Incorporating within-host diversity in phylogenetic analyses for detecting clusters of new HIV infections

August Guang[1,2], Mark Howison[3], Lauren Ledingham[4], Matthew D'Antuono[4], Philip A. Chan[4], Charles Lawrence[5], Casey W. Dunn[6], Rami Kantor[4]

[1] Center for Computational Biology of Human Disease, Brown University, Providence, RI, USA

[2] Center for Computation and Visualization, Brown University, Providence, RI, USA

[3] Research Improving People's Lives, Providence, RI, USA

[4] Division of Infectious Diseases, The Alpert Medical School, Brown University, Providence, RI, USA

[5] Division of Applied Mathematics, Brown University, Providence, RI, USA

[6] Department of Ecology and Evolutionary Biology, Yale University, New Haven, CT, USA

Abstract

Background: Phylogenetic analyses of HIV sequences obtained in clinical care are increasingly applied to detect clustering of new HIV infections and inform public health interventions to disrupt transmission. Conventional approaches summarize within-host HIV diversity with only a single consensus sequence per individual of only the *pol* gene, obtained from Sanger or next-generation sequencing (NGS).

Methods: We evaluated the potential benefits of incorporating diversity into phylogenetic cluster inference for all newly-HIV-diagnosed individuals within six months at the largest HIV center in Rhode Island. We compared Sanger- and NGS-derived *pol* and near-whole genome consensus sequences to an alternate *sample profiling* approach that incorporates deep-sequenced NGS data to capture within-host diversity.

Results: The space of phylogenies inferred through *profile sampling* reveals that consensus-inferred point estimates might not adequately represent them, both in topology and branch length. Cluster inference differed between Sanger- and NGS-derived sequences, and across gene regions; and only 30% of clusters were detected unanimously.

Discussion: *Profile sampling* can incorporate within-host HIV diversity captured in NGS-derived deep sequencing into phylogenetic analyses. This additional information might improve the robustness of cluster detection and has the potential to aid in better understanding and disrupting transmission.

Background

Public health officials and providers are interested in inferring transmission links between individuals with HIV to inform and improve HIV treatment and prevention []. In the absence of reliable patient contact histories, phylogenetic analysis of HIV sequence data can and has been be used to infer transmission clusters [1], under the assumption that two individuals sharing a most recent common ancestor in a phylogeny are more likely to share an epidemiological link in the real, unobservable transmission network. The application of phylogenetic analysis and cluster inference techniques in public health interventions to disrupt transmission was delineated as one of the four key pillars for achieving the Department of Health and Human Services' recently-announced plan for ending the HIV epidemic in the US [2].

While historically the phylogenetic informativeness of the *pol* region of the HIV genome was suggested and contested [3, 4], its use is now widespread in phylogenetic analysis and cluster inference, often due to the availability of *pol* sequences from guideline-recommended routine clinical drug resistance testing, typically performed by commercial Sanger sequencing [5]. In a recent review of HIV cluster inference, 98 out of 105 (93%) analyzed the *pol* region [6].

The increasing availability of NGS technology has led to longer and deeper sequencing of HIV, and data sets more routinely cover nearly the whole genome with thousands or more reads at each site [7]. Recent evidence suggests improvements in both phylogenetic analysis and cluster inference from near-whole genome HIV sequences obtained with NGS. For example, Yebra *et al.* [8] found that the accuracy of phylogenetic reconstruction and cluster inference on simulated sequences improved with longer genomic regions (with the best accuracy from a *gag-pol-env* concatenation). Novitsky *et al.* [9] similarly studied the effects on cluster inference of using longer genomic regions from real near-whole genome publicly available Sanger sequences, and found that the proportion of sequences in clusters increased with longer sequence regions.

While the potential advantages of longer sequences in inferring transmission have been demonstrated, the advantages of deeper sequencing have not yet been fully investigated, and whether it can improve HIV molecular clustering inference is unknown. In part this is because researchers often rely on consensus sequences that discard all but the majority variant at each site, since most phylogenetic methods require a single fully resolved sequence for each individual included in the phylogeny. Accordingly, researchers studying HIV transmission summarize the within-host HIV variation present in NGS data sets with a consensus sequence. Beyond the application of phylogenetic methods to HIV sequences, this consensus approach carries an underlying statistical assumption of *low relative entropy* [10]. In the context of HIV, this assumption is that a consensus sequence can adequately capture all of the relevant information

about within-host variation available in a deeply-sequenced NGS data set. Few previous studies of HIV transmission dynamics have accounted for within-host variation with coalescent evolutionary models [11, 12], but such models still assume a consensus sequence as the observed data.

In this study, we examine if and to what extent the incorporation of within-host variation available from deeply-sequenced NGS data improves consensus-sequence-based phylogenetic inference in a set comprising all newly HIV-diagnosed individuals during a six month time period from the Immunology Center, the largest HIV center in Rhode Island, USA. We present a new approach called *profile sampling* that uses the within-host variation in the NGS data to assess the robustness of the consensus approach for phylogenetic analysis and cluster inference.

Methods

Data collection and sequencing

HIV-1 *pol* Sanger sequences (HXB2 positions 2253-3554), available through clinical care, were collected from the 37 adults ($\geq$18 years) newly-diagnosed with HIV during the first six months of 2013 and treated at The Miriam Hospital Immunology Center in Providence, Rhode Island, USA. Patients at this Center represent approximately 80

In addition, blood specimens were obtained from consenting participants and processed to isolate peripheral blood mononuclear cells (PMBC) and plasma. Using both Sanger sequencing and NGS, near-whole genome viral sequences were obtained from one of those compartments; plasma for participants with detectable viral load and PBMC for participants with undetectable viral load or unsuccessful plasma genotyping. Total nucleic acids were extracted fand an in-house genotyping assay was used to generate the near-whole genome based on previously published methods [13, 14]. For each sample, two cDNA templates were generated by SuperscriptIII First Strand Synthesis System (Thermofisher, Carlsbad, CA), followed by eight separate nested PCR reactions; these eight amplicons span the near-whole genome of HIV. Final amplicon products were sequenced by the Sanger method using 3100 Genetic Analyzer (Applied Biosystems, Foster City, CA) and by NGS using Nextera XT DNA Library Prep chemistry (Illumina, San Diego, CA) to generate multiplexed libraries for Illumina's MiSeq platform with 250 base paired-end reads. Sanger consensus sequences were generated manually using Sequencher version 5.2.4 (Gene Codes, Ann Arbor, MI) to confirm degenerate nucleotides. NGS data were processed and demultiplexed using the BaseSpace cloud application (Illumina, San Diego, CA).

Profile sampling

We introduce a new approach for incorporating within-host variation into phylogenetic analysis, called *profile sampling*. We start by aligning each individual's NGS reads using the hivmmer pipeline [15], which we extended to support near-whole genome HIV data and to perform codon-aware alignment within each gene (pending release as hivmmer version 0.3.0). A key feature of this pipeline is its use of profile hidden Markov models (HMMs) to model and align collections of HIV sequences. Profile HMMs have been used in many kinds of biological sequence analyses and are particularly well-suited to modeling variation in populations of sequences [16]. Briefly, hivmmer performs quality control and error correction in overlapping regions of the read pairs using PEAR version 0.9.11 [17], translates them into each possible reading frame, aligns them in amino acid space to profile HMMs of all group M reference sequences from the Los Alamos National Lab HIV Database [18] using the profile HMM alignment tool HMMER version 3.1b2 [19], and produces a codon frequency table across the near-whole HIV genome. We refer to this resulting codon frequency table as the individual's HIV *profile*.

For each patient, we construct a population of fully-resolved sequences by sampling codons at each site in the genome using the codon frequencies from the *profile*. The collection of sampled sequences captures the empirical distribution of within-host variation at the codon level. We then sample 500 sequences from each individual's *profile*, and collate the sampled sequences into 500 *profile-sampled* data sets, each having one sampled sequence per individual. We subsequently construct the 500 fully-resolved sequences per individual in order to use existing phylogenetic methods, because there is currently no published method to our knowledge for inferring a phylogeny directly from the profile representation of the individual's within-host variation.

We estimate the within-host diversity for individuals as the average percent difference in nucleotides across all pairwise comparisons of their 500 profile-sampled nucleotide sequences. Variation in within-host diversity could be due to a variety of biological factors, such as virus mutation rate, effective viral population size, and time since infection, as well as technical factors such as sequencing depth and sequencing error rates. These pairwise differences are calculated using the Hamming distance [20] (also called the p-distance in the HIV genetic diversity literature [6, 21]).

Phylogenetic inference

For profile sampling, we perform phylogenetic inference of near-whole genome sequences (wgs) on each of the 500 profile-sampled data sets by estimating a multiple sequence alignment with mafft version 7.313 [22] and a maximum-likelihood phylogeny with the GTRCAT model and 100 rapid bootstrap replicates using RAxML version 8.2.12 [23].

In addition to profile sampling, we infer two phylogenies with the same tools and parameters, one from the NGS consensus sequences and one from Sanger consensus sequences. We perform cluster inference on these phylogenies using Cluster Picker [24] with thresholds of 80% bootstrap support and 4.5% genetic distance. Analysis source code is available from `https://github.com/kantorlab/hiv-profile-sampling`.

Lastly, we also perform phylogenetic inference on three clinically relevant sub-genomic regions: the protease and reverse transcriptase regions at the beginning of the *pol* gene ("prrt"), the *int* gene, and the *env* gene. The prrt and *int* regions are routinely sequenced in clinical care to detect drug resistance mutations and inform clinical choices of anti-retroviral therapy. The *env* region is sequenced to genotypically infer viral tropism and co-receptor usage.

We investigate the impact of within-host diversity on phylogenetic topology and evolutionary distance estimates. To examine the variation in topology, we first calculate the pairwise geodesic distance [25, 26] among the 500 phylogenies from the profile samples, as well as the phylogenies from the NGS and Sanger consensus sequences. Then we perform multi-dimensional scaling (MDS) on the resulting distance matrix to visualize the topological space in two dimensions. Next, to examine the variation in estimated evolutionary distance, we sum the branch lengths within each phylogeny across all branches and across only the tip branches, and visualize the distribution of these branch length sums. Finally, we examine the clusters that are detected in phylogenies of NGS consensus sequences versus Sanger consensus sequences, and across the four genomic regions. We use the proportion of times a cluster appears across the 500 *profile-sampled* phylogenies as a measure of support for that cluster. We refer to this value as the *cluster support* and note that it is similar to the use of bootstrap support for evaluating the robustness of an individual phylogeny's topology, but extends that idea to evaluating the robustness of cluster detection using within-host sequence variation.

Results

Profile sampling estimates of within-host diversity

Figure 1 shows the estimated within-host percent diversity in each examined genomic region across individuals, ordered by *env*, which we expected *a priori* to be the most variable region. The largest estimated diversity is in *env* for individual MC28 (4.0%), and *env* has the overall largest range in estimated diversity (0.2% to 4.0%). The other regions have ranges of 0.2% to 1.9% (prrt), 0.1% to 2.0% (*int*) and 0.2% to 2.6% (wgs). Such within-host estimations are not feasible via the conventional consensus Sanger or NGS approaches.

Phylogenetic estimates are sensitive to within-host diversity

Figure 2 demonstrates multi-dimensional scaling on the profile-sampled phylogenies and the phylogenies from the NGS and Sanger consensus sequences. The *profile-sampling* approach reveals a multi-modal topological space in which the phylogenies inferred from the consensus sequences can be outliers. The prrt region displays four topological clusters, and the NGS and Sanger consensus phylogenies are located in different *profile-sampled* clusters. There are two *profile-sampled* clusters in *int*, and the phylogeny of NGS consensus sequences lies at the center of one of the clusters, while the phylogeny of Sanger consensus sequences is an outlier. The phylogenies based on consensus sequences lie at opposite ends of the primary *profile-sampled* MDS axis in *env*, and of the secondary *profile-sampled* MDS axis in the wgs region.

Figure 3 shows the distribution of branch length sums across the compared phylogenies. Overall, the estimates are larger in *env* and the wgs regions, and smaller when restricting to only the tip branches. In some cases, the consensus phylogenies provide an adequate summary of the distribution (as in phylogeny of Sanger consensus sequences for all branches in *env*, or the phylogeny of NGS consensus sequences for tip branches in *env*). In other cases, the consensus phylogenies have estimates that are outliers in the distribution (as in both consensus phylogenies for either all or tip branches in the commonly-used prrt region).

Taken together, the heterogeneity between the phylogenetic results from profile sampling and the consensus-inferred point estimates demonstrate that within-host virus sequence diversity impacts the inference of virus phylogeny across patients, and that the consensus approach for collapsing within-host sequence variation can obscure both the magnitude and effect of these impacts. For example, the variation can result in multi-modal phylogenetic results and the point estimates from the phylogenies of consensus sequences can be outliers.

Inferred clusters differ by sequencing method and genomic region

The phylogeny of NGS consensus sequences detected eight clusters in prrt, nine in *int*, four in *env*, and six in wgs (Figure 4). Some clusters were robustly detected across all regions (e.g. the cluster (MC25, MC26, MC52), which had support between 98% and 100%), while other clusters were detected at lower *cluster support*, and were not detected consistently across regions (e.g. cluster (MC14, MC59), which only appeared in the *int* and wgs phylogenies, with *cluster support* of 60% to 62%).

The phylogeny of Sanger consensus sequences detected eight clusters in prrt, seven in *int*, six in *env*, and seven in wgs (Figure 5). As with the phylogeny of NGS consensus sequences, the support and consistency of cluster detection

varied across regions. Cluster detection was mostly consistent with the NGS consensus inference, except for the cluster (MC17, MC21), which only occurred in *env* in the Sanger consensus, and had zero *cluster support*. Three clusters were consistently detected across NGS and Sanger, and across all regions: (MC23, MC24); (MC25, MC26, MC52); and (MC27, MC28). Only 30% of clusters were detected unanimously. Figure 6 provides a visual summary of the differences in cluster detection between Sanger- and NGS-derived sequences and across genomic regions.

Discussion

In a dataset of all newly HIV-diagnosed individuals within six months at the largest HIV center in Rhode Island, USA, the deeper sequencing provided by NGS, represented by our introduced *profile sampling*, captured within-host diversity more effectively than previous generations of data acquisition methods. These results impact phylogenetic inference across individuals and suggest that the consensus approach is discarding potentially-relevant information that is present in the NGS data. This additional information has the potential to improve the robustness of cluster detection.

The diversity ranges we estimated are comparable to prior studies on within-host diversity [27, 28].

Although collapsing within-host diversity into a single consensus sequence simplifies downstream analyses, results presented here demonstrate that it discards potentially-relevant biological results and can mislead downstream resulting epidemiological consequences. For example, public health activities triggered by phylogenetic inference of HIV molecular clustering to inform and improve HIV prevention and treatment interventions can be affected [29]. In our data, clusters vary in their *profile-sampling* support, and consensus approaches can detect clusters that have low cluster support from deep-sequenced data. Public health activities may be better justified in scenarios where the clusters triggering them are determined to have high cluster support from deep-sequenced data. As data acquisition increasingly shifts to NGS approaches, it is important to compare results from new methods to the conventional Sanger *pol* consensus sequences that have been used since the first molecular characterizations of HIV sequence diversity.

Much of the enthusiasm about shifting from Sanger sequencing to NGS has been due to reducing costs and the ability to collect data on the entire HIV genome rather than a few genes. Our results suggest that much of the benefit of NGS may also be due to its greater depth of sequencing, which captures more variation in virus sequences within patients. This benefit can only be realized, though, if this variation is propagated to phylogenetic analyses, such as by the *profile-sampling* method introduced here, rather than being collapsed to a consensus sequence, as is conventionally done.

Our approach presented here allows for the consideration of within-host variation using standard off-the-shelf tools, just by propagating variation further in the analyses. Rather than collapse variation into consensus sequences prior to phylogenetic analysis, we suggest to create a profile that captures that variation, perform multiple phylogenetic analyses on sequences sampled from the profile, and then summarize the phylogenetic analyses. Future tools could incorporate the variation directly into the phylogenetic inference process itself 10.

In our comparison of cluster inference across genomic regions, we found that fewer clusters were detected overall in *env* and wgs compared to prrt and *int*. Prior studies of clustering from Sanger consensus sequences present mixed results on the prevalence of clustering across genomic regions. Some studies have found concordant clustering across *gag-env* [30] and *gag-pol-env* [31, 32], while others found fewer clusters in *pol* than in *env* [33], or fewer clusters in *gag-env* than in *pol* [34]. The additional information available in deep-sequenced NGS data, along with the cluster support measures provided by profile sampling, could help resolve these differences, as suggested here. In our data, not only were more clusters detected in the *pol* region, but those clusters had higher cluster support as measured by deep-sequenced NGS data.

One limitation of our study is the small number of participants and the short time-frame patients were enrolled in. The participants enrolled here represent a dense temporal sampling, and comprise all newly HIV-diagnosed individuals in a six month period at the largest HIV center in Rhode Island, USA, in which 80% of the state's people with HIV are cared for. The overall size of the HIV epidemic in Rhode Island was estimated as 2,396 individuals in 2016 [35], but NGS data for this population are not currently available beyond those presented here. In future work, we will apply *profile sampling* to larger NGS data sets, to assess cluster inference concordance between Sanger and NGS data, especially for clusters larger than three individuals, which was the largest cluster size here.

Our construction of HIV profiles from NGS data is also limited by the accuracy of the NGS assays themselves. The codon frequencies in the profiles may be biased measures of the true within-host diversity because of biases in PCR amplification. Sequencing protocols such as Primer ID [36] have been introduced to reduce and correct for these biases, and should be considered in the future.

In conclusion, the true HIV transmission network is unknown, but phylogenetic analysis and cluster inference are promising tools for aiding clinicians and public health officials in better understanding and in disrupting it [2]. Current phylogenetic approaches do not fully utilize the information on within-host diversity available in deep-sequenced, near-whole genome NGS data. As NGS data sets are increasingly available and become more representative of the HIV epidemics, we suggest that the additional information they measure has the potential to improve the robustness of HIV molecular cluster inference, the impact of which needs to be further investigated.

Funding

This work was supported by the National Institutes of Health [R01AI136058, K24AI134359, P30AI042853, P20GM109035] and a Brown University DEANS Award.

Acknowledgments

We thank Dr. Mia Coetzer for assistance with next-generation sequencing and feedback on an earlier draft of the manuscript. This research was conducted using computational resources and services at the Center for Computation and Visualization, Brown University.

References

1. Leitner, T, Escanilla, D, Franzen, C, Uhlen, M, and Albert, J. Accurate reconstruction of a known HIV-1 transmission history by phylogenetic tree analysis. Proceedings of the National Academy of Sciences 1996;93:10864–10869.
2. Fauci, AS, Redfield, RR, Sigounas, G, Weahkee, MD, and Giroir, BP. Ending the HIV Epidemic: A Plan for the United States. JAMA 2019;321:844.
3. Hué, S, Clewley, JP, Cane, PA, and Pillay, D. HIV-1 pol gene variation is sufficient for reconstruction of transmissions in the era of antiretroviral therapy. AIDS 2004;18:719–728.
4. Stürmer, M, Preiser, W, Gute, P, Nisius, G, and Doerr, HW. Phylogenetic analysis of HIV-1 transmission: pol gene sequences are insufficient to clarify true relationships between patient isolates. AIDS 2004;18:2109–2113.
5. Panel on Antiretroviral Guidelines for Adults and Adolescents. Guidelines for the Use of Antiretroviral Agents in Adults and Adolescents with HIV. Department of Health and Human Services. Available at `http://aidsinfo.nih.gov/contentfiles/lvguidelines/AdultandAdolescentGL.pdf`. Accessed 9-Feb-2020.
6. Hassan, AS, Pybus, OG, Sanders, EJ, Albert, J, and Esbjörnsson, J. Defining HIV-1 transmission clusters based on sequence data: AIDS 2017;31:1211–1222.
7. Voelkerding, KV, Dames, SA, and Durtschi, JD. Next-Generation Sequencing: From Basic Research to Diagnostics. Clinical Chemistry 2009;55:641–658.

8. Yebra, G, Hodcroft, EB, Ragonnet-Cronin, ML, et al. Using nearly full-genome HIV sequence data improves phylogeny reconstruction in a simulated epidemic. Scientific Reports 2016;6:39489.

9. Novitsky, V, Moyo, S, Lei, Q, DeGruttola, V, and Essex, M. Importance of Viral Sequence Length and Number of Variable and Informative Sites in Analysis of HIV Clustering. AIDS Research and Human Retroviruses 2015;31:531–542.

10. Guang, A, Zapata, F, Howison, M, Lawrence, CE, and Dunn, CW. An Integrated Perspective on Phylogenetic Workflows. Trends in Ecology & Evolution 2016;31:116–126.

11. Giardina, F, Romero-Severson, EO, Albert, J, Britton, T, and Leitner, T. Inference of Transmission Network Structure from HIV Phylogenetic Trees. PLOS Computational Biology 2017;13:e1005316.

12. Romero-Severson, E, Skar, H, Bulla, I, Albert, J, and Leitner, T. Timing and Order of Transmission Events Is Not Directly Reflected in a Pathogen Phylogeny. Molecular Biology and Evolution 2014;31:2472–2482.

13. Nadai, Y, Eyzaguirre, LM, Constantine, NT, et al. Protocol for Nearly Full-Length Sequencing of HIV-1 RNA from Plasma. PLoS ONE 2008;3:e1420.

14. Di Giallonardo, FD, Töpfer, A, Rey, M, et al. Full-length haplotype reconstruction to infer the structure of heterogeneous virus populations. Nucleic Acids Research 2014;42:e115–e115.

15. Howison, M, Coetzer, M, and Kantor, R. Measurement error and variant-calling in deep Illumina sequencing of HIV. Bioinformatics 2019;35:2029–2035.

16. Eddy, SR. What is a hidden Markov model? Nature Biotechnology 2004;22:1315–1316.

17. Zhang, J, Kobert, K, Flouri, T, and Stamatakis, A. PEAR: a fast and accurate Illumina Paired-End reAd mergeR. Bioinformatics 2014;30:614–620.

18. Los Alamos National Lab. HIV Database. Available at `http://www.hiv.lanl.gov/`. Accessed 9-Feb-2020.

19. Eddy, SR. Accelerated Profile HMM Searches. PLOS Computational Biology 2011;7:e1002195.

20. Allam, O, Samarani, S, and Ahmad, A. Hammering out HIV-1 incidence with Hamming distance: AIDS 2011;25:2047–2048.

21. Maldarelli, F, Kearney, M, Palmer, S, et al. HIV Populations Are Large and Accumulate High Genetic Diversity in a Nonlinear Fashion. Journal of Virology 2013;87:10313–10323.

22. Katoh, K and Standley, DM. MAFFT Multiple Sequence Alignment Software Version 7: Improvements in Performance and Usability. Molecular Biology and Evolution 2013;30:772–780.

23. Stamatakis, A. RAxML Version 8: A tool for Phylogenetic Analysis and Post-Analysis of Large Phylogenies. Bioinformatics 2014:btu033.

24. Ragonnet-Cronin, M, Hodcroft, E, Hué, S, et al. Automated analysis of phylogenetic clusters. BMC Bioinformatics 2013;14:317.

25. Billera, LJ, Holmes, SP, and Vogtmann, K. Geometry of the Space of Phylogenetic Trees. Advances in Applied Mathematics 2001;27:733–767.

26. Owen, M and Provan, JS. A Fast Algorithm for Computing Geodesic Distances in Tree Space. IEEE/ACM Transactions on Computational Biology and Bioinformatics 2011;8:2–13.

27. Li, G, Piampongsant, S, Faria, NR, et al. An integrated map of HIV genome-wide variation from a population perspective. Retrovirology 2015;12:18.

28. Zanini, F, Brodin, J, Thebo, L, et al. Population genomics of intrapatient HIV-1 evolution. eLife 2015;4:e11282.

29. Peters, PJ, Pontones, P, Hoover, KW, et al. HIV Infection Linked to Injection Use of Oxymorphone in Indiana, 2014–2015. New England Journal of Medicine 2016;375:229–239.

30. Han, Z, Leung, TW, Zhao, J, et al. A HIV-1 heterosexual transmission chain in Guangzhou, China: a molecular epidemiological study. Virology Journal 2009;6:148.

31. English, S, Katzourakis, A, Bonsall, D, et al. Phylogenetic analysis consistent with a clinical history of sexual transmission of HIV-1 from a single donor reveals transmission of highly distinct variants. Retrovirology 2011;8:54.

32. Kaye, M, Chibo, D, and Birch, C. Phylogenetic Investigation of Transmission Pathways of Drug-Resistant HIV-1 Utilizing Pol Sequences Derived From Resistance Genotyping: JAIDS Journal of Acquired Immune Deficiency Syndromes 2008;49:9–16.

33. Kapaata, A, Lyagoba, F, Ssemwanga, D, et al. HIV-1 Subtype Distribution Trends and Evidence of Transmission Clusters Among Incident Cases in a Rural Clinical Cohort in Southwest Uganda, 2004–2010. AIDS Research and Human Retroviruses 2013;29:520–527.

34. Ndiaye, HD, Tchiakpe, E, Vidal, N, et al. HIV Type 1 Subtype C Remains the Predominant Subtype in Men Having Sex with Men in Senegal. AIDS Research and Human Retroviruses 2013;29:1265–1272.

35. Rhode Island Department of Health. HIV Progress [Internet]. Available from: https://health.ri.gov/data/hiv/. [Accessed 12 Dec 2019]. 2019.

36. Jabara, CB, Jones, CD, Roach, J, Anderson, JA, and Swanstrom, R. Accurate sampling and deep sequencing of the HIV-1 protease gene using a Primer ID. Proceedings of the National Academy of Sciences 2011;108:20166–20171.

Figure 1: Intra-patient genetic diversity (defined as the average percent difference across all pairwise comparisons of the 500 profile-sampled nucleotide sequences for a patient) is highest in *env* for most individuals, and lies within the range of previously reported values.

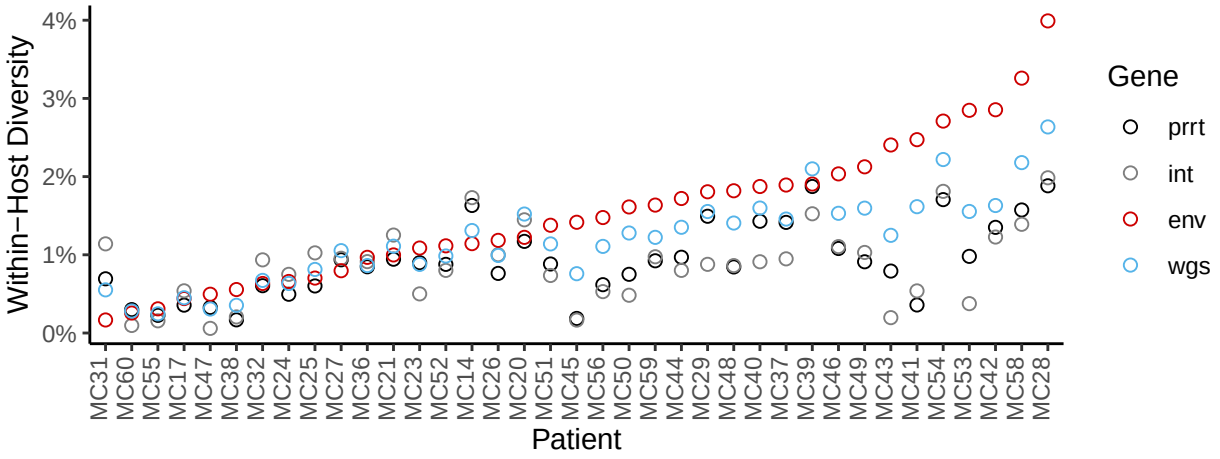

Figure 2: Multi-dimensional scaling of pairwise geodesic distance among maximum-likelihood phylogenies from the profile-sampling approach show that the space of phylogenies inferred for the prrt and int regions are multi-modal. The phylogenies from consensus sequences (black dots) are point estimates that do not capture the full variation in phylogenies that can be inferred from deeply-sequenced NGS data.

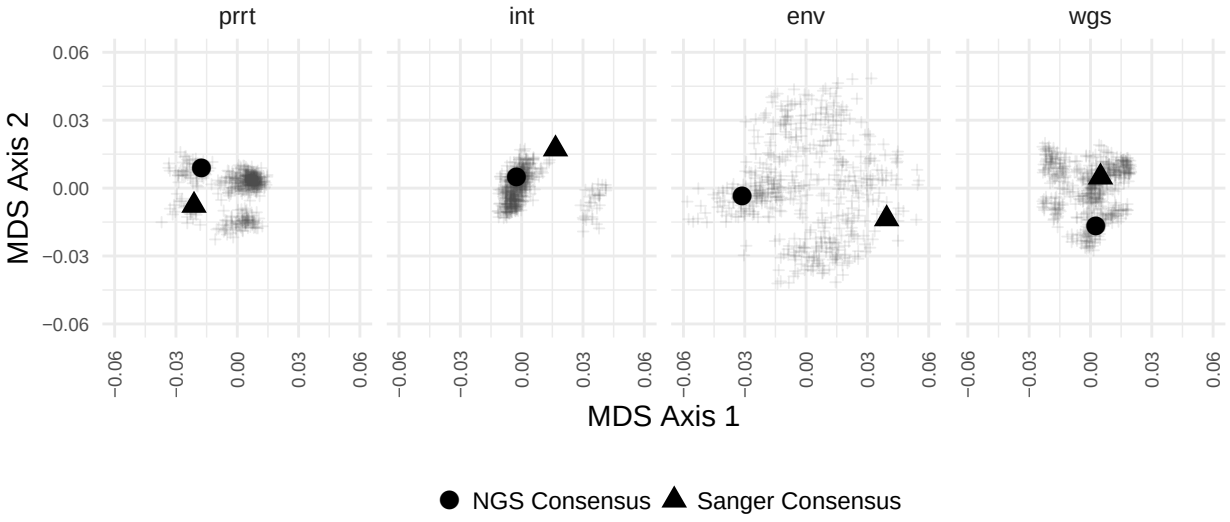

Figure 3: The total branch length in each of the profile-sampled phylogenies also varies. The phylogenies from consensus sequences (black dots) can lie at extreme values within these distribution, both when considering the lengths across all branches and the lengths across only the branches at the tips.

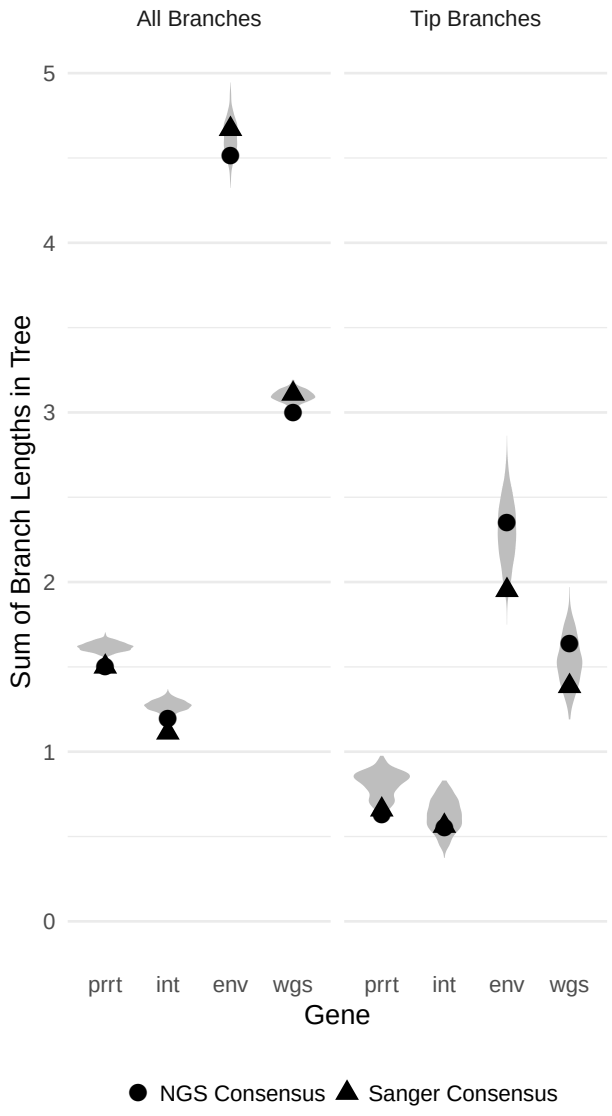

Figure 4: Clusters (vertical colored bars) inferred from the phylogenies of NGS consensus sequences differ across genomic regions. The largest number of clusters was inferred from the int region, and the smallest number from the env region. Profile sampling provides a bootstrapped measure of cluster support (annotation to bars).

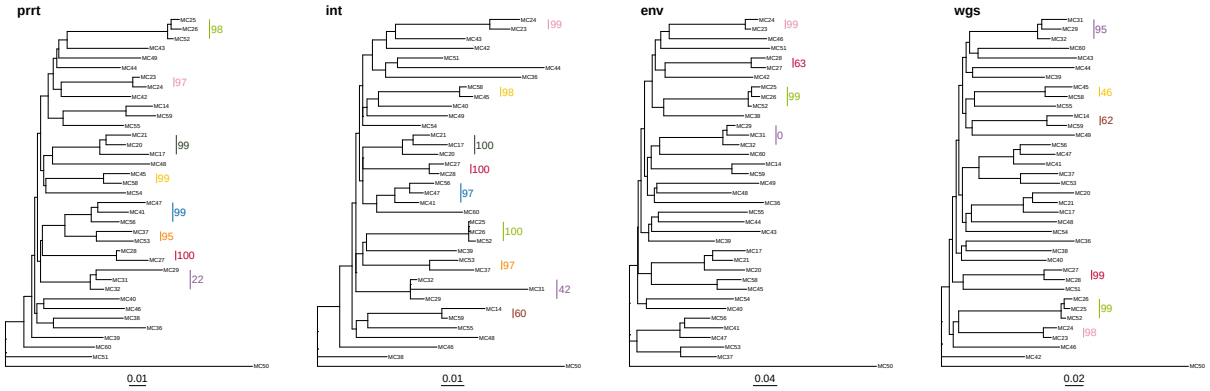

Figure 5: Clusters (vertical colored bars) inferred from the phylogenies of Sanger consensus sequences differ across genomic regions. The largest number of clusters was inferred from the int region, and the smallest number from the env region. Profile sampling provides a bootstrapped measure of cluster support (annotation to bars).

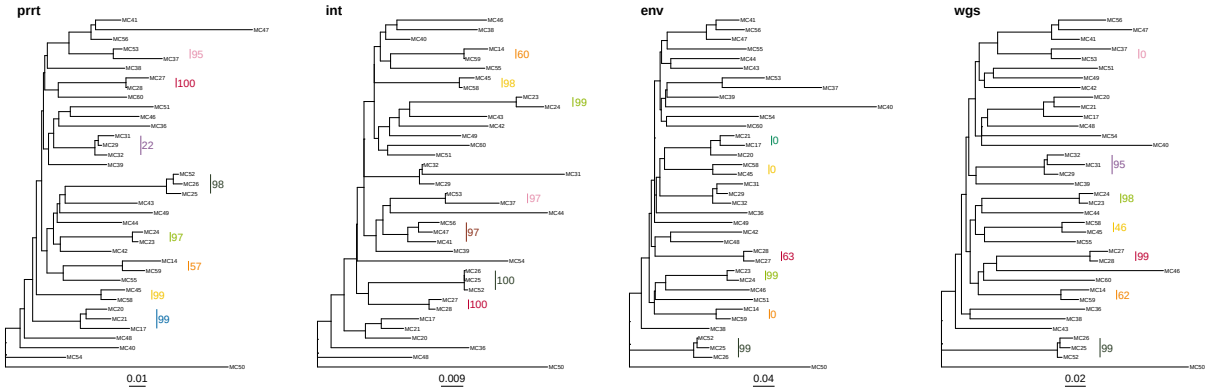

Figure 6: Summary of clusters identified by Sanger versus NGS consensus sequences across genomic regions. Numeric values indicate bootstrapped cluster support from the profile sampling method. A blank cell indicates that the cluster was not detected in that consensus sequence and genomic region.