Contents lists available at ScienceDirect

EBioMedicine

journal homepage: www.elsevier.com/locate/ebiom

EBioMedicine
Published by THE LANCET

Detection of HIV transmission hotspots in British Columbia, Canada: A novel framework for the prioritization and allocation of treatment and prevention resources

Angela McLaughlin[a,b,c], Paul Sereda[a], Natalia Oliveira[a], Rolando Barrios[a,b], Chanson J. Brumme[a], Zabrina L. Brumme[a,d], Julio S.G. Montaner[a,e], Jeffrey B. Joy[a,e,*]

[a] *British Columbia Centre for Excellence in HIV/AIDS, University of British Columbia Department of Medicine, 608-1081 Burrard Street, Vancouver, BC, V6Z 1Y6, Canada*
[b] *School of Population and Public Health, University of British Columbia, Vancouver, BC, Canada*
[c] *Bioinformatics, University of British Columbia, Vancouver, BC, Canada*
[d] *Faculty of Health Sciences, Simon Fraser University, Burnaby, BC, Canada*
[e] *Division of Infectious Diseases, Department of Medicine, University of British Columbia, Vancouver, BC, Canada*

ARTICLE INFO

Article history:
Received 26 June 2019
Accepted 14 September 2019
Available online 15 October 2019

Keywords:
HIV
Phylogenetics
Epidemiology
Transmission
Predictive model

ABSTRACT

Background: Identifying populations at high risk of HIV transmission is critical for prioritizing treatment and prevention resources and achieving the UNAIDS 90-90-90 Targets.
Methods: HIV transmission rates can be estimated from phylogenetic trees as viral lineage-level diversification rates. To identify HIV-1 transmission foci in British Columbia, Canada, we inferred diversification rates from phylogenetic trees of 36 271 HIV-1 sequences from 9630 anonymized individuals. Diversification rates were combined with sociodemographic and clinical data, then aggregated by patients' area of residence to predict the distribution of new HIV cases between 2008 and 2018. The predictive power of the model was compared with a phylogenetically uninformed model.
Findings: Aggregated diversification rate measures were predictive of new HIV cases in the subsequent year after adjusting for prevalent and incident cases in the previous year. For every one-unit increase in the mean of the top five diversification rates, the number of new HIV cases increased by on average 1·38-fold (95% CI, 1·28–1·49). In a blind prediction of 2018 cases, diversification rate improved the model's specificity by 12%, accuracy by 9%, top 20 agreement by 100%, and correlation of predicted and observed values by 162% relative to a model that incorporated epidemiological data alone.
Interpretation: By predicting the distribution of future HIV cases, a combined phylogenetic and epidemiological approach identifies hotspots where public health resources are needed most.
Funding: Canadian Institutes of Health Research, University of British Columbia, Public Health Agency of Canada, Genome Canada, Genome BC, Michael Smith Foundation for Health Research, and BC Centre for Excellence in HIV/AIDS.

Research in context

Evidence before this study

We searched PubMed for articles published between January 1996 and December 2018, to limit the search to transmission modeling in the HAART era, in any language that applied phylogenetic methods to infer HIV transmission history using the search terms "transmission", "phylogenetic", and "HIV". The large majority of studies applied some form of genetic clustering to identify sub-populations exposed to higher than average transmission rates. Although HIVTRACE is the most commonly applied clustering method using a threshold for pairwise TN93 genetic distances, there is no strong consensus in the field that this clustering method is the most appropriate. Other clustering methods identify epidemiologically linked sequences using patristic (tree) distance, apply subtree pruning, or have minimum bootstrap support requirements. Previous studies have compared the performance of clustering methods applied to both empirical and simulated epi-

* Corresponding author.
E-mail address: jjoy@cfenet.ubc.ca (J.B. Joy).

https://doi.org/10.1016/j.ebiom.2019.09.026

demics. Clustering methods varied in their ability to distinguish differences in sampling and transmission within the population. Further, the membership, size, and growth of inferred clusters differed between methods.

A limited number of studies have explored alternative methods such as a Markov-modulated Poisson Process and a multitype birth-death branching model process, respectively, to infer the evolution of HIV transmission rates within a phylogeny. However, these methods rely on strong assumptions that transmission rates evolve as discrete states, which is unlikely given the heterogeneity of individuals' risk activities, time to diagnosis, and travel patterns. There remains room for improvement in resolving HIV transmission activity in order to prioritize treatment and prevention services most efficiently.

Lineage-level diversification rate has been previously applied to quantify species' evolutionary relatedness in a conservation biology framework, yet we found no studies that applied diversification rate to quantify the evolutionary relatedness (and thus epidemiological connectedness) of viral lineages across the history of the HIV epidemic, or for any other pathogens for that matter.

Added value of this study

We combined viral diversification rate with clinical and epidemiological data to predict where and how many new HIV cases would arise in the future within British Columbia. This is a novel and pragmatic approach, which lent greater predictive power than using epidemiological data alone. We have demonstrated that geographically-aggregated viral diversification rate is a robust proxy for transmission rate that is without biases to clustering thresholds, while still protecting patient confidentiality. Viral diversification rate could also be employed to model the within- and between-host evolution of other pathogens, as well as the HIV epidemic in resource-limited settings.

Implications of all the available evidence

Spatiotemporal modeling of historic HIV transmission and epidemiological dynamics predicted the future trajectory of the HIV epidemic. Such predictive models should be utilized by public health authorities to target areas for prioritized treatment and prevention services. Future evaluations of the relative predictive power and sampling sensitivities of clustering, diversification rates, and combined methods are required.

1. Introduction

Although highly active antiretroviral therapy (HAART) has led to sustained decreases in HIV-related morbidity, mortality, and incidence [1], HIV transmission foci remain, even in developed countries [2,3]. Identifying areas at high risk of ongoing HIV transmission is critical for the efficient allocation of HIV prevention and treatment services to meet the UNAIDS 90-90-90 Targets [4]. Population-level HIV drug resistance genotyping datasets, collected for strain surveillance and during routine clinical care, provide a unique molecular window into viral transmission [3]. Indeed, phylogenetic inference of HIV transmission hot spots from existing resistance genotyping datasets represents a cornerstone of the United States' new strategy to end HIV [5].

For rapidly evolving pathogens like HIV, a well-informed molecular phylogeny can provide an estimate of the between-host transmission tree [6,7]. Phylogenetic clustering of HIV sequences is commonly used to identify groups of individuals implicated with higher transmission rates, information that is in turn used to prioritize HIV prevention services [2,8,9]. However, no standard cluster detection method has been agreed upon and cluster membership can vary dramatically between methods [10,11].

Alternatively, the transmission history of each virus in a phylogenetic tree can be estimated by its lineage-level diversification rate, defined as a tip's splitting rate across the entire tree path, weighted from tip to root [12,13]. Since transmission of HIV to a new host is equivalent to the formation of a new lineage, diversification rates inferred from between-host phylogenies can approximate historical transmission rates. In previous work, we reported that treatment experienced lineages displayed dramatically reduced HIV diversification rates, providing an independent validation of Treatment as Prevention® (TasP®) strategies [14]. Aggregating viral diversification rates by patients' geography of residence should illuminate areas with concentrated transmission activity.

Using longitudinal data collected from participants in the British Columbia Centre for Excellence in HIV/AIDS (BC-CfE) Drug Treatment Program (DTP), we modeled the spatiotemporal distribution of HIV lineage-level diversification rate in BC to predict where new HIV cases arise over time. We hypothesized that a model incorporating geographically-aggregated viral diversification rates would predict the location of new HIV cases significantly better than one incorporating clinical and epidemiological characteristics alone.

2. Methods

2.1. Study setting and participants

Between May 1996 and March 29 2018, 9630 HIV-infected individuals had drug resistance testing performed through the BC-CfE's Drug Treatment Program (DTP). Established in 1992, the DTP is an open treatment and research cohort that provides all medically eligible HIV-infected British Columbians with access to personalized HAART and related laboratory monitoring at no cost [15]. All HIV-infected BC residents are eligible whether or not they access HAART. Baseline and follow up plasma samples are routinely collected for viral load monitoring and HIV drug resistance testing; follow up continues until death or emigration from BC. Ethical approval for this study was granted by the University of British Columbia - Providence Health Care Research Ethics Board (H17-01812). No further samples were collected nor were additional patients recruited.

2.2. Clinical, epidemiological, and demographic variables

In addition to sequences, available data for this analysis included sample collection date; treatment initiation date; treatment regimen; date of first viral load; physician estimated date of seroconversion; plasma viral load (HIV RNA copies/mL); HIV subtype classifications (generated using the SCUEAL algorithm) [16]; ethnicity; birth year; sex at birth; sex; self-reported risk factors (injection drug use, men who have sex with men, heterosexual contact, any receipt of blood product or exposure to blood risk, other risk exposure); having ever tested positive for hepatitis C infection; having ever had acquired immune deficiency syndrome (AIDS); if applicable date of mortality and cause of death; local health authority, health authority, postal code, census tract (CT), and census metropolitan area (CMA) of patient residence; and forward sortation area (first three digits of postal code) of physician requesting test. Baseline CD4 counts were not available for this analysis. The viral load assay detection limit changed several times prior to 2008, so for consistency, only data from 2008 onwards was analyzed in the final model. The detection limit over that period was 40–10 000 000 copies viral RNA/mL plasma using a Roche COBAS HIV-1 Ampliprep Taqman assay [17]. To suit the longitudinal nature of the analysis, patient data was carried over year to year until new data was available (ie more recent viral load), unless participants were removed due to death or migration out of province. HIV sequences and patient information were stored in a secure Oracle database in access-restricted facilities at the BCCfE. Patient data were de-identified and doubly anonymized with randomly generated 6-character identifiers. No document was created linking identifiers to the patient.

2.3. Phylogenetic inference

A total of 36 271 HIV resistance genotype tests from 9630 people living with HIV (PLHIV) were completed (mean, 3·77; median, 2; range 1–46 per individual). Genotypes comprised HIV *protease* and partial *reverse transcriptase* sequences, hereafter referred to as partial *pol*. See Supplementary Appendix for details on the sequence data. Sequences were aligned to the HXB2 reference genome (GenBank Accession #K03455) using MAFFT version 7.310 [18]. The alignment was visualized and curated in AliView V1.17.1 [19]. Insertions and deletions relative to HXB2, as well as amino acids corresponding to WHO recognized drug resistance mutation sites, were removed from the alignment [20]. A set of shuffled bootstrap alignments were generated to infer 100 approximate maximum likelihood phylogenetic trees (in units of substitutions per site), implemented in FastTree2.1 [21]. Subsequently, trees were pruned to include patients' oldest samples and then rooted using root-to-tip regression in the R package ape version 5.0 [22]. Time interval trees for each study year were generated, which only included sequences from that year or earlier, such that each year the trees only had tips added for newly infected (or newly connected with care) participants (Fig. S3).

2.4. Diversification rate

For each tip on a rooted bifurcating (time interval bootstrap) tree, the viral lineage-level diversification rate (DR_i) was calculated as the reciprocal sum of N_i branch lengths (l_j) from tip i to the root, with each consecutive edge (j) down-weighted by a factor of 1/2 [12].

$$Diversification\ Rate_i = Equal\ Splits_i^{-1} = \left(\sum_{j=1}^{N_i} \frac{l_j}{2^{j-1}}\right)^{-1}$$

The annual change in diversification rate for each lineage over time was calculated as the difference between the lineage's mean diversification rate across bootstraps for a given year and the preceding year, with a value of zero for the first year a patient entered the cohort. A lineage's diversification rate would increase if a new virus was added to the tree in close phylogenetic proximity. Diversification rate, its distribution over time, and the robustness of estimates to different tree-building algorithms are detailed in the Supplementary Appendix.

2.5. Predictive modeling

Patient attributes and lineage-level diversification rates were aggregated by their census tract of residence. Census tracts are small geographic areas that usually have a population under 10 000, which represent subdivisions of census agglomerations with core populations of ≥ 50 000 in the previous census [23]. Further details of patient attributes, aggregated variables, and geographic analysis are presented in the Supplementary Appendix. The modeling outcome was the number of new HIV cases in each census tract in the subsequent year, as estimated by date of first detectable viral load (see Supplementary Appendix for outcome validity discussion). The viral load assay detection limit changed several times prior to 2008, so for consistency, only data from 2008 onwards was analyzed in the final model. Since the outcome a discrete count with a large percentage of zero values and some overdispersion (Fig. S2), a zero-inflated negative binomial (ZINB) model was deemed most appropriate [24]. The ZINB model has two parts: a binomial model to predict whether there were greater than zero new cases and a negative-binomial model to predict the number of cases if cases were greater than zero. Exponentiated coefficients in each model were interpreted as adjusted odds ratios and adjusted relative risks, respectively. A hold out cross-validation supervised machine learning algorithm was applied to train and test the model (Fig. S7). The data was segregated into a training subset (2009, 2010, 2012, 2013, 2015, and 2016) and a testing subset of interspersed years (2008, 2011, and 2014). The 2017 dataset (2017 predictors, 2018 new cases) was reserved for a final blind prediction and the full dataset for 2018 was obtained after the final model was selected. To assess how informative diversification rate measures were in predicting new HIV cases, the full ZINB model was reduced to a nested model excluding any phylogenetic measures (referred to as reduced ZINB). The goodness-of-fit of these models, a full zero-inflated Poisson (ZIP), and a full Hurdle model were compared in Table S3. For both the testing dataset and the blind prediction dataset, the predictive fit of the full ZINB and reduced ZINB were compared in terms of their Pearson's correlation coefficient of predicted and observed values (for observed values greater than zero); sensitivity; specificity; positive predictive value; negative predictive value; accuracy; and top 20 agreement, which is the proportion of census tracts predicted as being among the top twenty for highest number of new HIV cases that were observed (Fig. 3).

2.6. Spatial autocorrelation

In order to assess whether there was significant spatial autocorrelation in our dataset, we calculated a global Moran's I for the training subset of the outcome data as well as for the final model residuals, then tested whether it was significantly greater than expected using a Monte Carlo simulation [25]. Spatial weights were equally divided among neighbors that shared at least one vertex with each other. To further evaluate if the inclusion of spatial autocorrelation in the model would improve the goodness of fit or predictive power, we built a Bayesian hierarchical zero-inflated Poisson model with conditionally autoregressive (CAR) priors using the R package, CARBayes [26]. CAR priors apply a binary neighbor-based matrix, which is appropriate for irregular lattice data, as in the case of census tract level spatial data [27,28]. The CAR priors were modeled using a single set of random effects, as described by Leroux et al. [29] A Markov Chain Monte Carlo simulation (burn in: 100 000; chain length: 300 000; thinning level 30) was applied to search the posterior space. The goodness of fit and posterior coefficient estimates were compared to the final ZINB model.

3. Results

3.1. Study population

Of the 13 431 cumulative Drug Treatment Program participants enrolled up to March 29 2018, 9630 (72%) individuals were represented in the HIV sequence database and 6944 (52% overall) of them reported their baseline census tract of residence or postal code and were included in the final analysis (Table 1). The analyzed population was comparable to the DTP population with sequences in most regards including the distribution of sexes (male: 81·3% among DTP participants with sequences, 83·3% among analyzed population); dominant HIV subtype (subtype B: 92·4%, 92·5%); median baseline age (38, 38); and median baseline diversification rate (52·5 (subs/site)$^{-1}$, 52·0 (subs/site)$^{-1}$). The study

Table 1
Characteristics of the Drug Treatment Program cohort with sequences compared to the study population restricted to those who have reported a census tract of residence.

		DTP cohort with sequences			Study population		
Parameter	Characteristic	N*	Total	%	N*	Total	%
Total participants		9630	9630	100	6944	6944	100
Sex	**Male**	8861	7208	81·3	6634	5525	83·3
	Female	8861	1592	18·0	6634	1053	15·9
	Male-Female**	8861	51	0·6	6634	47	0·7
	Female-Male**	8861	10	0·1	6634	9	0·1
Sex at Birth	**Male**	8861	7259	81·9	6634	5572	84·0
	Female	8861	1602	18·1	6634	1062	16·0
Subtype	**B**	9630	8895	92·4	6944	6420	92·5
	C	9630	361	3·7	6944	254	3·7
	AE	9630	163	1·7	6944	118	1·7
	A	9630	85	0·9	6944	59	0·8
	AG	9630	60	0·6	6944	43	0·6
	Other	9630	34	0·4	6944	26	0·4
	D	9630	32	0·3	6944	24	0·3
Health Authority	**Vancouver Coastal**	8459	4923	58·2	6589	4217	64·0
	Fraser	8459	1894	22·4	6589	1581	24·0
	Vancouver Island	8459	858	10·1	6589	475	7·2
	Interior	8459	478	5·7	6589	191	2·9
	Northern	8459	306	3·6	6589	125	1·9
Self-reported risk factors***	**Injection drug use**	5500	3175	57·7	3948	2146	54·4
	Men who have sex with men	6874	3487	50·7	5051	2783	55·1
	Heterosexual contact	6874	2201	32·0	5051	1562	30·9
	Other	6874	308	4·5	5051	214	4·2
	Exposure to blood products	6874	225	3·3	5051	161	3·2
Self-reported ethnicity***	**White**	5753	3944	68·6	4284	2998	70·0
	First Nations	5753	1275	22·2	4284	811	18·9
	Asian	5753	435	7·6	4284	378	8·8
	Black	5753	247	4·3	4284	192	4·5
	Hispanic	5753	221	3·8	4284	176	4·1
Clinical illness	**Ever had hepatitis C virus**	8318	3268	39·3	6247	2265	36·3
	Ever had AIDS	8861	1074	12·1	6634	824	12·4
	Died	7231	1527	21·1	5283	966	18·3
Baseline measures		**N***	**Median**	**IQR**	**N***	**Median**	**IQR**
	Baseline diversification rate (subs/site)$^{-1}$	9630	52·5	34·4–89·1	6944	52	34·3–88·3
	Baseline log10 (viral load) $\log_{10}$(viral RNA/mL plasma)	9630	4·5	3·8–5·0	6944	5·3	4·5–7·0
	Baseline age	9624	38	31–46	6916	38	31–45

* N = the number of participants who provided information for a given characteristic.

** Identifying as transgender was self-reported. Male-Female indicates an individual who was born a male and transitioned to a female, and Female-Male indicates an individual who was born a female and transitioned to a male.

*** Patients may report multiple risk factors and ethnicities.

population underrepresented participants from the Vancouver Island (10·1%, 7·2%), Interior (5·7%, 2·9%), and Northern (3·6%, 1·9%) Health Authorities, and First Nations people (22·2%, 18·9%). Additionally, the median baseline viral load was somewhat lower in the DTP population with sequences than among those who reported a census tract (4·5, 5·3 $\log_{10}$ copies/mL). Although patient characteristics for the entire Drug Treatment Program population were not available for comparison to the study group, estimated risk exposure characteristics of the prevalent HIV population in 2014 from the British Columbia Centre for Disease Control were comparable to the study population in terms of, for instance, the percent of PLHIV who identified as MSM (BCCDC, 49%; study population, 55.1%), with the caveat that we are comparing different years [30].

3.2. Diversification rate

The distribution of lineage-level diversification rates across a representative approximate maximum likelihood tree from 2018 is presented in Fig. 1. Lineages associated with rapid and extensive branching are typified by high diversification rates. We showed that our diversification rate calculations were robust to tree-building methods and HIV genomic region (Supplementary Appendix). Although there was a slight increase in both mean and median diversification rate over time attributable to sampling from a larger tree (see Supplementary Appendix), the population-level distribution of diversification rates did not significantly change across the study period (Kruskal-Wallis, $p = 0{\cdot}385$).

3.3. Predictive modeling

The unadjusted relationships of aggregated variables with the outcome were calculated to inform their model inclusion (Table S2). Among the highest Pearson correlation coefficients were new HIV cases in the previous year ($r = 0{\cdot}74$, $p < 0{\cdot}001$), the total number of PLHIV with changes in diversification rate $\geq$ 1 (subs/site)$^{-1}$ ($r = 0{\cdot}72$, $p < 0{\cdot}001$), the sum of PLHIV ($r = 0{\cdot}69$, $p < 0{\cdot}001$), and the mean of the top 5 diversification rates ($r = 0{\cdot}48$, $p < 0{\cdot}001$). After identifying a final ZINB model, a likelihood ratio test confirmed that it fit the data significantly better than its Poisson equivalent ($p < 0{\cdot}001$), Hurdle equivalent ($p < 0{\cdot}001$), and the reduced ZINB without phylogenetic measures ($p < 0{\cdot}001$) (Table S3). Further, the AIC, BIC, and log-likelihood values for the aforementioned models were all lower for the full ZINB model.

The mean of the top five ln(diversification rates) was significant in both the binomial and negative binomial parts of the model (Table 2). The adjusted odds of a census tract having greater than zero new HIV cases in the subsequent year was 3·10 (95% CI, 1·61–5·97) times higher for every one-unit increase in the mean of the top five ln(diversification rates). In census tracts with greater than zero new HIV cases, for every one-unit increase in the mean of the

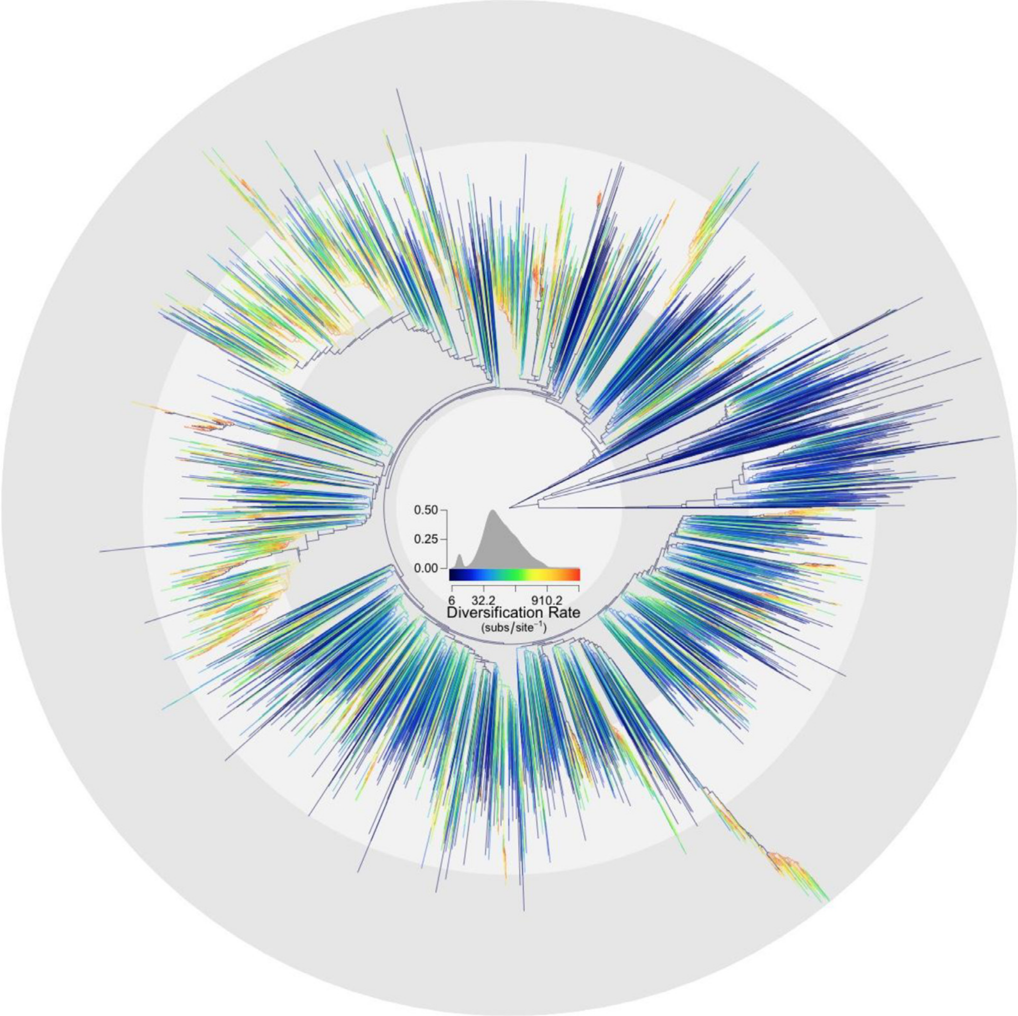

Fig. 1. A representative bootstrap approximate maximum likelihood phylogenetic tree of baseline HIV sequences available for all participants ($n = 9630$) as of March 2018 with branches colored by the lineage-level diversification rate. Cooler colors represent slower diversification rates while warmer colors represent faster diversification. Grey concentric rings qualitatively distinguish lineages that have diverged the most from the root.

Table 2
The final zero-inflated negative binomial (ZINB) predictive model is a two-part model composed of a binomial model to predict whether there were greater than zero new cases, and a negative-binomial model to predict the number of cases, if cases were greater than zero.

Binomial {0, >0} model	Adjusted odds ratio	95% CI	p-value
Total PLHIV	0·81	0·74–0·88	<0·001
Mean of top five ln(diversification rate)	3·10	1·61–5·97	<0·001

Negative binomial {count if >0} model	Adjusted relative risk	95% CI	p-value
Total new cases	1·07	1·03–1·11	<0·001
Mean of top five ln(diversification rate)	1·38	1·28–1·49	<0·001
Total # PLHIV with change in diversification rate ≥ 1	1·10	1·07–1·13	<0·001

top five ln(diversification rates) in that tract, the adjusted risk of new HIV cases in the subsequent year was 1·38 (95% CI, 1·28–1·49) times higher. Moreover, for every additional PLHIV with change in diversification rate ≥ 1 (subs/site)$^{-1}$ in a census tract, the adjusted risk of new HIV cases in the subsequent year was 1·10 (95% CI, 1·07–1·13) times higher. The spatiotemporal distribution of new HIV cases in downtown Vancouver across the study period was heterogeneous and was collectively predicted by the distribution of significant predictor variables (Fig. 2).

Comparing the predictive fit of the ZINB model to its reduced equivalent for the testing dataset, the mean Pearson correlation coefficient of predicted and observed values for the full ZINB model of 0·77 was 17% higher than that of the reduced ZINB model at 0·66 (Fig. 3; Table S4). Interestingly, the reduced ZINB model had a 14% higher mean sensitivity than the full ZINB model (0·75, 0·66), but the full ZINB model had an 11% higher mean specificity (0·68, 0·61) due to the reduced ZINB producing more false positives. The mean accuracy was the same for both models (0·67). The top 20 agreement (proportion of census tracts predicted to be among the top 20 highest number of new HIV cases that were observed) was 28% higher for the full ZINB model (0.68) compared to the reduced model (0.53).

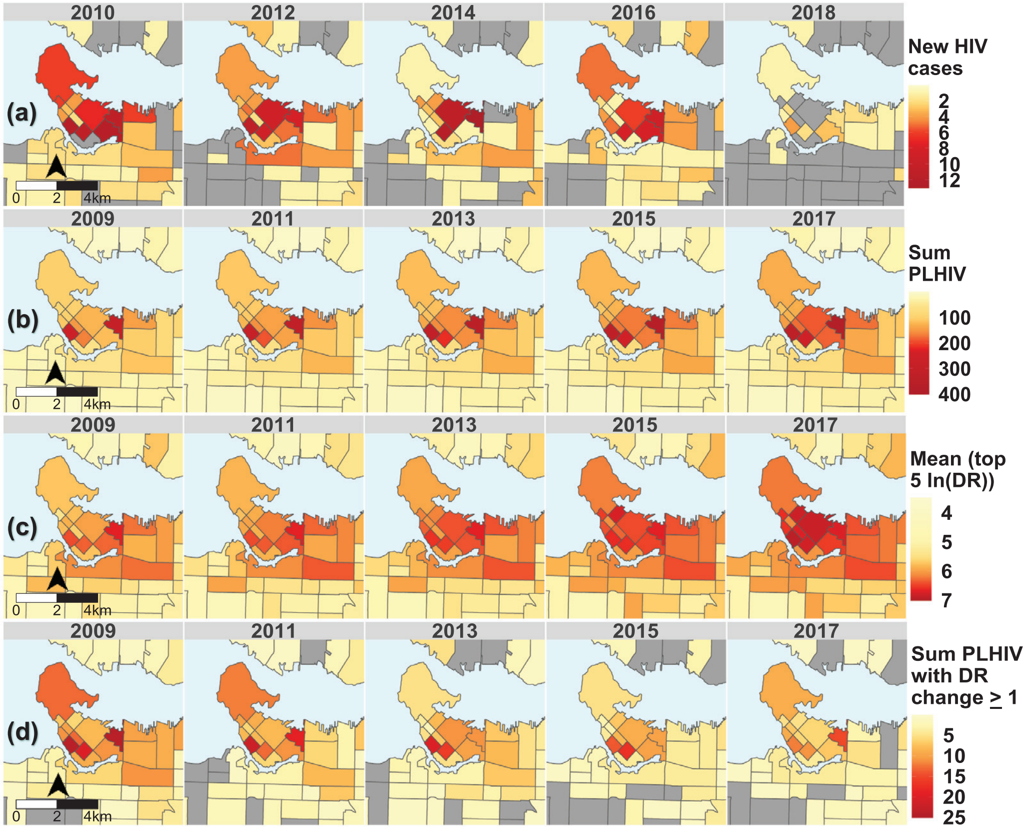

Fig. 2. A combination of epidemiological and phylogenetic variables predicted the spatiotemporal distribution of new HIV cases in downtown Vancouver, BC in the subsequent year. Only predictor values for odd study years between 2008 and 2017 with corresponding new HIV cases in the subsequent years were shown for conciseness, however even years were also included in the analysis. The outcome, (**a**) total new HIV cases, was collectively predicted by (**b**) the total prevalent cases of PLHIV, (**c**) the mean of the top five ln(diversification rates, DR), and (**d**) the sum of PLHIV with annual changes in diversification rate (DR) ≥1 (subs/site)$^{-1}$. Grey census tracts have values of zero.

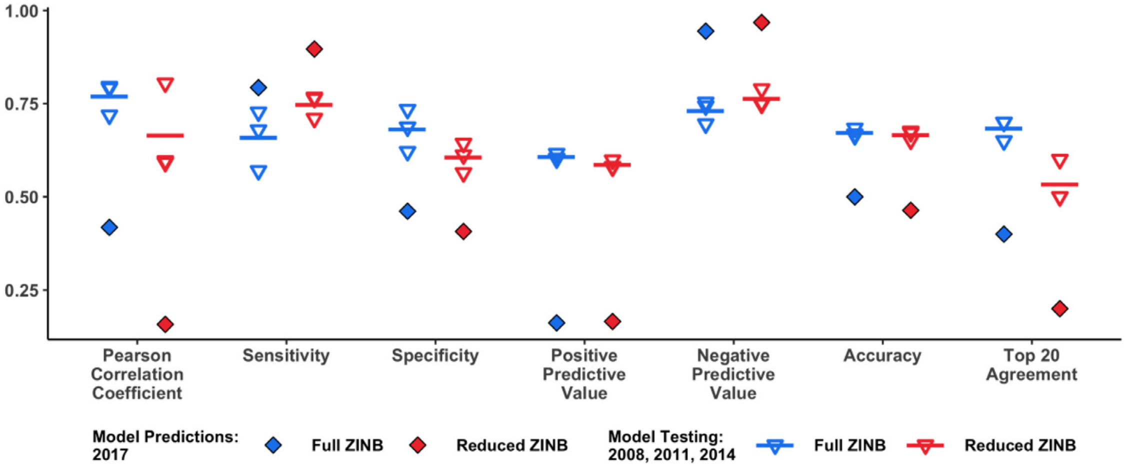

Fig. 3. A comparison of the predictive power of the full ZINB and ZINB without phylogenetic measures (reduced ZINB). Hollow triangles represent testing data subset values and are summarized by their mean, while the diamonds illustrate the predictive values for the blind prediction of 2018 new HIV cases based on 2017 predictors. Criteria considered for predictive power include the Pearson's correlation coefficient (for observed cases greater than zero to remove the effect of zero-inflation); sensitivity; specificity; positive predictive value; negative predictive value; accuracy; and the top 20 agreement.

Of the 201 new HIV cases with sequences in 2018, only 64 had reported census tracts or postal codes in BC by February 4 2019, a markedly lower percent than previous years (Table S5). Regardless, observed new HIV cases in 2018 were compared to blind predictions from the model using 2017 predictors (Fig. 3, Fig. S6). The Pearson correlation coefficient between observed and predicted number of HIV cases for the full ZINB model was markedly better than the reduced model by 162% (0·42, 0·16). The sensitivity for both models was high, although the reduced model had a 14% higher sensitivity of 0·90 compared to 0·79 in the full model. The full ZINB model had a 12% higher specificity than the reduced model (0·46, 0·41), a 9% higher accuracy than the reduced ZINB model (0·50, 0·46), and a 100% higher top 20 agreement (0.4, 0.2).

3.4. Spatial autocorrelation

A Markov Chain permutation test for global Moran's I revealed that although there was significant spatial autocorrelation within the outcome ($p<0{\cdot}001$), the final ZINB model residuals were not significantly spatially autocorrelated ($p=0{\cdot}086$). To further investigate if explicit incorporation of spatial autocorrelation could improve the model, a Bayesian hierarchical zero-inflated Poisson model with conditionally autoregressive (CAR) priors was constructed. The log likelihood of the CAR model (−1718) was poorer than that of the final ZINB model (−1547) and the AIC value of the CAR model (3108) was nearly indistinguishable from the final ZINB model (3109). Coefficient estimates from the CAR model (Table S6) overlapped greatly with the final ZINB model, further corroborating that the exclusion of a spatial autocorrelation term in the final ZINB model did not artificially inflate coefficient errors.

4. Discussion

Our findings revealed that geographically-aggregated HIV lineage-level diversification rate, supplemented with clinical and epidemiological data, better predicted the number and location of future new HIV cases across BC than clinical and epidemiological data alone. Models that predict where new HIV cases will arise are valuable to public health authorities for prioritizing treatment and prevention services to areas of greatest need.

Multiple measures of aggregated diversification rate were significantly correlated with new HIV cases in the subsequent year. After adjusting for prevalent and incident cases in the previous year, the mean of the top five diversification rates and the number of individuals with large annual changes in diversification rate were both predictive of new HIV cases. This supports the utility of lineage-level diversification rates as estimates of HIV transmission and suggests that the communities where actively transmitting PLHIV reside tend to be those where new HIV cases arise. Though initially counterintuitive, increases in the prevalent PLHIV population of a census tract decreased the risk of new HIV cases in the subsequent year. This was likely because the census tracts with the highest prevalence were well known to public health officials and already had significant prevention and treatment programs in place. Additionally, there may be some element of transmission saturation whereby areas with previously high incidence may now have fewer susceptible people in the population who have never been infected, reducing the local transmission rate. Counter to previous studies [17,31], we did not find that any measures of community viral load were predictive of where new HIV cases arose after correcting for other variables, however this could be because our study group was limited to those with resistance tests and reported census tracts.

The full model exhibited an improved correlation of observed and predicted new HIV cases and a higher specificity in both the testing and blind prediction data relative to the reduced model, whereas the reduced model had a higher sensitivity than the full model. There is a trade-off between the two models, as the reduced model detected more true positives, but also more false positives, than the full model. Including diversification rate improved the positive predictive rate, or the proportion of true positives among all detected positives. In the case of allocating limited public health care resources to geographic communities, identifying false positive areas, that in reality will not have any new HIV cases arise within them, as hot spots would waste limited resources. The specificity, in this case, is arguably more important because it permits public health authorities to sift out low priority areas (true negatives) to focus on high priority areas without wasting any resources. While the differences in the models' predictive powers were modest, they demonstrated that the inclusion of diversification rate measures was informative to the number and location of new HIV cases in the subsequent year. Further optimization of variable selection could improve its predictive power further. The 2018 blind predictions were somewhat hindered by the low percent of new HIV cases with sequences who also had census tract information. The dataset we were working with lacked information related to patients' sexual behavior, such as marital status, frequency of sexual activity, or consistency of condom use. Incorporating these factors into the model could have further improved the model's predictive power, however their inclusion could also threaten to over parameterize the model.

Spatial autocorrelation is a concern when fitting models to spatial data, as the assumption of independent observations may not be met and can lead to overestimates of statistical significance [32]. The Moran's tests suggested that while the number of new HIV cases was spatially autocorrelated between adjacent communities, this spatial dependence was accounted for in the model through our included predictors. By including the number of new cases in the previous year as a predictor, for example, we informed the model where values were likely to be similar. A brief comparison of the final ZINB model to a Bayesian hierarchical model that accounted for spatial autocorrelation revealed that the models did not differ greatly in their goodness of fit or coefficient estimates. Future improvements to the model could consider more complex measurements of adjacency and proximity.

Restricting analyses to individuals who resided in a census tract could theoretically introduce geographic sampling bias by excluding those who were in small rural communities not covered by census tracts and those who had insecure housing. However, we found that participants with reported census tracts were comparable to the DTP group with sequences available. A greater sampling bias was likely derived from those who were disconnected from care entirely, as undiagnosed PLHIV might disproportionately contribute to HIV transmission [33]. Individuals in well-serviced communities with strong health seeking behaviors were more likely to be connected with care, diagnosed, and have an associated viral sequence. In 2014, an estimated 83% of PLHIV in BC were diagnosed [4]; nevertheless efforts should be made to engage hard-to-reach individuals in care in a way that is non-threatening, inclusive, and feasible. Furthermore, we assumed that our study population did not migrate between census tracts within BC during the study period. Although this assumption is surely not true for all individuals, our dataset was restricted to the geography of patient residence at the first viral load test. Migration between census tracts would add noise to the model, leading to an underestimation of the added predictive power of phylogenetic metrics.

Other attempts have been made to quantify the evolution of pathogen transmission rates across phylogenies [34,35]. These methods fundamentally differ from the methodology described here as these tree-wide estimates assume that the transmission rates evolve as a discrete state according to either a Markov-modulated Poisson process [34] or a multi-type birth–death branching model process [35], in contrast to our lineage-level estimates, which do not rely on assumptions about the evolution of rates.

It is imperative that researchers in genomic epidemiology mitigate the risk of patient identification [36]. In this study, protection of patients' privacy and confidentiality was achieved by merging census tracts until no fewer than three PLHIV resided in the final merged census tract at any point in time. Another reason for aggregating multiple census tracts is that rates computed using small populations at risk are relatively unstable [37].

In future work, we intend to validate the model on other HIV-afflicted geographies with lower sampling coverage, although the model would likely have to be re-trained to be applicable to widely different localized epidemics. Thus, the model's external validity lies in its premise, while the relationship between geographically-aggregated phylogenetic, clinical, and epidemiological data and the resulting number and location of new HIV cases will likely differ somewhat by geography. Simulating an epidemic could also be useful to assess the effect of down-sampling on the distribution and interpretation of diversification rates. Further, calculating the lineage-level diversification rates of individuals within clusters could increase the resolution of existing prioritization schemes. Lineage-level diversification rate may also represent a useful metric for transmission of other rapidly-evolving pathogens.

Viral lineage-level diversification rates approximate HIV transmission and when aggregated by individuals' geography of residence revealed high priority transmission hot spots. Phylogenetic analyses of pre-existing HIV genotyping data can refine predictions of where public health resources will have the greatest impact.

Funding sources

The authors have received funding from Canadian Institutes of Health Research, University of British Columbia, British Columbia Health Officers Council, Public Health Agency of Canada, Genome Canada, Genome BC, Michael Smith Foundation for Health Research, and BC Centre for Excellence in HIV/AIDS. Besides the BC Centre for Excellence, none of the funding sources were involved in data analysis or interpretation, nor patient recruitment. The collection of data for the Drug Treatment Program is overseen by the BC Centre for Excellence in HIV/AIDS, Providence Health Care, University of British Columbia, and Simon Fraser University representatives. The corresponding author, JBJ, and AM had full access to all data and had final responsibility for the decision to submit for publication.

Author contributions

JBJ and AM conceived and designed the study. With the guidance of JBJ, AM conducted the literature search. The initial study proposal was presented to RB, PS, NO, and CJB, who helped to refine the scope of the analysis and clarify internal data availability. CJB, ZLB, and the clinical lab team at the BC-CfE generated the sequence data, while the Drug Treatment Program team headed by RB assembled the accompanying demographic, risk factor, and clinical patient data. PS, RB, and NO processed and fulfilled the data request. NO helped to clarify the data dictionary and ran preliminary data checks. AM and JBJ developed the methods, analyzed the data, and interpreted the findings. AM wrote the original draft of the manuscript with the guidance of JBJ. AM, JBJ, JSGM, PS, and ZLB contributed to manuscript review and editing.

Declaration of competing interest

AM was supported with a CIHR Canada Graduate Scholarship Masters award, a UBC Faculty of Medicine Dorothy Helmer award, and the Dr. Ken Benson Memorial award through the Health Officers Council of BC. ZLB is supported by a Scholar Award from the Michael Smith Foundation for Health Research, and has also received grants, paid to her institution, by the Canadian Institutes of Health Research (PJT-148621 and PJT-159625) and the US National Institutes of Health (R21A127029). JBJ is funded by a Genome Canada, Genome BC Bioinformatics and Computational Biology grant - 287PHY, and the Public Health Agency of Canada. JSGM is supported with grants paid to his institution by the British Columbia Ministry of Health and by the US National Institutes of Health (r01da036307); he has also received limited unrestricted funding, paid to his institution, from Abbvie, Bristol-Myers Squibb, Gilead Sciences, Janssen, Merck, and ViiV Healthcare. RB, NO, PS, and CJB declare no competing interests.

Acknowledgments

We thank our many colleagues at the BC Centre for Excellence in HIV/AIDS, particularly the clinical laboratory, bioinformatics pipeline, and Drug Treatment Program teams. We would also like to thank P. Richard Harrigan for his contributions to the laboratory and study inception. All inferences, opinions, and conclusions drawn in this paper are those of the authors, and do not reflect the opinions of the funders.

Supplementary material

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.ebiom.2019.09.026.

References

[1] Montaner JSG, Lima VD, Harrigan PR, Lourenço L, Yip B, Nosyk B, et al. Expansion of HAART coverage is associated with sustained decreases in HIV/AIDS morbidity, mortality and HIV transmission: the "HIV treatment as prevention" experience in a Canadian setting. Sluis-Cremer N, editor. PLoS ONE 2014;9(2):e87872.
[2] Poon A, Gustafson R, Daly P, Zerr L, Demlow SE. Near real-time monitoring of hiv transmission hotspots from routine HIV genotyping: an implementation case study. The Lancet HIV 2016;3(5):e231–8.
[3] Poon AFY, Joy JB, Woods CK, Shurgold S, Colley G, Brumme CJ, et al. The impact of clinical, demographic and risk factors on rates of HIV transmission: a population-based phylogenetic analysis in British Columbia, Canada. J Infect Dis 2015;211(6):926–35.
[4] Lima VD, St-Jean M, Rozada I, Shoveller JA, Nosyk B, Hogg RS, et al. Progress towards the United Nations 90-90-90 and 95-95-95 targets: the experience in British Columbia, Canada. J Int AIDS Soc. 3rd ed. 2017;20(3):e25011.
[5] Fauci AS, Redfield RR, Sigounas G, Weahkee MD, Giroir BP. Ending the HIV epidemic. JAMA. Am Med Assoc 2019;321(9):844–5.
[6] Holmes EC, Nee S, Rambaut A, Garnett GP, Harvey PH. Revealing the history of infectious disease epidemics through phylogenetic trees. Philosophical Transactions of the Royal Society B: biological Sciences. Roy. Soc. 1995;349(1327):33–40.
[7] Leitner T, Escanilla D, Franzen C, Uhlen M, Albert J. Accurate reconstruction of a known HIV-1 transmission history by phylogenetic tree analysis. Proc Natl Acad Sci USA. Natl Acad Sci 1996;93(20):10864–9.
[8] Wertheim JO, Murrell B, Mehta SR, Forgione LA, Kosakovsky Pond SL, Smith DM, et al. Growth of HIV-1 molecular transmission clusters in New York City. J Infect Dis 2018;5 e1000590–11.
[9] Little SJ, Kosakovsky Pond SL, Anderson CM, Young JA, Wertheim JO, Mehta SR, et al. Using HIV networks to inform real time prevention interventions. Harrigan PR, editor. PLoS ONE Public Library of Science 2014;9(6):e98443–8.
[10] Poon AFY. Impacts and shortcomings of genetic clustering methods for infectious disease outbreaks. Virus Evol 2016;2(2) vew031–9.
[11] Ratmann O, Hodcroft EB, Pickles M, Cori A, Hall M, Lycett S, et al. Phylogenetic tools for generalized HIV-1 epidemics: findings from the PANGEA-HIV methods comparison. Mol Biol Evol 2017;34(1):185–203.
[12] Jetz W, Thomas GH, Joy JB, Hartmann K, Mooers AO. The global diversity of birds in space and time. Nature 2012;491(7424):444–8.
[13] Redding DW, Mooers AO. Incorporating evolutionary measures into conservation prioritization. Conserv Biol. 2006;20(6):1670–8.
[14] Joy JB, Liang R, McCloskey RM, Nguyen T, Brumme CJ, Colley G, et al. Phylogenetically estimated HIV diversification rates reveal prevention of HIV-1 by antiretroviral therapy. J Int Aids Soc. Geneva, Switzerland; 2015;16 :1–1.
[15] Hogg RS, Rhone SA, Yip B, Sherlock C, Conway B, Schechter MT, et al. Antiviral effect of double and triple drug combinations amongst HIV-infected adults: lessons from the implementation of viral load-driven antiretroviral therapy. AIDS 1998;12(3):279–84.
[16] Kosakovsky Pond SL, Posada D, Stawiski E, Chappey C, Poon AFY, Hughes G, et al. An evolutionary model-based algorithm for accurate phylogenetic breakpoint mapping and subtype prediction in HIV-1. Fraser C, editor. PLoS Comput Biol. 2009;5(11) e1000581–21.
[17] Montaner JS, Lima VD, Barrios R, Yip B, Wood E, Kerr T, et al. Association of highly active antiretroviral therapy coverage, population viral load, and yearly new HIV diagnoses in British Columbia, Canada: a population-based study. The Lancet. Elsevier Ltd; 2010;376(9740):532–9.
[18] Katoh K, Toh H. Recent developments in the MAFFT multiple sequence alignment program. Brief. Bioinformatics 2008;9(4):286–98.
[19] Larsson A. AliView: a fast and lightweight alignment viewer and editor for large datasets. Bioinformatics 2014;30(22):3276–8.
[20] Rhee S-Y, Blanco JL, Jordan MR, Taylor J, Lemey P, Varghese V, et al. Geographic and temporal trends in the molecular epidemiology and genetic mechanisms of transmitted HIV-1 drug resistance: an Individual-Patient- and Sequence-Level meta-analysis. Carr A, editor. PLoS Med. 2015;12(4):e1001810–29.
[21] Price MN, Dehal PS, Arkin AP. FastTree 2–approximately maximum-likelihood trees for large alignments. Poon AFY, editor. PLoS ONE 2010;5(3):e9490.

[22] Paradis E, Claude J, Strimmer K. APE: analyses of phylogenetics and evolution in R language. Bioinformatics 2004;20(2):289–90.

[23] Statistics Canada. Dictionary, census of population, 2016 - Census tract (CT) [Internet]. 2016 [cited 2019 Jan 3]. pp. 1–2. Available from: https://www12.statcan.gc.ca/census-recensement/2016/ref/dict/geo013-eng.cfm.

[24] Hu M-C, Pavlicova M, Nunes EV. Zero-Inflated and hurdle models of count data with extra zeros: examples from an HIV-Risk reduction intervention trial. Am J Drug Alcoh Abuse 2nd ed. 2011;37(5):367–75.

[25] Bivand RS, Wong DWS. Comparing implementations of global and local indicators of spatial association. Springer Berlin Heidelberg. TEST 2018;27(3):716–748.

[26] Lee D. CARBayes: an r package for Bayesian spatial modeling with conditional autoregressive priors. J Stat Softw 2013;55(13):1–24.

[27] Mets KD, Armenteras D, Dávalos LM. Spatial autocorrelation reduces model precision and predictive power in deforestation analyses. Ecosphere 2017;8(5) e01824–18.

[28] Ver Hoef JM, Hanks EM, Hooten MB. On the relationship between conditional (CAR) and simultaneous (SAR) autoregressive models. Spat Stat 2018;25:68–85.

[29] Leroux BG, Lei X, Breslow N. Estimation of disease rates in small areas: a new mixed model for spatial dependence. Statistical models in epidemiology, the environment, and clinical trials. Halloran ME, Berry D, editors; 1999. New York, NY.

[30] BC Centre for Disease Control. HIV in British Columbia: Annual Surveillance Report 2016. Retrieved from http://www.bccdc.ca/search?k=hiv%20annual%20report. 2018 Jun 26:1–46.

[31] Wood E, Kerr T, Marshall BDL, Li K, Zhang R, Hogg RS, et al. Longitudinal community plasma HIV-1 RNA concentrations and incidence of HIV-1 among injecting drug users: prospective cohort study. BMJ 2009;338:b1649.

[32] Musenge E, Chirwa TF, Kahn K, Vounatsou P. Bayesian analysis of zero inflated spatiotemporal HIV/TB child mortality data through the INLA and SPDE approaches: applied to data observed between 1992 and 2010 in rural North East South Africa. Int J Appl Earth Observ Geoinform Elsevier B.V. 2013;22:86–98.

[33] Xia Q, Wiewel EW, Torian LV. Revisiting the methodology of measuring HIV community viral load. J Acquir Immune Defic Syndr 2013;63(2):e82–4.

[34] McCloskey RM, Poon AFY. A model-based clustering method to detect infectious disease transmission outbreaks from sequence variation. Kosakovsky Pond SL, editor. PLoS Comput Biol. 2017;13(11):e1005868–17.

[35] 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.

[36] Gilbert M, Swenson L, Unger D, Scheim A, Grace D. Need for robust and inclusive public health ethics review of the monitoring of HIV phylogenetic clusters for HIV prevention. The Lancet HIV. Elsevier Ltd 2016;3(10):e461.

[37] Dark SJ, Bram D. The modifiable areal unit problem (MAUP) in physical geography. Prog Phys Geogr 2016;31(5):471–9.