Title

Optimizing ancestral trait reconstruction of large HIV Subtype C datasets through multiple-trait subsampling

Permalink

https://escholarship.org/uc/item/7q26v0ts

Journal

Virus Evolution, 9(2)

ISSN

2057-1577

Authors

Li, Xingguang
Trovão, Nídia S
Wertheim, Joel O
et al.

Publication Date

2023-12-01

DOI

10.1093/ve/vead069

Peer reviewed

Virus Evolution, 2023, **9(2)**, 1–11
DOI: https://doi.org/10.1093/ve/vead069
Advance Access Publication 22 November 2023
Research Article

Optimizing ancestral trait reconstruction of large HIV Subtype C datasets through multiple-trait subsampling

Xingguang Li,[1,2,†,‡] Nídia S. Trovão,[3,†,§] Joel O. Wertheim,[4] Guy Baele,[5,**] and Adriano de Bernardi Schneider[6,***,*]

[1]Ningbo No.2 Hospital, Ningbo 315010, China, [2]Ningbo Institute of Life and Health Industry, University of Chinese Academy of Sciences, Ningbo 315000, China, [3]Division of International Epidemiology and Population Studies, Fogarty International Center, National Institutes of Health, 31 Center Dr, Bethesda, MA 20892, USA, [4]Department of Medicine, University of California, La Jolla, San Diego, CA 92093, USA, [5]Department of Microbiology, Immunology and Transplantation, Rega Institute, KU Leuven, Leuven BE-3000, Belgium and [6]Genomics Institute, University of California Santa Cruz, Santa Cruz, CA 95064, USA

[†]These authors contributed equally to this work.

[‡]https://orcid.org/0000-0002-3470-2196

[§]https://orcid.org/0000-0002-2106-1166

[**]https://orcid.org/0000-0002-1915-7732

[***]https://orcid.org/0000-0001-7487-266X

[*]Corresponding author: E-mail: adeberna@ucsc.edu

Abstract

Large datasets along with sampling bias represent a challenge for phylodynamic reconstructions, particularly when the study data are obtained from various heterogeneous sources and/or through convenience sampling. In this study, we evaluate the presence of unbalanced sampled distribution by collection date, location, and risk group of human immunodeficiency virus Type 1 Subtype C using a comprehensive subsampling strategy and assess their impact on the reconstruction of the viral spatial and risk group dynamics using phylogenetic comparative methods. Our study shows that a most suitable dataset for ancestral trait reconstruction can be obtained through subsampling by all available traits, particularly using multigene datasets. We also demonstrate that sampling bias is inflated when considerable information for a given trait is unavailable or of poor quality, as we observed for the trait risk group. In conclusion, we suggest that, even if traits are not well recorded, including them deliberately optimizes the representativeness of the original dataset rather than completely excluding them. Therefore, we advise the inclusion of as many traits as possible with the aid of subsampling approaches in order to optimize the dataset for phylodynamic analysis while reducing the computational burden. This will benefit research communities investigating the evolutionary and spatio-temporal patterns of infectious diseases.

Keywords: HIV Subtype C; multiple-trait subsampling; ancestral trait reconstruction; phylogenetic comparative methods; subsampling approaches.

Introduction

Large sequencing efforts have greatly increased the availability of genomic data of infectious agents or pathogens in public databases (Sheng et al. 2021). This data availability has led to the development of novel methods to speed up molecular epidemiological analyses of these datasets (Turakhia et al. 2021; McBroome et al. 2022). Yet, although these tools aim to solve the problem of data processing, they do not resolve the issue of data representativeness by themselves, as seen through the presence of sampling bias in large databases, resulting in datasets with a skewed distribution of certain traits not truly representing the population diversity. Genetic databases, such as GenBank (Sayers 2022) and the Global Initiative on Sharing All Influenza Data (Elbe and Buckland-Merrett 2017), are used as repositories for genomic data, which are often deposited at the moment of submission of a manuscript to peer-reviewed journals, with a few notable exceptions such as the genomic data deposited during the Ebola outbreak in West Africa (Arias et al. 2016), the severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2)/coronavirus disease 2019 pandemic (Roncoroni et al. 2021; Furuse 2021), and throughout the seasonal influenza virus surveillance efforts to inform vaccine composition (Hay and McCauley 2018). In public databases, sampling bias can be seen through the random deposit of samples in the database in an unintended way (sequences being deposited as project-dependent and not population-dependent) that does not reflect a fair representation of the true population, resulting in some traits (i.e. genetic diversity, location, and populations at greater risk of HIV acquisition [PGRHA]) of the target population having a lower or higher sampling probability than others compared to their actual prevalence (Faria et al. 2014; Popejoy and Fullerton 2016; Viana et al. 2022; He et al. 2012).

Sampling bias is a persistent concern when performing phylogeographic inference (De Maio et al. 2015; Kalkauskas et al. 2021). Apart from an increase in taxon sampling having been shown to aid in the reduction of phylogenetic error (Zwickl and Hillis 2002), several software applications target the reduction of size and redundancy for the purpose of phylogenetic analysis (Menardo et al. 2018) or the increase in phylogenetic diversity while reducing data set size (Minh, Klaere, and von Haeseler 2009). Sequencing errors and insufficient representation from large datasets can result in incorrect phylogenetic inferences, impeding accurate downstream conclusions (Vakulenko, Deviatkin, and Lukashev 2019), potentially affecting efforts such as lineage tracing (Elliott et al. 2020; Turakhia et al. 2020).

Sampling bias can, for example, lead to incorrect inference of ancestral locations and migration rates from oversampled regions, leading to spurious results that may affect public policy in the response of an epidemic (De Maio et al. 2015). The presence of sampling bias is challenging for all currently available phylogeographic models, and mitigating such bias might require large data set sizes and the incorporation of associated metadata in those models (Layan et al. 2023). Subsampling is typically a strategy to mitigate any biases present in a dataset and thus to improve the representativeness of the actual patterns of the epidemics. However, a recent study has shown that such subsampling strategies do not consistently improve (discrete) phylogeographic inference at intermediate levels of sampling bias and that the improvement is dependent on the actual migration model (Layan et al. 2023). The major purpose of subsampling is to make phylogenetic and phylodynamic analyses of very large genetic datasets computationally tractable, such as those for the human immunodeficiency virus (HIV) (Faria et al. 2014), seasonal influenza viruses (Bedford et al. 2015), and SARS-CoV-2 (Chakraborty et al. 2021).

Representativeness is multi-dimensional, in the sense that a single genomic dataset does not only consist of genomic data but multiple underlying metadata layers (traits) which when combined allow for a comprehensive view of the population represented by the dataset. Studies focusing on the spatio-temporal dynamics of pathogens often tend to subsample based on location, particularly for the challenging discrete (location) trait reconstruction analysis (Faria et al. 2014). However, even when the goal is to purely reconstruct the pathogen's spatial spread, including more traits during the subsampling process might improve the representativeness of the actual underlying patterns of the epidemics and lead to more accurate results.

Among the existing large genomic data repositories, the Los Alamos National Laboratory (LANL) HIV Sequence Database (https://www.hiv.lanl.gov) is one of the most widely used databases for HIV research. In addition to genomic data, the database contains metadata associated with the viral genetic sequences, including records of the collection date and sampling country along with PGRHA information for certain samples, thus making it an ideal database to evaluate sampling bias on multiple traits associated with the samples.

HIV-1 Subtypes B and C have the largest number of sequences recorded in the LANL HIV Sequence Database (https://www.hiv.lanl.gov/content/sequence/HIV/mainpage.html). As of 3 August 2022, there were 535,995 and 152,290 records for HIV-1 Subtypes B and C, respectively. Subtype B is the most widespread HIV-1 variant accounting for approximately 11% of all infections worldwide (Junqueira and Almeida 2016) and has been extensively studied including in the phylodynamic context (Hong et al. 2020). Despite there being studies addressing the global evolution and spatio-temporal patterns of HIV-1 Subtype C (Novitsky et al. 2010), to the best of our knowledge, the present study is the first one to address the potential influence of sampling bias on the accuracy of such reconstructions within Subtype C.

Many studies tend to focus on genomic analyses for specific regions; however, the regional transmission dynamics might not fully represent the overall evolutionary and spatio-temporal patterns of the disease worldwide. In this study, we evaluate the effect of dataset subsampling based on a combination of traits (date, country, and PGRHA) on phylogenetic inference and subsequent downstream analysis, such as ancestral trait reconstruction and phylogeographic inference. We subsample and analyze two large HIV-1 Subtype C datasets with sequences collected globally obtained from LANL, encompassing near-complete genome and partial *pol* gene, with associated metadata. We find that subsampling using a combination of genetic sequence and metadata traits yields more phylogenetic results closer to patterns in the original dataset than the usual subsampling based on a single metadata trait while being more computationally efficient.

Materials and methods

Sequence dataset compilation

All available near-complete genome sequences (HXB2 Genome Position 790–9417, with minimum fragment length of 6,000 bp) and partial *pol* sequences (HXB2 Genome Positions 2200–3500, with minimum fragment length of 600 bp) of HIV-1 Subtype C with known sampling dates and geographic information were retrieved from the LANL HIV Sequence Database (https://www.hiv.lanl.gov) on 26 March 2021. Problematic sequences, as defined by LANL, were removed, and only one sequence per patient was selected before download. The sequence quality was analyzed using the Quality Control tool from the LANL site, and all genotype assignments were confirmed using the Recombinant Identification Program v3.0 (Siepel et al. 1995). Hypermutation analysis was performed using Hypermut v2.0 (Rose and Korber 2000). The two final datasets include 1,221 publicly available near-complete genome sequences of HIV-1 Subtype C (full1221) with known sampling year (1986–2019) and locations (32 countries) and 34,229 publicly available partial *pol* sequences of HIV-1 Subtype C (pol34229) with known sampling year (1986–2019) and locations (106 countries). For both full1221 and pol34229 datasets, we grouped PGRHA into six categories: male who have sex with male (SM), people who inject drugs (PI), heterosexual (SH), mother-to-baby (MB), not recorded (NR), and other (OT), as described at LANL (https://www.hiv.lanl.gov/content/sequence/HIV/data_dictionary/data_dictionary.html).

Molecular sequence analyses

The full1221 and pol34229 datasets were processed separately. Multiple sequence alignments of the two datasets (full1221 and pol34229) were obtained using MAFFT v7.427 (Katoh and Standley 2013) under an automatic algorithm and subsequently adjusted manually in BioEdit v7.2.5 (Alzohairy 2011). Next, we excluded sequences with more than 50 per cent gaps as well as duplicate sequences, defined as having the same collection date, country, PGRHA, and nucleotide sequence. This resulted in a full genome dataset of 1,210 sequences (full), and a *pol* gene dataset comprising 33,859 sequences (pol).

Subsampling was performed using SAMPI (J. L. Cherry, unpublished; https://github.com/jlcherry/SAMPI) to obtain a homogeneous collection of samples using the variables country, PGRHA, and year while maintaining a manageable dataset size lower than 1000 sequences for computational efficiency. Three subsets with repetitions for full genomes and *pol* gene were assembled: (CP) country, PGRHA, and year, (C) country and year, and (P) PGRHA and year. This resulted in the following datasets: fullCP (ten sequences per date, country, and PGRHA, $n = 626$ sequences), fullC (ten sequences per date and country, $n = 562$ sequences), fullP (ten sequences per date and PGRHA, $n = 393$ sequences), polCP (one sequence per date, country, and PGRHA, $n = 986$ sequences), polC (one sequence per date and country, $n = 698$ sequences), and polP (seven sequences per date and PGRHA, $n = 727$ sequences). We selected a higher number of sequences per date and PGRHA for polC given the smaller number states of PGRHA and with the objective of having a subsampled dataset of similar size to other datasets. All subsampling processes were performed using the following order of preferences: (1) uniformity of the number of sequences temporally, (2) completeness of collection date, (3) sequences with fewer number of gaps indivisible by three, (4) sequences with fewer ambiguous nucleotides, and finally (5) sequences longer in length. To examine the reproducibility of the datasets and analysis, we performed three independent repetitions of each subsampling strategy.

Multiple iterations of maximum-likelihood (ML) phylogenetic reconstruction using RAxML v8.2.12 (Stamatakis 2014) under a GTR + Γ4 + I nucleotide substitution model with 1,000 bootstrap replicates were performed, with removal of outlier sequences—those with incongruent sampling dates and root-to-tip genetic divergence—via the TempEst software package v1.5.3 (Rambaut et al. 2016). This resulted in full genome datasets with 626 (fullCP), 562 (fullC), and 393 (fullP) sequences and partial *pol* gene datasets with 986 (polCP), 698 (polC), and 727 (polP) sequences.

Phylogenetic reconstruction

ML phylogenetic reconstruction was performed for the original datasets and their subsampling replicates (full, fullCP, fullC, fullP, polCP, polC, and polP) using RAxML v8.2.12 (Stamatakis 2014) under the GTR + Γ4 + I nucleotide substitution model with 1,000 bootstrap replicates. Due to the large size of the pol dataset ($n = 33{,}859$ sequences), we were unable to use RAxML v8.2.12 (Stamatakis 2014) to reconstruct the ML phylogeny tree, and thus, ML phylogenetic reconstruction was performed using a more time-efficient algorithm, IQ-TREE v2.1.2, with the GTR + F + R10 substitution model (Nguyen et al. 2015). In addition, we used FigTree v1.4.4 (http://tree.bio.ed.ac.uk/software/figtree/) to visualize and annotate the phylogenetic trees with geographic location and PGRHA.

Phylogenetic tree comparison

In order to understand how subsampling affected the tree topology among the shared taxa among all trees for pol and for the full genome datasets, we compared the tree topologies using the ClusteringInfoDist metric (see below), which provides a similarity score between trees with the same sequences as tips. To this end, we first extracted subtrees from each dataset containing the intersection of the taxa present in all trees using *ybyra_pruner.py* from the 'YBYRA' package (Machado 2015). Then, we separately compared each subtree set from the near-complete genome and partial *pol* gene using the 'ClusteringInfoDist' function from the 'TreeDist' package as implemented in R (Bogdanowicz and Giaro 2012; Lin, Rajan, and Moret 2012; Smith 2020).

The 'ClusteringInfoDist' function performs better than other metrics of comparison such as Robinson-Foulds, Quarted and Path in quantifying tree similarity across different tree disturbances, such as the move length of a taxon within the tree, number of tips moved, tree spaces, and degenerate datasets (Smith 2020, 2022). The ClusteringInfoDist algorithm calculates a normalized tree similarity and distance measures based on the amount of phylogenetic or clustering information that two trees hold in common, where a lower value corresponds to trees that are topologically more similar, with a zero distance corresponding to identical trees. The normalization process on ClusteringInfoDist allows for a better comparison between the results of analyses coming from distinct datasets (i.e. results from pol comparisons vs results from full comparisons). We calculated the average and mean and performed a two-tailed distribution *t*-test assuming two samples of unequal variance (heteroscedastic) in order to identify statistical significance (<0.01) between the ClusterInfoDist values for the groups of subsampled trees.

Transmission networks

We explored the robustness of the subsampling method by generating transmission networks based on our phylogenetic reconstructions for all datasets. To do this, we employed the parsimony ancestral reconstruction method available in StrainHub v1.1.2 (de Bernardi Schneider et al. 2020). This allowed us to depict the dynamics and connectivity between each trait of interest, including 'country' and 'PGRHA'.

Through the visualization of these disease transmission networks, we hoped to gain a deeper understanding of the disease's behavior and the impact of each network node on disease spread. This included identifying whether a single node within the network acted as a super spreader or if the disease spread was evenly balanced among all nodes.

An important note is the use of the maximum parsimony model in our analysis. Although we acknowledge that this may not be optimal for transmission and phylogeography inference due to the non-uniqueness of the most parsimonious solution and the exponential growth of solutions with the number of traits, it was selected given the vast amount of data to be analyzed.

We drew attention to the influence that subsampling can have on downstream analyses by demonstrating its impact. Moreover, StrainHub does offer compatibility with phylogenetic trees arising from Bayesian inference output, albeit only for the maximum credibility tree analysis (de Bernardi Schneider et al. 2020).

StrainHub generates a transmission network based on character state changes in metadata, such as collection location, mapped on the phylogenetic tree. The nodes of this transmission network represent the relationship of the ancestral and descendant states of the pathogen sequences (e.g. changes in geography, host shifts, and among PGRHA) (de Bernardi Schneider et al. 2020). We evaluated to what extent subsampling interfered with the structure of the networks by comparing the networks indirectly through the centrality metrics of each network (Rodrigues 2019). Metadata were extracted from the sequence headers, and geographic coordinates were extracted from latlong.net. We ranked the datasets' metadata (country and PGRHA) by degree centrality and source hub ratio (SHR). Degree centrality is defined as the number of edges a trait state has within the network, meaning that the higher the degree, the more connected the state is to other states. The estimates associated with SHR, a score that ranges from 0 to 1, indicate a sink or source behavior of a particular state, respectively (hub has a SHR = 0.5), as implemented in StrainHub (de Bernardi Schneider et al. 2020). We also calculated the Pearson product-moment correlation coefficient for all pairs of trait states for all original and subsampled full and pol datasets to understand how subsampling affects the overall transmission network structure.

Results

Subsampling

In the full dataset, genome sequences collected in South Africa (ZA; 49.7 per cent; 601/1,210) and Zambia (ZM; 8.1 per cent; 219/1,210) and from NR (68.6 per cent; 830/1,210) and SH (26.9 per cent; 325/1,210) PGRHA groups are over-represented compared to the numbers for other countries and PGRHA groups (Fig. 1).

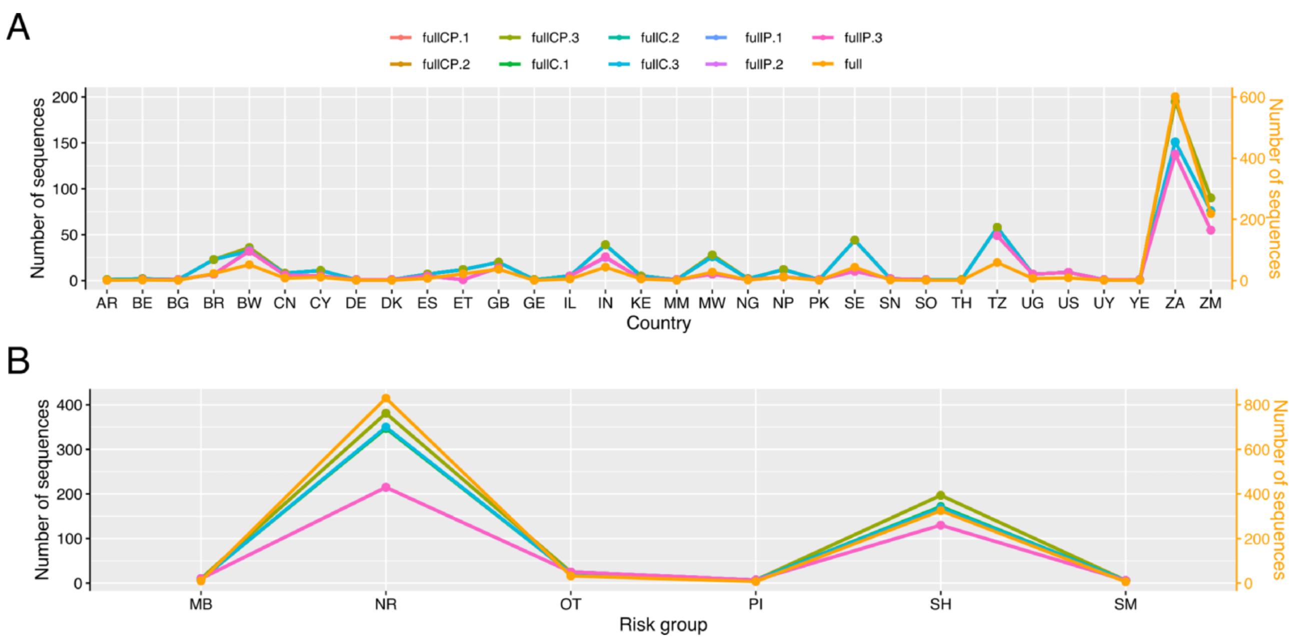

Figure 1. Sampling distributions of metadata traits for the full and subsampled datasets of HIV-1 Subtype C. (A) Country distribution for the full and subsampled datasets. The distribution of the original dataset shows a disproportionate amount of samples sampled from BR, BW, IN, MW, SE, TZ, ZA, and ZM. (B) PGRHA distribution for the full and subsampled datasets. The distribution of the data shows a large amount of missing data (labeled NR) and higher amount of SH in comparison to other PGRHA. The number of sequences for the full dataset is labeled on the right y-axis.

Our subsampling strategy resulted in datasets with the following reduced sequence counts (average between three repetitions of subsampled datasets) for ZA and ZM: 31.2 per cent (195/626) and 14.4 per cent (90/626) in the fullCP datasets, 26.9 per cent (151/562) 192 and 13.5 per cent (76/562) in the fullC datasets, and 34.9 per cent (137.3/393) and 14.0 per cent (55/393) in the fullP datasets. Similarly, the subsampling by PGRHA results in datasets with the following NR and SH genome sequence counts: 60.9 per cent (381/626) and 31.5 per cent (197/626) in the fullCP datasets, 62.1 per cent (349/562) and 30.2 per cent (170/562) in the fullC datasets, and 54.7 per cent (215/393) and 33.1 per cent (130/393) in the fullP datasets, respectively.

In the pol datasets (Fig. 2), the partial *pol* gene sequences collected in ZA (51.1 per cent [17,312/33,859]) and India (IN; 8.6 per cent [2,922/33,859]) and from the NR PGRHA (90.4 per cent [30,615/33,859]) are over-represented. After subsampling, the average between three repetitions of subsampled datasets for the partial *pol* gene sequences obtained in ZA and IN account for 6.8 per cent (67/986) and 5.7 per cent (56/986) in the polCP datasets, 4.0 per cent (28/698) and 3.3 per cent (23/698) in the polC datasets, and 24.4 per cent (177.3/727) and 13.9 per cent (101.3/727) in the polP datasets, respectively. Likewise, the subsampled partial *pol* gene sequences collected from the NR PGRHA now account for 62.7 per cent (618/986) from the polCP datasets, 75.2 per cent (524.7/698) from the polC datasets, and 26.5 per cent (193/727) from the polP datasets.

The subsampling of the polP datasets yield a different country composition in comparison to other datasets (46 or 47 of 105 countries), given the large amount of data and the subsampling method that did not include country as a subsampling trait. For downstream analyses, we compared only the intersection of data between each dataset, i.e. pol (105 countries) vs polCP (105 countries); polCP (105 countries) vs polC (105 countries); pol, polCP, or polC (46 or 47 of 105 countries) vs polP (46 or 47 countries).

Tree comparisons

For the full datasets (Fig. 3A), the topologies of the subtrees subsampled by country (fullC; average = 26.24) are the closest in similarity to that of the original dataset (full). Nevertheless, fullCP datasets (average = 28.06) have very close values to fullC, with fullP (average = 33.94) being the most distant datasets to full. For the *pol* gene dataset (Fig. 3B), the topologies of all subsampled subtrees are mostly equidistant to the original pol dataset (polCP average = 34.92; polC average = 34.87; polP average = 36.02). We estimated that the datasets were not significantly different across full and pol subsamplings, with the exception of fullCP vs fullP (*P*-value = 0.009) and fullC vs fullP (*P*-value = 0.0005). Nevertheless, we observe overall similar values across subsamplings for both full and pol datasets (small variance across subsamplings; fullCP = 2.47; fullC = 0.72; fullP = 0.90; polCP = 1.10; polC = 2.49; and polP = 3.57).

Transmission networks

We generated transmission networks for all full and pol datasets. We observed that there was limited variation of the correlation of the degree centrality metric between country and PGRHA with the original dataset across repetitions of the same subsampling strategy or across the three subsampling strategies for full and pol datasets, with the exception of polP (Fig. 4 and Supplementary Figs. S1 and S2). This means that the degree of connectivity of each country or PGRHA node in the overall transmission network is maintained irrespective of the subsampling strategy. Despite the overall maintenance of the country and PGRHA node importance, their behaviors (i.e. sink/hub or source of disease), assessed using the SHR estimate, varies with the subsampling strategy employed. We observed that the correlation of the SHR with the original dataset for the country trait is highest for fullCP and fullC and lowest for fullP. This pattern is also observed for the pol

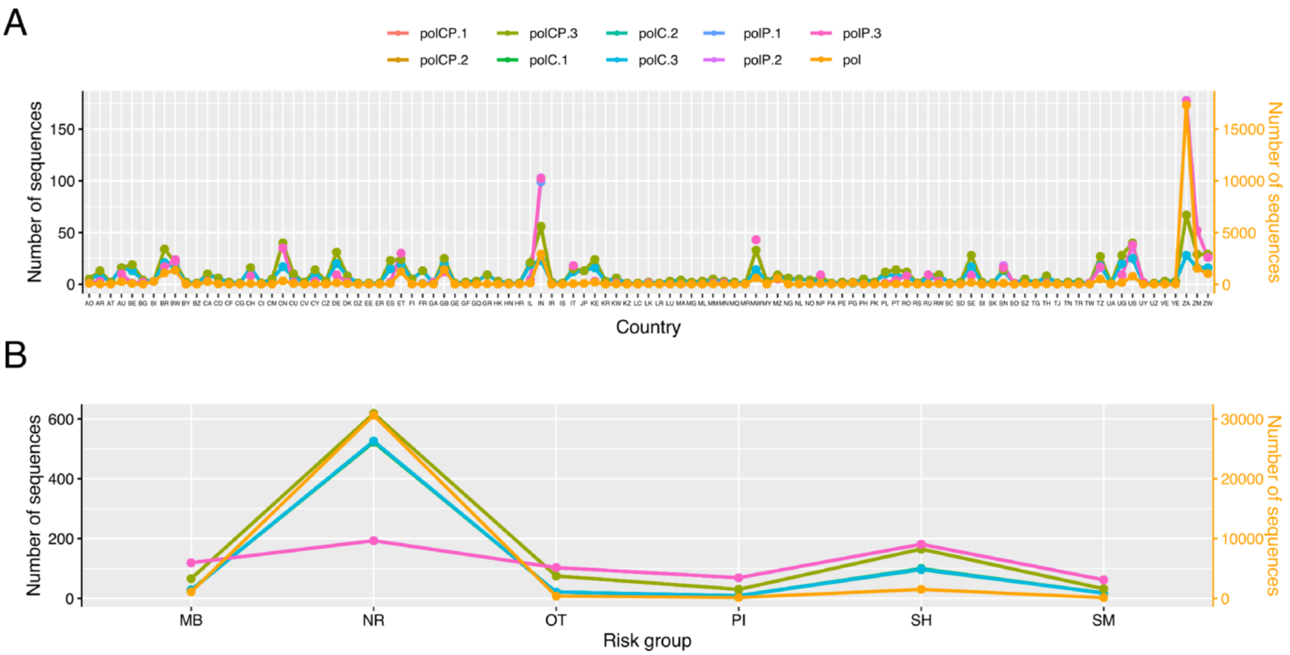

Figure 2. Sampling distributions of metadata traits for the pol and subsampled datasets of HIV-1 Subtype C. (A) Country distribution for the pol and subsampled datasets. The distribution of the original dataset shows a larger amount of samples sampled from BR, BW, ET, GB, IN, MW, MZ, TZ, US, ZM, and ZW, with a disproportionate amount of samples from ZA. (B) PGRHA distribution for the pol and subsampled datasets. The distribution of the data shows a large amount of missing data (labeled NR) and slightly higher amount of SH in comparison to other PGRHA. The number of sequences for the pol dataset is labelled on the right y-axis.

subsamplings, even though the overall SHR correlations with the original dataset are lower than those for the full dataset.

Interestingly, the correlation of SHR for the PGRHA trait is higher for both fullCP and fullC than for fullP. This might suggest that for the full dataset, even when subsampling is done solely using country, we obtain a distribution of samples such that they are similar to the overall PGRHA trait structure of the original full dataset (Fig. 5). However, the limited information for PGRHA yields results that do not represent the overall PGRHA trait structure in the original dataset. This behavior is not observed in the pol subsampling, where polCP and polP have the highest SHR correlation with the original dataset, agreeable with the fact that both strategies include information for PGRHA, whereas polC yields the lowest correlation with the original PGRHA transmission network.

Looking within the transmission networks geographically, in the full genome dataset, the countries with the highest degree centrality are (in order of high- to low-degree centrality) ZA, Sweden (SE), ZM, Tanzania (TZ), and the United Kingdom (GB) (Fig. 5A). However, GB does not rank among the top five countries in the fullCP and fullC datasets, where it is replaced by Botswana (BW). Moreover, SE does not rank among the top five countries in the fullP dataset, but it does include GB and BW. In the pol datasets, given the larger number of countries, we elected to display the top nine countries for each dataset. The countries with the highest degree centrality are (in order of high- to low-degree centrality) ZA, ZM, GB, Ethiopia (ET), Zimbabwe (ZW), IN, United States (US), TZ, and Burundi (BI). The polCP datasets best represent the pol dataset's top ten, with only two countries being replaced (GB and BI by SE and Malawi [MW]), with these two countries being replaced as Numbers 8 and 9. The polC dataset replaced three of the top nine countries from the original dataset (IN, TZ, BI with Australia [AU], Germany, and Israel), and polP replaced three countries as well (GB, TZ, and BI with AU, BW, and MW) (Fig. 5B).

We have also summarized the spatial transmission dynamics among the original full and pol datasets and their respective subsampled dataset. For the full datasets (Supplementary Fig. S3), even though there was some variation across replicates, we observed an overall similar pattern characterized by viral dissemination from Africa to Europe and vice versa, as well as from Africa do the US and South Asia. The pol dataset (Supplementary Fig. S4) depicts more complex spatio-temporal dynamics that included the viral movements described for the full datasets, plus introductions from North America to Asia, Europe to South America, and Africa to Oceania. The patterns observed for polP were sparser, likely due to the reduced number of countries included for this dataset compared to other pol datasets.

The transmission network among PGRHA shares a similar result across all full datasets, with SH and NR having the highest mean degree centrality, therefore contributing to the highest number of connections within the network, and OT, PI, MB, and SM contributing less (Fig. 5C). For the pol datasets, we see similar results with NR and SH having the highest for pol, polCP, and polC and a more elevated degree centrality for OT, PI, MB, and SM, with polP having a high degree centrality on all PGRHA with an almost uniform distribution (Fig. 5D).

The summarized PGRHA transmission dynamics among the original full and pol datasets and their respective subsampled dataset highlight the results described in the paragraph above. In both full and pol datasets (Supplementary Figs. S5 and S6), we reconstructed similar PGRHA dynamics with NR acting as the main source among PGRHA followed by the SH group, irrespective of subsampling. In the full datasets, we estimated the large majority of transmission dynamics occurring from NR to SH. In the pol datasets, the viral transmissions were mostly from NR to SH, but we also estimated a substantial proportion of viral seeding from NR to MB. The patterns observed for polP were unlike those observed for any other dataset, with no particular PGRHA standing

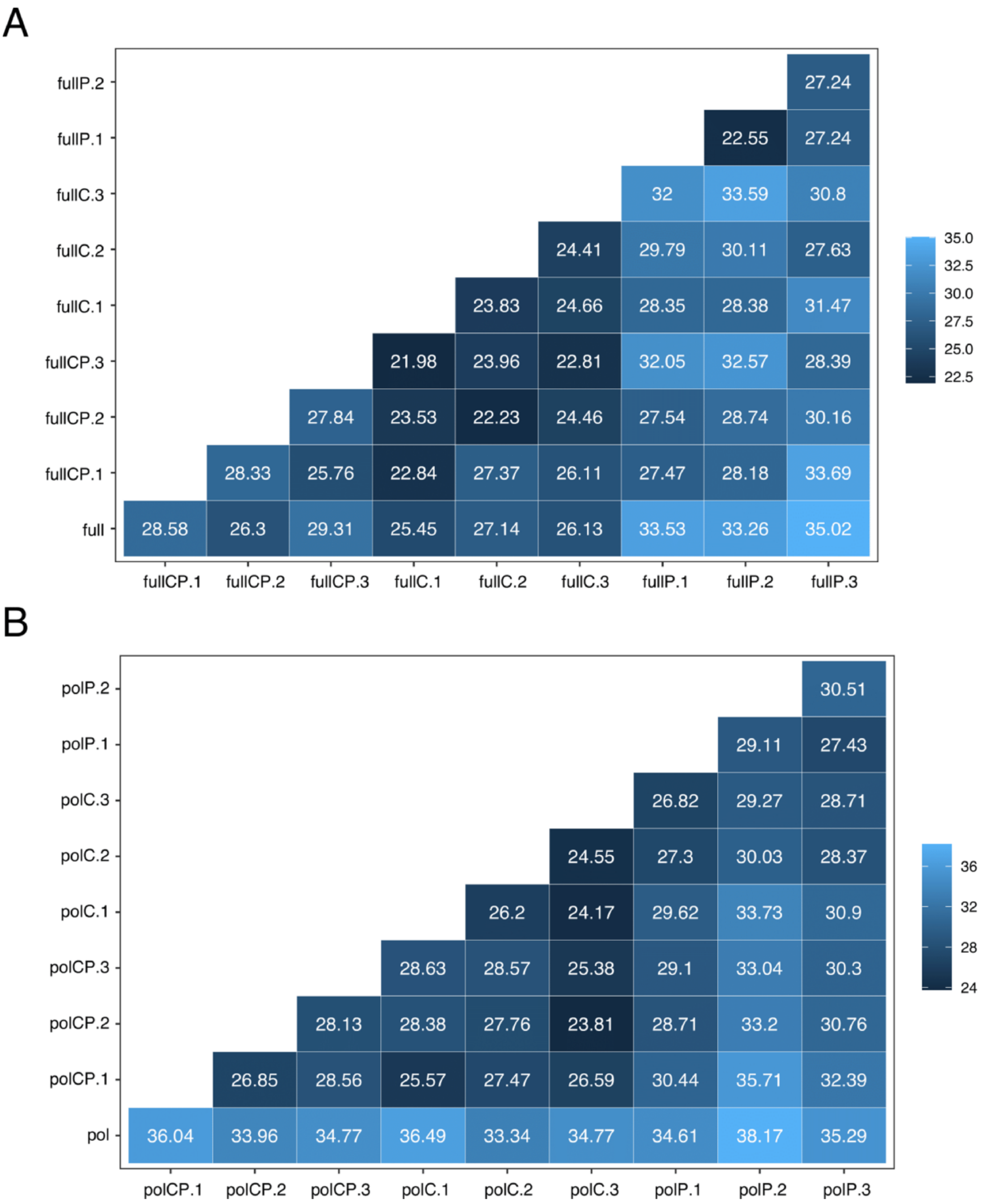

Figure 3. Cluster info distance comparison of the phylogenetic topologies of the full (A) and pol (B) with their respective subsampled dataset subtrees. Zero cluster info distance equals identical trees. The topologies of the full and pol subsampled dataset subtrees are overall similar to that of their respective original datasets.

out in terms of viral source or sink. This is again likely due to the reduced number of countries included in this dataset compared to other pol datasets.

Discussion

In this study, we investigated HIV-1 Subtype C evolutionary and spatio-temporal dynamics while subsampling the genetic data to decrease the sequence counts from oversampled traits. Subsampling was performed in order to mitigate biases introduced during the sampling of PGRHA, as well as of countries through time. To this end, we compiled comprehensive sequence datasets of full genomes and the *pol* gene region and revealed that both datasets contained inherent biases irrespective of the trait studied, as observed through the heterogeneous distribution of the datasets (Figs. 1 and 2). We could not compare the subsampled dataset distributions to the real population case estimates, which are impossible to obtain. The available epidemiological curves are not desegregated by HIV-1 type and are biased by time and spatial surveillance coverage and effectiveness (https://cdn.who.int/media/docs/default-source/hq-hiv-hepatitis-and-stis-library/key-facts-hiv-2021-26july2022.pdf? sfvrsn=8f4e7c93_5). For these reasons, we assumed for this study that an unbiased sampling should follow a near-uniform distribution.

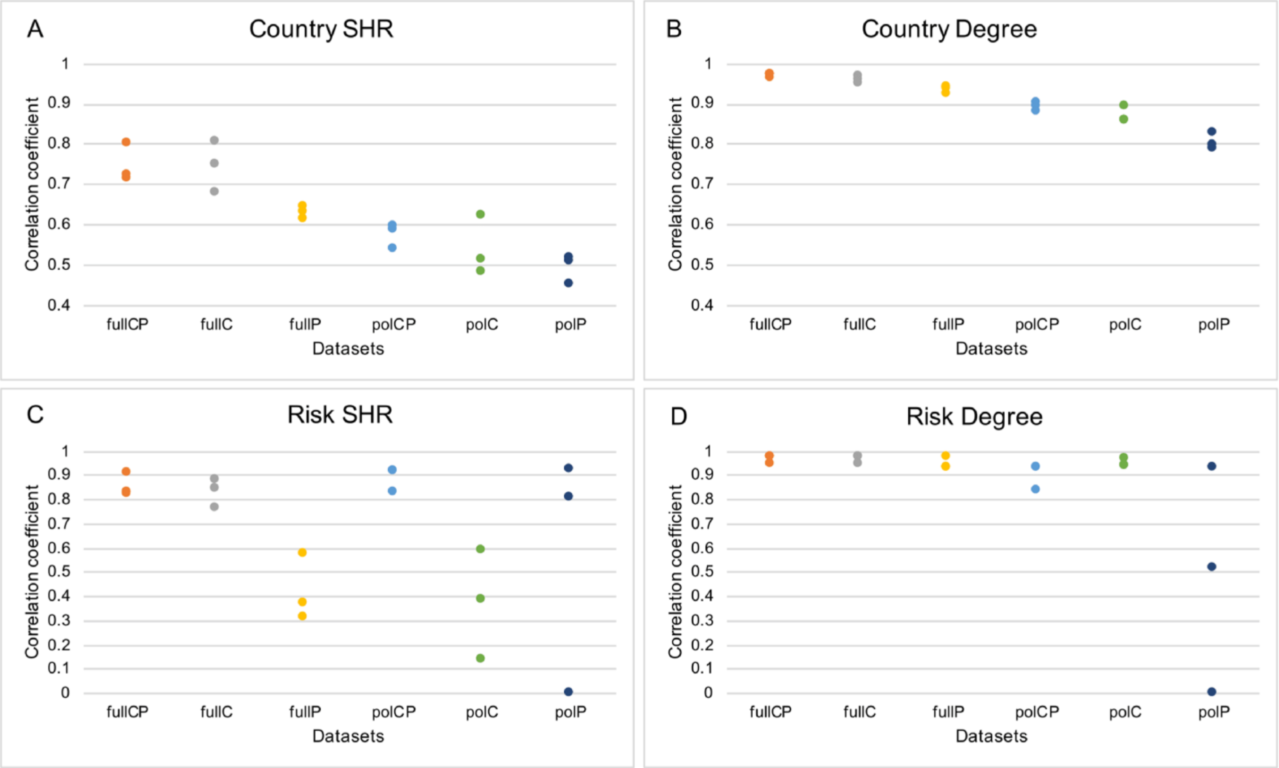

Figure 4. Correlation of SHR and degree centrality metrics as a proxy for transmission network structures of HIV-1 Subtype C for the full and pol with respective subsampled datasets by date, country, and PGRHA. The correlation of the transmission network is higher (thus, similar structures) as the correlation coefficient approximates to 1. (A) Estimates of similarity of the spatial transmission network structure for all subsampled datasets based on SHR metric; (B) estimates of similarity of the spatial transmission network structure of HIV-1 Subtype C for all subsampled datasets based on the degree centrality metric; (C) estimates of similarity of the PGRHA transmission network structure of HIV-1 Subtype C for all subsampled datasets based on SHR metric; (D) estimates of similarity of the PGRHA transmission network structure of HIV-1 Subtype C for all subsampled datasets based on the degree centrality metric. The degree of connectivity of each country or PGRHA node in the overall transmission network is generally maintained irrespective of the subsampling strategy; however, there is a discrepancy of the country and PGRHA node behaviors as indicated by the varying SHR per subsampling strategy.

Sampling strategies and procedures like subsampling methods can help address varying trait representativeness in the metadata associated with genomic datasets. There are many studies that apply subsampling methods in an attempt to correct for bias on sampling date and location. For instance, studies targeting the early spread and epidemic ignition of HIV-1 in humans (Faria et al. 2014), studies investigating the spatial history of HIV-1 Subtype B in the US (Hong et al. 2020), and studies exploring the rapid epidemic expansion of the SARS-CoV-2 Omicron variant in southern Africa (Viana et al. 2022). Nevertheless, these studies do not comprehensively examine the effects of varying representativeness of traits and its implications on phylodynamic reconstruction. Here, we have observed that a more comprehensive subsampling strategy that includes as many traits as possible (date, location, and PGRHA) yields the best result in retaining the original dataset properties, as demonstrated by the high similarities of the transmission networks between the HIV-1 Subtype C full and pol and the fullCP and polCP datasets, respectively (Fig. 4). Furthermore, studies that take into account sampling bias are often limited to a single replicate of a particular subsampling method (Faria et al. 2014; Nasir et al. 2022; Okoh et al. 2022; Viana et al. 2022). We have demonstrated that this likely does not have harmful implications for the interpretation of the results as there is little variation of the overall tree topology across subsampling replicates (Fig. 3), as well as in the ancestral trait reconstruction (Fig. 4).

Comparing the tree topologies of the original full and pol datasets with their respective subsampled datasets allows uncovering which subsampling strategies best represent the original structure and whether that structure is punctuated by a particular trait. Our analyses indicate that both full and pol along with their subsampled datasets present comparable variability in tree topology among the subsets both in terms of country and PGRHA. However, there are inherent limitations in both datasets as observed by the majority of sequences labeled as NR (PGRHA) in all datasets.

Our analyses indicate that there is a slightly stronger signal in the full dataset for location as shown by the smaller distances across the original dataset and those subsampled using the location trait, whereas the pol dataset seems to hold the same level of information for both country and PGRHA traits, indicating a more balanced dataset. The comparable ClusterInfoDist metric across datasets and respective subsample repetitions suggest that irrespective of the subsampling strategy the overall structure of the original topology is maintained given the similar values across all comparisons (Fig. 3). We can assume that the reason behind the original full and pol datasets having a slightly lower degree of similarity, as measured through ClusterInfoDist, to the datasets subsampled by PGRHA (fullP and polP) might be that the datasets are mainly driven by location, regardless of the skewed sampling distribution in certain geographical locations, such as IN and ZA (Figs. 1 and 2), which may be in part due to the predominance of the HIV-1 Subtype C in these regions (Gartner et al. 2020). In

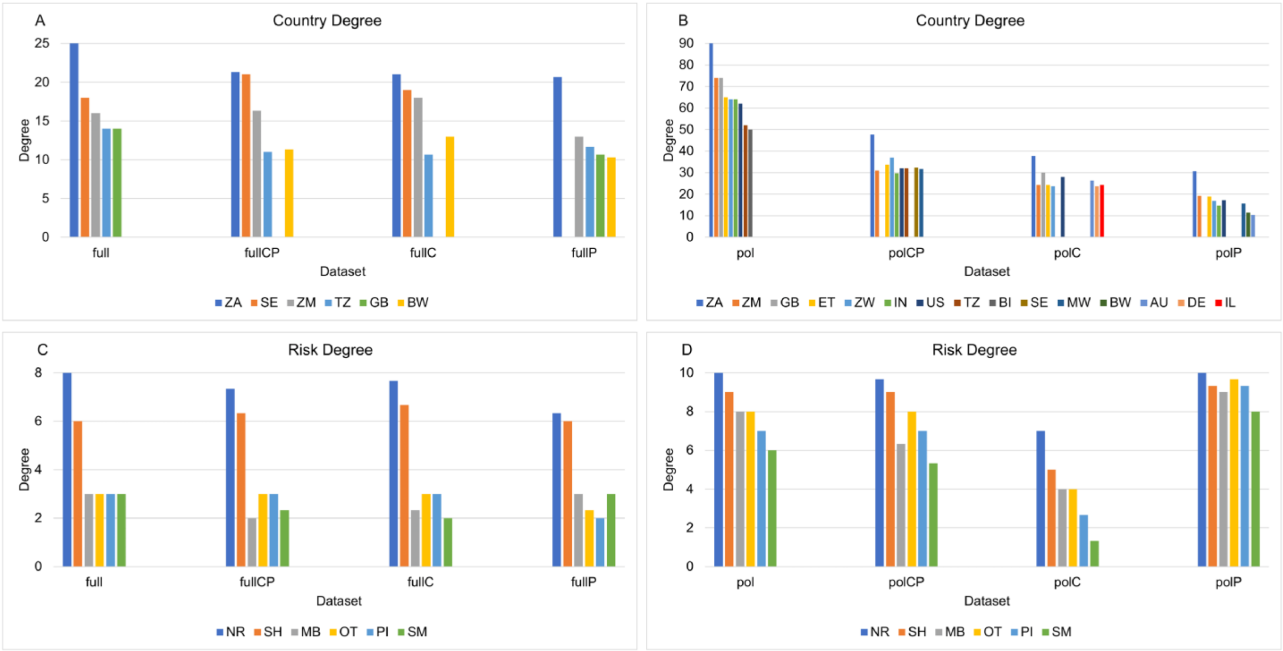

Figure 5. Distribution of the average degree centrality among all subsamples for the top clusters in the pol and full subsets. (A) Degree centrality for top five countries (country trait) for full subsets; (B) degree centrality for top nine countries (country trait) for pol subsets; (C) degree centrality for all PGRHA for full subsets; (D) degree centrality for all PGRHA trait for pol subsets. Including country and PGRHA as subsampling traits yields the most consistent results for both traits' transmission networks. The distribution of degree centrality among the nodes of the networks evaluated shows that the datasets subsampled by PGRHA and country or solely by country result in patterns similar to those for the original pol and full datasets. The top countries in terms of degree centrality are mostly conserved across the full datasets, with wider variance observed in the pol datasets likely due to the larger number of locations for the country trait in these datasets.

this scenario, subsampling solely by PGRHA would have a stronger effect on the tree topology, possibly due to the different behavior of the PGRHA within each country. In addition to that, subsampling by PGRHA produces a distinct outcome from subsampling by country or by both country and PGRHA, most likely owing to the large presence of missing data labeled as NR.

Consequently, the limited information for PGRHA in both full and pol datasets produces inconsistent transmission networks. Overall, these results indicate that the full subsampled datasets produce transmission networks that have higher correlations to the original dataset for both country and PGRHA and yield more comprehensive evolutionary histories. This result demonstrates once more how multigene datasets provide higher accuracy in phylogenetic analysis despite lower dataset sizes (Rokas and Carroll 2005).

Geographically, the top countries for the spread of HIV-1 Subtype C, as measured by degree centrality of the nodes within the transmission network on both full and pol datasets, are in line with the previous studies (Gartner et al. 2020). The most prominent PGRHA in both original full and pol networks is SH, as seen by the larger sampling of heterosexual individuals in these datasets (Figs. 1 and 2), which likely represents the current state of the HIV-1 Subtype C epidemic at global scale (Brown and Peerapatanapokin 2019). In pol, we also observed MB as a major PGRHA acting as a transmission source. NR seems to be largely connected to SH in both datasets, indicating that the vast majority of non-reported PGRHA may belong to SH (Supplementary Figs. S5 and S6). The behavior of PGRHA is expected to be dependent on regional norms (Rhodes and Simic 2005; Ordonez and Marconi 2012; Wyatt et al. 2012); thus, the lack of coverage of locations in polP may be the reason why these results diverge considerably from those of other datasets. Additionally, the countries excluded from polP might be those that have a stronger signal for the dynamics observed in other datasets, namely transmission events from NR to SH and from NR to MB. Therefore, as expected, including both country and PGRHA as subsampling traits yields results for both country and PGRHA transmission networks more consistent with the patterns observed for the original pol dataset.

Our attempts to mitigate bias by employing multiple subsampling strategies are not without limitations. For instance, since they rely on the metadata available for the genetic data, we might not address biases created by unknown factors. In this HIV-1 Subtype C study, most of the metadata regarding PGRHA is NR, and NR accounted for 68.6 per cent and 90.4 per cent of sequences in the full and pol datasets, respectively. Besides, some of the metadata may be mislabeled, such as reports of SM due to HIV/acquired immunodeficiency syndrome-related stigmatization and discrimination, as reported in the previous studies (Zai et al. 2020). We hypothesize that the metadata associated with risk groups might be artificially biased toward 'NR'. Therefore, this may have affected the accuracy of the phylodynamic reconstructions and ancestral trait reconstruction. We recommend that the researchers complete this metadata field as much as possible when submitting sequence information. Both sequence data and associated metadata are critical to gain more detailed insights into the evolutionary and spatio-temporal patterns of HIV-1 Subtype C and other pathogens. Therefore, more reporting and sharing of data in an open and real-time fashion is needed for an effective public health response.

Comparing the original and subsampled datasets to epidemiological data could be a solution to the present issue in sampling. However, this type of data also often suffers from biases, including those created by under-sampling in low- and middle-income countries or are not documented particularly for the early dynamics of the epidemics (Dudas et al. 2021; Zeller et al. 2021). We made an effort to obtain retrospective epidemiological data documenting the number of patients infected with HIV-1 Subtype C, but the epidemiological data does not report nor is sorted by subtype, which further complicated this endeavor. Additionally, HIV/AIDS being a disease associated with severe stigma could lead to case reports that do not accurately represent the overall circulation patterns (Chen et al. 2017; Zai et al. 2020).

When comparing SAMPI with other subsampling methods, we acknowledge that most subsampling tools aim to produce datasets that have fewer biases of their traits. Therefore, we anticipate that different subsampling tools would not significantly affect the results of our study. Furthermore, popular phylodynamic reconstruction methods rely on discrete trait analysis that are quite sensitive to sampling biases in a similar fashion, where oversampled traits would likely be inferred as sources and undersampled traits would be inferred as sinks in the transmission network. Thus, there is a need to employ careful subsampling strategies before venturing in these types of phylodynamic reconstructions. Alternatively, structured coalescent models allow reconstruction of transmission dynamics that is almost insensitive to sampling bias. However, these methods are excessively computationally expensive and therefore are limited to research questions that require smaller dataset sizes De Maio et al. (2015).

Even though we here offer a detailed approach to reduce inherent biases and further optimize ancestral trait reconstruction by subsampling large datasets, there are other procedures to account for issues with sampling, including careful research and surveillance design, simulations, and weighted methods based on metrics such as prevalence (He et al. 2012; Leon, Jauffret-Roustide and Le Strat, 2015; Clark et al. 2018; Gunduz and Aydin 2021; McArdle et al. 2021; Yang 2022). New methodological developments enable phylogeographic inferences that are not affected by sampling bias (De Maio et al. 2015) but currently do not scale well with the increasing number of sequences and locations, hence make analysis of large data sets computationally challenging.

The analysis pipeline proposed is fast and publicly available at GitHub. SAMPI (J. L. Cherry, unpublished; https://github.com/jlcherry/SAMPI) is a pathogen-agnostic subsampling tool that can be freely used to study other infectious diseases in a computationally efficient manner (Nasir et al. 2022; Trovão et al. 2022; Trovao et al. 2023). Furthermore, this study sheds light on how to analyze and subsample large public datasets and further investigate the impact of subsampling, which can also highlight the importance of specific traits that are highly correlated with transmission networks. The potential challenges and limitations are mainly associated with the quality of the dataset in terms of available metadata, the extent of the biases, as well as the sequencing quality.

Conclusion

In summary, we address the challenges of working with large datasets and sampling bias using a subsampling approach based on date, country, and PGRHA. We evaluate how this approach can mitigate sampling bias while maintaining the properties of the original datasets and computationally optimize data analyses based on the available metadata. We also highlight the importance of rigorously recording metadata in addition to the genetic sequences. This study systematically evaluates strategies to optimize ancestral trait reconstruction in HIV-1 Subtype C and will be helpful for future phylodynamic analysis of this virus, as well as serve as a reference to the study of other pathogens.

Supplementary data

Supplementary data is available at *Virus Evolution* online.

Data availability

The datasets used in this study, which are sourced from public databases as detailed in the Methods section of the article, will be made available upon reasonable request to the corresponding author.

Acknowledgements

A.d.B.S. acknowledges support from the California Department of Public Health (Contract No. 20-11088), and the National Institutes of Health National Institute of Allergy and Infectious Diseases (Grant Number AI135992) to A.d.B.S. and J.O.W.. G.B. acknowledges support from the Internal Funds KU Leuven (Grant No. C14/18/094) and the Research Foundation—Flanders ('Fonds voor Wetenschappelijk Onderzoek—Vlaanderen', G0E1420N, G098321N).

Author contributions

X.L., N.S.T., and A.d.B.S. conceived, designed the study, and drafted the manuscript. X.L., N.S.T., and A.d.B.S. analyzed the data. X.L., N.S.T., J.O.W., G.B., and A.d.B.S. interpreted the data and provided critical comments. All authors reviewed and approved the final manuscript.

Conflict of interest The authors declare no competing interests.

References

Alzohairy, A. (2011) 'BioEdit: An Important Software for Molecular Biology', *GERF Bulletin of Biosciences*, 2: 60–1.

Arias, A. et al. (2016) 'Rapid Outbreak Sequencing of Ebola Virus in Sierra Leone Identifies Transmission Chains Linked to Sporadic Cases', *Virus Evolution*, 2: vew016.

Bedford, T. et al. (2015) 'Global Circulation Patterns of Seasonal Influenza Viruses Vary with Antigenic Drift', *Nature*, 523: 217–20.

Bogdanowicz, D., and Giaro, K. (2012) 'Matching Split Distance for Unrooted Binary Phylogenetic Trees', *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 9: 150–60.

Brown, T., and Peerapatanapokin, W. (2019) 'Evolving HIV Epidemics: The Urgent Need to Refocus on Populations with Risk', *Current Opinion in HIV and AIDS*, 14: 337–53.

Chakraborty, C. et al. (2021) 'Evolution, Mode of Transmission, and Mutational Landscape of Newly Emerging SARS-CoV-2 Variants', *mBio*, 12: 10–1128.

Chen, X. et al. (2017) 'First Description of Two New HIV-1 Recombinant Forms CRF82_cpx and CRF83_cpx among Drug Users in Northern Myanmar', *Virulence*, 8: 497–503.

Clark, S. J. et al. (2018) 'Hyak Mortality Monitoring System: Innovative Sampling and Estimation Methods—Proof of Concept by Simulation', *Global Health, Epidemiology and Genomics*, 3: 1–14.

de Bernardi Schneider, A. et al. (2020) 'StrainHub: A Phylogenetic Tool to Construct Pathogen Transmission Networks', *Bioinformatics*, 36: 945–7.

De Maio, N. et al. (2015) 'New Routes to Phylogeography: A Bayesian Structured Coalescent Approximation', *PLoS Genetics*, 11: e1005421.

Dudas, G. et al. (2021) 'Emergence and Spread of SARS-CoV-2 Lineage B.1.620 with Variant of Concern-Like Mutations and Deletions', *Nature Communications*, 12: 5769.

Elbe, S., and Buckland-Merrett, G. (2017) 'Data, Disease and Diplomacy: GISAID's Innovative Contribution to Global Health', *Global Challenges*, 1: 33–46.

Elliott, I. et al. (2020) 'Oxford Nanopore MinION Sequencing Enables Rapid Whole Genome Assembly of *Rickettsia typhi* in a Resource-Limited Setting', *American Journal of Tropical Medicine & Hygiene*, 102: 408–14.

Faria, N. R. et al. (2014) 'HIV Epidemiology. The Early Spread and Epidemic Ignition of HIV-1 in Human Populations', *Science*, 346: 56–61.

Furuse, Y. (2021) 'Genomic Sequencing Effort for SARS-CoV-2 by Country during the Pandemic', *International Journal of Infectious Diseases: IJID: Official Publication of the International Society for Infectious Diseases*, 103: 305–7.

Gartner, M. J. et al. (2020) 'Understanding the Mechanisms Driving the Spread of Subtype C HIV-1', *EBioMedicine*, 53: 102682.

Gunduz, N., and Aydin, C. (2021) 'Optimal Bandwidth Estimators of Kernel Density Functionals for Contaminated Data', *Journal of Applied Statistics*, 48: 2239–58.

Hay, A. J., and McCauley, J. W. (2018) 'The WHO Global Influenza Surveillance and Response System (GISRS)—A Future Perspective', *Influenza and Other Respiratory Viruses*, 12: 551–7.

He, X. et al. (2012) 'A Comprehensive Mapping of HIV-1 Genotypes in Various Risk Groups and Regions across China Based on a Nationwide Molecular Epidemiologic Survey', *PLoS One*, 7: e47289.

Hong, S. L. et al. (2020) 'In Search of Covariates of HIV-1 Subtype B Spread in the United States—A Cautionary Tale of Large-Scale Bayesian Phylogeography', *Viruses*, 12: 182.

Junqueira, D. M., and Almeida, S. E. (2016) 'HIV-1 Subtype B: Traces of a Pandemic', *Virology*, 495: 173–84.

Kalkauskas, A. et al. (2021) 'Sampling Bias and Model Choice in Continuous Phylogeography: Getting Lost on a Random Walk', *PLoS Computational Biology*, 17: e1008561.

Katoh, K., and Standley, D. M. (2013) 'MAFFT Multiple Sequence Alignment Software Version 7: Improvements in Performance and Usability', *Molecular Biology and Evolution*, 30: 772–80.

Layan, M. et al. (2023) 'Impact and Mitigation of Sampling Bias to Determine Viral Spread: Evaluating Discrete Phylogeography through CTMC Modeling and Structured Coalescent Model Approximations', *Virus Evolution*, 9: vead010.

Leon, L., Jauffret-Roustide, M., and Le Strat, Y. (2015) 'Design-Based Inference in Time-Location Sampling', *Biostatistics*, 16: 565–79.

Lin, Y., Rajan, V., and Moret, B. M. (2012) 'A Metric for Phylogenetic Trees Based on Matching', *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 9: 1014–22.

Machado, D. J. (2015) 'YBYRA Facilitates Comparison of Large Phylogenetic Trees', *BMC Bioinformatics*, 16: 1–4.

McArdle, C. E. et al. (2021) 'Findings from the Hispanic Community Health Study/Study of Latinos on the Importance of Sociocultural Environmental Interactors: Polygenic Risk Score-by-Immigration and Dietary Interactions', *Front Genetics*, 12: 1–15.

McBroome, J. et al. (2022) 'Identifying SARS-CoV-2 Regional Introductions and Transmission Clusters in Real Time', *Virus Evolution*, 8: veac048.

Menardo, F. et al. (2018) 'Treemmer: A Tool to Reduce Large Phylogenetic Datasets with Minimal Loss of Diversity', *BMC Bioinformatics*, 19: 1–8.

Minh, B. Q., Klaere, S., and von Haeseler, A. (2009) 'Taxon Selection under Split Diversity', *Systemic Biology*, 58: 586–94.

Nasir, A. et al. (2022) 'Evolutionary History and Introduction of SARS-CoV-2 Alpha VOC/B.1.1.7 in Pakistan through International Travelers', *Virus Evolution*, 8: Veac020.

Nguyen, L. T. et al. (2015) 'IQ-TREE: A Fast and Effective Stochastic Algorithm for Estimating Maximum-Likelihood Phylogenies', *Molecular Biology and Evolution*, 32: 268–74.

Novitsky, V. et al. (2010) 'HIV-1 Subtype C Phylodynamics in the Global Epidemic', *Viruses*, 2: 33–54.

Okoh, O. S. et al. (2022) 'Epidemiology and Genetic Diversity of SARS-CoV-2 Lineages Circulating in Africa', *iScience*, 25: 103880.

Ordonez, C. E., and Marconi, V. C. (2012) 'Understanding HIV Risk Behavior from a Sociocultural Perspective', *Journal of AIDS & Clinical Research*, 3: e108.

Popejoy, A. B., and Fullerton, S. M. (2016) 'Genomics Is Failing on Diversity', *Nature*, 538: 161–4.

Rambaut, A. et al. (2016) 'Exploring the Temporal Structure of Heterochronous Sequences Using TempEst (Formerly Path-O-Gen)', *Virus Evolution*, 2: vew007.

Rhodes, T., and Simic, M. (2005) 'Transition and the HIV Risk Environment', *British Medical Journal*, 331: 220–3.

Rodrigues, F. A. (2019) *In A Mathematical Modeling Approach from Nonlinear Dynamics to Complex Systems*. Macau, E. E. N., (ed.), pp. 177–96. Cham, Switzerland: Springer International Publishing.

Rokas, A., and Carroll, S. B. (2005) 'More Genes or More Taxa? The Relative Contribution of Gene Number and Taxon Number to Phylogenetic Accuracy', *Molecular Biology and Evolution*, 22: 1337–44.

Roncoroni, M. et al. (2021) 'A SARS-CoV-2 Sequence Submission Tool for the European Nucleotide Archive', *Bioinformatics*, 37: 3983–5.

Rose, P. P., and Korber, B. T. (2000) 'Detecting Hypermutations in Viral Sequences with an Emphasis on G→A Hypermutation', *Bioinformatics*, 16: 400–1.

Sayers, E. W. et al. (2022) 'GenBank', *Nucleic Acids Research*, 50: D161–4.

Sheng, J. et al. (2021) 'COVID-19 Pandemic in the New Era of Big Data Analytics: Methodological Innovations and Future Research Directions', *British Journal of Management*, 32: 1164–83.

Siepel, A. C. et al. (1995) 'A Computer Program Designed to Screen Rapidly for HIV Type 1 Intersubtype Recombinant Sequences', *AIDS Research and Human Retroviruses*, 11: 1413–6.

Smith, M. R. (2020) 'Information Theoretic Generalized Robinson-Foulds Metrics for Comparing Phylogenetic Trees', *Bioinformatics*, 36: 5007–13.

Smith, M. R. (2022) 'Robust Analysis of Phylogenetic Tree Space', *Systematic Biology*, 71: 1255–70.

Stamatakis, A. (2014) 'RAxML Version 8: A Tool for Phylogenetic Analysis and Post-Analysis of Large Phylogenies', *Bioinformatics*, 30: 1312–3.

Trovao, N. S. et al. (2023) 'Evolutionary and Spatiotemporal Analyses Reveal Multiple Introductions and Cryptic Transmission of SARS-CoV-2 VOC/VOI in Malta', *Microbiology Spectrum*, e0153923.

Trovão, N. S. et al. (2022) 'Evolution of Influenza A Virus Hemagglutinin H1 and H3 across Host Species', *bioRxiv*.

Turakhia, Y. et al. (2020) 'Stability of SARS-CoV-2 Phylogenies', *PLoS Genetics*, 16: e1009175.

Turakhia, Y. et al. (2021) 'Ultrafast Sample Placement on Existing tRees (Usher) Enables Real-time Phylogenetics for the SARS-CoV-2 Pandemic', *Nature Genetics*, 53: 809–16.

Vakulenko, Y., Deviatkin, A., and Lukashev, A. (2019) 'The Effect of Sample Bias and Experimental Artefacts on the Statistical Phylogenetic Analysis of Picornaviruses', *Viruses*, 11: 1032.

Viana, R. et al. (2022) 'Rapid Epidemic Expansion of the SARS-CoV-2 Omicron Variant in Southern Africa', *Nature*, 603: 679–86.

Wyatt, G. E. et al. (2012) ‘Are Cultural Values and Beliefs Included in U.S. Based HIV Interventions?’, *Preventive Medicine*, 55: 362–70.

Yang, H. et al. (2022) ‘Association between Natural/Built Campus Environment and Depression among Chinese Undergraduates: Multiscale Evidence for the Moderating Role of Socioeconomic Factors after Controlling for Residential Self-Selection’, *Frontiers in Public Health*, 10: 844541.

Zai, J. et al. (2020) ‘Tracing the Transmission Dynamics of HIV-1 CRF55_01B’, *Scientific Reports*, 10: 5098.

Zeller, M. et al. (2021) ‘Emergence of an Early SARS-CoV-2 Epidemic in the United States’, *Cell*, 184: 4939–52.e15.

Zwickl, D. J., and Hillis, D. M. (2002) ‘Increased Taxon Sampling Greatly Reduces Phylogenetic Error’, *Systematic Biology*, 51: 588–98.