Robustness of phylogenetic analysis for detecting clusters of new HIV infections

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

[1] Computing and Information Services, Brown University, Providence, RI, USA

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

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

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

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

Abstract

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

Methods: We evaluate the robustness of the consensus approach and the potential benefits of considering the within-host diversity in phylogenetic analysis and cluster inference for all newly HIV-diagnosed individuals in the first half of 2013 at the largest HIV center in Rhode Island, USA. We compare Sanger and NGS-derived *pol* and near-whole genome consensus sequences to an alternate approach that samples many sequences per individual from a profile hidden Markov model of their NGS data.

Results: The space of phylogenies inferred through sampling is multi-modal, suggesting that a consensus-inferred phylogeny is not an appropriate summary of within-host variation. Cluster inference differs in phylogenies from consensus sequences from Sanger and NGS data, and across gene regions.

Discussion: The choice of sequencing and summarization methods affects the detection of clusters, and should be considered carefully in public health applications of phylogenetic analysis to disrupt HIV transmission.

Background

Clinicians are generally interested in inferring transmission links between HIV-infected individuals to improve HIV treatment and prevention. In the absence of reliable patient contact histories, phylogenetic analysis of viral sequence data can 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 a transmission link in the real, unobservable transmission network. The application of phylogenetic analysis and cluster inference techniques in public health interventions to disrupt transmission was recently identified 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 inference power of the *pol* region of the HIV genome was initially contested [3, 4], its use is now widespread in phylogenetic and transmission cluster analyses, often due to the availability of *pol* sequences from routine clinical genotyping by commercial Sanger sequencing. In a recent meta study, Novitsky *et al.* (forthcoming) found that 102 out of 107 published studies of HIV cluster inference between 2016 and 2019 used the *pol* region alone.

The increasing availability of NGS technology has led to longer and deeper sequencing of HIV, and data sets that cover nearly the whole genome with thousands or more reads at each site. 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.* [5] 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.* [6] similarly studied the effects on cluster inference of using longer genomic regions from real near-whole genome Sanger sequences, and found that the proportion of sequences in clusters increased with longer sequence regions.

Most phylogenetic methods require a single fully resolved sequence for each individual included in the phylogeny. Accordingly, researchers often summarize the within-host variation present in NGS data sets with a consensus sequence. In the larger context of phylogenetic methods (e.g. beyond their application to HIV sequences), this consensus approach carries an underlying statistical assumption of *low relative entropy* [7]. In the context of HIV, tihs assumption is that the consensus sequence adequately captures most of the relevant information about the distribution of sequences that might be constructed from the within-host variation given the NGS data set. Some studies of HIV transmission dynamics have accounted for this variation with coalescent within-host evolutionary models [8, 9], but such models still assume a consensus sequence as the observed data.

In this study, we examine whether this assumption underlying the consensus approach holds on a data set comprising

all newly HIV-diagnosed individual in the first half of 2013 from the largest HIV center in Rhode Island, USA. We present a new analytical method called *profile sampling* that uses the within-host variation in the NGS data to assess the robustness of phylogenetic analysis and cluster inference when using the consensus approach.

Methods

Our study and data collection were approved by Lifespan's Institutional Review Board. We collected viral sequences from patients newly diagnosed with HIV during 2013 and treated at The Miriam Hospital Immunology Center in Providence, Rhode Island, USA. Inclusion criteria were: (i) HIV-infected adults, 18 years of age or older; (ii) diagnosed with HIV during 2013; and (iii) available *pol* sequence from routine drug resistance testing. We obtained whole genome viral sequences for the 37 included individuals using both Sanger and NGS. Patient identifiers were removed and all analysis was conducted with de-identified sequence data.

We introduce a new approach for incorporating within-host variation into phylogentic analysis, called *profile sampling*. We start by aligning each individual's NGS reads to all reference database of all sequences from the Los Alamos National Lab HIV Database with the probabilistic multiple sequence aligner HMMER version 3.1b2 [10], using a variant of the *hivmmer* pipeline [11] that has been extended to support near-whole genome HIV data. Next, we convert the resulting multiple sequence alignment to a profile hidden Markov model (pHMM). Because pHMMs are probabilistic representations of alignments, we can sample genomic sequences from each individual's pHMM.

sequences further allows us to estimate the probability of a molecular transmission 228 cluster given the intra-patient variation. To do so, we first infer a multiple sequence 229 alignments and phylogenetic tree for each sampled genomic sequence. Then, we 230 compute the bootstrap support [37] for each sampled tree and use a 99cutoff to identify transmission clusters. Phylogenetic support value thresholds are the 232 most common approach used to identify transmission clusters [9], hence our decision to 233 apply it to our workflow. The number of times a particular transmission cluster is 234 identified in the set of phylogenetic tree samples divided by the number of samples is 235 taken as the probability of that transmission cluster given the intra-patient variation. 236 For the profile-sampling approach we use the well-established pHMM alignment tool 243 HMMER 3.1b2 [38] to build, align and sample from the pHMMs.

For both approaches we estimate multiple alignments across individuals using mafft version 7.305b [39] and maximum-likelihood phylogenies with bootstrap replicates using RAxML version 8.2.10.

Results

Profile sampling captures inter-host diversity

Profile sampling allows us to visualize the degree of polymorphism within a patient (Figure 1) and sample genomic sequences in proportion to the within-host variation. It can also build a consensus sequence for easy comparison of new approaches to established methods. By visualizing the degree of polymorphism, we can assess whether genomic variation exists at significantly high enough frequencies such that neither consensus genomes nor any other point estimator are sufficient summaries.

Figure 1 shows substantial variation exists in the HIV profiles of all patients. Across all patients, we found that 5.63 percent of sites in the HIV genome are polymorphic, with 11.4 percent of those sites in the gag region, 21.4 percent of those sites in the pol region, and 29.5 percent of those sites in the env region of the genome. Here we define polymorphism to mean that a particular site has greater than 1

Phylogenetic estimates are sensitive to inter-host diversity

Given that previous studies on individuals infected with HIV have shown high intra-patient variation, and given that we found similar results in our set of data (Figure 1), we investigated its impact on downstream topological variation in the phylogenetic tree. The results show that trees inferred from the profile-sampled sequences do not cluster around any of the trees inferred from a genome point estimate (Figure 2).

When and MDS plot is generated for only the profile-sampled trees (Figure 3), they appeared to form two clusters. This is likely due to zoom on the profile-sampled trees, as can be seen from the axes scales for Figure 2 and Figure 3. Supertrees of the two clusters reveal that most of the clades in the two subtrees are the same, with only patients out of 37 having a different topological placement on the two trees (Figure 4). This suggests that distinct clusters of phylogenetic trees exist given the intra-patient variation, but that the differences between them are not large. However, the consensus genome appears to be an inadequate estimate summary for the intra-patient variation, as can be seen in Figure 5.

Inferred clusters differ by sequencing method and genomic region

Discussion

The true HIV transmission network is often unknown, but phylogenetic trees of patient viral sequences can provide information on transmission clusters that identify links between patients [1] as well as relevant features of transmission dynamics. [22, 23]. Current phylogenetic approaches to transmission inference utilize a single summary consensus sequence per patient, which ignores the extensive intra-patient viral diversity [6, 18].

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. 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.
6. 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.
7. 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.
8. 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.

9. 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.

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

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

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 the env region for most individuals, and lies within the range of previously reported values.

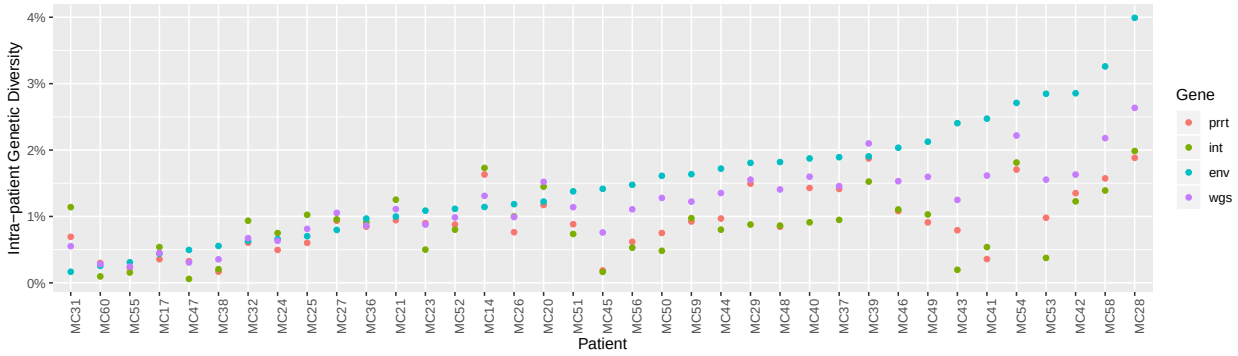

Figure 2: Multi-dimensional scaling of pairwise geodesic distance among maximum-likelihood phylogenies from the profile-sampling approach show that the space of phylogenies inferred for the prrt and int regions are multi-modal. The phylogenies from consensus sequences (black dots) are point estimates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

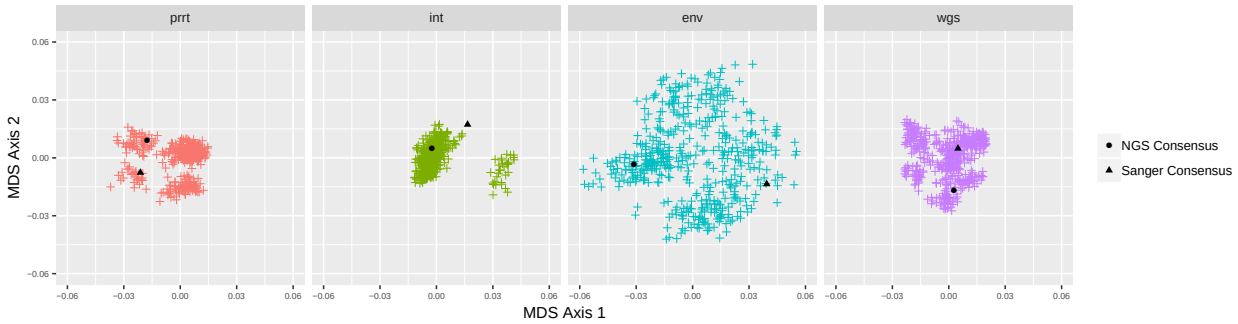

Figure 3: The total branch length in each of the profile-sampled phylogenies also varies. The phylogenies from consensus sequences (black dots) can lie at extreme values within these distribution, both when considering the lengths across all branches and the lengths across only the branches at the tips.

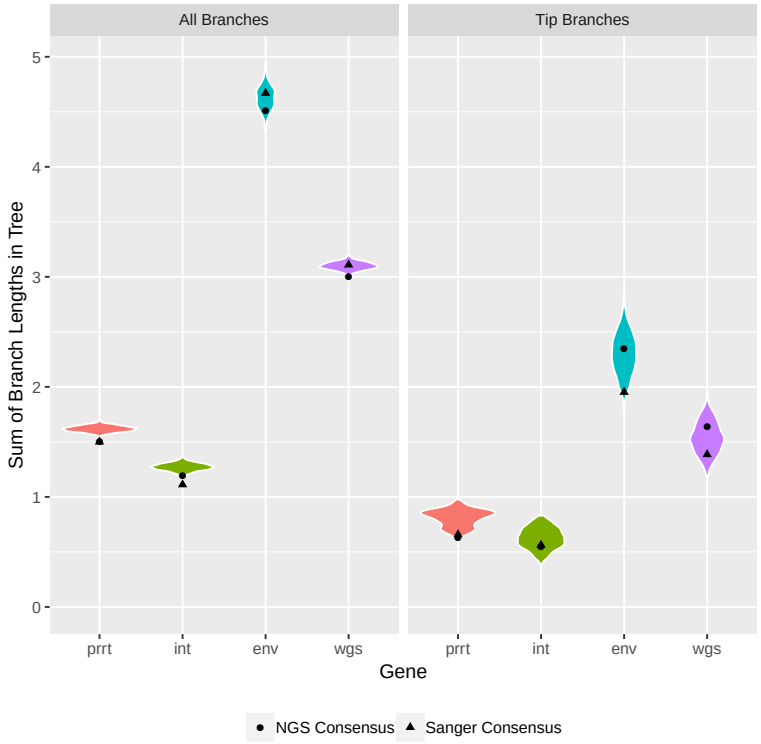

Figure 4: Clusters (vertical colored bars) inferred from the phylogenies of NGS consensus sequences differ across genomic regions. The largest number of clusters was inferred from the int region, and the smallest number from the env region. Profile sampling provides a bootstrapped measure of cluster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

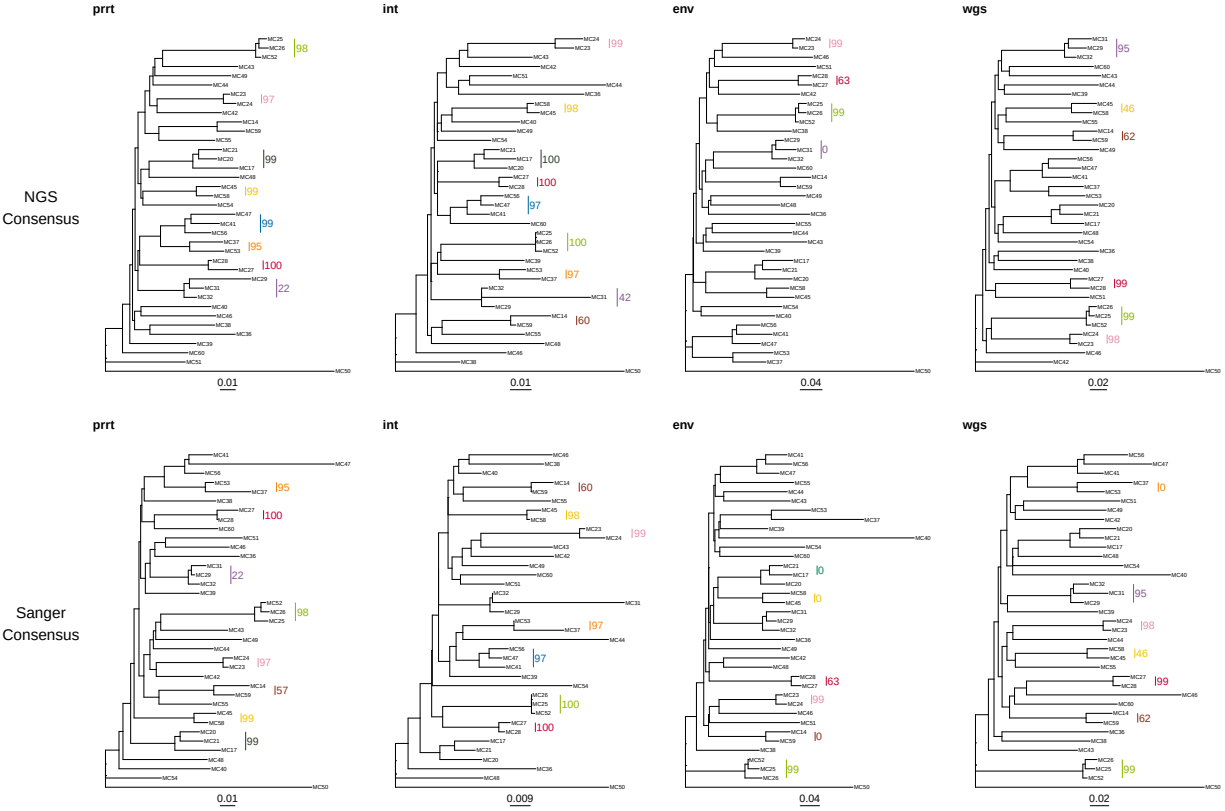

Figure 5: Summary of clusters identified by Sanger versus NGS consensus sequences across genomic regions. Numeric values indicate bootstrapped cluster support from the profile sampling method.