PLOS COMPUTATIONAL BIOLOGY

RESEARCH ARTICLE

Full-Length Envelope Analyzer (FLEA): A tool for longitudinal analysis of viral amplicons

Kemal Eren[1], Steven Weaver[2], Robert Ketteringham[3], Morné Valentyn[3], Melissa Laird Smith[4,5], Venkatesh Kumar[6], Sanjay Mohan[6], Sergei L. Kosakovsky Pond[2], Ben Murrell[6,7]*

1 Bioinformatics and Systems Biology, University of California San Diego, La Jolla, CA, USA, **2** Department of Biology and Institute for Genomics and Evolutionary Medicine, Temple University, Philadelphia, PA, USA, **3** Division of Medical Virology, Department of Pathology, Institute of Infectious Diseases and Molecular Medicine, University of Cape Town, Cape Town, Western Cape, South Africa, **4** Department of Genetics and Genomic Sciences, Icahn School of Medicine at Mount Sinai, New York, NY, USA, **5** Icahn Institute of Genomics and Multiscale Biology, Icahn School of Medicine at Mount Sinai, New York, NY, USA, **6** Division of Infectious Diseases and Global Public Health, Department of Medicine, University of California San Diego, La Jolla, CA, USA, **7** Department of Microbiology, Tumor and Cell Biology, Karolinska Institutet, Stockholm, Sweden

* benjamin.murrell@ki.se

Citation: Eren K, Weaver S, Ketteringham R, Valentyn M, Laird Smith M, Kumar V, et al. (2018) Full-Length Envelope Analyzer (FLEA): A tool for longitudinal analysis of viral amplicons. PLoS Comput Biol 14(12): e1006498. https://doi.org/10.1371/journal.pcbi.1006498

Editor: Timothée Poisot, Universite de Montreal, CANADA

Received: January 4, 2018

Accepted: September 10, 2018

Published: December 13, 2018

Data Availability Statement: All data are available from the NCBI Sequence Read Archive under BioProject PRJNA320111 (https://www.ncbi.nlm.nih.gov/bioproject/PRJNA320111/). Processed data are available at http://flea.murrell.group/view/P018.

Funding: Research reported here was supported by the National Institutes of Health: the National Institute Of Allergy And Infectious Diseases under award numbers R00AI120851, UM1AI068618, and R01AI120009, the National Institute on Drug

Abstract

Next generation sequencing of viral populations has advanced our understanding of viral population dynamics, the development of drug resistance, and escape from host immune responses. Many applications require complete gene sequences, which can be impossible to reconstruct from short reads. HIV *env*, the protein of interest for HIV vaccine studies, is exceptionally challenging for long-read sequencing and analysis due to its length, high substitution rate, and extensive indel variation. While long-read sequencing is attractive in this setting, the analysis of such data is not well handled by existing methods. To address this, we introduce FLEA (Full-Length Envelope Analyzer), which performs end-to-end analysis and visualization of long-read sequencing data. FLEA consists of both a pipeline (optionally run on a high-performance cluster), and a client-side web application that provides interactive results. The pipeline transforms FASTQ reads into high-quality consensus sequences (HQCSs) and uses them to build a codon-aware multiple sequence alignment. The resulting alignment is then used to infer phylogenies, selection pressure, and evolutionary dynamics. The web application provides publication-quality plots and interactive visualizations, including an annotated viral alignment browser, time series plots of evolutionary dynamics, visualizations of gene-wide selective pressures (such as *dN*/*dS*) across time and across protein structure, and a phylogenetic tree browser. We demonstrate how FLEA may be used to process Pacific Biosciences HIV *env* data and describe recent examples of its use. Simulations show how FLEA dramatically reduces the error rate of this sequencing platform, providing an accurate portrait of complex and variable HIV *env* populations. A public instance of FLEA is hosted at http://flea.datamonkey.org. The Python source code for the FLEA pipeline can be found at https://github.com/veg/flea-pipeline. The client-side application is available at https://github.com/veg/flea-web-app. A live demo of the P018 results can be found at http://flea.murrell.group/view/P018.

Abuse under R33DA041007, and the National Institute of General Medical Sciences under U01GM110749, and in part by the University of California, San Diego Center for AIDS Research (P30AI036214). KE was supported in part by R21AI115701. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

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

Author summary

Viral populations constantly evolve and diversify. In this article we introduce a method, FLEA, for reconstructing and visualizing the details of evolutionary changes. FLEA specifically processes data from sequencing platforms that generate reads that are long, but error-prone. To study the evolutionary dynamics of entire genes during viral infection, data is collected via long-read sequencing at discrete time points, allowing us to understand how the virus changes over time. However, the experimental and sequencing process is imperfect, so the resulting data contain not only real evolutionary changes, but also mutations and other genetic artifacts caused by sequencing errors. Our method corrects most of these errors by combining thousands of erroneous sequences into a much smaller number of unique consensus sequences that represent biologically meaningful variation. The resulting high-quality sequences are used for further analysis, such as building an evolutionary tree that tracks and interprets the genetic changes in the viral population over time. FLEA is open source, and is freely available online.

This is a *PLOS Computational Biology* Software paper.

Introduction

Next generation sequencing (NGS) has become an invaluable tool for studying HIV and other rapidly evolving viruses by providing direct high resolution measurements of viral genetic diversity within the host. NGS has been used to study immune escape [1–7], drug resistance [3, 7–12], transmission bottlenecks [3, 13–15], population structure and dynamics [2, 3, 16–22], tropism dynamics [23], and multiplicity of infection [24]. It is also used in clinical virology [25, 26]. For reviews of the promises and challenges of NGS applications in virology, see [27, 28], [29], and [30].

Full-length sequences can resolve features that are difficult to assemble from short sequences [8, 31]. For instance, Pacific Biosciences SMRT sequences were able to resolve 1.5 kb *msg* isoforms from *Pneumocystis jirovecii*, but reads from a 454 instrument could not be assembled correctly [31]. For tracking evolutionary patterns in viral populations, accurately resolving these features provides a more accurate history of the population, which becomes especially relevant when epistatic interactions and linkage between mutations effect phenotypic changes in the pathogen [32–34]. For example, studies of HIV *env* frequently use functional assays to measure the potency with which a given antibody or donor serum neutralizes a specific *env* strain [35], which requires knowing the full *env* sequence.

We have developed a pipeline for handling long read HIV *env* sequencing data from within-host viral populations: the Full-Length Envelope Analyzer (FLEA). FLEA addresses the specific challenges posed by large volumes of such data, e.g., using the sequencing protocols we previously described in Laird Smith *et al* [36], which also contains an overview of a prototype of FLEA. Here we describe the full pipeline and experimentally demonstrate its ability to resolve populations of closely related variants. FLEA uses state-of-the-art tools and methods at every step and can be accessed through a web browser or on a high-performance cluster. FLEA is readily extensible to other genes and systems.

FLEA has recently been used by the authors in two high-profile studies. In [37], we describe how FLEA was used to process PacBio HIV *env* data from a clinical trial of monoclonal antibody 10-1074. For sequences sampled before and after therapy, FLEA reveals that prior to antibody therapy low-frequency *env* variants were present with mutations that typically confer resistance to 10-1074. Additionally, when resistance emerges, it emerges multiple times, exploiting many different resistance pathways. FLEA was also used to characterize the longitudinal *env* population that drove development of a broadly neutralizing antibodies against the apex of the *env* trimer, sampled from donor PC64 from the Protocol C primary infection cohort [38].

There exist dozens of standalone pipelines developed for analyzing HIV and related sequence data, including longitudinal samples [4, 9, 12, 39]. However, it was necessary to develop a new tool due to HIV *env*'s extensive natural indel variation and the high rate of indels in long PacBio reads, which are especially problematic when any spurious indel in the 2.6kb *env* amplicon corrupts the reading frame, rendering the sequence uninterpretable. Previous analysis [36] determined that, for PacBio amplicon sequencing of an Env clone (for the set of sequencing and filtering conditions employed therein), 4 out of 5 errors are indels, and these occur more commonly in long homopolymer runs, with the per-base error rate ranging from 1 in 300 for a homopolymer of length 2, but up to nearly 1 in 50 for a homopolymer of length 6. On average, the per-base error rate was around 1 in 200, yielding an average of roughly 13 errors per 2.6kb sequence.

With HIV *env*, the common strategy of mapping reads to a reference fails because the diversity in variable regions of *env*, predominantly driven by extensive long insertions and deletions, means that these regions in sampled reads lack homology to those in any heterologous reference sequence, causing alignment-to-reference strategies to fail. Instead, FLEA relies on a fine-grained cluster-and-consensus strategy to remove spurious indels from reads. The task is related to Liang *et al.* (2016), but, rather than distinguishing a small number of variants at 81-91% identity, we must distinguish potentially hundreds of variants that differ by only a handful of bases.

The main contribution of the FLEA pipeline, therefore, is the reconstruction of a population, including accurate inference of relative frequencies of closely-related minority variants, from SMRT sequences alone. In addition, it performs many useful analyses on this population, such as alignment, phylogenetic reconstruction, and selection inference, and provides interactive visualizations for the results. The full pipeline is available as an online resource, or for local installation.

Design and implementation

Pipeline

The input to FLEA is a set of FASTQ files from the PacBio RS-II or Sequel. Each set corresponds to one time point, containing circular consensus sequence (CCS) reads, which can be obtained using the "Reads of Insert" protocol on PacBio's SMRTportal or SMRTanalysis tools. Upon completion, the FLEA pipeline produces results as JSON (Javascript Object Notation) files, a standard format for machine (and human-) readable structured data. The logic of FLEA is implemented in Nextflow [40], a workflow framework for deploying parallel pipelines to clusters and clouds.

FLEA consists of multiple sub-pipelines, as shown in Fig 1. Details of the quality and consensus pipelines are depicted in Fig 2. Together, these two pipelines take error-prone CCS reads and convert them into unique high-quality consensus sequences. The alignment pipeline generates a multiple sequence alignment, which is used by multiple methods in the analysis pipeline.

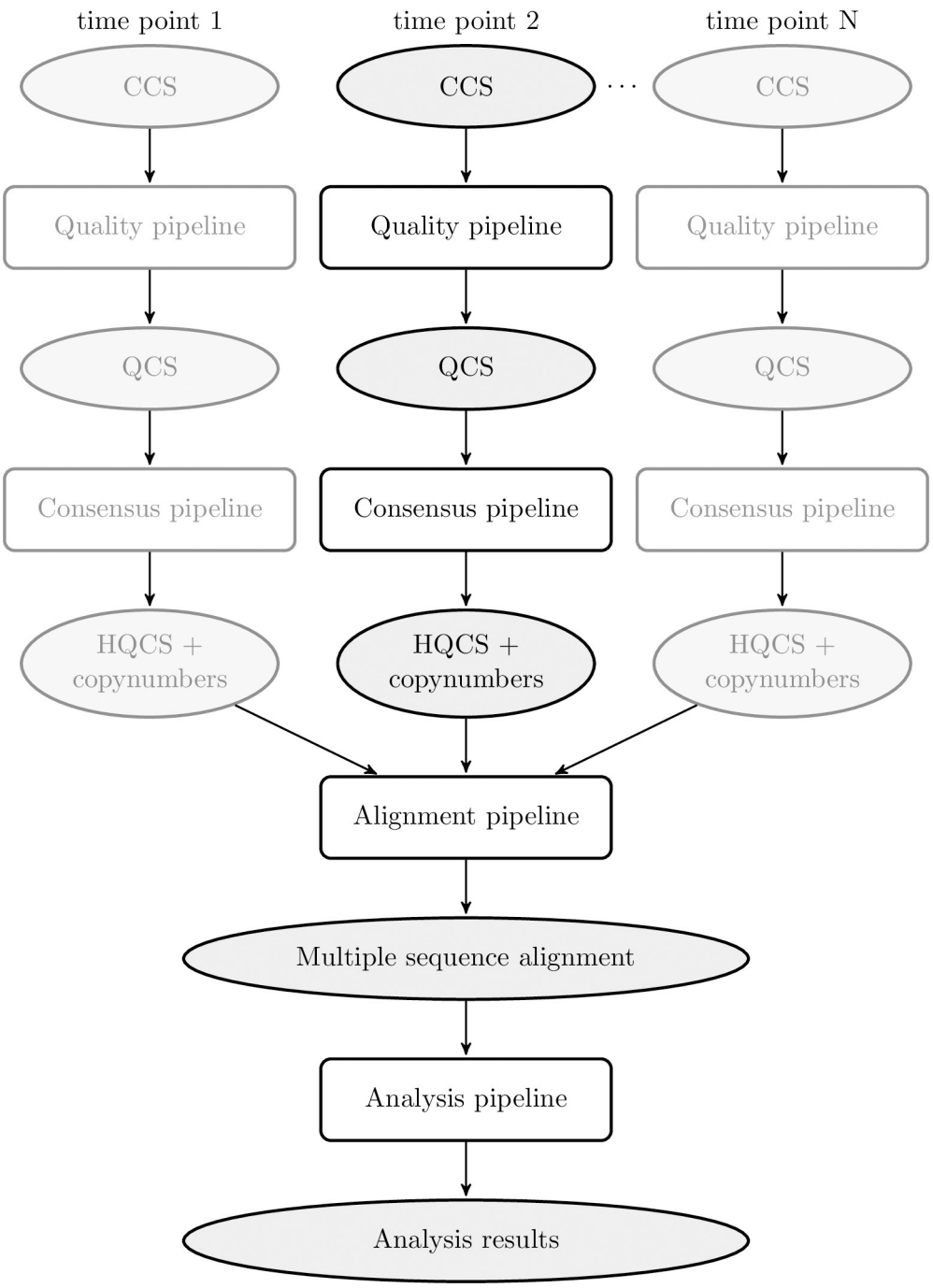

Fig 1. Overview of the FLEA pipeline, broken into conceptual sub-pipelines. The *Quality* and *Consensus* sub-pipelines process each time point separately. Duplicate steps in other time points are grayed out. CCS stands for "circular consensus sequences"; QCS for "quality-controlled sequences", and HQCS for "high-quality consensus sequences".

https://doi.org/10.1371/journal.pcbi.1006498.g001

Quality assurance sub-pipeline

The first steps remove low quality reads and filter out common sequencing artifacts. Parameters given in these steps were chosen for full-length HIV envelope sequences from the RS-II or

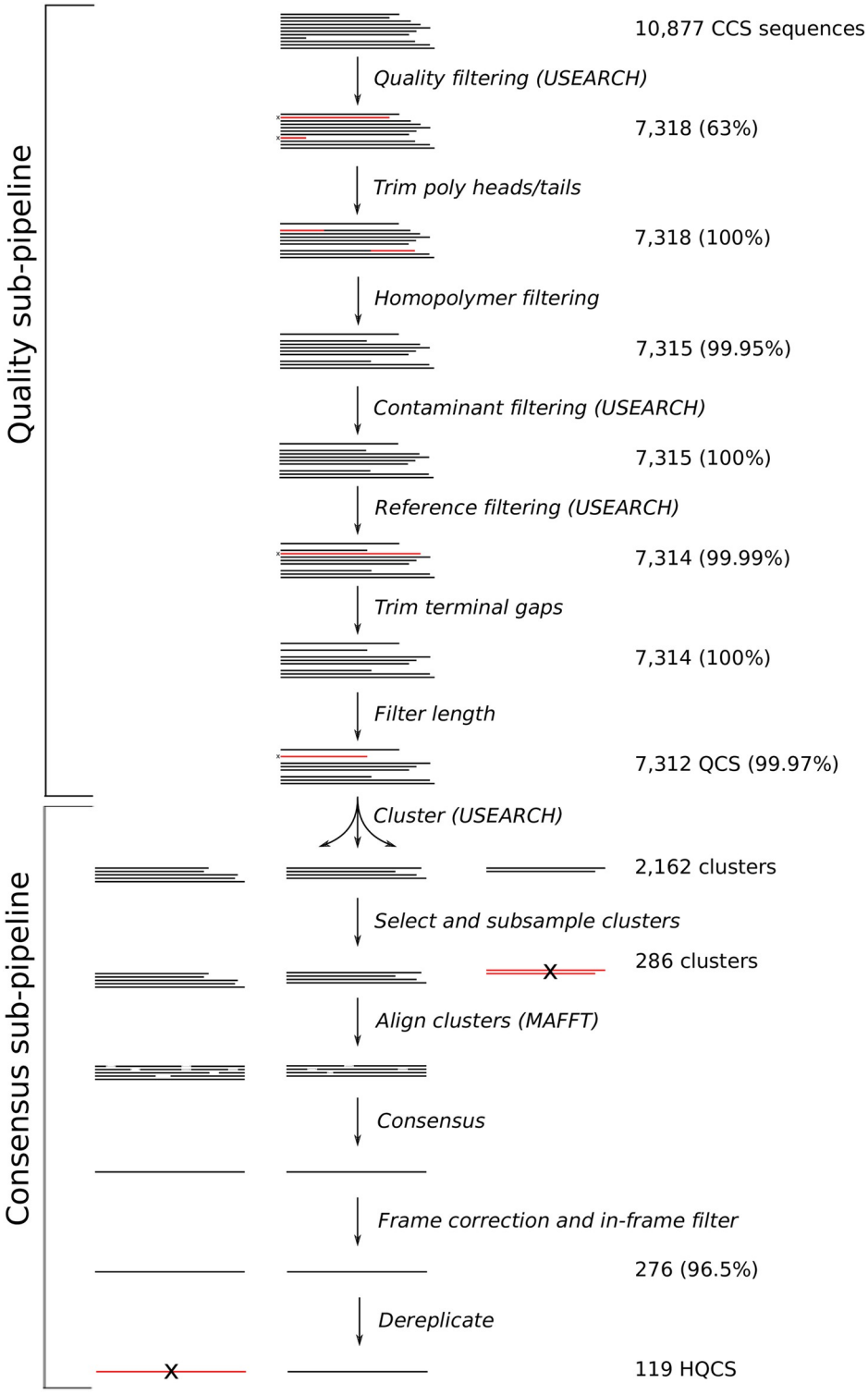

Fig 2. Quality and consensus sub-pipelines. These steps are repeated independently on each time point. Numbers are reported from the analysis of sequences from the first time point (V03) of donor P018, which is three months post infection. Percentages give the fraction of sequences retained after filtering. Tasks indicate whether they use third-party tools USEARCH or MAFFT.

https://doi.org/10.1371/journal.pcbi.1006498.g002

Sequel platforms. Other reads with different properties (error rates, error models, lengths, homopolymer distributions, etc.) likely require different parameters. All steps are run independently per time point.

1. **Filter by error rate**. The input FASTQ files contain Phred scores for each base, encoding the probabilities of incorrect base calls. USEARCH [41] is used to remove reads with an expected error rate greater than 1%, computed as the mean of the per-base error probabilities.

2. **Trim heads/tails**. A fraction of reads from the Laird Smith *et al.* sequencing protocol contain poly-A or poly-T heads or tails (cause unknown), which can be hundreds of bases long and sometimes contain a small number of other bases.
These heads and tails are trimmed with a hidden Markov model (Fig 3) implemented in Pomegranate [42]. The emission probabilities of the model were fixed, and the transitions trained using Baum-Welch. The Viterbi path for each sequence is computed, and bases emitted by head and tail nodes are removed.

3. **Filter long runs**. Reads with homonucleotide runs longer than 16 bases are discarded. This length was chosen to be twice the length of the longest such run in the LANL HIV database [43].

4. **Filter contaminants and trim reads**. Sample contamination can introduce non-native sequences that interfere with subsequent analyses, so these contaminants must identified and discarded. USEARCH is used to compare reads to a contaminant database and a reference database using `usearch_global`. Alignments returned from querying the database are then used to trim reads to the gene boundaries. Trimming terminal insertions is vital for the accuracy of downstream tasks, such as length filtering and clustering.
The contaminant database contains HXB2 and NL4-3 *env*, each ubiquitous in labs working with *env* sequences and a common source of sample contamination. Reads that match with $\geq$ 98% identity are discarded. Since a 1% error rate cutoff was earlier used, this parameter conservatively ensures that these contaminants are almost certainly identified.
The reference database contains thirty-eight sequences representing the major HIV Group M subtypes from the LANL sequence database [43]. Reads with $\leq$ 70% identity to every sequence in the reference database are discarded. This cutoff is chosen to retain reads

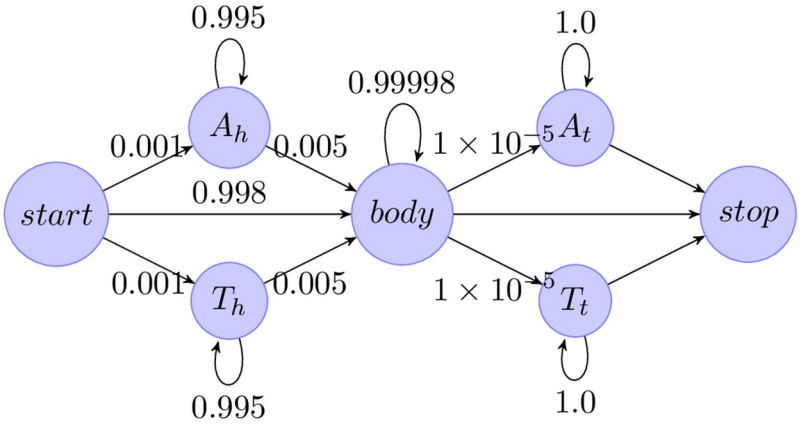

Fig 3. Hidden Markov model used for trimming poly-A and poly-T heads and tails. *A* head and tail states have a small ($p = 0.01$) probability to emit non-A bases, and similarly for *T*. The *body* state emits all four bases with equal probability. The *start*, and *stop* states emit nothing.

https://doi.org/10.1371/journal.pcbi.1006498.g003

remotely similar to HIV Group M while excluding contaminants such as human or bacterial genome reads. If a sample is from SIV, or from a non group-M HIV+ donor, then more appropriate reference sequences should be added to the database.

5. **Filter by length**. By default, sequences shorter than 90% or longer than 110% of the length of the reference sequence are discarded. However, sequences with large deletions are frequently observed in HIV. These likely represent replication incompetent envelopes, and their reduced length can cause them to be dramatically oversampled due to PCR length bias. Users who want to include these species in their analyses should modify these parameters.

Reads that pass this quality assurance phase have low expected error rates and no homonucleotide runs, are within 70% identity of at least one reference sequence, are (after trimming) no more than 10% different in length than a reference sequence, and do not match the contaminant database. We refer to these sequences as quality-controlled sequences (QCS).

Consensus sub-pipeline for variant identification. Even for highly diverse populations, unique reads in a sequencing run outnumber the true unique variants, predominantly due to sequencing errors. The problem is far more significant in long reads than in short reads, precluding the use of amplicon denoising strategies used to reduce error rates in short read sequencing [44]. To accommodate this effect, the next phase of the `FLEA` pipeline clusters and combines QCS reads, attempting to infer the true variants in each time point. It also attempts to detect and correct frameshift errors.

All of the following tasks are run separately for each time point, yielding sets of unique in-frame consensus sequences. We refer to these sequences as high-quality consensus sequences (HQCS).

1. **Cluster**. `USEARCH` is used with the `cluster_fast` command to generate clusters with 99% nucleotide identity. This parameter approximates the 1% error cutoff used in the error rate filtering step, so that pairwise distances of sequences in the same cluster are consistent with the sequencing error. `cluster_fast` runs in a single pass, so it is sensitive to input order. Sequences are sorted from lowest to highest quality according to expected error rate; experiment suggests that this order yields better results (Tables A, B, and C in S1 Text).

2. **Select and subsample clusters**. Clusters with fewer than three members are discarded, because they are too small to de-noise by majority consensus. Clusters with more than 50 members are subsampled to the top 50 with the lowest expected error rate to speed up the multiple sequence alignment step.

3. **Align and consensus**. `MAFFT` [45] is used to align each cluster. The consensus sequence of each alignment is computed.

4. **Frame correction** In-frame consensus sequences from all time points are collected into a `USEARCH` database for frame correction. `usearch_global` is then used to align each out-of-frame sequence to its top hit. The nucleotide alignment is used to correct incomplete codons: short insertions (1 or 2 base pairs) are discarded, and single deletions are replaced with the aligned base. Sequences with longer insertions or deletions are discarded. All changes are logged, so that the user can identify the sequences that have been corrected.

5. **Uniqueness** Non-unique consensus sequences are dereplicated using `usearch --fastx_uniques`.

6. **Copy numbers** The number of sequences per cluster provides an estimate of the relative abundance of that HQCS in the population. Those numbers are further augmented by adding

sequences orphaned by cluster filtering and HQCS dereplication. `usearch_global` is used to assign each QCS to its nearest HQCS. The number of sequences accrued by each HQCS is interpreted as its copy number.

Alignment sub-pipeline. The HQCSs from all time points are combined into a single file, translated to protein sequences, and aligned using `MAFFT`. A Python script then transfers the gaps from each aligned protein sequence to the corresponding nucleotide sequences to produce a codon-level nucleotide multiple sequence alignment of all unique variants from all time points.

Analysis sub-pipeline. The analyses used in `FLEA` take as input the two outputs of the alignment phase: a codon multiple sequence alignment of all unique HQCS sequences from all time points, and their associated copy numbers. These data are used for the following analyses.

1. **Time point metrics**. `HyPhy` [46] scripts are used to compute evolutionary metrics (total, *dN*, and *dS* divergence and diversity) and phenotypic metrics (protein length, potential N-linked glycosylation sites, isoelectric point) for each annotated region (e.g., V1, MPER) in the amplicon for each time point.
2. **Early consensus**. The early consensus is inferred by taking the copy-number-weighted codon consensus of the codon-aligned HQCSs from the earliest time point. By including gaps, this consensus sequence is already aligned with the rest of the multiple sequence alignment. This strategy is acceptable for primary infection studies from single founders with very low early diversity, in which case the consensus sequence should closely match the most recent common ancestor.
3. **Reference coordinates**. `MAFFT` is used to assign HXB2 [47] coordinates to the gapped early consensus sequence, which are then transferred to the full multiple sequence alignment.
4. **Infer phylogeny**. A maximum-likelihood phylogenetic tree is inferred with FastTree2 [48, 49] under the general time reversible model. The tree is rooted on the early consensus sequence.
5. **Ancestral sequence reconstruction**. `HyPhy` is used to infer ancestral sequences at the internal nodes of the phylogeny, using joint maximum likelihood reconstruction and the generalised time-reversible model (GTR) [50].
6. **Multidimensional scaling**. A distance matrix is computed for all HCQC sequences using the Tamura Nei 93 distance [51]. Metric multidimensional scaling [52] (implemented in `scikit-learn` [53]) is used to find a two-dimensional embedding of the sequences that approximates their pairwise distances. This embedding with this evolutionary distance is meant to show relationships that cannot be represented in a phylogenetic tree, because of recombination.
7. **FUBAR**. Site-specific selection rates are inferred using `FUBAR` [54], implemented in `HyPhy`.
8. **Position-specific changes**. Entropy and Jensen-Shannon divergence are computed for each position in each time point.

The results of these analyses are provided to the user in an interactive web application, described next.

Web application

The FLEA web app is built using modern web design principles. It consists of two parts: a Javascript client-side app, written using the Ember.js [55] framework, and a server-side REST (REpresentational State Transfer) service for serving JSON-formatted data. There are two main benefits to using this decoupled pattern for scientific web applications. First, the client-side code only needs to be downloaded once, at the start of the session. The data are requested from the server and cached as needed. Once everything is loaded, the visualizations run entirely in the browser with no delays for page loads. Second, the REST service may be reused by other apps and third-party tools.

The web app presents the results of the FLEA analysis as a series of interactive visualizations. The report is organized into the following sections.

Multidimensional scaling. A two dimensional embedding of the HQCSs is visualized as a bubble plot, showing changes in population structure over time, as shown in Fig 4. This visualization has been especially useful for investigating populations with superinfection, or with multiple founders, where aggressive recombination between vastly different *env* variants precludes the use of phylogenies.

Evolutionary trajectory. The evolutionary trajectory viewer plots evolutionary and phenotypic metrics for each time point and multiple regions in the amplicon, giving a high-level

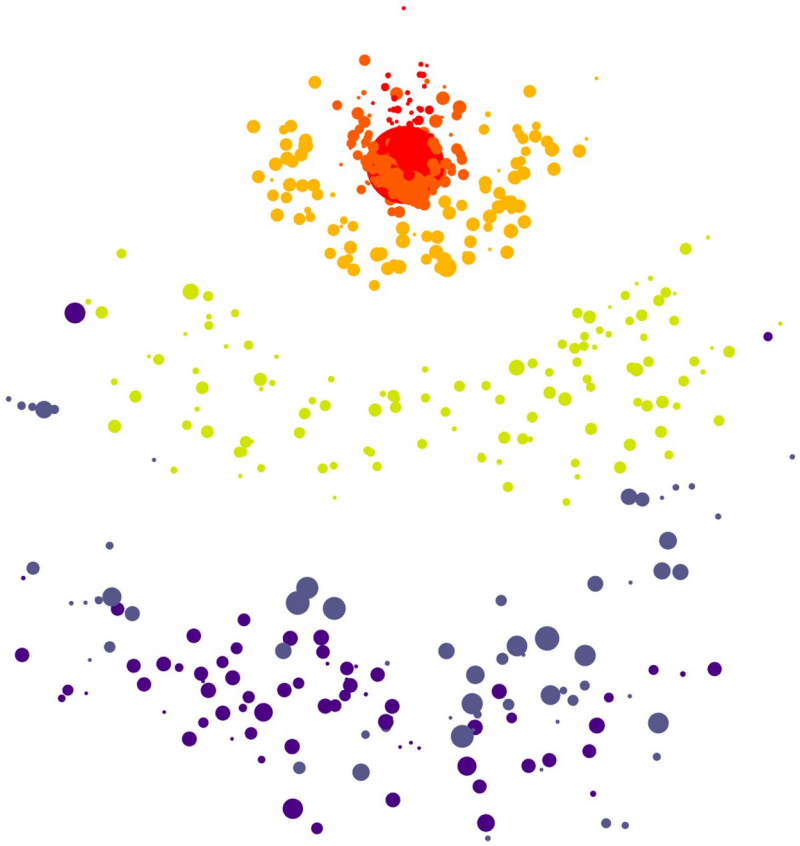

Fig 4. Screenshot of the multidimensional scaling plot. The embedding in two dimensions preserves pairwise evolutionary distances between HQCSs. Node area is proportional to copy number, and color corresponds to time point. The increasing genetic diversity of the population is visible as time goes on.

https://doi.org/10.1371/journal.pcbi.1006498.g004

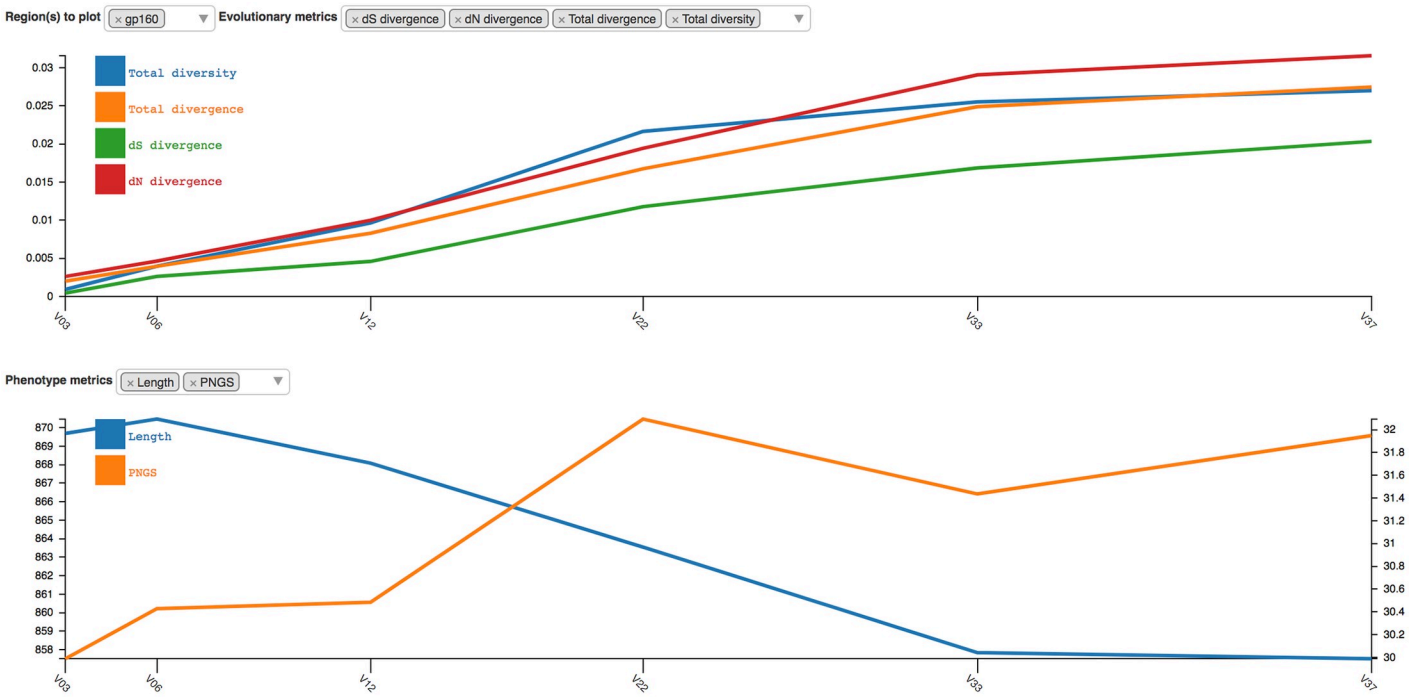

Fig 5. Screenshot of the evolutionary trajectory report. Four evolutionary metrics (*dS* divergence, *dN* divergence, total divergence, and total diversity) and two phenotype metrics (length and possible N-linked glycosylation sites) are shown for gp160.

https://doi.org/10.1371/journal.pcbi.1006498.g005

overview of population dynamics over time. Fig 5 shows the plot for the entire gp160 region of HIV Env, which is generated with the `D3.js` plotting library [56].

Sequences. The multiple sequence alignment of all the HQCS sequences is the foundation for all subsequent analyses. It is displayed in the amino acid sequences viewer, which contains a custom alignment browser and an interactive motif dynamics plot, as shown in Fig 6.

Protein structure. The protein structure viewer maps evolutionary metrics to an interactive three-dimensional structure of the protein, customized from PDB ID `5FUU`, a recently resolved cryo-EM structure [57], and rendered using `pv` [58]. Missing residues are rendered as spheres which are positioned by Bézier curve interpolation. *dN*/*dS* ratios, Jensen Shannon divergence, and entropy may all be mapped to the protein structure, as shown in Fig 7. The same metrics are also plotted in one dimension for each time point, as shown in Fig 8. The protein visualization interacts with the sequence viewer by showing alignment positions and highlighting the residues in the selected sequence motif.

Trees. The tree viewer renders a tree browser with `phylotree.js` [59], as shown in Fig 9. Leaf nodes are scaled to the copy number of their sequence. The tree zoom level, layout, and coloring is interactively modifiable. Motifs selected in the sequence viewer are mapped to the tree. Ancestral nodes are colored by motif, allowing inferred changes to be tracked through the entire phylogeny.

Results

The entire pipeline was run on HIV *env* reads from donor P018, which are available from the NCBI Sequence Read Archive under BioProject PRJNA320111, and were sequenced as part of [36] on the RS-II instrument, using the older generation P5/C3 PacBio sequencing chemistry. The full dataset contains 58,468 CCS reads. The reads are split across six time points, which

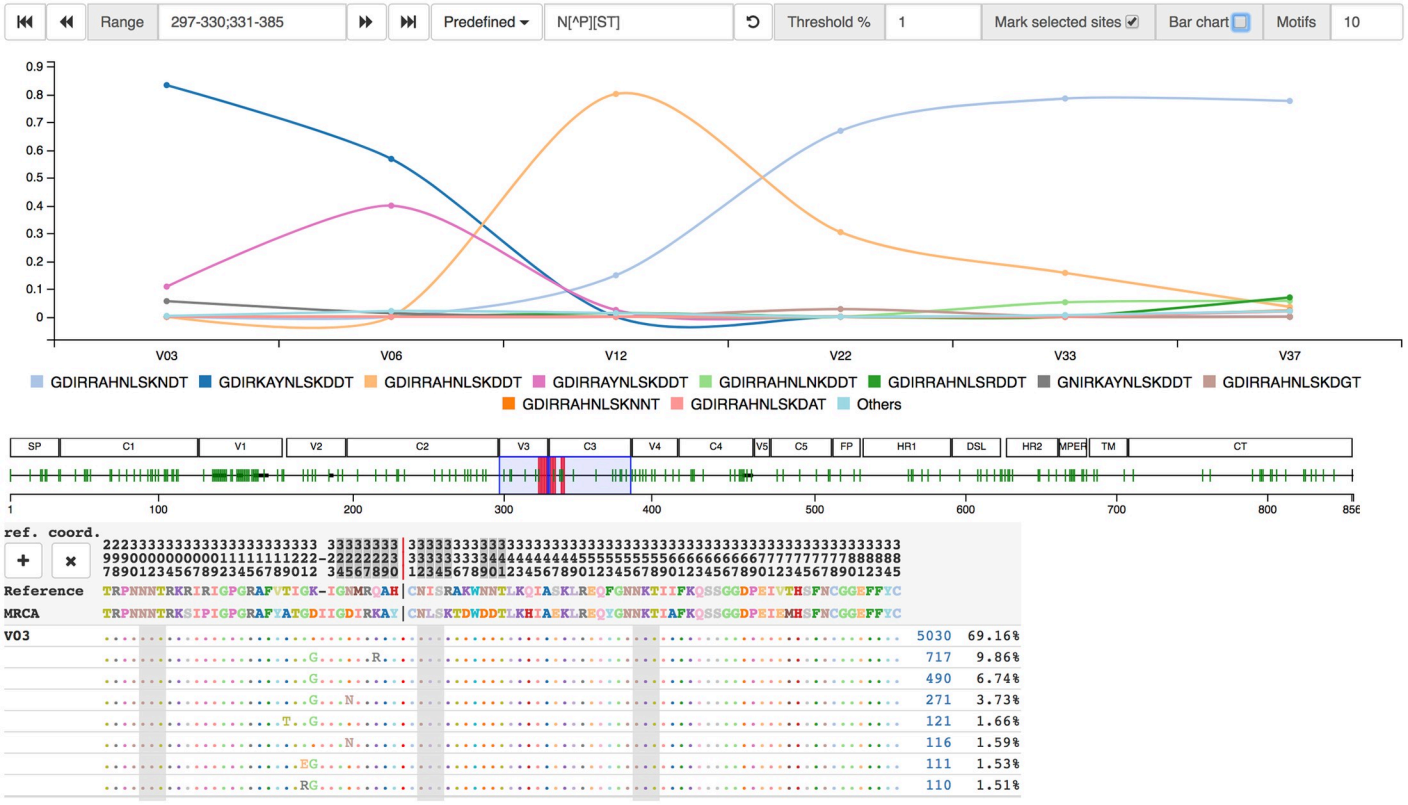

Fig 6. Screenshot of amino acid sequences viewer. Sequences are grouped by identity, with aggregate copy number and population percentage shown to the right. An overview of the amplicon, optionally annotated with region names, provides fast access to different locations of the alignment. Selecting columns of the alignment interactively updates the amino acid dynamics plot, showing the dynamics of the selected motif over time. In this case, the trajectory shows changes in the N332 glycan supersite. Sites inferred by FUBAR to be undergoing positive selection are selectable.

https://doi.org/10.1371/journal.pcbi.1006498.g006

are coded as V03, V06, V12, V22, V33, and V37, where *Vx* corresponds to a visit *x* months post infection. The number of reads per time point ranges from 7,530 in V33 to 11,806 in V06.

Results on simulated data

The true sequences and copy numbers are not known for the P018 data. In order to assess the accuracy of our inferred sequence population, we used the HQCSs from a previous FLEA run to simulate a gold standard dataset on which to assess the FLEA pipeline.

The simulation procedure starts with the HQCSs and copy numbers from the FLEA results on P018, then augments them with additional mutated sequences to create a gold standard set of templates. Mutated sequences were added because our clustering strategy may artificially merge similar templates. For each template, noisy reads with a SMRT-style error profile were sampled. Full details of the simulation process appear in the supporting information. These simulated reads were sent through the FLEA pipeline, both with and without frame correction.

The resulting QCS and HQCS sequences were compared to the ground truth using Earth Mover's Distance (EMD), using normalized copy numbers for the population weights and edit distance for the distance matrix. The fully constrained EMD has units that can be directly interpreted as the average change per nucleotide necessary to transform one sequence population into another. We also calculate two variants of EMD for further insight into how well the

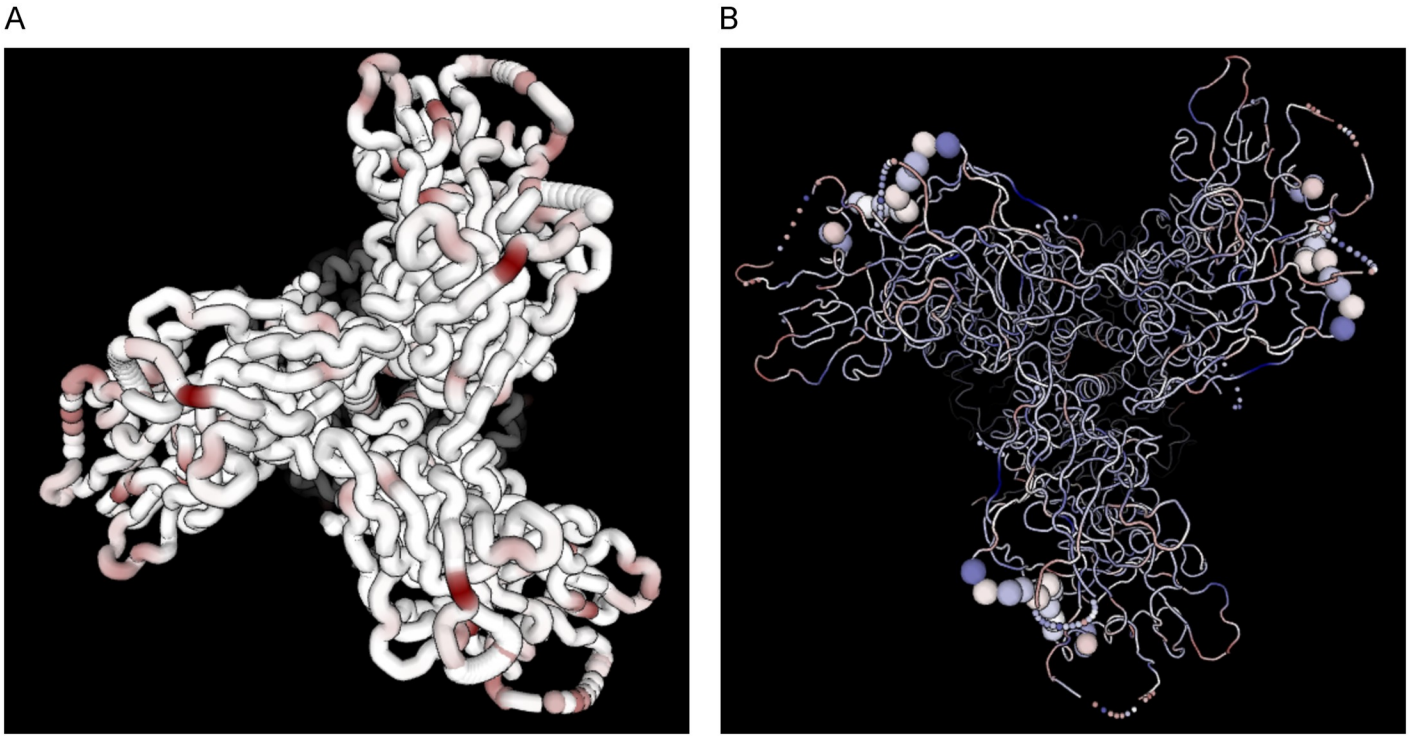

Fig 7. Screenshots of the interactive three-dimensional Env structure, colored according to JS divergence (left) and *dN*/*dS* values (right). Positions imputed to be undergoing more positive selection ($dN/dS > 1$) are darker red, and positions undergoing more purifying selection ($dN/dS < 1$) are darker blue. The right structure also shows motif positions highlighted in the sequence viewer.

https://doi.org/10.1371/journal.pcbi.1006498.g007

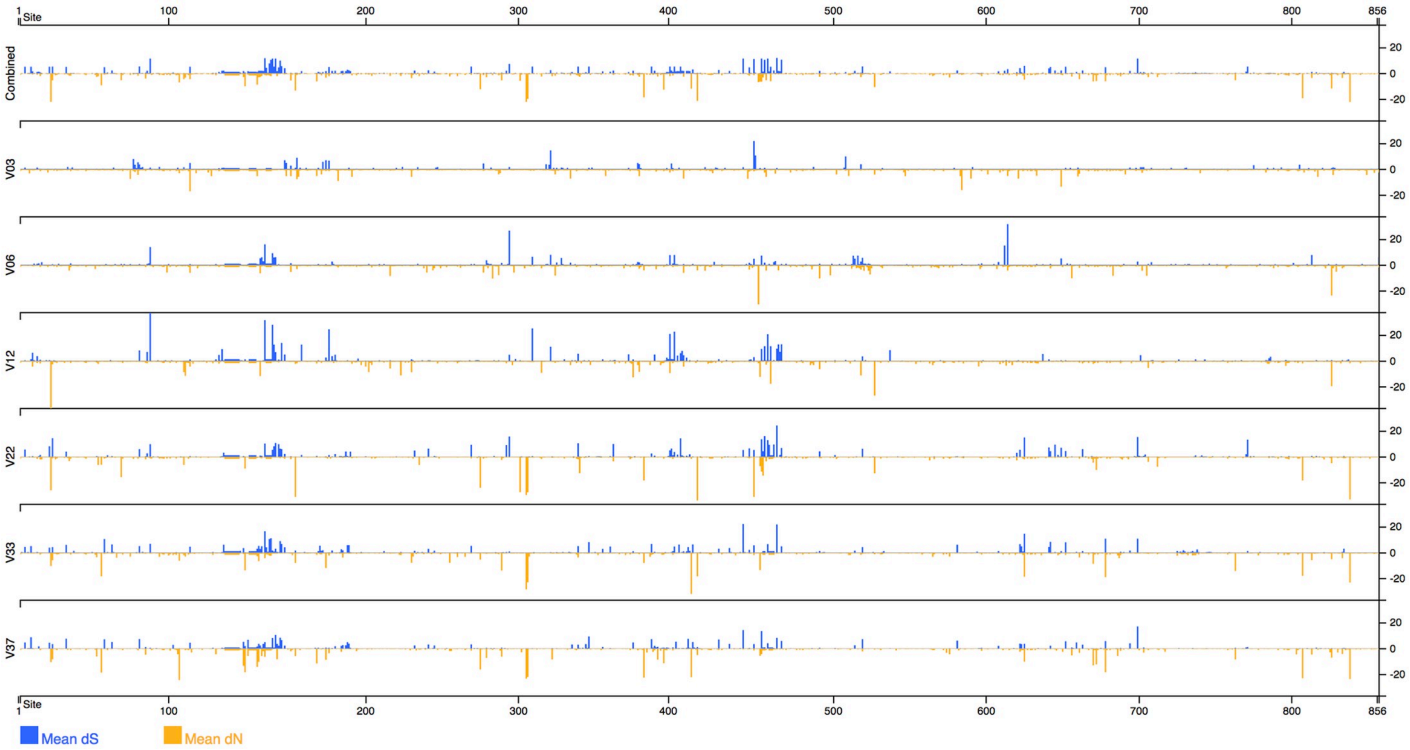

Fig 8. Screenshot of *dN*/*dS* values mapped to protein positions and separated by time point.

https://doi.org/10.1371/journal.pcbi.1006498.g008

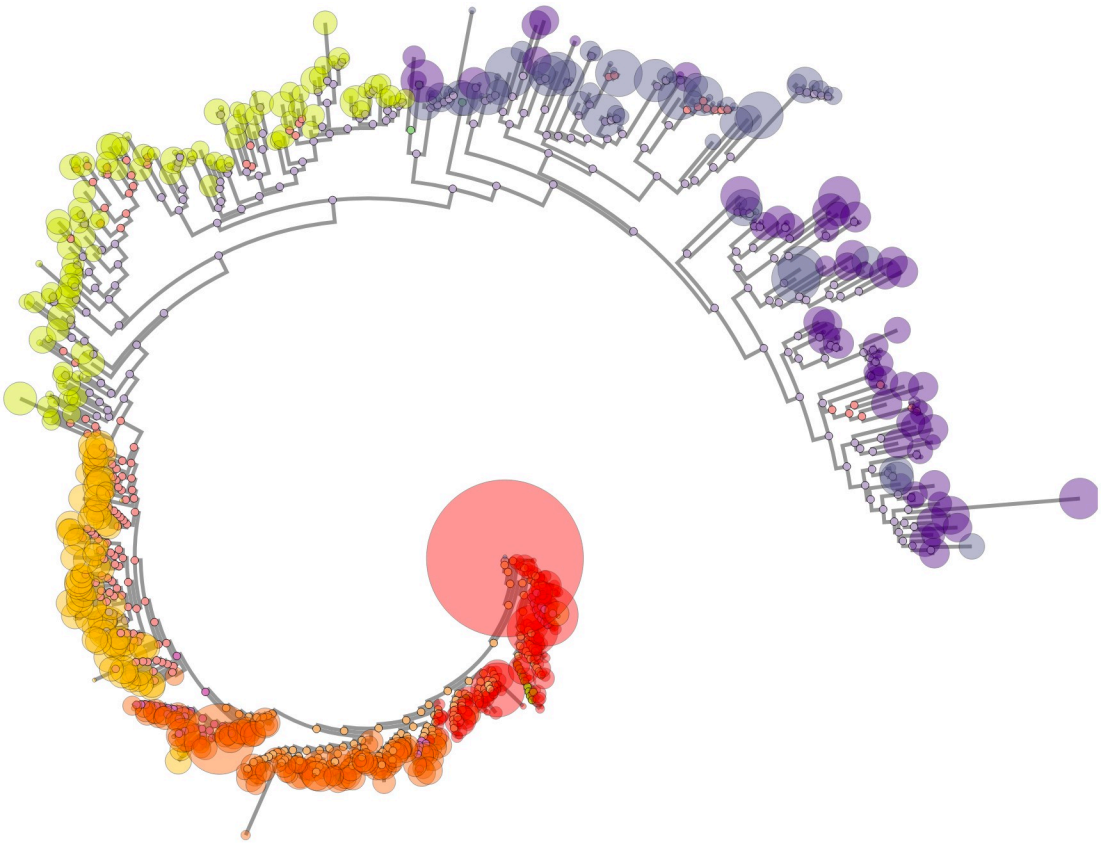

Fig 9. Screenshot of the phylogenetic tree viewer. Leaf node size corresponds to sequence copy number. Node color corresponds to time point. Since ancestral sequences have been inferred, ancestral nodes are colored according to the selected motif, which in this case is the N332 glycan supersite.

https://doi.org/10.1371/journal.pcbi.1006498.g009

inferred population *B* estimates the sequences in the ground truth population *A*. EMD_{FP} removes the constraint on *A*, allowing any amount of flow from *A* to *B*. It is a measure of false positives because it grows when *B* contains extra sequences distant from any in *A*. Similarly, EMD_{FN} removes the constraint on *B*. It grows when *B* fails to recapitulate sequences in *A*, and therefore is a measure of false negatives.

To see the effect of sequencing runs of different depths, the experiment was repeated for 300, 1,000, 3,000, and 10,000 reads per time point. The results, which appear in Table 1, show the benefit of FLEA's approach of reducing sequence errors via clustering and consensus. The QCS sequences, although they have few false negatives ($EMD_{FN} = 0.0782$) for $n = 10,000$, are dominated by false positives ($EMD_{FP} = 8.3$). However, adding the consensus sub-pipeline virtually eliminates false positives ($EMD_{FP} = 0.0336$), at the cost of only a 2.4x increase in false negatives, for a 8.6x improvement in EMD to 1.0549. The frame correction step further improve both EMD_{FP} and EMD_{FN} because it turns false positives into true positives.

The full-length *env* sequencing protocol yields approximately 10,000 reads per run; the P018 data averaged 9,744 reads per time point. Therefore, these results with $n = 10,000$ suggest that FLEA is capable of taking a full sequencing run of CCS reads from a diverse viral population with an average of 9.56 errors per sequence and inferring HQCSs with an average of 1.01 errors per sequence, which corresponds to an average error rate of 0.038%. Moreover, these error rates are mostly caused by low-abundance sequences in both the true population and the

Table 1. EMD metrics for various numbers of reads, averaged across all time points. "mean errors" gives the average number of errors in the reads, estimated from the simulated Phred scores.

n	mean errors	consensus type	*EMD*	EMD_{FP}	EMD_{FN}
300	9.63	QCS	12.3769	8.3418	2.8956
		HQCS	7.1570	0.4050	5.4271
		HQCS (corrected)	6.4752	0.3020	4.5533
1000	9.63	QCS	10.5433	8.3686	1.2551
		HQCS	2.8279	0.0610	1.1453
		HQCS (corrected)	2.7557	0.0666	1.0405
3000	9.6	QCS	9.5053	8.2837	0.3908
		HQCS	1.6432	0.0146	0.4322
		HQCS (corrected)	1.5168	0.0045	0.2925
10000	9.56	QCS	9.0734	8.3080	0.0782
		HQCS	1.0549	0.0336	0.1735
		HQCS (corrected)	1.0146	0.0073	0.1463

https://doi.org/10.1371/journal.pcbi.1006498.t001

inferred FLEA sequences. Fig 10 shows that FLEA perfectly recovers all sequences from all time points that account for at least 1.6% of the population. An in-depth breakdown of the false negatives appears in Table D and Fig. M in S1 Text.

Results on real data: Donor P018

FLEA was run directly on the P018 sequences, and the results are summarized here. The full results of this run are available to view at http://flea.murrell.group/view/P018.

Fig 2 shows the number of sequences from the V03 time point that make it to each stage of the quality and consensus pipelines. At three months post infection, the majority amino-acid sequence variant is shared by 52.1% of the population, and the next most common variants accounts for just 8.66%. This relative lack of diversity is consistent with early infection

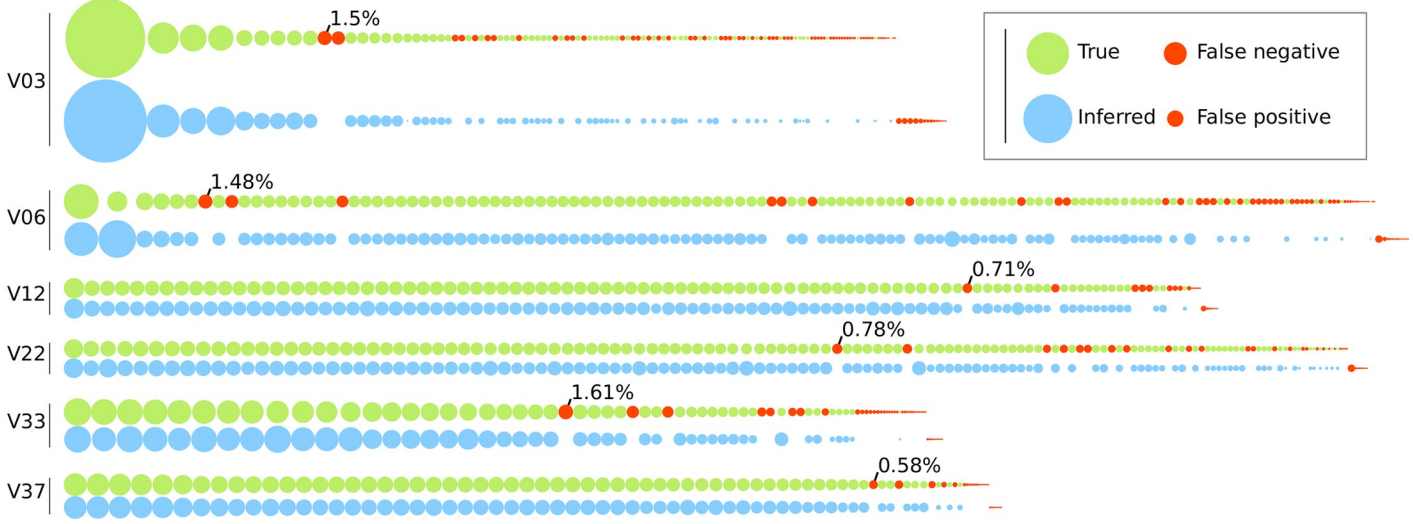

Fig 10. Comparison of true sequence abundances versus copy numbers inferred by FLEA for each time point of the simulated P018 data. Each node represents one sequence, with the area denoting its relative abundance in the population. The true population (top) is colored green. For each true sequence, the matching HQCS sequences appears below it in blue. Red nodes denote false negatives and positives. The most common false negative for each time point is annotated with its abundance.

https://doi.org/10.1371/journal.pcbi.1006498.g010

dynamics. By 37 months post infection there is much more diversity: the most common variant accounts for only 3.96% of the population.

Donor P018 shows signs of potential N332 glycan specificity, as shown by the motif trajectories in Fig 6. The glycan supersite, centered around N332 in V3, is a common target for broadly-neutralizing antibodies [60] because these sites are often conserved, so mutations in these regions are associated with escape [61]. A year into sampling (V12), mutations 328R and 330H dominate, and the majority of sequences also contain 339N from 22 months (V22) onwards.

The error rates of PacBio CCS sequences are usefully predicted by the QV scores provided by the instrument [36]. We show (Fig. G through Fig. L in S1 Text that the effective number of bases that are corrected in each CCS read (as measured by the difference between that CCS and the HQCS to which it contributes) was extremely well predicted (Spearman's *rho* from 0.69 to 0.76) by the QV scores. This result is especially encouraging given that our pipeline does not currently exploit these QV scores beyond the initial filtering step. Further, PacBio sequences have higher indel than substitution rates, and this was recapitulated in the number of corrected indels vs substitutions, although this ratio appeared to vary from one time point to the next.

Availability and future directions

A public instance of FLEA is hosted at http://flea.datamonkey.org. The Python source code for the FLEA pipeline can be found at https://github.com/veg/flea-pipeline. The client-side application is available at https://github.com/veg/flea-web-app. A live demo of the P018 results can be found at http://flea.murrell.group/view/P018, with an explanatory page at: http://murrell.group/FLEAexplained/.

The FLEA pipeline analyzes longitudinal full-length *env* sequences and provides visualizations of the results. Using simulations, we show that FLEA is capable of inferring accurate HIV *env* consensus sequences and population frequencies. Despite each CCS read containing an average of ten errors, our approach distinguishes variants that differ by as little as one base from an amplicon with high indel variation. It uses those high-quality consensus sequences to generate a codon-aware multiple sequence alignment of all time points, estimate ancestral sequences, infer the phylogenetic tree, and perform many other population-level analyses with high accuracy. These results are presented in a visualization suite that is highly general and applicable to many related sequencing problem.

While we provide a web application that should suffice for sequencing most standard Env samples from HIV-1 group M, we recommend that those who frequently engage in such sequencing, or who wish to sequence less straightforward samples (eg. SIV or SHIV), install FLEA locally. This provides a range of customization and tuning options, such as the filtering parameters and the set of reference sequences.

While our USEARCH-based clustering and consensus strategy for denoising long PacBio amplicons performs well when error rates are $< 1\%$, there is a clear need for more sophisticated long-read de-noising algorithms that exploit the additional depth of lower quality reads that we currently discard. This will be especially beneficial for longer PacBio amplicons, because the CCS read quality distribution degrades with length. For example, while we can currently obtain around 15,000 CCS reads $< 1\%$ from a P6/C4 RS-II run of our 2.6kb *env* amplicon; this read count drops to $\sim 1,000$ for full-length 9kb HIV genomes. Additionally, FLEA does sometimes erroneously collapse sequences from very similar templates, and more sophisticated approaches to amplicon denoising could likely improve upon this.

Both the pipeline and client-side visualizations are under development, with many improvements planned, including a novel clustering algorithm that reduces false positives and a novel consensus algorithm that uses quality scores and performs frame correction. We plan to integrate epitope prediction into the FLEA pipeline and add appropriate visualizations for the case when users have IC_{50} values available for their sequences. Finally, FLEA will be expanded to support other amplicons.

Supporting information

S1 Text.
(PDF)

Author Contributions

Conceptualization: Kemal Eren, Sergei L. Kosakovsky Pond, Ben Murrell.

Data curation: Kemal Eren, Melissa Laird Smith, Ben Murrell.

Funding acquisition: Sergei L. Kosakovsky Pond, Ben Murrell.

Investigation: Kemal Eren, Ben Murrell.

Methodology: Kemal Eren, Sergei L. Kosakovsky Pond, Ben Murrell.

Project administration: Ben Murrell.

Software: Kemal Eren, Steven Weaver, Robert Ketteringham, Morné Valentyn, Venkatesh Kumar, Sanjay Mohan, Sergei L. Kosakovsky Pond.

Supervision: Sergei L. Kosakovsky Pond, Ben Murrell.

Validation: Kemal Eren.

Visualization: Kemal Eren.

Writing – original draft: Kemal Eren, Ben Murrell.

Writing – review & editing: Kemal Eren, Steven Weaver, Melissa Laird Smith, Venkatesh Kumar, Sanjay Mohan, Sergei L. Kosakovsky Pond, Ben Murrell.

References

1. DeLeon O, Hodis H, O'Malley Y, Johnson J, Salimi H, Zhai Y, et al. Accurate predictions of population-level changes in sequence and structural properties of HIV-1 Env using a volatility-controlled diffusion model. PLOS Biology. 2017 04; 15(4):1–38. https://doi.org/10.1371/journal.pbio.2001549
2. Fischer W, Ganusov VV, Giorgi EE, Hraber PT, Keele BF, Leitner T, et al. Transmission of single HIV-1 genomes and dynamics of early immune escape revealed by ultra-deep sequencing. PLOS ONE. 2010 08; 5(8):1–15. https://doi.org/10.1371/journal.pone.0012303
3. Henn MR, Boutwell CL, Charlebois P, Lennon NJ, Power KA, Macalalad AR, et al. Whole genome deep sequencing of HIV-1 reveals the impact of early minor variants upon immune recognition during acute infection. PLOS Pathogens. 2012 03; 8(3):1–14. https://doi.org/10.1371/journal.ppat.1002529
4. Leung P, Bull R, Lloyd A, Luciani F. A bioinformatics pipeline for the analyses of viral escape dynamics and host immune responses during an infection. BioMed Research International. 2014; 2014. https://doi.org/10.1155/2014/680249
5. McCloskey RM, Liang RH, Harrigan PR, Brumme ZL, Poon AFY. An evaluation of phylogenetic methods for reconstructing transmitted HIV variants using longitudinal clonal HIV sequence data. Journal of Virology. 2014 June; 88(11):6181–6194. https://doi.org/10.1128/JVI.00483-14 PMID: 24648453
6. Pandit A, de Boer RJ. Reliable reconstruction of HIV-1 whole genome haplotypes reveals clonal interference and genetic hitchhiking among immune escape variants. Retrovirology. 2014 Jul; 11(1):56. https://doi.org/10.1186/1742-4690-11-56 PMID: 24996694

7. Tsibris AMN, Korber B, Arnaout R, Russ C, Lo CC, Leitner T, et al. Quantitative deep sequencing reveals dynamic HIV-1 escape and large population shifts during CCR5 antagonist therapy in vivo. PLOS ONE. 2009 05; 4(5):1–12. https://doi.org/10.1371/journal.pone.0005683
8. Huang DW, Raley C, Jiang MK, Zheng X, Liang D, Rehman MT, et al. Towards better precision medicine: PacBio Single-molecule long reads resolve the interpretation of HIV drug resistant mutation profiles at explicit quasispecies (haplotype) level. Journal of data mining in genomics & proteomics. 2016 January; 7(1).
9. Huber M, Metzner KJ, Geissberger FD, Shah C, Leemann C, Klimkait T, et al. MinVar: A rapid and versatile tool for HIV-1 drug resistance genotyping by deep sequencing. Journal of Virological Methods. 2017; 240:7–13. https://doi.org/10.1016/j.jviromet.2016.11.008 PMID: 27867045
10. Mukherjee R, Jensen ST, Male F, Bittinger K, Hodinka RL, Miller MD, et al. Switching between Raltegravir resistance pathways analyzed by deep sequencing. AIDS. 2011 Oct; 25(16):1951–1959. https://doi.org/10.1097/QAD.0b013e32834b34de PMID: 21832937
11. Svarovskaia ES, Martin R, McHutchison JG, Miller MD, Mo H. Abundant drug-resistant NS3 mutants detected by deep sequencing in HCV-infected patients undergoing NS3 protease inhibitor monotherapy. Journal of Clinical Microbiology. 2012;. https://doi.org/10.1128/JCM.00838-12 PMID: 22837328
12. Gianella S, Delport W, Pacold ME, Young JA, Choi JY, Little SJ, et al. Detection of minority resistance during early HIV-1 infection: natural variation and spurious detection rather than transmission and evolution of multiple viral variants. Journal of Virology. 2011;. https://doi.org/10.1128/JVI.02582-10
13. Varble A, Albrecht RA, Backes S, Crumiller M, Bouvier NM, Sachs D, et al. Influenza A virus transmission bottlenecks are defined by infection route and recipient host. Cell Host & Microbe. 2014; 16(5):691–700. https://doi.org/10.1016/j.chom.2014.09.020
14. Bull RA, Luciani F, McElroy K, Gaudieri S, Pham ST, Chopra A, et al. Sequential bottlenecks drive viral evolution in early acute Hepatitis C virus infection. PLOS Pathogens. 2011 09; 7(9):1–14. https://doi.org/10.1371/journal.ppat.1002243
15. Wang GP, Sherrill-Mix SA, Chang KM, Quince C, Bushman FD. Hepatitis C virus transmission bottlenecks analyzed by deep sequencing. Journal of Virology. 2010; 84(12):6218–6228. https://doi.org/10.1128/JVI.02271-09 PMID: 20375170
16. Gianella S, Kosakovsky Pond SL, Oliveira MF, Scheffler K, Strain MC, De la Torre A, et al. Compartmentalized HIV rebound in the central nervous system after interruption of antiretroviral therapy. Virus Evolution. 2016; 2(2):vew020. https://doi.org/10.1093/ve/vew020 PMID: 27774305
17. Poon AFY, Swenson LC, Bunnik EM, Edo-Matas D, Schuitemaker H, van't Wout AB, et al. Reconstructing the dynamics of HIV evolution within hosts from serial deep sequence data. PLOS Computational Biology. 2012 11; 8(11):1–11. https://doi.org/10.1371/journal.pcbi.1002753
18. Kortenhoeven C, Joubert F, Bastos AD, Abolnik C. Virus genome dynamics under different propagation pressures: reconstruction of whole genome haplotypes of West Nile viruses from NGS data. BMC Genomics. 2015 Feb; 16(1):118. https://doi.org/10.1186/s12864-015-1340-8 PMID: 25766117
19. Mangul S, Wu NC, Mancuso N, Zelikovsky A, Sun R, Eskin E. Accurate viral population assembly from ultra-deep sequencing data. Bioinformatics. 2014; 30(12):i329–i337. https://doi.org/10.1093/bioinformatics/btu295 PMID: 24932001
20. Skums P, Mancuso N, Artyomenko A, Tork B, Mandoiu I, Khudyakov Y, et al. Reconstruction of viral population structure from next-generation sequencing data using multicommodity flows. BMC Bioinformatics. 2013 Jun; 14(9):S2. https://doi.org/10.1186/1471-2105-14-S9-S2 PMID: 23902469
21. Wu X, Zhou T, Zhu J, Zhang B, Georgiev I, Wang C, et al. Focused evolution of HIV-1 neutralizing antibodies revealed by structures and deep sequencing. Science. 2011;. https://doi.org/10.1126/science.1207532
22. Yin L, Liu L, Sun Y, Hou W, Lowe AC, Gardner BP, et al. High-resolution deep sequencing reveals biodiversity, population structure, and persistence of HIV-1 quasispecies within host ecosystems. Retrovirology. 2012 Dec; 9(1):108. https://doi.org/10.1186/1742-4690-9-108 PMID: 23244298
23. Sede MM, Moretti FA, Laufer NL, Jones LR, Quarleri JF. HIV-1 tropism dynamics and phylogenetic analysis from longitudinal ultra-deep sequencing data of CCR5- and CXCR4-using variants. PLOS ONE. 2014 07; 9(7):1–14. https://doi.org/10.1371/journal.pone.0102857
24. Pacold ME, Pond SLK, Wagner GA, Delport W, Bourque DL, Richman DD, et al. Clinical, virologic, and immunologic correlates of HIV-1 intraclade B dual infection among men who have sex with men. AIDS (London, England). 2012 January; 26(2):157–165. https://doi.org/10.1097/QAD.0b013e32834dcd26
25. Capobianchi MR, Giombini E, Rozera G. Next-generation sequencing technology in clinical virology. Clinical Microbiology and Infection. 2013; 19(1):15–22. https://doi.org/10.1111/1469-0691.12056 PMID: 23279287

26. Quiñones-Mateu ME, Avila S, Reyes-Teran G, Martinez MA. Deep sequencing: Becoming a critical tool in clinical virology. Journal of Clinical Virology. 2014; 61(1):9–19. https://doi.org/10.1016/j.jcv.2014.06.013 PMID: 24998424

27. Leung P, Eltahla AA, Lloyd AR, Bull RA, Luciani F. Understanding the complex evolution of rapidly mutating viruses with deep sequencing: Beyond the analysis of viral diversity. Virus Research. 2017; 239:43–54. https://doi.org/10.1016/j.virusres.2016.10.014 PMID: 27888126

28. Vincent AT, Derome N, Boyle B, Culley AI, Charette SJ. Next-generation sequencing (NGS) in the microbiological world: How to make the most of your money. Journal of Microbiological Methods. 2017; 138:60–71. https://doi.org/10.1016/j.mimet.2016.02.016 PMID: 26995332

29. McElroy K, Thomas T, Luciani F. Deep sequencing of evolving pathogen populations: applications, errors, and bioinformatic solutions. Microbial Informatics and Experimentation. 2014 Jan; 4(1):1. https://doi.org/10.1186/2042-5783-4-1 PMID: 24428920

30. Beerenwinkel N, Zagord O. Ultra-deep sequencing for the analysis of viral populations. Current Opinion in Virology. 2011; 1(5):413–418. https://doi.org/10.1016/j.coviro.2011.07.008 PMID: 22440844

31. Rhoads A, Au KF. PacBio sequencing and its applications. Genomics, Proteomics & Bioinformatics. 2015; 13(5):278–289. https://doi.org/10.1016/j.gpb.2015.08.002

32. Gupta A, Adami C. Strong selection significantly increases epistatic interactions in the long-term evolution of a protein. PLOS Genetics. 2016 03; 12(3):1–23. https://doi.org/10.1371/journal.pgen.1005960

33. Parera M, Perez-Alvarez N, Clotet B, Martínez MA. Epistasis among deleterious mutations in the HIV-1 protease. Journal of Molecular Biology. 2009; 392(2):243–250. https://doi.org/10.1016/j.jmb.2009.07.015 PMID: 19607838

34. Weinreich DM. High-throughput identification of genetic interactions in HIV-1. Nature Genetics. 2011 May; 43(5):398–400. https://doi.org/10.1038/ng.820 PMID: 21522176

35. Sarzotti-Kelsoe M, Bailer RT, Turk E, li Lin C, Bilska M, Greene KM, et al. Optimization and validation of the TZM-bl assay for standardized assessments of neutralizing antibodies against HIV-1. Journal of Immunological Methods. 2014; 409:131–146. https://doi.org/10.1016/j.jim.2013.11.022 PMID: 24291345

36. Laird Smith M, Murrell B, Eren K, Ignacio C, Landais E, Weaver S, et al. Rapid sequencing of complete env genes from primary HIV-1 samples. Virus Evolution. 2016; 2(2):vew018. https://doi.org/10.1093/ve/vew018 PMID: 29492273

37. Caskey M, Schoofs T, Gruell H, Settler A, Karagounis T, Kreider E, et al. Antibody 10-1074 suppresses viremia in HIV-1-infected individuals. Nature Medicine. 2017 1;. https://doi.org/10.1038/nm.4268 PMID: 28092665

38. Landais E, Murrell B, Briney B, Murrell S, Rantalainen K, Berndsen ZT, et al. HIV envelope glycoform heterogeneity and localized diversity govern the initiation and maturation of a V2 apex broadly neutralizing antibody lineage. Immunity. 2017; 47(5):990–1003.e9. https://doi.org/10.1016/j.immuni.2017.11.002 PMID: 29166592

39. Liang M, Raley C, Zheng X, Kutty G, Gogineni E, Sherman BT, et al. Distinguishing highly similar gene isoforms with a clustering-based bioinformatics analysis of PacBio single-molecule long reads. BioData Mining. 2016 Apr; 9(1):13. https://doi.org/10.1186/s13040-016-0090-8 PMID: 27051465

40. Di Tommaso P, Chatzou M, Floden EW, Barja PP, Palumbo E, Notredame C. Nextflow enables reproducible computational workflows. Nature Biotechnology. 2017 Apr; 35:316. https://doi.org/10.1038/nbt.3820 PMID: 28398311

41. Edgar RC. Search and clustering orders of magnitude faster than BLAST. Bioinformatics. 2010; 26(19):2460–2461. https://doi.org/10.1093/bioinformatics/btq461 PMID: 20709691

42. Jacob Schreiber. Pomegranate;. Software download. Available from: https://github.com/jmschrei/pomegranate.

43. Foley BT, Leitner TK, Apetrei C, Hahn B, Mizrachi I, Mullins J, et al. HIV Sequence Compendium 2017. Los Alamos National Lab. (LANL), Los Alamos, NM (United States); 2017.

44. Edgar RC, Flyvbjerg H. Error filtering, pair assembly and error correction for next-generation sequencing reads. Bioinformatics. 2015; 31(21):3476–3482. https://doi.org/10.1093/bioinformatics/btv401 PMID: 26139637

45. Katoh K, Misawa K, Kuma K, Miyata T. MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. Nucleic Acids Research. 2002; 30(14):3059–3066. https://doi.org/10.1093/nar/gkf436 PMID: 12136088

46. Pond SLK, Frost SDW, Muse SV. HyPhy: hypothesis testing using phylogenies. Bioinformatics. 2005; 21(5):676–679. https://doi.org/10.1093/bioinformatics/bti079 PMID: 15509596

47. Ratner L, Haseltine W, Patarca R, Livak KJ, Starcich B, Josephs SF, et al. Complete nucleotide sequence of the AIDS virus, HTLV-III. Nature. 1985 Jan; 313:277. https://doi.org/10.1038/313277a0 PMID: 2578615

48. Price MN, Dehal PS, Arkin AP. FastTree: Computing large minimum evolution trees with profiles instead of a distance matrix. Molecular Biology and Evolution. 2009; 26(7):1641–1650. https://doi.org/10.1093/molbev/msp077 PMID: 19377059

49. Price MN, Dehal PS, Arkin AP. FastTree 2—Approximately maximum-likelihood trees for large alignments. PLOS ONE. 2010 03; 5(3):1–10. https://doi.org/10.1371/journal.pone.0009490

50. Tavaré S. Some probabilistic and statistical problems in the analysis of DNA sequences. Lectures on Mathematics in the Life Sciences. 1986; 17(2):57–86.

51. Tamura K, Nei M. Estimation of the number of nucleotide substitutions in the control region of mitochondrial DNA in humans and chimpanzees. Molecular Biology and Evolution. 1993; 10(3):512–526. https://doi.org/10.1093/oxfordjournals.molbev.a040023 PMID: 8336541

52. Torgerson WS. Multidimensional scaling: I. Theory and method. Psychometrika. 1952; 17:401–419. https://doi.org/10.1007/BF02288916

53. David Cournapeau. scikit-learn;. Software download. Available from: https://scikit-learn.org.

54. Murrell B, Moola S, Mabona A, Weighill T, Sheward D, Kosakovsky Pond SL, et al. FUBAR: A Fast, Unconstrained Bayesian AppRoximation for Inferring Selection. Molecular Biology and Evolution. 2013; 30(5):1196–1205. https://doi.org/10.1093/molbev/mst030 PMID: 23420840

55. Ember Core Team. Ember.js;. Software download. Available from: https://emberjs.com/.

56. Mike Bostock, Jason Davies, Jeffrey Heer, Vadim Ogievetsky, and community. D3.js;. Software download. Available from: http://d3js.org/.

57. Lee JH, Ozorowski G, Ward AB. Cryo-EM structure of a native, fully glycosylated, cleaved HIV-1 envelope trimer. Science. 2016; 351(6277):1043–1048. https://doi.org/10.1126/science.aad2450 PMID: 26941313

58. Marco Biasini. pv;. Software download. Available from: http://biasmv.github.io/pv/.

59. Sergei L Kosakovsky Pond. phylotree.js;. Software download. Available from: https://github.com/veg/phylotree.js.

60. 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 Pathogens. 2016 01; 12(1):1–22. https://doi.org/10.1371/journal.ppat.1005369

61. Deshpande S, Patil S, Kumar R, Hermanus T, Murugavel KG, Srikrishnan AK, et al. HIV-1 clade C escapes broadly neutralizing autologous antibodies with N332 glycan specificity by distinct mechanisms. Retrovirology. 2016 Aug; 13(1):60. https://doi.org/10.1186/s12977-016-0297-2 PMID: 27576440