Viral Genetic Linkage Analysis in the Presence of Missing Data

Citation
Liu, Shelley H., Gabriel Erion, Vladimir Novitsky, and Victor De Gruttola. 2015. "Viral Genetic Linkage Analysis in the Presence of Missing Data." PLoS ONE 10 (8): e0135469. doi:10.1371/journal.pone.0135469. http://dx.doi.org/10.1371/journal.pone.0135469.

Published Version
doi:10.1371/journal.pone.0135469

Permanent link
http://nrs.harvard.edu/urn-3:HUL.InstRepos:22856840

RESEARCH ARTICLE

Viral Genetic Linkage Analysis in the Presence of Missing Data

Shelley H. Liu[1]*, Gabriel Erion[2], Vladimir Novitsky[3], Victor De Gruttola[1]

1 Department of Biostatistics, Harvard School of Public Health, Boston, Massachusetts, United States of America, **2** School of Engineering and Applied Sciences, Harvard University, Cambridge, Massachusetts, United States of America, **3** Department of Immunology and Infectious Diseases, Harvard School of Public Health, Boston, Massachusetts, United States of America

* shelleyliu@fas.harvard.edu

Citation: Liu SH, Erion G, Novitsky V, Gruttola VD (2015) Viral Genetic Linkage Analysis in the Presence of Missing Data. PLoS ONE 10(8): e0135469. doi:10.1371/journal.pone.0135469

Editor: Luis Menéndez-Arias, Centro de Biología Molecular Severo Ochoa (CSIC-UAM), SPAIN

Received: January 28, 2015

Accepted: July 23, 2015

Published: August 24, 2015

Data Availability Statement: All HIV-1 sequences are available from GenBank database. The accession numbers for the 370 HIV-1 Subtype C sequences are: KF374125, KF374117, KF374118, KF374119, KF374120, KF374121, KF374122, KF374123, KF374126, KF374127, KF374128, KF374130, KF374134, KF374135, KF374136, KF374137, KF374138, KF374141, KF374142, KF374177, KF374149, KF374150, KF374151, KF374153, KF374156, KF374157, KF374159, KF374160, KF374161, KF374163, KF374164, KF374165, KF374166, KF374167, KF374168, KF374169, KF374170, KF374171, KF374172, KF374174, KF374175, KF374178, KF374179,

Abstract

Analyses of viral genetic linkage can provide insight into HIV transmission dynamics and the impact of prevention interventions. For example, such analyses have the potential to determine whether recently-infected individuals have acquired viruses circulating within or outside a given community. In addition, they have the potential to identify characteristics of chronically infected individuals that make their viruses likely to cluster with others circulating within a community. Such clustering can be related to the potential of such individuals to contribute to the spread of the virus, either directly through transmission to their partners or indirectly through further spread of HIV from those partners. Assessment of the extent to which individual (incident or prevalent) viruses are clustered within a community will be biased if only a subset of subjects are observed, especially if that subset is not representative of the entire HIV infected population. To address this concern, we develop a multiple imputation framework in which missing sequences are imputed based on a model for the diversification of viral genomes. The imputation method decreases the bias in clustering that arises from informative missingness. Data from a household survey conducted in a village in Botswana are used to illustrate these methods. We demonstrate that the multiple imputation approach reduces bias in the overall proportion of clustering due to the presence of missing observations.

Introduction

Targeting HIV prevention interventions to high-risk groups [1] can be aided by analyses of HIV viral genetic linkage. In particular, investigation of viral genetic sequences of study participants in community-based studies can reveal the viral strains propagating within and across communities and the characteristics of people infected with these strains. Numerous studies have already demonstrated the use of viral sequences from HIV-infected participants to investigate patterns of phylogenetic linkage, which provides information about patterns of HIV transmission dynamics [2–22].

KF374181, KF374183, KF374184, KF374186, KF374188, KF374189, KF374191, KF374192, KF374193, KF374194, KF374195, KF374196, KF374198, KF374200, KF374201, KF374202, KF374203, KF374204, KF374205, KF374206, KF374207, KF374208, KF374209, KF374210, KF374211, KF374212, KF374213, KF374214, KF374215, KF374216, KF374217, KF374221, KF374223, KF374224, KF374252, KF374230, KF374231, KF374233, KF374234, KF374237, KF374238, KF374239, KF374241, KF374242, KF374243, KF374244, KF374245, KF374246, KF374248, KF374249, KF374250, KF374253, KF374255, KF374256, KF374257, KF374258, KF374260, KF374262, KF374264, KF374267, KF374268, KF374269, KF374270, KF374271, KF374273, KF374275, KF374276, KF374277, KF374278, KF374279, KF374282, KF374283, KF374287, KF374289, KF374291, KF374292, KF374293, KF374294, KF374295, KF374298, KF374299, KF374300, KF374301, KF374302, KF374303, KF374304, KF374305, KF374306, KF374307, KF374309, KF374310, KF374312, KF374314, KF374315, KF374318, KF374319, KF374320, KF374321, KF374322, KF374323, KF374325, KF374326, KF374327, KF374330, KF374331, KF374332, KF374335, KF374337, KF374339, KF374341, KF374343, KF374345, KF374347, KF374349, KF374350, KF374351, KF374352, KF374353, KF374354, KF374355, KF374356, KF374357, KF374358, KF374359, KF374361, KF374362, KF374363, KF374364, KF374365, KF374366, KF374367, KF374369, KF374370, KF374371, KF374372, KF374373, KF374375, KF374376, KF374377, KF374379, KF374380, KF374382, KF374383, KF374384, KF374385, KF374386, KF374387, KF374388, KF374391, KF374394, KF374397, KF374398, KF374399, KF374400, KF374404, KF374405, KF374407, KF374408, KF374409, KF374411, KF374413, KF374415, KF374419, KF374420, KF374424, KF374425, KF374428, KF374429, KF374430, KF374431, KF374432, KF374434, KF374436, KF374437, KF374438, KF374439, KF374440, KF374441, KF374442, KF374444, KF374446, KF374449, KF374451, KF374452, KF374453, KF374457, KF374463, KF374464, KF374467, KF374470, KF374474, KF374475, KF374476, KF374477, KF374478, KF374479, KF374480, KF374481, KF374485, KF374488, KF374489, KF374490, KF374492, KF374494, KF374496, KF374501, KF374502, KF374504, KF374505, KF374506, KF374507, KF374508, KF374509, KF374510, KF374511, KF374512, KF374513, KF374518, KF374521, KF374522, KF374523, KF374525, KF374526, KF374528, KF374531, KF374532, KF374533, KF374535,

In this paper, we consider data collected from a household survey in a village in Botswana that was done in conjunction with a pilot study of combination HIV prevention. Nonresponse is a major concern in these types of studies; contributing factors include work patterns, such as that of migrant workers and other highly mobile individuals, temporary absence of the individuals, and migration [23]. Because certain demographic groups, especially males, are systematically under-represented in the sample, methods to control for bias from missing data are required to make valid inferences about genetic linkage. Missing data may arise by happenstance or by design; consideration of the potential for bias in estimation of probabilities of linkage is especially important when the pattern of missingness may be related to characteristics of observed and missing observations—a pattern is that is referred to as informative missingness [24, 25]. We propose methods to adjust for the presence of informatively missing data in viral genetic linkage analysis.

We investigate viral linkage though examination of clustering, which is a measure of relatedness among viruses sampled from different people. One metric of interest is the overall proportion of clustering in the community, i.e., the proportion of prevalent HIV cases that are phylogenetically linked to other residents in the same community. In this study our definition of clustering is based on the pairwise distance between viral sequences; we define two individuals to be linked when this distance is below a specified threshold. Test statistics can help determine whether certain demographic groups are more likely to be clustered, which aids in understanding HIV transmission patterns—especially if sampled viral sequences include new infections. Linkage analyses can also reveal the overall proportion of clustering in the dataset, which provides information about the extent to which individuals are being infected by strains circulating inside their community or by strains outside of the community. Information regarding whether HIV transmissions are mainly occurring within or between communities can help reveal virus transmission dynamics, as well as identify subgroups of the population that are driving the epidemic [4, 6, 8, 13–17, 22, 26–30]. In the setting where an intervention is being provided, changes in the overall proportion of clustering within a village receiving the intervention over time are informative regarding the efficacy of the interventions; if it is working, there should be a decrease in virologically linked cases over time.

As mentioned above, a major problem arises in linkage analyses when the sample of viral sequences is incomplete; this situation may occur because intended subjects cannot be found or refuse participation. Bias arises even if observations are missing completely at random. Because viral linkage is based on the clustering of pairs of individuals, random missing links in the viral transmission chain will cause underestimation in the proportion of clustering. Since observing a link requires the presence of both linked sequences, removal of a subject removes all links to that subject. Hence, the number of missing links between sequences can increase more rapidly than does the number of missing subjects. Therefore, unbiased estimation of the true proportion of sequences that cluster requires information about both the total population size (the number of missing subjects plus the number of observed sequences), and characteristics that impact the probability that a sequence is missing.

When observations are informatively missing (i.e. viral genetic information is more likely to be missing for individuals of certain demographic groups), additional sources of bias are introduced. The fact that the presence of linkage depends on each sequence in the database creates a challenge for proper handling of missing data. Adjustment for missing data can be at the test statistic level [31] or at the viral sequence level. Carnegie et al. 2014 showed that the probability that individual sequences are linked can be consistently estimated using the observed data, and the probability that groups of sequences are linked can be estimated as well under the assumption that pairs of sequences are uncorrelated. The authors also developed a resampling approach for estimating the proportions of linkage that accounts for correlation between pairs

KF374536, KF374538, KF374539, KF374543, KF374546, KF374547, KF374548, KF374549, KF374550, KF374553, KF374554, KF374555, KF374558, KF374560, KF374561, KF374562, KF374563, KF374564, KF374565, KF374569, KF374570, KF374572, KF374573, KF374575, KF374576, KF374579, KF374580, KF374581, KF374585, KF374587, KF374588, KF374589, KF374590, KF374591, KF374592, KF374593, KF374594, KF374595, KF374596, KF374597, KF374598, KF374601, KF374604, KF374606, KF374607, KF374608, KF374609, KF374610, KF374611, KF374612, KF374613, KF374617, KF374618, KF374619, KF374620, KF374623, KF374626, KF374627, KF374112, KF374628, KF374629, KF374630, KF374631, KF374633, KF374634, KF374636, KF374638, KF374639, KF374640, KF374641, KF374642, KF374645, KF374646, KF374647, KF374648, KF374649, KF374650, KF374651, KF374652, KF374654, KF374655, KF374656, KF374658, KF374660, KF374661, KF374663, KF374664, KF374665, KF374668, KF374669, KF374670, KF374671, KF374672, KF374673, KF374674, KF374675, KF374676, KF374677, KF374678. The accession number for single HIV-1 subtype A1 sequence is KP334131.

Funding: NIH Grant number R37 51164 (VG). THE NIH had no role in study design, data collection or analysis. The Mochudi Prevention Project in Botswana (VN) was supported and funded by NIH grant R01 AI083036, an HIV Prevention Program for Mochudi, Botswana.

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

of individuals. Here, we study a similar problem but use a multiple imputation approach framework that imputes missing sequences at the translated amino acid level, which is applicable in a very broad array of analyses that require imputed sequences, including phylogenetic analyses. In our investigation we focus on translated amino acid sequences as we are interested in non-synonymous mutations; previous research has shown that almost all site mutations within protein-coding regions are non-synonymous mutations [32]. Our focus here is on the use of imputed-complete datasets to investigate the proportion of overall HIV clustering in the targeted population. This paper investigates sensitivity of estimation for incomplete data, and the extent to which the proposed multiple imputation approach reduces bias in viral linkage estimates.

Materials and Methods

The data we consider arises from a pilot study that provided information to aid in the design of a combination prevention cluster-randomized trial in Botswana [33]. The treatment modalities under investigation include testing of all consenting village subjects, provision of treatment for patients with viral loads above 10,000 copies/mL, and male circumcision for all HIV-uninfected men. The pilot study was undertaken in Mochudi, Botswana in 2011-2013, and provided household surveys, HIV testing for all consenting household members not previously diagnosed with HIV infection, laboratory assessment of viral load and CD4 lymphocyte count for HIV+ participants, and ascertainment of antiretroviral treatment status. The study was conducted according to the principles expressed in the Declaration of Helsinki, and was approved by the Health Research and Development Committee (HRDC) of the Republic of Botswana, and the Office of Human Research Administration (OHRA) of the Harvard School of Public Health. All adult study subjects provided written informed consent for participation in the study; all minor study subjects provided written informed assent, and each minor's guardian provided written informed consent for their participation in the study. All available data were used for this paper; the 371 HIV-1 sequences analyzed in this paper are from patients enrolled in the first phase of the pilot study that was completed in Spring of 2012. HIV-1 subtyping revealed that 370 sequences were subtype C, and one sequence was HIV-1 subtype A1. Of the 371 subjects in the data, 287 were female and 84 were male, among whom 124 females and 24 males were categorized as young (<35 years). Young females accounted for 43% of females and young males accounted for 29% of males. For all virological samples, the V1C5 codon-based alignment was generated as described elsewhere [26] using muscle [34] in MEGA5 [35].

An estimate of the overall proportion of clustering, or the proportion of sequences that link to at least one other sequence, was calculated using the pairwise distance matrix of translated amino acid residues from phylogenetic analysis, which quantifies the degree of similarity between any two sequences. Several methods exist for calculating the distance matrix. These include Jukes-Cantor [36], Dayhoff [37], and JTT [38], among others. These methods assume different relationships among transition probabilities and different equilibrium frequencies. For simplicity of illustration, we define pairwise distance as the number of absolute amino acid changes between any two sequences divided by the total sequence length, but the viral sequence imputation method would work for other distance measures. D_{i_j} represents the pairwise distance between individuals i and j, and N is the total number of individuals in the dataset. If D_{i_j} is less than the threshold, then individuals i and j are considered to be clustered. Moreover, if D_{i_j} is less than the threshold for at least one individual j, where $i \neq j$, then individual i is considered to cluster. We represent this by introducing an indicator variable, I_i, for individual i such that:

$$I_i = \begin{cases} 1 & \text{if } D_{i_j} \text{ is less than the threshold for at least one individual j, where } i \neq j \\ 0 & \text{Otherwise} \end{cases}$$

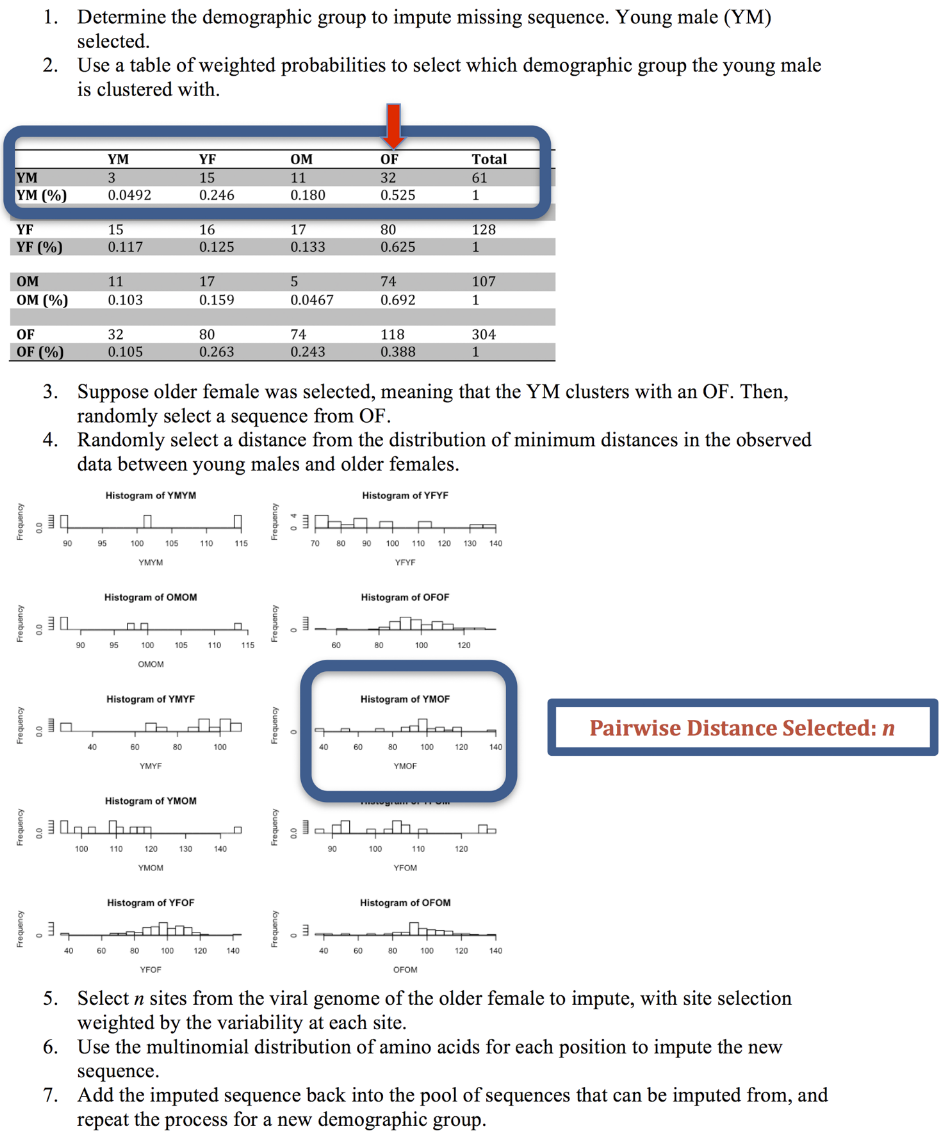
Overall proportion of clustering: $\frac{1}{N}\sum_{i=1}^{N} I_i$

We consider two thresholds for illustration of the sequence imputation method: 0.10 and 0.15. These thresholds were chosen based on the distribution of the V1C5 pairwise distances of the translated amino acids. Our methods apply for any threshold; appropriate choice of a clustering threshold is an active research area and can be affected by many factors, such as sampling density [39].

Our sequence imputation approach follows the multiple imputation statistical paradigm for dealing with informative missing data [24, 25]. Multiple imputation is a statistical method that allows for valid inferences from incomplete data, by creating imputed-complete datasets. We use the following formulas to calculate the mean proportion of clustering across imputed-complete datasets, the variance, standard deviation and standard error of the mean.

Following the notation of Rubin 1996, the m imputed-complete datasets have corresponding m estimated statistics $[Q_{.1}, \ldots, Q_{.m}]$ and variance-covariance matrices $[U.1, \ldots, U.m]$. The repeated-imputation estimate is $Q_m = \sum_{l=1}^{m} Q_{.l}/m$, the associated variance-covariance of Q_m is $T_m = U_m + (m + 1)/mB_m$, where $U_m = \sum_{l=1}^{m} U_{.l}/m$ is the within-imputation variability, which is calculated by bootstrap as described in [40]. A collection of bootstrapped alignments is generated by resampling the sites of the imputed alignment with replacement. Each bootstrapped alignment has the same number of sequences as the imputed alignment, but the column of nucleotides representing a given site in the imputed alignment may be present multiple times or not at all in the bootstrapped alignment. $B_m = 1/(m-1)\sum_{I=1}^{m}(Q_{.I} - Q_m)(Q_{.I} - Q_m)'$ is the between-imputation variability.

The following algorithm illustrates the sequence imputation mechanism, which is illustrated in Fig 1. Here we assume that demographic categorization is possible for everyone in the population, regardless of whether a sequence is available. *(1)* For each missing sequence, determine the demographic group, of which there are four: young males, young females, older males or older females. *(2)* Use a table of weighted probabilities to select the demographic group with which the subject whose sequences is missing is clustered. The weighted probabilities are the probabilities, calculated using observed sequences, that a sequence in one demographic group has its nearest neighbor in another demographic group. These probabilities are defined by the distribution of minimum distances, which is the minimum number of absolute amino acid changes from one sequence to the next. *(3)* Randomly select a viral sequence from this demographic group. *(4)* From the distribution of minimum distances in the observed data between the demographic groups of a subject with a missing sequence and the randomly selected sequence, select a distance, n, at random. This distance is the number of absolute amino acid changes between two sequences. *(5)* Select n sites from the viral genome of the sequence in *(3)* to modify. Site selection is weighted by the variability of each site, based on the marginal frequency of the most common amino acid at each site. Marginal frequencies for each amino acid site were computed using the Biopython package, which was also used for alignment parsing [41]. *(6)* Impute the amino acids for each of the n positions based on assumption that these amino acids are multi-nomially distributed. *(7)* Add the imputed sequence back into the pool of sequences that can be imputed from.



1. Determine the demographic group to impute missing sequence. Young male (YM) selected.
2. Use a table of weighted probabilities to select which demographic group the young male is clustered with.

	YM	YF	OM	OF	Total
YM	3	15	11	32	61
YM (%)	0.0492	0.246	0.180	0.525	1
YF	15	16	17	80	128
YF (%)	0.117	0.125	0.133	0.625	1
OM	11	17	5	74	107
OM (%)	0.103	0.159	0.0467	0.692	1
OF	32	80	74	118	304
OF (%)	0.105	0.263	0.243	0.388	1

3. Suppose older female was selected, meaning that the YM clusters with an OF. Then, randomly select a sequence from OF.
4. Randomly select a distance from the distribution of minimum distances in the observed data between young males and older females.

Pairwise Distance Selected: *n*

5. Select *n* sites from the viral genome of the older female to impute, with site selection weighted by the variability at each site.
6. Use the multinomial distribution of amino acids for each position to impute the new sequence.
7. Add the imputed sequence back into the pool of sequences that can be imputed from, and repeat the process for a new demographic group.

Fig 1. Schematic of the sequence-imputation method.

doi:10.1371/journal.pone.0135469.g001

Applications

Simulation Study

A simulation study demonstrated the ability of the model to reduce the biases arising from informative missing data. The study was based on incomplete datasets that were created by deleting sequences from the available data, and then using the multiple imputation procedure to estimate rates of clustering for the entire observed database of 371 sequences. Deletions of $m = (10, 20, \ldots, 200)$ subjects were made. We refer to the sample of $N = 371$ sequences as the observed data set; the observed data with m deletions that yield $N = 371 - m$ sequences as the incomplete-observed data; and the incomplete-observed dataset plus the imputed values as the imputed-complete data.

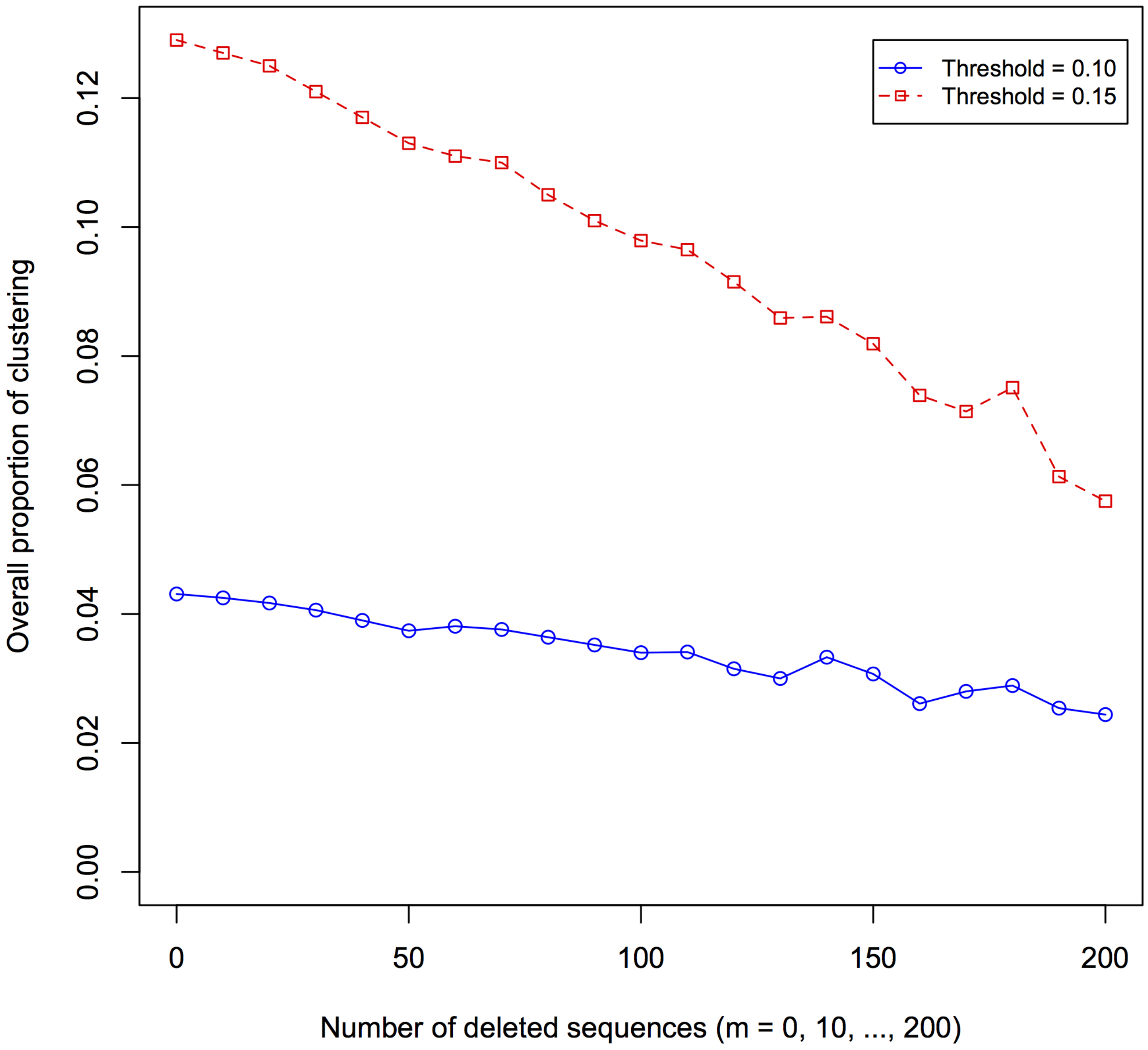


Fig 2. Decrease in overall proportion of clustering with deletion of sequences. Clustering at each number of deletions is averaged over 100 different random deletions of the same number of sequences.

doi:10.1371/journal.pone.0135469.g002

Fig 2 demonstrates that as the number of deleted sequences increases, the proportion of overall clustering decreases. The sequences were deleted selectively; weighting was by gender, with sequences of females deleted ten times more often than that of males. Because of the dearth of young males in the observed dataset, the male:female deletion ratio was chosen to preserve sequences of young males in the incomplete-observed datasets when *m* was large. The proportion of subjects who cluster with others is naturally larger with a higher threshold (0.15 instead of 0.10).

Table 1 shows that the adjustment using multiple imputation reduces—and in some cases nearly eliminates—the bias resulting from missing data. We consider incomplete-observed datasets created by deleting *m* = 50, 100, 200 sequences from the observed dataset and calculate

Table 1. Estimated overall proportion of clustering for observed, incomplete-observed and imputed-complete datasets.

Thres	Clus, Obs	# Del	Clus, In-Obs	Clus, Impute-Com	Cov, In-Obs	Cov, Impute-Com
0.1	0.043	50	0.039	0.043	83%	94%
		100	0.036	0.043	66%	89%
		200	0.024	0.037	30%	61%
0.15	0.13	50	0.11	0.13	48%	91%
		100	0.099	0.13	16%	85%
		200	0.059	0.10	1.4%	54%

Thres stands for threshold; *Clus* for clustering; *Obs* for observed data; *# Del* for number of deleted sequences; *In-Obs* for ncomplete-observed data; *Impute-Com* for imputed-complete data; *Cov* stands for coverage. Clustering assessed at the 0.10 and 0.15 thresholds, with varying number of deleted sequences (m = 50, 100, or 200).

doi:10.1371/journal.pone.0135469.t001

the mean proportion of clustering for the thresholds of 0.10 and 0.15. The variance was calculated as described in the Methods section. The table also provides the coverage percentages for the 95% confidence intervals on estimates from deletion and imputation. The confidence intervals were obtained by bootstrapping standard errors. For each incomplete-observed dataset, in which *m* sequences had been deleted, ten imputed-complete datasets were constructed to calculate the imputed-complete estimates and standard errors. This process was repeated over 100 different subset deletions (incomplete datasets with *m* deletions) in order to calculate coverage. Coverage was defined as the proportion of times that 100 repetitions over 100 different subset deletions yielded a 95% confidence interval that included the clustering value obtained from the complete data. As the amount of missing data increases, the performance of the multiple imputation estimates and the coverage worsens, but the bias is always considerably decreased and coverage improved compared to the results for the incomplete-observed data.

Fig 3 demonstrates the bias-correcting effect of the multiple imputation. Deleting 100 sequences leads to noticeably lower clustering. For the incomplete-dataset, at the 0.10 threshold, overall proportion of clustering is 0.030 [0.020, 0.039]; at the 0.15 threshold, overall proportion of clustering is 0.10 [0.095, 0.11]. The multiple imputation method substantially reduces bias and results in larger 95% confidence intervals. For the imputed-complete dataset, at the 0.10 threshold, the overall proportion of clustering is 0.040 [0.016, 0.064]; at the 0.15 threshold, the overall proportion of clustering is 0.13 [0.094, 0.16].

We investigate the performance of the sequence imputation method for two scenarios for sampling data missing not at random (MNAR) in Table 2. For the first, sequences are deleted from the observed-complete dataset in proportion to the number of other sequences with which they cluster, plus one. If, for example, a sequence clusters with 9 other sequences, it will be ten times more likely to be deleted than a sequence which did not cluster with any other sequences. Under the second scenario, sequences are deleted in a manner inversely proportional to their number of clustering partners. For each incomplete-observed dataset, in which $m = 100$ sequences were deleted, ten imputed-complete datasets were constructed to calculate the imputed-complete estimates and standard errors. This process was repeated over 1000 different subset deletions (incomplete datasets with $m = 100$ deletions) in order to calculate coverage. The results demonstrate that serious violation of assumptions regarding the nature of the missingness process greatly impact results. As expected, clustering estimates from the

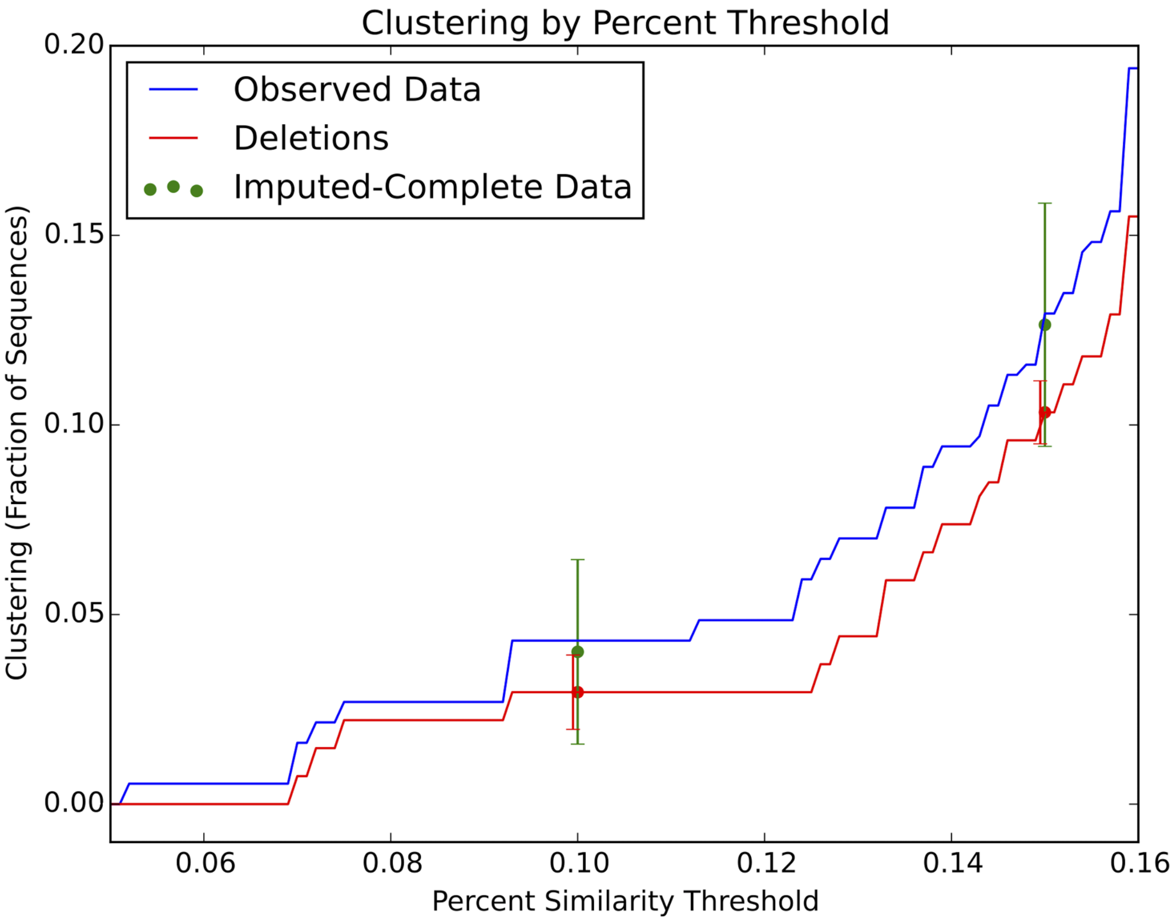


Fig 3. Demonstrating the corrective effect of a representative imputation. Deleting 100 sequences leads to noticeably lower clustering. Imputation substantially improves estimates, and results in slightly larger error bars.

doi:10.1371/journal.pone.0135469.g003

Table 2. Estimated overall proportion of clustering for observed, incomplete-observed and imputed-complete datasets under MNAR.

Deletion weight	Clus, Obs	Clus, In-Obs	Clus, Impute-Com	Cov, In-Obs	Cov, Impute-Com
Proportional	0.043	0.018	0.022	8%	42%
Inversely proportional		0.043	0.053	80%	93%

Clustering assessed at the 0.10 threshold, with m = 100 deletions. *Clus* stands for clustering; *Obs* for observed data; *In-Obs* for ncomplete-observed data; *Impute-Com* for imputed-complete data; *Cov* stands for coverage.

doi:10.1371/journal.pone.0135469.t002

imputed-complete dataset under proportional deletions are much smaller than the analogous estimates under inversely proportional deletions.

Mochudi Pilot Study

We applied our methods to adjust for the fact that the observed Mochudi pilot dataset is an incomplete, biased sample of the population of interest. The focus of estimation is the true proportion of the clustering in the entire population targeted by the Mochudi pilot study. To improve estimation, we impute missing sequences to create an imputed-population dataset. Although we do not have exact information on the number of missing observations, we do have information about the age-gender structure for Botswana as a whole as well as information about HIV-1 prevalence by age and gender in Mochudi. While we cannot be certain that the age-gender structure for Mochudi resembles that of Botswana as a whole, the former is likely to be much closer to the structure for the country than is the observed sample from Mochudi, given its notable under-representation of males. The 2013 CIA World Factbook provides an age-gender breakdown for Botswana [42], which we divide into four categories: young males/females (15-35 years) and older males/females (> 35 years). In order to obtain the estimated proportion of HIV-1 infected people in each age-gender category in Mochudi, we multiply the proportion of people in each age-gender category for Botswana as a whole by the estimated prevalence in those categories, obtained from the Mochudi pilot study [43]. From these calculations we can estimate the proportions of HIV-infected subjects that are in each age-category in Mochudi. We then add imputed sequences to our observed dataset so that the proportions of subjects in each age-gender category of our augmented sample matches those based on our estimates. The estimated proportions are: young males (0.097), older males (0.292), young females (0.295), older females (0.317). To achieve these proportions, we impute sequences for 84 older males, 0 older females, 24 young males, and 25 young females; a total of 133 sequences are added to our observed dataset of 371 sequences. We note that while this database may not be complete, it should be considerably closer to complete than is the observed sample, as older women were most likely to participate in the Mochudi pilot [43]. Furthermore, the imputation adjusts for the under-sampling of males.

Fig 4 shows the effect of applying our viral sequence imputation method to the observed sequence data. In the observed sample (N = 371), the proportion of overall clustering was estimated to be 0.043 [0.034, 0.053] at the 0.10 clustering threshold and 0.13 [0.11, 0.14] at the 0.15 clustering threshold. In the imputed-population dataset (N = 504), the proportions of overall clustering at the two thresholds were 0.062 [0.038, 0.086] and 0.15 [0.12, 0.18], respectively. We note that for both the 0.10 and 0.15 thresholds, the upper confidence interval for the clustering in the observed data excludes the point estimate for the imputed-population dataset. The clustering estimates for the imputed-population data reflect a lower bound as there are likely more missing sequences than the 133 we imputed; if some older females are missing, then additional imputed sequences in all categories would be required.

Discussion

To assess clustering and viral linkage in the presence of missing data, we propose an approach that combines a traditional missing data framework with a model for viral diversification after transmission. In the setting we consider, our sequence imputation method corrects for biases in viral genetic linkage analyses in the presence of informatively missing data. The performance of the multiple imputation approach appears to depend on the proportion of data that are missing. We implemented simulation studies in which our procedures are used to make

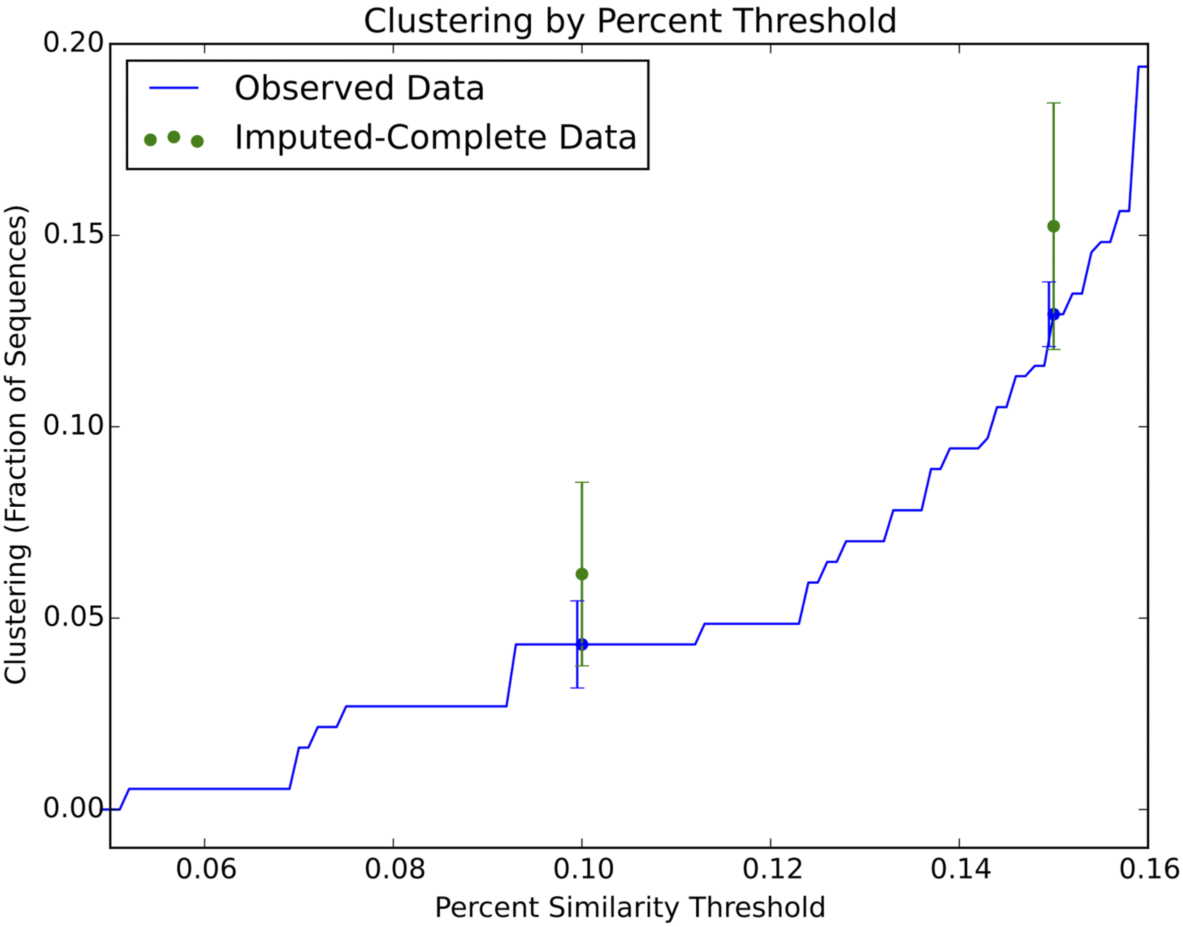


Fig 4. Estimating the true proportion of clustering in the Mochudi population. Treating the observed Mochudi data (N = 371) as a biased sample from the population, and imputing the imputed-population dataset (N = 504) based on Botswana and Mochudi-specific age and gender breakdowns to create a database with the same demographic structure as Botswana as a whole. Point estimates and error bars for the 0.10 and 0.15 thresholds of clustering.

doi:10.1371/journal.pone.0135469.g004

inferences on the incomplete-observed dataset, and compared the results with those from the complete-observed dataset.

In our simulation study, the proposed method performed well when the proportion missing is less than 30%, but underestimated the proportion of sequences that are clustered when the proportion missing is over 50%. One possible reason for this bias is in the distribution of minimum pairwise distances, from which we randomly select the number of amino acid changes to impute from the selected sequence. Bias arises from the fact that this distribution is based on observed distances from an incomplete sample. The distribution of minimum distances is increasingly shifted to the right with greater number of deletions because of the smaller number of links in the incomplete data. To investigate this issue, we conducted a sensitivity analysis as follows: we used the distribution of minimum distances from the full observed dataset rather than from the incomplete dataset in the sequence imputation approach. This approach, however, did not significantly reduce the bias. Next, we investigated the sensitivity to use of a kernel density estimator for estimating the distribution of minimum distances, compared to using the ad-hoc method. The results in S1 Table do not suggest any major difference in results when using kernel density estimation compared to the approach described above. We also conducted a sensitivity analysis to determine if the proportions of sequences deleted by gender affected the imputed-complete clustering estimates in our simulation study. We simulated male:female sequence deletion ratios ranging from 1:1 to 1:10 (M:F); the latter was the ratio used in simulation studies throughout the paper. S2 Table does not show a substantial impact of the deletion ratios on the performance of our method. Bias may also arise from limitations in the model for viral diversification—future studies based on clonal rather than population sequences may provide more information for this model. The lower coverage noted when the proportion of data that is missing exceeds 50% likely reflects not only lower amount of information used to estimate the quantities of importance (distances between genotypes), but also more reliance on generating sequences that may be several transmissions away from someone with an observed sequence.

To impute sequences requires a number of assumptions. One regards the covariation between amino acid positions of the *gp120* region; we assumed independence, but other choices are possible. A second assumption is that the clustered sequences have similar processes of diversification as those of the sequences of the larger population from which these clustered sequences are sampled. In addition, we assume that sequences are missing at random given observed characteristics on the missing sequences (e.g. age and sex of the hosts).

In this paper, we do not provide detailed guidance on choice of clustering threshold. The sequence imputation method is applicable to a diverse range of distance matrix definitions, and we present a simple distance matrix definition for ease of illustration. In practice, thresholds will be specific both to the distance matrix definition and the study objective. For example, in assessing whether two individuals are likely to be fairly closely connected in the same transmission chain, it might be preferable to use a lower threshold. By contrast, assessing whether two individuals are infected by strains circulating within a given community or by strains circulating only outside of the community, one might prefer a higher threshold. We recommend investigation at a range of thresholds as shown above, although we selected two thresholds to demonstrate coverage.

In the simulated imputed-population dataset, in which we apply our method to adjust for the bias and incompleteness of the observed Mochudi pilot dataset, our estimates of clustering have limitations that arise from the sampling of the pilot study population in Mochudi [43]. We solely adjust for biases that result when the probability of sampling depends only on age category and gender. We note that we cannot exclude the possibility that probabilities of sampling depend on other factors as well. Nonetheless, we demonstrate that the sequence imputation method used to adjust clustering estimates can greatly improve estimation if these assumptions are met. The method could easily accommodate other factors that impact probability of sampling, when they are known and measured..

In summary, viral genetic linkage analyses has been shown to be useful in making inferences about transmission patterns, but analyses can yield biased results unless the impact of incomplete sampling of populations of interest is properly taken into account.

Supporting Information

S1 Table. Estimated overall proportion of clustering for observed, incomplete-observed and imputed-complete datasets using kernel density estimation for the distribution of minimum distances. Clustering assessed at the 0.10 threshold, with m = 100 deletions. Coverage is calculated as described in Table 2. Ratio of male:female deletions was 10:1. *Clus* stands for clustering; *Obs* for observed data; *In-Obs* for ncomplete-observed data; *Impute-Com* for imputed-complete data; *Cov* stands for coverage.
(PDF)

S2 Table. Estimated overall proportion of clustering for observed, incomplete-observed and imputed-complete datasets at varying ratios of male:female sequence deletions. Clustering assessed at the 0.10 threshold, with m = 100 deletions. Coverage is calculated as described in Table 2. *Clus* stands for clustering; *Obs* for observed data; *In-Obs* for ncomplete-observed data; *Impute-Com* for imputed-complete data; *Cov* stands for coverage.
(PDF)

Author Contributions

Conceived and designed the experiments: VG SHL GE. Performed the experiments: SHL GE VN. Analyzed the data: GE SHL. Contributed reagents/materials/analysis tools: VN. Wrote the paper: SHL VG.

References

1. DeGruttola E, Smith D, Little S, Miller V. Developing and Evaluating Comprehensive HIV Infection Control Strategies: Issues and Challenges. Clin Infect Dis. 2010;Suppl 3(50):S102–7. doi: 10.1086/651480
2. Little SJ, Kosakovsky Pond SL, Anderson CM, Young JA, Wertheim JO, Mehta SR, et al. Using HIV networks to inform real time prevention interventions. PLoS One. 2014; 9(6):e98443. Available from: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC4047027/pdf/pone.0098443.pdf doi: 10.1371/journal.pone.0098443 PMID: 24901437
3. Vrancken B, Rambaut A, Suchard MA, Drummond A, Baele G, Derdelinckx I, et al. The genealogical population dynamics of HIV-1 in a large transmission chain: bridging within and among host evolutionary rates. PLoS Comput Biol. 2014; 10(4):e1003505. Available from: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3974631/pdf/pcbi.1003505.pdf doi: 10.1371/journal.pcbi.1003505 PMID: 24699231
4. Volz EM, Ionides E, Romero-Severson EO, Brandt MG, Mokotoff E, Koopman JS. HIV-1 transmission during early infection in men who have sex with men: a phylodynamic analysis. PLoS Med. 2013; 10 (12):e1001568; discussion e1001568. Available from: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3858227/pdf/pmed.1001568.pdf doi: 10.1371/journal.pmed.1001568 PMID: 24339751
5. Volz EM, Koelle K, Bedford T. Viral phylodynamics. PLoS Comput Biol. 2013; 9(3):e1002947. Available from: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3605911/pdf/pcbi.1002947.pdf doi: 10.1371/journal.pcbi.1002947 PMID: 23555203
6. Volz EM, Koopman JS, Ward MJ, Brown AL, Frost SD. Simple epidemiological dynamics explain phylogenetic clustering of HIV from patients with recent infection. PLoS Comput Biol. 2012; 8(6):e1002552. Available from: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3386305/pdf/pcbi.1002552.pdf doi: 10.1371/journal.pcbi.1002552 PMID: 22761556
7. Volz EM, Kosakovsky Pond SL, Ward MJ, Leigh Brown AJ, Frost SD. Phylodynamics of infectious disease epidemics. Genetics. 2009; 183(4):1421–1430. Available from: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2787429/pdf/GEN18341421.pdf doi: 10.1534/genetics.109.106021 PMID: 19797047
8. Leigh Brown AJ, Lycett SJ, Weinert L, Hughes GJ, Fearnhill E, Dunn DT. Transmission network parameters estimated from HIV sequences for a nationwide epidemic. J Infect Dis. 2011; 204(9):1463–1469. Available from: http://jid.oxfordjournals.org/content/204/9/1463.full.pdf doi: 10.1093/infdis/jir550
9. Novitsky V, Wang R, Lagakos S, Essex M. HIV-1 Subtype C Phylodynamics in the Global Epidemic. Viruses. 2010; 2(1):33–54. Available from: http://www.mdpi.com/1999-4915/2/1/33/pdf doi: 10.3390/v2010033 PMID: 21994599
10. Wertheim JO, Kosakovsky Pond SL, Little SJ, De Gruttola V. Using HIV transmission networks to investigate community effects in HIV prevention trials. PLoS One. 2011; 6(11):e27775. Available from: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3218056/pdf/pone.0027775.pdf doi: 10.1371/journal.pone.0027775 PMID: 22114692
11. Wertheim JO, Leigh Brown AJ, Hepler NL, Mehta SR, Richman DD, Smith DM, et al. The global transmission network of HIV-1. J Infect Dis. 2014; 209(2):304–313. Available from: http://jid.oxfordjournals.org/content/209/2/304.long doi: 10.1093/infdis/jit524
12. Wertheim JO, Scheffler K, Choi JY, Smith DM, Kosakovsky Pond SL. Phylogenetic relatedness of HIV-1 donor and recipient populations. J Infect Dis. 2013; 207(7):1181–1182. Available from: http://jid.oxfordjournals.org/content/207/7/1181.full.pdf doi: 10.1093/infdis/jit021
13. Bezemer D, van Sighem A, Lukashov VV, van der Hoek L, Back N, Schuurman R, et al. Transmission networks of HIV-1 among men having sex with men in the Netherlands. Aids. 2010; 24(2):271–282. doi: 10.1097/QAD.0b013e328333ddee PMID: 20010072

14. Bezemer D, Faria NR, Hassan A, Hamers RL, Mutua G, Anzala O, et al. HIV Type 1 transmission networks among men having sex with men and heterosexuals in Kenya. AIDS Res Hum Retroviruses. 2014; 30(2):118–126. Available from: http://online.liebertpub.com/doi/pdfplus/10.1089/aid.2013.0171 doi: 10.1089/aid.2013.0171 PMID: 23947948

15. Brenner BG, Roger M, Routy JP, Moisi D, Ntemgwa M, Matte C, et al. High rates of forward transmission events after acute/early HIV-1 infection. J Infect Dis. 2007; 195(7):951–959. Available from: http://jid.oxfordjournals.org/content/195/7/951.full.pdf doi: 10.1086/512088

16. Brenner BG, Roger M, Stephens D, Moisi D, Hardy I, Weinberg J, et al. Transmission clustering drives the onward spread of the HIV epidemic among men who have sex with men in Quebec. J Infect Dis. 2011; 204(7):1115–1119. Available from: http://jid.oxfordjournals.org/content/204/7/1115.full.pdf doi: 10.1093/infdis/jir468

17. Brenner B, Wainberg MA, Roger M. Phylogenetic inferences on HIV-1 transmission: implications for the design of prevention and treatment interventions. Aids. 2013; 27(7):1045–1057. Available from: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3786580/pdf/nihms509196.pdf doi: 10.1097/QAD.0b013e32835cffd9 PMID: 23902920

18. Kouyos RD, von Wyl V, Yerly S, Boni J, Taffe P, Shah C, et al. Molecular epidemiology reveals long-term changes in HIV type 1 subtype B transmission in Switzerland. J Infect Dis. 2010; 201(10):1488–1497. Available from: http://jid.oxfordjournals.org/content/201/10/1488.full.pdf doi: 10.1086/651951

19. Leventhal GE, Gunthard HF, Bonhoeffer S, Stadler T. Using an epidemiological model for phylogenetic inference reveals density dependence in HIV transmission. Mol Biol Evol. 2014; 31(1):6–17. Available from: http://mbe.oxfordjournals.org/content/31/1/6.full.pdf doi: 10.1093/molbev/mst172 PMID: 24085839

20. Stadler T, Bonhoeffer S. Uncovering epidemiological dynamics in heterogeneous host populations using phylogenetic methods. Philos Trans R Soc Lond B Biol Sci. 2013; 368(1614):20120198. Available from: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3678323/pdf/rstb20120198.pdf doi: 10.1098/rstb.2012.0198

21. Hughes GJ, Fearnhill E, Dunn D, Lycett SJ, Rambaut A, Leigh Brown AJ. Molecular phylodynamics of the heterosexual HIV epidemic in the United Kingdom. PLoS Pathog. 2009; 5(9):e1000590. Available from: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2742734/pdf/ppat.1000590.pdf doi: 10.1371/journal.ppat.1000590 PMID: 19779560

22. Lewis F, Hughes GJ, Rambaut A, Pozniak A, Leigh Brown AJ. Episodic sexual transmission of HIV revealed by molecular phylodynamics. PLoS Med. 2008; 5(3):e50. Available from: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2267814/pdf/pmed.0050050.pdf doi: 10.1371/journal.pmed.0050050 PMID: 18351795

23. Ziraba A, Madise N, Matilu M, Zulu E, Kebaso J, Khamadi V, et al. The effect of participant nonresponse on HIV prevalence estimates in a population-based survey in two informal settlements in Nairobi city. Population Health Metrics. 2010; 22(8).

24. Rubin D. Multiple Imputation after 18+ years. JASA. 1996; 91(434):437–89.

25. Rubin D. Multiple Imputation for Nonresponse in Surveys. New York: John Wiley & Sons, Inc; 1987.

26. Novitsky V, Bussmann H, Logan A, Moyo S, van Widenfelt E, Okui L, et al. Phylogenetic relatedness of circulating HIV-1C variants in Mochudi, Botswana. PLoS One. 2013; 8(12):e80589. Available from: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3859477/pdf/pone.0080589.pdf doi: 10.1371/journal.pone.0080589 PMID: 24349005

27. Grabowski MK, Lessler J, Redd AD, Kagaayi J, Laeyendecker O, Ndyanabo A, et al. The role of viral introductions in sustaining community-based HIV epidemics in rural Uganda: evidence from spatial clustering, phylogenetics, and egocentric transmission models. PLoS Med. 2014; 11(3):e1001610. Available from: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3942316/pdf/pmed.1001610.pdf doi: 10.1371/journal.pmed.1001610 PMID: 24595023

28. Delatorre E, Bello G. Phylodynamics of HIV-1 subtype C epidemic in east Africa. PLoS One. 2012; 7(7):e41904. doi: 10.1371/journal.pone.0041904 PMID: 22848653

29. Faria N, Sigaloff K, van de Vijver D, Tatem A, Pineda A, Wallis C, et al. Migration of HIV-1 Subtypes in East Africa Is Associated With Proximity To Highway Corridor. Abstract 225. 2014;CROI(2014):Boston, MA.

30. Chia J, Aghokeng A, Guichet E, Ayouba A, Ahuka-Mundeke S, Vidal N, et al. Ongoing Cross-Species Transmission of Simian Retroviruses and High HIV Prevalence in Cameroon. Abstract 226. 2014;CROI(2014):Boston, MA.

31. Carnegie NB, Wang R, Novitsky V, De Gruttola V. Linkage of viral sequences among HIV-infected village residents in Botswana: estimation of linkage rates in the presence of missing data. PLoS Comput Biol. 2014; 10(1):e1003430. Available from: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3886896/pdf/pcbi.1003430.pdf doi: 10.1371/journal.pcbi.1003430 PMID: 24415932

32. Gao F, Chen Y, Levy N, Conway J, Kepler T, Hui H. Unselected Mutations in the Human Immunodeficiency Virus Type 1 Genome Are Mostly Nonsynonymous and Often Deleterious. JVI. 2004; 5 (78):2426–2433. doi: 10.1128/JVI.78.5.2426-2433.2004

33. Wang R, Goyal R, Lei Q, Essex M, De Gruttola V. Sample size considerations in the design of cluster randomized trials of combination HIV prevention. Clin Trials. 2014; 11(309).

34. Edgar RC. MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res. 2004; 32(5):1792–1797. Available from: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC390337/pdf/gkh340.pdf doi: 10.1093/nar/gkh340 PMID: 15034147

35. Tamura K, Peterson D, Peterson N, Stecher G, Nei M, Kumar S. MEGA5: Molecular Evolutionary Genetics Analysis using Maximum Likelihood, Evolutionary Distance, and Maximum Parsimony Methods. Mol Biol Evol. 2011; 28:2731–2739. doi: 10.1093/molbev/msr121 PMID: 21546353

36. Jukes T, Cantor C. Evolution of Protein Molecules. New York: Academic Press; 1969.

37. Schwarz R, Dayhoff M. Matrices for detecting distant relationships. In: Atlas of protein sequences. National Biomedical Research Foundation; 1979. p. 353–58.

38. Jones DT, Taylor WR, Thornton JM. The rapid generation of mutation data matrices from protein sequences. Comput Appl Biosci. 1992; 8(3):275–282. PMID: 1633570

39. Novitsky V, Moyo S, Lei Q, DeGruttola V, Essex M. Impact of Sampling Density on the Extent of HIV Clustering. AIDS Res Hum Retroviruses. 2014 Oct;Available from: http://www.ncbi.nlm.nih.gov/pubmed/25275430 doi: 10.1089/aid.2014.0173

40. Felsenstein J. Confidence limits on phylogenies: an approach using the bootstrap. Evolution. 1985;p. 783–791. doi: 10.2307/2408678

41. Cock PJ, Antao T, Chang JT, Chapman BA, Cox CJ, Friedberg I, et al. Biophython: freely available Python tools for computational molecular biology and bioinformatics. Bioinformatics. 2009; 25 (11):1422–3. doi: 10.1093/bioinformatics/btp163 PMID: 19304878

42. Central Intelligence Agency. The World Factbook. Africa: Botswana; 2013. Available from: https://www.cia.gov/library/publications/the-world-factbook/geos/bc.html

43. Novitsky V. personal communication.