synonymous/synonymous substitution rates, the proportion of long versus short branching-events, ψ, which is a correlate of rate heterogeneity, and the occurrence of branching-events, η, which is a correlate of transition-transversion rates [40,41]. Spectral density profile summary statistics λ^*,ψ, and η were estimated for phylogenies reconstructed from empirical and simulated sequences. Simulated alignments were iteratively down-sampled 100 times to a random set of 10 sequences. Divergence, pairwise distance, and phylogenetic metrics were calculated on each downsampled alignment.

Subtype C sequence antigen prediction

A phylogeny for subtype C sequences was constructed with IQ-TREE 2 [78] based on the model with the lowest BIC identified with ModelFinder [79]. Divergence from the majority consensus was computed for each sequence and pairwise distances were computed for all sequences.

For the subtype C consensus, exposed residues (i.e., accessible to antibodies) were defined as we previously described [27] and B-cell epitopes were predicted with Bepi-Pred 2.0 [80] using an epitope prediction threshold of 0.5. The number of polymorphic sites among simulated sequences corresponding to predicted B-cell epitopes were quantified.

To select candidate antigen sequences, simulated sequences were filtered by those that had a non-consensus residue (with respect to the subtype C consensus) in at least three key Ab epitopes (VRC01, CAP256-VRC26.25, PGT121, 10E8, and 35O22). Of these sequences with minimal variability, the two maximally divergent sequences simulated by F_1 and two by F_2 were selected as candidate antigens. Maximally divergent sequences were selected to cover as much genetic space as possible within the simulated minimal divergence.

Structure prediction and visualization

The structure of one subunit of the Env-trimer for subtype C consensus (the seed sequence) and four in silico-derived antigenic sequences were predicted with ColabFold [81]. The alignment was prepared using MMseqs2 [82] and the structure prediction was carried out with AlphaFold2 [47]. Before feeding to the ColabFold, the signal peptide and the sequence after the transmembrane helix were removed from the sequence. The structure figure is rendered by PyMol (https://pymol.org/). If a substitution is between a pair of highly similar residues (RK, QE, QN, ED, DN, TS, SA, VI, IL, LM, and FY), the residue is colored yellow; other changes are colored red.

Statistical analysis

Shapiro's normality test was used to determine if data were normally distributed. If data were normally distributed, pairwise comparisons were made using a Student's t-test; and otherwise using a Mann-Whitney U test. Two-sample Kolmogorov-Smirnov tests were used to compare distributions. Statistical tests were only used to compare empirical data but not simulated data. Comparisons with simulated data were made by assessing inclusion/exclusion of values within the 95% confidence intervals of simulated data.

Supporting information

S1 Fig. Highlighter plots of 6 RV217 infections with multiple founder variants. For each individual, a highlighter plot is constructed from sequences sampled during acute infection using the consensus as the master sequence. The number of days post-diagnosis at which each sequence was sampled is listed to the right of each plot.
(TIF)

S2 Fig. Phylogenies of 6 RV217 infections with multiple founder variants. For each individual, a phylogeny constructed from sequences sampled during acute infection and rooted on the majority consensus sequence.
(TIF)

S3 Fig. Mismatched sites between seeding, training, and simulated sequences. Correlation plot of the percentage of mismatched non-gapped sites between the consensus of the seeding alignment, simulated alignment, and training alignment for sequences simulated under each RV217 multi-founder infection.
(TIF)

S1 File. Candidate antigen sequences simulated from a subtype C Env consensus sequence and trained on pooled founder alignments sampled from six multi-founder acute

infections in RV217.
(TXT)

Acknowledgments

We are indebted to the RV217 participants and clinical team. We also thank Julie Ake, Merlin Robb and Sandhya Vasan.

The views expressed are those of the authors and should not be construed to represent the positions of the U.S. Army, the Department of Defense, or the Department of Health and Human Services.

Author Contributions

Conceptualization: Eric Lewitus, Morgane Rolland.

Data curation: Jennifer Hoang, Yifan Li.

Formal analysis: Eric Lewitus, Morgane Rolland.

Investigation: Eric Lewitus, Yifan Li, Hongjun Bai.

Methodology: Eric Lewitus, Yifan Li, Hongjun Bai, Morgane Rolland.

Resources: Jennifer Hoang, Yifan Li.

Software: Eric Lewitus.

Visualization: Eric Lewitus, Hongjun Bai.

Writing – original draft: Eric Lewitus, Morgane Rolland.

Writing – review & editing: Eric Lewitus, Jennifer Hoang, Yifan Li, Hongjun Bai, Morgane Rolland.

References

1. Roberts JD, Bebenek K, Kunkel TA. The accuracy of reverse transcriptase from HIV-1. Science. 1988; 242(4882):1171–3. Epub 1988/11/25. https://doi.org/10.1126/science.2460925 PMID: 2460925.
2. Smyth RP, Davenport MP, Mak J. The origin of genetic diversity in HIV-1. Virus Res. 2012; 169(2):415–29. Epub 2012/06/26. https://doi.org/10.1016/j.virusres.2012.06.015 PMID: 22728444.
3. Cuevas JM, Geller R, Garijo R, Lopez-Aldeguer J, Sanjuan R. Extremely High Mutation Rate of HIV-1 In Vivo. PLoS Biol. 2015; 13(9):e1002251. Epub 2015/09/17. https://doi.org/10.1371/journal.pbio.1002251 PMID: 26375597; PubMed Central PMCID: PMC4574155.
4. Richman DD, Wrin T, Little SJ, Petropoulos CJ. Rapid evolution of the neutralizing antibody response to HIV type 1 infection. Proc Natl Acad Sci U S A. 2003; 100(7):4144–9. Epub 2003/03/20. https://doi.org/10.1073/pnas.0630530100 PMID: 12644702; PubMed Central PMCID: PMC153062.
5. Wei X, Decker JM, Wang S, Hui H, Kappes JC, Wu X, et al. Antibody neutralization and escape by HIV-1. Nature. 2003; 422(6929):307–12. Epub 2003/03/21. https://doi.org/10.1038/nature01470 PMID: 12646921.
6. Tomaras GD, Yates NL, Liu P, Qin L, Fouda GG, Chavez LL, et al. Initial B-cell responses to transmitted human immunodeficiency virus type 1: virion-binding immunoglobulin M (IgM) and IgG antibodies followed by plasma anti-gp41 antibodies with ineffective control of initial viremia. J Virol. 2008; 82(24):12449–63. Epub 2008/10/10. https://doi.org/10.1128/JVI.01708-08 PMID: 18842730; PubMed Central PMCID: PMC2593361.
7. Mikell I, Sather DN, Kalams SA, Altfeld M, Alter G, Stamatatos L. Characteristics of the earliest cross-neutralizing antibody response to HIV-1. PLoS Pathog. 2011; 7(1):e1001251. Epub 2011/01/21. https://doi.org/10.1371/journal.ppat.1001251 PMID: 21249232; PubMed Central PMCID: PMC3020924.
8. Gray ES, Madiga MC, Hermanus T, Moore PL, Wibmer CK, Tumba NL, et al. The neutralization breadth of HIV-1 develops incrementally over four years and is associated with CD4+ T cell decline and high

viral load during acute infection. J Virol. 2011; 85(10):4828–40. Epub 2011/03/11. https://doi.org/10.1128/JVI.00198-11 PMID: 21389135; PubMed Central PMCID: PMC3126191.

9. Simek MD, Rida W, Priddy FH, Pung P, Carrow E, Laufer DS, et al. Human immunodeficiency virus type 1 elite neutralizers: individuals with broad and potent neutralizing activity identified by using a high-throughput neutralization assay together with an analytical selection algorithm. J Virol. 2009; 83 (14):7337–48. Epub 2009/05/15. https://doi.org/10.1128/JVI.00110-09 PMID: 19439467; PubMed Central PMCID: PMC2704778.

10. Hraber P, Seaman MS, Bailer RT, Mascola JR, Montefiori DC, Korber BT. Prevalence of broadly neutralizing antibody responses during chronic HIV-1 infection. AIDS. 2014; 28(2):163–9. Epub 2013/12/24. https://doi.org/10.1097/QAD.0000000000000106 PMID: 24361678; PubMed Central PMCID: PMC4042313.

11. Sather DN, Armann J, Ching LK, Mavrantoni A, Sellhorn G, Caldwell Z, et al. Factors associated with the development of cross-reactive neutralizing antibodies during human immunodeficiency virus type 1 infection. J Virol. 2009; 83(2):757–69. Epub 2008/11/07. https://doi.org/10.1128/JVI.02036-08 PMID: 18987148; PubMed Central PMCID: PMC2612355.

12. Piantadosi A, Panteleeff D, Blish CA, Baeten JM, Jaoko W, McClelland RS, et al. Breadth of neutralizing antibody response to human immunodeficiency virus type 1 is affected by factors early in infection but does not influence disease progression. J Virol. 2009; 83(19):10269–74. Epub 2009/07/31. https://doi.org/10.1128/JVI.01149-09 PMID: 19640996; PubMed Central PMCID: PMC2748011.

13. Moore PL, Gray ES, Wibmer CK, Bhiman JN, Nonyane M, Sheward DJ, et al. Evolution of an HIV glycan-dependent broadly neutralizing antibody epitope through immune escape. Nat Med. 2012; 18 (11):1688–92. Epub 2012/10/23. https://doi.org/10.1038/nm.2985 PMID: 23086475; PubMed Central PMCID: PMC3494733.

14. Klein F, Diskin R, Scheid JF, Gaebler C, Mouquet H, Georgiev IS, et al. Somatic mutations of the immunoglobulin framework are generally required for broad and potent HIV-1 neutralization. Cell. 2013; 153 (1):126–38. Epub 2013/04/02. https://doi.org/10.1016/j.cell.2013.03.018 PMID: 23540694; PubMed Central PMCID: PMC3792590.

15. Wibmer CK, Bhiman JN, Gray ES, Tumba N, Abdool Karim SS, Williamson C, et al. Viral escape from HIV-1 neutralizing antibodies drives increased plasma neutralization breadth through sequential recognition of multiple epitopes and immunotypes. PLoS Pathog. 2013; 9(10):e1003738. Epub 2013/11/10. https://doi.org/10.1371/journal.ppat.1003738 PMID: 24204277; PubMed Central PMCID: PMC3814426.

16. Sather DN, Carbonetti S, Malherbe DC, Pissani F, Stuart AB, Hessell AJ, et al. Emergence of broadly neutralizing antibodies and viral coevolution in two subjects during the early stages of infection with human immunodeficiency virus type 1. J Virol. 2014; 88(22):12968–81. Epub 2014/08/15. https://doi.org/10.1128/JVI.01816-14 PMID: 25122781; PubMed Central PMCID: PMC4249098.

17. Gao F, Bonsignori M, Liao HX, Kumar A, Xia SM, Lu X, et al. Cooperation of B cell lineages in induction of HIV-1-broadly neutralizing antibodies. Cell. 2014; 158(3):481–91. Epub 2014/07/30. https://doi.org/10.1016/j.cell.2014.06.022 PMID: 25065977; PubMed Central PMCID: PMC4150607.

18. Landais E, Huang X, Havenar-Daughton C, Murrell B, Price MA, Wickramasinghe L, et al. Broadly Neutralizing Antibody Responses in a Large Longitudinal Sub-Saharan HIV Primary Infection Cohort. PLoS Pathog. 2016; 12(1):e1005369. Epub 2016/01/15. https://doi.org/10.1371/journal.ppat.1005369 PMID: 26766578; PubMed Central PMCID: PMC4713061.

19. Rusert P, Kouyos RD, Kadelka C, Ebner H, Schanz M, Huber M, et al. Determinants of HIV-1 broadly neutralizing antibody induction. Nat Med. 2016; 22(11):1260–7. Epub 2016/11/01. https://doi.org/10.1038/nm.4187 PMID: 27668936.

20. Kouyos RD, Rusert P, Kadelka C, Huber M, Marzel A, Ebner H, et al. Tracing HIV-1 strains that imprint broadly neutralizing antibody responses. Nature. 2018; 561(7723):406–10. Epub 2018/09/12. https://doi.org/10.1038/s41586-018-0517-0 PMID: 30202088.

21. Luo S, Perelson AS. Competitive exclusion by autologous antibodies can prevent broad HIV-1 antibodies from arising. Proc Natl Acad Sci U S A. 2015; 112(37):11654–9. Epub 2015/09/02. https://doi.org/10.1073/pnas.1505207112 PMID: 26324897; PubMed Central PMCID: PMC4577154.

22. Powell RL, Kinge T, Nyambi PN. Infection by discordant strains of HIV-1 markedly enhances the neutralizing antibody response against heterologous virus. J Virol. 2010; 84(18):9415–26. Epub 2010/07/16. https://doi.org/10.1128/JVI.02732-09 PMID: 20631143; PubMed Central PMCID: PMC2937625.

23. Cortez V, Odem-Davis K, McClelland RS, Jaoko W, Overbaugh J. HIV-1 superinfection in women broadens and strengthens the neutralizing antibody response. PLoS Pathog. 2012; 8(3):e1002611. Epub 2012/04/06. https://doi.org/10.1371/journal.ppat.1002611 PMID: 22479183; PubMed Central PMCID: PMC3315492.

24. Sheward DJ, Marais J, Bekker V, Murrell B, Eren K, Bhiman JN, et al. HIV Superinfection Drives De Novo Antibody Responses and Not Neutralization Breadth. Cell Host Microbe. 2018; 24(4):593–9 e3. Epub 2018/10/03. https://doi.org/10.1016/j.chom.2018.09.001 PMID: 30269971; PubMed Central PMCID: PMC6185870.

25. Rolland M, Tovanabutra S, Dearlove B, Li Y, Owen CL, Lewitus E, et al. Molecular dating and viral load growth rates suggested that the eclipse phase lasted about a week in HIV-1 infected adults in East Africa and Thailand. PLoS Pathog. 2020; 16(2):e1008179. Epub 2020/02/07. https://doi.org/10.1371/journal.ppat.1008179 PMID: 32027734; PubMed Central PMCID: PMC7004303.

26. Dearlove B, Tovanabutra S, Owen CL, Lewitus E, Li Y, Sanders-Buell E, et al. Factors influencing estimates of HIV-1 infection timing using BEAST. PLoS Comput Biol. 2021; 17(2):e1008537. Epub 2021/02/02. https://doi.org/10.1371/journal.pcbi.1008537 PMID: 33524022; PubMed Central PMCID: PMC7877758.

27. Lewitus E, Townsley SM, Li Y, Donofrio GC, Dearlove BL, Bai H, et al. HIV-1 infections with multiple founders associate with the development of neutralization breadth. PLoS Pathog. 2022; 18(3): e1010369. Epub 2022/03/19. https://doi.org/10.1371/journal.ppat.1010369 PMID: 35303045; PubMed Central PMCID: PMC8967031.

28. Townsley SM, Donofrio GC, Jian N, Leggat DJ, Dussupt V, Mendez-Rivera L, et al. B cell engagement with HIV-1 founder virus envelope predicts development of broadly neutralizing antibodies. Cell Host Microbe. 2021. Epub 2021/03/05. https://doi.org/10.1016/j.chom.2021.01.016 PMID: 33662277.

29. Robb ML, Eller LA, Kibuuka H, Rono K, Maganga L, Nitayaphan S, et al. Prospective Study of Acute HIV-1 Infection in Adults in East Africa and Thailand. N Engl J Med. 2016; 374(22):2120–30. Epub 2016/05/19. https://doi.org/10.1056/NEJMoa1508952 PMID: 27192360; PubMed Central PMCID: PMC5111628.

30. Rolland M, Edlefsen PT, Larsen BB, Tovanabutra S, Sanders-Buell E, Hertz T, et al. Increased HIV-1 vaccine efficacy against viruses with genetic signatures in Env V2. Nature. 2012. Epub 2012/09/11. https://doi.org/10.1038/nature11519 [pii]. PMID: 22960785.

31. Lewitus E, Sanders-Buell E, Bose M, O'Sullivan AM, Poltavee K, Li Y, et al. RV144 vaccine imprinting constrained HIV-1 evolution following breakthrough infection. Virus Evol. 2021; 7(2):veab057. Epub 2021/09/18. https://doi.org/10.1093/ve/veab057 PMID: 34532060; PubMed Central PMCID: PMC8438874.

32. Marichannegowda MH, Mengual M, Kumar A, Giorgi EE, Tu JJ, Martinez DR, et al. Different evolutionary pathways of HIV-1 between fetus and mother perinatal transmission pairs indicate unique immune selection in fetuses. Cell Rep Med. 2021; 2(7):100315. Epub 2021/08/03. https://doi.org/10.1016/j.xcrm.2021.100315 PMID: 34337555; PubMed Central PMCID: PMC8324465.

33. Mishra N, Sharma S, Dobhal A, Kumar S, Chawla H, Singh R, et al. Broadly neutralizing plasma antibodies effective against autologous circulating viruses in infants with multivariant HIV-1 infection. Nat Commun. 2020; 11(1):4409. Epub 2020/09/04. https://doi.org/10.1038/s41467-020-18225-x PMID: 32879304; PubMed Central PMCID: PMC7468291.

34. Keele BF, Giorgi EE, Salazar-Gonzalez JF, Decker JM, Pham KT, Salazar MG, et al. Identification and characterization of transmitted and early founder virus envelopes in primary HIV-1 infection. Proc Natl Acad Sci U S A. 2008; 105(21):7552–7. Epub 2008/05/21. https://doi.org/10.1073/pnas.0802203105 PMID: 18490657; PubMed Central PMCID: PMC2387184.

35. Abrahams MR, Anderson JA, Giorgi EE, Seoighe C, Mlisana K, Ping LH, et al. Quantitating the multiplicity of infection with human immunodeficiency virus type 1 subtype C reveals a non-poisson distribution of transmitted variants. J Virol. 2009; 83(8):3556–67. Epub 2009/02/06. https://doi.org/10.1128/JVI.02132-08 PMID: 19193811; PubMed Central PMCID: PMC2663249.

36. Janes H, Herbeck JT, Tovanabutra S, Thomas R, Frahm N, Duerr A, et al. HIV-1 infections with multiple founders are associated with higher viral loads than infections with single founders. Nat Med. 2015; 21 (10):1139–41. Epub 2015/09/01. https://doi.org/10.1038/nm.3932 PMID: 26322580; PubMed Central PMCID: PMC4598284.

37. Tully DC, Ogilvie CB, Batorsky RE, Bean DJ, Power KA, Ghebremichael M, et al. Differences in the Selection Bottleneck between Modes of Sexual Transmission Influence the Genetic Composition of the HIV-1 Founder Virus. PLoS Pathog. 2016; 12(5):e1005619. Epub 2016/05/11. https://doi.org/10.1371/journal.ppat.1005619 PMID: 27163788; PubMed Central PMCID: PMC4862634.

38. Corey L, Gilbert PB, Juraska M, Montefiori DC, Morris L, Karuna ST, et al. Two Randomized Trials of Neutralizing Antibodies to Prevent HIV-1 Acquisition. N Engl J Med. 2021; 384(11):1003–14. Epub 2021/03/18. https://doi.org/10.1056/NEJMoa2031738 PMID: 33730454; PubMed Central PMCID: PMC8189692.

39. Le SQ, Gascuel O. An improved general amino acid replacement matrix. Mol Biol Evol. 2008; 25 (7):1307–20. Epub 2008/03/28. https://doi.org/10.1093/molbev/msn067 PMID: 18367465.

40. Lewitus E, Morlon H. Characterizing and Comparing Phylogenies from their Laplacian Spectrum. Syst Biol. 2016; 65(3):495–507. Epub 2015/12/15. https://doi.org/10.1093/sysbio/syv116 PMID: 26658901.

41. Lewitus E, Rolland M. A non-parametric analytic framework for within-host viral phylogenies and a test for HIV-1 founder multiplicity. Virus Evol. 2019; 5(2):vez044. Epub 2019/11/09. https://doi.org/10.1093/ve/vez044 PMID: 31700680; PubMed Central PMCID: PMC6826062.

42. Wu X, Yang ZY, Li Y, Hogerkorp CM, Schief WR, Seaman MS, et al. Rational design of envelope identifies broadly neutralizing human monoclonal antibodies to HIV-1. Science. 2010; 329(5993):856–61. Epub 2010/07/10. https://doi.org/10.1126/science.1187659 PMID: 20616233; PubMed Central PMCID: PMC2965066.

43. Doria-Rose NA, Bhiman JN, Roark RS, Schramm CA, Gorman J, Chuang GY, et al. New Member of the V1V2-Directed CAP256-VRC26 Lineage That Shows Increased Breadth and Exceptional Potency. J Virol. 2016; 90(1):76–91. Epub 2015/10/16. https://doi.org/10.1128/JVI.01791-15 PMID: 26468542; PubMed Central PMCID: PMC4702551.

44. Julien JP, Sok D, Khayat R, Lee JH, Doores KJ, Walker LM, et al. Broadly neutralizing antibody PGT121 allosterically modulates CD4 binding via recognition of the HIV-1 gp120 V3 base and multiple surrounding glycans. PLoS Pathog. 2013; 9(5):e1003342. Epub 2013/05/10. https://doi.org/10.1371/journal.ppat.1003342 PMID: 23658524; PubMed Central PMCID: PMC3642082.

45. Soto C, Ofek G, Joyce MG, Zhang B, McKee K, Longo NS, et al. Developmental Pathway of the MPER-Directed HIV-1-Neutralizing Antibody 10E8. PLoS One. 2016; 11(6):e0157409. Epub 2016/06/15. https://doi.org/10.1371/journal.pone.0157409 PMID: 27299673; PubMed Central PMCID: PMC4907498.

46. Huang J, Kang BH, Pancera M, Lee JH, Tong T, Feng Y, et al. Broad and potent HIV-1 neutralization by a human antibody that binds the gp41-gp120 interface. Nature. 2014; 515(7525):138–42. Epub 2014/09/05. https://doi.org/10.1038/nature13601 PMID: 25186731; PubMed Central PMCID: PMC4224615.

47. Jumper J, Evans R, Pritzel A, Green T, Figurnov M, Ronneberger O, et al. Highly accurate protein structure prediction with AlphaFold. Nature. 2021; 596(7873):583–9. Epub 2021/07/16. https://doi.org/10.1038/s41586-021-03819-2 PMID: 34265844; PubMed Central PMCID: PMC8371605.

48. Pitisuttithum P, Gilbert P, Gurwith M, Heyward W, Martin M, van Griensven F, et al. Randomized, double-blind, placebo-controlled efficacy trial of a bivalent recombinant glycoprotein 120 HIV-1 vaccine among injection drug users in Bangkok, Thailand. J Infect Dis. 2006; 194(12):1661–71. Epub 2006/11/17. https://doi.org/10.1086/508748 PMID: 17109337.

49. Flynn NM, Forthal DN, Harro CD, Judson FN, Mayer KH, Para MF, et al. Placebo-controlled phase 3 trial of a recombinant glycoprotein 120 vaccine to prevent HIV-1 infection. J Infect Dis. 2005; 191(5):654–65. Epub 2005/02/03. https://doi.org/10.1086/428404 PMID: 15688278.

50. Buchbinder SP, Mehrotra DV, Duerr A, Fitzgerald DW, Mogg R, Li D, et al. Efficacy assessment of a cell-mediated immunity HIV-1 vaccine (the Step Study): a double-blind, randomised, placebo-controlled, test-of-concept trial. Lancet. 2008; 372(9653):1881–93. Epub 2008/11/18. https://doi.org/10.1016/S0140-6736(08)61591-3 PMID: 19012954; PubMed Central PMCID: PMC2721012.

51. Rerks-Ngarm S, Pitisuttithum P, Nitayaphan S, Kaewkungwal J, Chiu J, Paris R, et al. Vaccination with ALVAC and AIDSVAX to prevent HIV-1 infection in Thailand. N Engl J Med. 2009; 361(23):2209–20. Epub 2009/10/22. https://doi.org/10.1056/NEJMoa0908492 PMID: 19843557.

52. Gray GE, Allen M, Moodie Z, Churchyard G, Bekker LG, Nchabeleng M, et al. Safety and efficacy of the HVTN 503/Phambili study of a clade-B-based HIV-1 vaccine in South Africa: a double-blind, randomised, placebo-controlled test-of-concept phase 2b study. Lancet Infect Dis. 2011; 11(7):507–15. Epub 2011/05/17. https://doi.org/10.1016/S1473-3099(11)70098-6 PMID: 21570355; PubMed Central PMCID: PMC3417349.

53. Hammer SM, Sobieszczyk ME, Janes H, Karuna ST, Mulligan MJ, Grove D, et al. Efficacy trial of a DNA/rAd5 HIV-1 preventive vaccine. N Engl J Med. 2013; 369(22):2083–92. Epub 2013/10/09. https://doi.org/10.1056/NEJMoa1310566 PMID: 24099601; PubMed Central PMCID: PMC4030634.

54. Gray GE, Bekker LG, Laher F, Malahleha M, Allen M, Moodie Z, et al. Vaccine Efficacy of ALVAC-HIV and Bivalent Subtype C gp120-MF59 in Adults. N Engl J Med. 2021; 384(12):1089–100. Epub 2021/03/25. https://doi.org/10.1056/NEJMoa2031499 PMID: 33761206; PubMed Central PMCID: PMC7888373.

55. Chung AW, Ghebremichael M, Robinson H, Brown E, Choi I, Lane S, et al. Polyfunctional Fc-effector profiles mediated by IgG subclass selection distinguish RV144 and VAX003 vaccines. Sci Transl Med. 2014; 6(228):228ra38. Epub 2014/03/22. https://doi.org/10.1126/scitranslmed.3007736 PMID: 24648341.

56. Mdluli T, Jian N, Slike B, Paquin-Proulx D, Donofrio G, Alrubayyi A, et al. RV144 HIV-1 vaccination impacts post-infection antibody responses. PLoS Pathog. 2020; 16(12):e1009101. Epub 2020/12/09.

https://doi.org/10.1371/journal.ppat.1009101 PMID: 33290394; PubMed Central PMCID: PMC7748270.

57. Ren X, Li Y, Liu X, Shen X, Gao W, Li J. Computational Identification of Antigenicity-Associated Sites in the Hemagglutinin Protein of A/H1N1 Seasonal Influenza Virus. PLoS One. 2015; 10(5):e0126742. Epub 2015/05/16. https://doi.org/10.1371/journal.pone.0126742 PMID: 25978416; PubMed Central PMCID: PMC4433265.

58. Steinbruck L, McHardy AC. Inference of genotype-phenotype relationships in the antigenic evolution of human influenza A (H3N2) viruses. PLoS Comput Biol. 2012; 8(4):e1002492. Epub 2012/04/26. https://doi.org/10.1371/journal.pcbi.1002492 PMID: 22532796; PubMed Central PMCID: PMC3330098.

59. Neher RA, Bedford T, Daniels RS, Russell CA, Shraiman BI. Prediction, dynamics, and visualization of antigenic phenotypes of seasonal influenza viruses. Proc Natl Acad Sci U S A. 2016; 113(12):E1701–9. Epub 2016/03/10. https://doi.org/10.1073/pnas.1525578113 PMID: 26951657; PubMed Central PMCID: PMC4812706.

60. Kratsch C, Klingen TR, Mumken L, Steinbruck L, McHardy AC. Determination of antigenicity-altering patches on the major surface protein of human influenza A/H3N2 viruses. Virus Evol. 2016; 2(1): vev025. Epub 2016/10/25. https://doi.org/10.1093/ve/vev025 PMID: 27774294; PubMed Central PMCID: PMC4989879.

61. Luksza M, Lassig M. A predictive fitness model for influenza. Nature. 2014; 507(7490):57–61. Epub 2014/02/28. https://doi.org/10.1038/nature13087 PMID: 24572367.

62. Korber B, Gaschen B, Yusim K, Thakallapally R, Kesmir C, Detours V. Evolutionary and immunological implications of contemporary HIV-1 variation. Br Med Bull. 2001; 58:19–42. Epub 2001/11/21. https://doi.org/10.1093/bmb/58.1.19 PMID: 11714622.

63. Gaschen B, Taylor J, Yusim K, Foley B, Gao F, Lang D, et al. Diversity considerations in HIV-1 vaccine selection. Science. 2002; 296(5577):2354–60. Epub 2002/06/29. https://doi.org/10.1126/science.1070441 PMID: 12089434.

64. Rolland M. HIV-1 phylogenetics and vaccines. Curr Opin HIV AIDS. 2019; 14(3):227–32. Epub 2019/03/30. https://doi.org/10.1097/COH.0000000000000545 PMID: 30925535.

65. Dearlove B, Lewitus E, Bai H, Li Y, Reeves DB, Joyce MG, et al. A SARS-CoV-2 vaccine candidate would likely match all currently circulating variants. Proc Natl Acad Sci U S A. 2020; 117(38):23652–62. Epub 2020/09/02. https://doi.org/10.1073/pnas.2008281117 PMID: 32868447; PubMed Central PMCID: PMC7519301.

66. Rolland M, Jensen MA, Nickle DC, Yan J, Learn GH, Heath L, et al. Reconstruction and function of ancestral center-of-tree human immunodeficiency virus type 1 proteins. J Virol. 2007; 81(16):8507–14. Epub 2007/06/01. https://doi.org/10.1128/JVI.02683-06 PMID: 17537854; PubMed Central PMCID: PMC1951385.

67. Fischer W, Perkins S, Theiler J, Bhattacharya T, Yusim K, Funkhouser R, et al. Polyvalent vaccines for optimal coverage of potential T-cell epitopes in global HIV-1 variants. Nat Med. 2007; 13(1):100–6. Epub 2006/12/26. https://doi.org/10.1038/nm1461 PMID: 17187074.

68. Barouch DH O'Brien KL, Simmons NL, King SL, Abbink P, Maxfield LF, et al. Mosaic HIV-1 vaccines expand the breadth and depth of cellular immune responses in rhesus monkeys. Nat Med. 2010; 16(3):319–23. Epub 2010/02/23. https://doi.org/10.1038/nm.2089 PMID: 20173752; PubMed Central PMCID: PMC2834868.

69. Barouch DH, Tomaka FL, Wegmann F, Stieh DJ, Alter G, Robb ML, et al. Evaluation of a mosaic HIV-1 vaccine in a multicentre, randomised, double-blind, placebo-controlled, phase 1/2a clinical trial (APPROACH) and in rhesus monkeys (NHP 13–19). Lancet. 2018; 392(10143):232–43. Epub 2018/07/27. https://doi.org/10.1016/S0140-6736(18)31364-3 PMID: 30047376; PubMed Central PMCID: PMC6192527.

70. Letourneau S, Im EJ, Mashishi T, Brereton C, Bridgeman A, Yang H, et al. Design and pre-clinical evaluation of a universal HIV-1 vaccine. PLoS One. 2007; 2(10):e984. Epub 2007/10/04. https://doi.org/10.1371/journal.pone.0000984 PMID: 17912361; PubMed Central PMCID: PMC1991584.

71. Rolland M, Nickle DC, Mullins JI. HIV-1 group M conserved elements vaccine. PLoS Pathog. 2007; 3(11):e157. Epub 2007/12/07. https://doi.org/10.1371/journal.ppat.0030157 PMID: 18052528; PubMed Central PMCID: PMC2098811.

72. Gaiha GD, Rossin EJ, Urbach J, Landeros C, Collins DR, Nwonu C, et al. Structural topology defines protective CD8(+) T cell epitopes in the HIV proteome. Science. 2019; 364(6439):480–4. Epub 2019/05/03. https://doi.org/10.1126/science.aav5095 PMID: 31048489; PubMed Central PMCID: PMC6855781.

73. Moyo N, Wee EG, Korber B, Bahl K, Falcone S, Himansu S, et al. Tetravalent Immunogen Assembled from Conserved Regions of HIV-1 and Delivered as mRNA Demonstrates Potent Preclinical T-Cell

Immunogenicity and Breadth. Vaccines (Basel). 2020; 8(3). Epub 2020/07/10. https://doi.org/10.3390/vaccines8030360 PMID: 32640600; PubMed Central PMCID: PMC7563622.

74. Jardine J, Julien JP, Menis S, Ota T, Kalyuzhniy O, McGuire A, et al. Rational HIV immunogen design to target specific germline B cell receptors. Science. 2013; 340(6133):711–6. Epub 2013/03/30. https://doi.org/10.1126/science.1234150 PMID: 23539181; PubMed Central PMCID: PMC3689846.

75. Rose PP, Korber BT. Detecting hypermutations in viral sequences with an emphasis on G—> A hypermutation. Bioinformatics. 2000; 16(4):400–1. Epub 2000/06/27. https://doi.org/10.1093/bioinformatics/16.4.400 PMID: 10869039.

76. Katoh K, Standley DM. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. Mol Biol Evol. 2013; 30(4):772–80. Epub 2013/01/19. https://doi.org/10.1093/molbev/mst010 PMID: 23329690; PubMed Central PMCID: PMC3603318.

77. Schliep KP. phangorn: phylogenetic analysis in R. Bioinformatics. 2011; 27(4):592–3. Epub 2010/12/21. https://doi.org/10.1093/bioinformatics/btq706 PMID: 21169378; PubMed Central PMCID: PMC3035803.

78. Minh BQ, Schmidt HA, Chernomor O, Schrempf D, Woodhams MD, von Haeseler A, et al. IQ-TREE 2: New Models and Efficient Methods for Phylogenetic Inference in the Genomic Era. Mol Biol Evol. 2020; 37(5):1530–4. Epub 2020/02/06. https://doi.org/10.1093/molbev/msaa015 PMID: 32011700; PubMed Central PMCID: PMC7182206.

79. Kalyaanamoorthy S, Minh BQ, Wong TKF, von Haeseler A, Jermiin LS. ModelFinder: fast model selection for accurate phylogenetic estimates. Nat Methods. 2017; 14(6):587–9. Epub 2017/05/10. https://doi.org/10.1038/nmeth.4285 PMID: 28481363; PubMed Central PMCID: PMC5453245.

80. Jespersen MC, Peters B, Nielsen M, Marcatili P. BepiPred-2.0: improving sequence-based B-cell epitope prediction using conformational epitopes. Nucleic Acids Res. 2017; 45(W1):W24–W9. Epub 2017/05/05. https://doi.org/10.1093/nar/gkx346 PMID: 28472356; PubMed Central PMCID: PMC5570230.

81. Mirdita M, Schutze K, Moriwaki Y, Heo L, Ovchinnikov S, Steinegger M. ColabFold: making protein folding accessible to all. Nat Methods. 2022; 19(6):679–82. Epub 2022/06/01. https://doi.org/10.1038/s41592-022-01488-1 PMID: 35637307; PubMed Central PMCID: PMC9184281.

82. Mirdita M, Steinegger M, Soding J. MMseqs2 desktop and local web server app for fast, interactive sequence searches. Bioinformatics. 2019; 35(16):2856–8. Epub 2019/01/08. https://doi.org/10.1093/bioinformatics/bty1057 PMID: 30615063; PubMed Central PMCID: PMC6691333.