










Concordance of HIV transmission risk factors elucidated using viral diversification rate and phylogenetic clustering

Angela McLaughlin ^{1,2}, Paul Sereda ¹, Chanson J. Brumme ^{1,3},
Zabrina L. Brumme ^{1,4}, Rolando Barrios ¹, Julio S. G. Montaner ^{1,3} and
Jeffrey B. Joy ^{1,2,3,*}

¹British Columbia Centre for Excellence in HIV/AIDS, St. Paul's Hospital, 608-1081 Burrard Street, Vancouver, BC V6Z 1Y6, Canada; ²Department of Bioinformatics, University of British Columbia, Genome Sciences Centre, British Columbia Cancer Agency, 100-570 West 7th Avenue, Vancouver, BC V5Z 4S6, Canada; ³Division of Infectious Diseases, Department of Medicine, University of British Columbia, 452D, Heather Pavilion East, Vancouver General Hospital, 2733 Heather Street, Vancouver, BC V5Z 3J5, Canada and ⁴Faculty of Health Sciences, Simon Fraser University, Blusson Hall, Room 11300, 8888 University Drive, Burnaby, BC V5A 1S6, Canada

*Corresponding author. British Columbia Centre for Excellence in HIV/AIDS, University of British Columbia, Bioinformatics, 608-1081 Burrard Street, Vancouver, BC V6Z 1Y6, Canada. Tel: +1 (604) 368-5569; E-mail: jjoy@bccfe.ca
Received 24 May 2020; revised version accepted 14 September 2021

ABSTRACT

Background and objectives: Although HIV sequence clustering is routinely used to identify subpopulations experiencing elevated transmission, it over-simplifies transmission dynamics and is sensitive to methodology. Complementarily, viral diversification rates can be used to approximate historical transmission rates. Here, we investigated the concordance and sensitivity of HIV transmission risk factors identified by phylogenetic clustering, viral diversification rate, changes in viral diversification rate and a combined approach.

Methodology: Viral sequences from 9848 people living with HIV in British Columbia, Canada, sampled between 1996 and February 2019, were used to infer phylogenetic trees, from which clusters were identified and viral diversification rates of each tip were calculated. Factors associated with heightened transmission risk were compared across models of cluster membership, viral diversification rate, changes in diversification rate, and viral diversification rate among clusters.

Results: Viruses within larger clusters had higher diversification rates and lower changes in diversification rate than those within smaller clusters; however, rates within individual clusters, independent of size, varied widely. Risk factors for both cluster membership and elevated viral diversification rate included being male, young, a resident of health authority E, previous injection drug use, previous

hepatitis C virus infection or a high recent viral load. In a sensitivity analysis, models based on cluster membership had wider confidence intervals and lower concordance of significant effects than viral diversification rate for lower sampling rates.

Conclusions and implications: Viral diversification rate complements phylogenetic clustering, offering a means of evaluating transmission dynamics to guide provision of treatment and prevention services.

Lay Summary: Understanding HIV transmission dynamics within clusters can help prioritize public health resource allocation. We compared socio-demographic and clinical risk factors associated with phylogenetic cluster membership and viral diversification rate, a historical branching rate, in order to assess their relative concordance and sampling sensitivity.

KEYWORDS: molecular phylogenetics; phylogenetic clustering; molecular epidemiology; HIV-1; molecular evolution; transmission; risk factors

INTRODUCTION

Identifying risk factors for elevated HIV transmission rates is critical for effective allocation of limited public health resources for prevention and treatment. In the Canadian province of British Columbia (BC), where antiretroviral resistance testing has been standard-of-care since 1998 and a province-wide programme to expand access to HIV testing, care and treatment have been in place since 2010, it is estimated that in 2016, 84% of people living with HIV (PLHIV) were diagnosed, 85% of those diagnosed were on treatment and 93% of those on treatment were virally suppressed [1]. While projections suggested that BC is on track to meet the Joint United Nations Programme on HIV/AIDS 95–95–95 targets for 2030 [1], particular groups remain disproportionately affected by HIV/AIDS. For instance, in BC, men who have sex with men (MSM) have comprised most new HIV diagnoses in recent years [2].

Quantifying HIV transmission dynamics by reconstructing social or sexual networks via traditional epidemiological methods such as contact tracing are limited by the virus' long period of infectiousness and low per act transmission rate [3]. The substantial viral genetic diversity within and between PLHIV resulting from HIV's rapid evolution can be leveraged to build phylogenetic trees that approximate between-host transmission networks [4], though these trees generally cannot be used to infer transmission directionality [5].

Identifying clusters of closely related HIV sequences is commonly performed to detect subpopulations experiencing higher than average transmission rates [6–9]. A multitude of non-parametric methods is commonly applied to infer clusters from either genetic or phylogenetic tree distances, with varying criteria related to monophyly and bootstrap support (see Poon [10] for a more comprehensive review of clustering methods). Results however can vary widely by method [10–12]. For instance, the maximum pairwise genetic or phylogenetic distance threshold strongly affects which clusters are recovered from the true transmission network [10, 12]. Furthermore, commonly used phylogenetic clustering methods that assume clusters form clades [7, 13, 14] do not consistently recover sexual contact network communities [11]. Cluster detection has been criticized for more

readily detecting differences in subpopulation sampling rates, rather than transmission rates [10]. Recent parametric clustering methods have been developed that consider the evolution of transmission rates along phylogenies and thus distinguish transmission from sampling differences [15, 16]. However, they are reliant on strong assumptions about how transmission rates evolve along the tree, such as that branching (transmission) events are generated by a Poisson process controlled by a discrete evolving character [15], which does not necessarily account for stochasticity in individuals' behaviour affecting the probability of a transmission event. Generally, evaluating whether or not an individual is within a cluster oversimplifies within-cluster transmission dynamics, particularly if large clusters are common [17]. Moreover, while interpreting cluster membership is straightforward, it inherently ignores differences within groups.

We previously showed that a model incorporating geographically aggregated viral diversification rate was more predictive of the location and number of new HIV cases compared to models comprising epidemiological data alone [18]. Furthermore, viral diversification rates were significantly lower among PLHIV receiving antiretroviral treatment than those were not, validating the effectiveness of Treatment as Prevention [19].

We sought to investigate the relationship between viral diversification rate and clustering, as well as evaluate the concordance of risk factors associated with viral diversification rate and phylogenetic cluster membership using the clustering methodology applied by Poon *et al.* [6, 17]. Using HIV sequence data annotated with clinical and socio-demographic data for 8063 PLHIV in BC (after restricting to those alive on 4 February 2019), we inferred four models for the respective outcomes of individuals' 2018 viral diversification rate, individuals' annual change in viral diversification rate between 2018 and 2017, cluster membership and viral diversification rate among the clustered population. We then compared risk factors across the four models to test the hypothesis that risk factors for elevated viral diversification rate among PLHIV in BC would be broadly consistent with those identified in phylogenetic cluster analyses. A subsampling sensitivity analysis was also undertaken to evaluate each model strategy's robustness to inferring consistent effect sizes in scenarios with less sampling representation.

METHODOLOGY

Study setting and participants

The BC Centre for Excellence in HIV/AIDS (BC-CfE) Drug Treatment Programme (DTP) is an ongoing clinical registry that provides all PLHIV in BC with access to HIV care, HIV drug resistance genotyping and highly active antiretroviral therapy at no cost to the individual [20]. DTP and drug resistance testing eligibility are described in the [supplementary materials](#). Out of 13 307 DTP participants who ever had a detectable viral load between 30 May 1996 and 4 February 2019, 9848 participants had drug resistance testing done. HIV sequences and individual information were stored in secure, access-restricted facilities at the BC-CfE in a password-protected Oracle database. Individuals' data were de-identified and doubly anonymized. Ethical approval for this study was granted by the University of British Columbia—Providence Health Care Research Ethics Board (H17-01812). The HIV sequence data cannot be shared outside the BC-CfE as a condition of ethics approval.

Phylogenetic inference

A total of 37 304 clinical HIV resistance genotype tests from 9848 PLHIV were completed on samples collected between 30 May 1996 and 4 February 2019 at the BC-CfE (see workflow in [Supplementary Fig. S1](#)). Between 1 and 46 genotypic resistance tests were performed per individual (mean, 3.79; median, 2). Plasma samples were screened for drug resistance mutations by sequencing the HIV *protease* and partial *reverse transcriptase* genes, referred to as partial pol. Sequences were aligned using MAFFT version 7.310 [21] and visualized in AliView V1.17.1 [22], where insertions and deletions relative to HXB2 (GenBank #K03455), as well as amino acids corresponding to WHO recognized drug resistance mutation sites were removed prior to tree inference. A set of shuffled bootstrap alignments were generated to infer 100 approximate maximum likelihood phylogenetic trees with a general time-reversible substitution model and gamma-distributed rate variation in FastTree2.1 [23].

Cluster identification and characterization

HIV phylogenies with all available sequences were used to identify clusters defined by patristic distance cutoffs, as routinely applied by the BC phylogenetic surveillance programme [6, 17]. In order for viruses from different individuals to be linked in a cluster, the pairwise patristic distance between a participant's earliest HIV sequence and any sequence from a different individual must be < 0.02 substitutions per site (95th percentile of within-host patristic distances, [17]) in >90% of bootstrap phylogenies. Clusters were limited to comprise ≥ 5 members to

protect individuals' confidentiality [17], consistent with the methods approved by the research ethics board.

To further characterize the differences within the clustering subpopulation, the distribution of viral diversification rates and changes in viral diversification rate were compared among (i) those who did and did not cluster, and (ii) different cluster size ranges, using violin plots and non-parametric Kruskal–Wallis tests, followed by pairwise two-sided Wilcoxon rank-sum tests (i.e. Mann–Whitney) with a Bonferroni *P*-value adjustment to compare multiple cluster size ranges. Clusters were binned into cluster size categories (5–10, 11–20, 21–50, 51–100 and >100) based on the total number of PLHIV (without excluding those who had died) within them in 2018 ([Supplementary Fig. S3](#)). Viral diversification rates and changes in diversification rates were visualized for three representative clusters with newly diagnosed cases in 2018 using the Compound Spring Embedder Layout in Cytoscape v3.7.2 [24].

Viral diversification rates

The full bootstrap trees ($n=37\,304$) were pruned into 2017 ($n=9596$) and 2018 ($n=9832$) trees, which were generated by retaining only sequences collected prior to or during each year and then pruning to each individual's earliest sample ([Supplementary Fig. S1](#)). Trees were rooted based on root-to-tip regression of evolutionary divergence over time, fit using the lowest root mean square error via the rtt command in R package ape v5.0 [25, 26]. Although including all sampled viruses is useful to infer overall phylogenetic relationships between and within hosts, we retained only the earliest sample available from each individual to avoid artificially inflating diversification rate for individuals with more than one sample. For each tip on each bootstrap tree, the viral diversification rate was calculated and the mean diversification rate across 100 bootstrap trees was computed. As the reciprocal of the equal splits metric [27], the viral diversification rate for each tip on a rooted bifurcating tree is 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 $1 - 2$ [28]. See [Supplementary Fig. S12](#) for an example of calculating viral diversification rate.

$$\text{Viral Diversification Rate}_i = \text{Equal Splits}_i^{-1} = \left(\sum_{j=1}^{N_i} \frac{l_j}{2^{j-1}} \right)^{-1}$$

The annual change in diversification rate for each participant between 2018 and 2017 was calculated as the difference between their 2018 and 2017 diversification rates. Cases newly diagnosed in 2018 (222 of the 8063 participants included in the analysis) were assigned a change in diversification rate of zero to emphasize changes in transmission rate among the prevalent population. Distribution of log-transformed diversification

rate in 2018 for full dataset and subsampled sets are provided in [Supplementary Fig. S4](#).

Comparisons of transmission risk factors

Analyses were restricted to 8063 individuals alive at the time of data generation in February 2019 to focus on current transmission trends. The association of individuals' clinical and socio-demographic attributes with different estimates of transmission activity was compared. As both viral diversification rate and change in diversification rate are continuous outcomes with approximately lognormal distributions ([Supplementary Fig. S4](#)), log linear regression models were inferred and exponentiated coefficients were interpreted as adjusted relative risks (aRRs). Where cluster membership was the outcome, a multivariate logistic model was inferred and exponentiated coefficients were interpreted as adjusted odds ratios (aORs). The aRRs and aORs for reported risk factors were calculated relative to reporting 'no' for each risk factor and individuals could report multiple risk factors. In the combined modelling approach, risk factors associated with elevated viral diversification rates among the clustered subpopulation were evaluated, highlighting differences within clusters. The model output for viral diversification rate among the non-clustered subpopulation, as well as changes in diversification rate among the clustered and non-clustered subpopulations, were explored in the [supplementary material \(Supplementary Figs S5–S7\)](#).

Variables included in the analysis are sex at birth; age at the end of 2018 (categorized as 29 and under; 30–44; 45–59; and 60 and over), health authority (HA) of residence at enrolment (anonymized to letters A through E for regional confidentiality); risk factor exposures including identifying as MSM, having heterosexual exposure, identifying as people who inject drugs (PWID), previous hepatitis C virus (HCV) infection; and most recent plasma viral load (HIV RNA copies/mL). Individuals who had unreported heterosexual exposure were the same as those with unreported MSM exposure, thus only MSM unreported was included in the models.

Sampling sensitivity analysis

To characterize the robustness of modelling approaches to inferring consistent effect size estimates with lower sampling representation, we randomly subsampled 25, 50 and 75% of the 8063 individuals in the analysis. Clusters were identified anew using pruned trees with viral sequences (tips) collected from subsampled patients in the full tree for all bootstraps, resulting in respective trees with 7736 (20.1%), 15 443 (41.4%) and 23 308 (62.5%) tips. The 2018 and 2017 interval trees from the original analysis were pruned to only tips from subsampled patients to recalculate diversification rates, which resulted in

2018 trees with 20.5%, 40.9%, and 61.4% of tips present in the original 2018 tree, and 20.3%, 40.9%, and 61.3% of tips present in the original 2017 tree. Models were re-inferred for the 25%, 50%, and 75% subsamples ([Supplementary Figs S9–S11](#)). We quantified how well models using subsampled data recapitulated the full model by comparing for each parameter and each subsample whether the significance (yes or no) and effect direction (above or below one) of each coefficient was the same, the difference in mean effect sizes, the proportion of confidence interval (CI) overlap and the fold increase in CI width ([Fig. 4](#)). The distribution of viral diversification rates and cluster sizes for the full and subsampled datasets were compared in [Supplementary Figs S2 and S3](#).

RESULTS

Characteristics of phylogenetic clusters

Of the 8063 representative viruses from PLHIV in the study, 3343 (41.5%) were phylogenetically clustered ([Table 1](#)). There were 227 clusters identified, of which 143 clusters comprised 5–10 members, 5 had > 100 members and the largest cluster had 407 members, prior to excluding participants who had passed away ([Fig. 1A](#)). Clustering was associated with sex at birth (chi-squared test: $P < 0.001$) and age ($P < 0.001$). Younger age groups had a greater tendency to cluster than older age groups—those 30–44 represented 28.2% of the total clustered population compared to 22.9% of the total study population. Furthermore, clustering was dependent on individuals' HA of residence ($P < 0.001$). MSM were underrepresented in the clustering population ($P < 0.001$; 30.3% of clustering population; 37.1% of study population), while PWID made up a greater proportion of the clustering population ($P < 0.001$; 41.5%; 29.4%), as did those with a previous HCV infection ($P < 0.001$; 33.9%; 24.1%).

Relationship between viral diversification rate and clustering

Individuals who clustered tended to have viruses with higher diversification rates; however, viruses with high diversification rates were also found among individuals who did not cluster ([Supplementary Fig. S2](#)). The distributions of log-transformed viral diversification rates and changes in log diversification rate between individuals who did and did not cluster differed significantly ([Fig. 1B and C](#)). A Kruskal–Wallis rank-sum test confirmed that diversification rates were significantly higher among those in clusters ($P < 0.001$), while the annual changes in diversification rate were significantly lower among members of clusters ($P < 0.001$). Among those who clustered, log-transformed diversification rates varied widely from 2.34 to 7.19—compared



Table 1. Characteristics of the overall study population and the subset who were members of phylogenetic clusters

Characteristics	Study population (<i>n</i> = 8063)			Clustering population (<i>n</i> = 3343)			<i>P</i> -values	
	<i>n</i>	% of total	% of reported	<i>n</i>	% of total	% of reported		
Sex at birth	Female	1394	17.3	17.8	643	19.2	19.8	<0.001
	Male	6428	79.7	82.2	2602	77.8	80.2	–
	Unreported	241	3.0	–	98	2.9	–	–
Age category	60 and over	1956	24.3	24.4	634	19	19	<0.001
	45–59	3890	48.2	48.4	1594	47.7	47.8	–
	30–44	1845	22.9	23.0	943	28.2	28.3	–
	29 and under	341	4.2	4.2	163	4.9	4.9	–
	Unreported	31	0.4	–	9	0.3	–	–
Health authority	A	418	6.2	6.7	162	5.8	12.9	<0.001
	B	778	11.6	12.5	158	5.7	12.6	–
	C	1680	25.0	26.9	335	12	26.7	–
	D	4351	64.8	69.7	184	6.6	14.7	–
	E	263	3.9	4.2	667	23.9	53.1	–
	Unreported	573	8.5	–	1837	65.9	–	–
Men who have sex with men	No	2673	33.2	47.2	1421	42.5	58.4	<0.001
	Yes	2988	37.1	52.8	1012	30.3	41.6	–
	Unreported	2402	29.8	–	910	27.2	–	–
Heterosexual exposure	No	3761	46.6	66.4	1573	47.1	64.7	<0.001
	Yes	1900	23.6	33.6	860	25.7	35.3	–
	Unreported	2402	29.8	–	910	27.2	–	–
People who inject drugs	No	4015	49.8	62.9	1287	38.5	48.1	<0.001
	Yes	2371	29.4	37.1	1388	41.5	51.9	–
	Unreported	1677	20.8	–	668	20	–	–
Previous Hepatitis C infection	No	3187	39.5	62.1	986	29.5	46.5	<0.001
	Yes	1942	24.1	37.9	1133	33.9	53.5	–
	Unreported	2934	36.4	–	1224	36.6	–	–

Reported *P*-values were calculated using chi-squared contingency table tests.

to 2.32–6.03 among those who did not cluster. Annual changes in diversification rate were similarly variable among both individuals who did and did not cluster. Together, this suggests transmission levels vary both within and outside identified clusters.

Diversification rates and annual changes in diversification rate also differed significantly by cluster size category (Kruskal–Wallis: both $P < 0.001$; Fig. 1D and E). Individuals from smaller clusters had significantly lower viral diversification rates than individuals from all larger cluster size ranges (Wilcoxon tests with Bonferroni correction: all $P < 0.001$, except comparing clusters sized 11–20 to 51–100 and clusters sized 21–50 to those sized >100). Conversely, individuals within smaller clusters had significantly higher annual changes in diversification rate compared to individuals in larger clusters (all $P < 0.001$, except

comparing clusters sized 5–10 to 11–20 and clusters size 51–100 to >100).

Although larger clusters generally had higher viral diversification rates and lower changes in viral diversification rate than smaller clusters, Fig. 2 illustrates how viral diversification rates and changes in viral diversification rate can vary substantially within clusters, regardless of size. Regardless of cluster membership, individuals who were phylogenetically proximate to newly diagnosed cases in 2018 (<0.02 substitutions per site across >90% of bootstraps) had significantly higher changes in viral diversification rate and higher viral diversification rates compared to individuals who were not proximate to new cases, regardless of cluster membership (Kruskal–Wallis: both $P < 0.001$; Supplementary Fig. S13). We estimated there were 339 individuals paired to individuals whose first antiretroviral

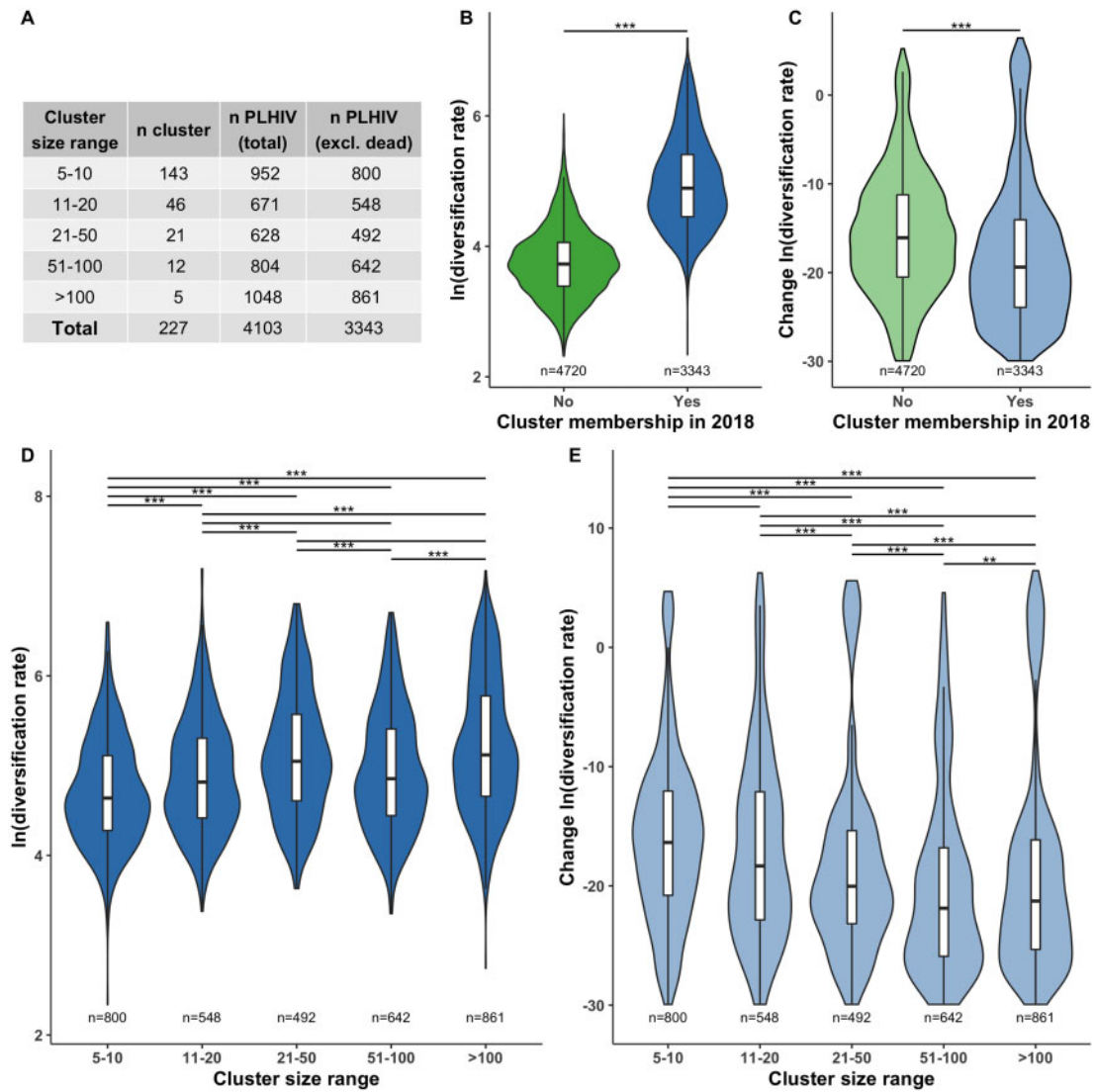


Figure 1. Diversification rates differ by phylogenetic cluster membership and size range. Clusters must contain a minimum of five members connected by pairwise patristic distances <0.02 substitutions/site and supported by >90% of bootstrap phylogenies. (A) Clusters were grouped by their number of members (size ranges), within which the number of clusters, total number of people living with HIV and number of people living with HIV alive in February 2019 were calculated. (B) Log-transformed viral diversification rates and (C) annual changes in log-transformed diversification rates differed based on cluster membership in 2018. (D) Log-transformed viral diversification rate within different cluster size ranges and (E) annual changes in log-transformed diversification rates also differed. Significance was assessed using non-parametric Kruskal–Wallis tests across groups, followed by pairwise Wilcoxon rank-sum tests with a Bonferroni correction for multiple comparisons, where * $P < 0.05$, ** $P < 0.01$ and *** $P < 0.001$. The lower and upper hinges in the boxplot correspond to the 25th and 75th percentiles, the whiskers extend 1.5 times the interquartile range from the hinge, and the middle bar represents the median

dispensation was in 2018 (taken as a proxy for new infection in 2018)—310 of 339 (91.4%) proximate individuals were in clusters.

Comparing risk factors for cluster membership, viral diversification rate, change in diversification rate and viral diversification rate among the clustered population

Similar risk factors were identified, albeit with varying effect sizes and CIs, through respective analyses of viral

diversification rate in 2018, annual changes in diversification rate between 2017 and 2018, cluster membership in 2018 and diversification rate among the clustered population (Fig. 3). Cluster membership aORs were consistently further from 1 with wider CIs than viral diversification rate aRRs.

Men were at a significantly higher adjusted risk than women in all models except viral diversification rate among clustered, with an aOR of 1.27 (95% CI: 1.10–1.46) in the cluster membership analysis, an aRR of 1.10 (1.05–1.16) in the diversification rate analysis and an aRR of 1.05 (1.01–1.09) in the change in

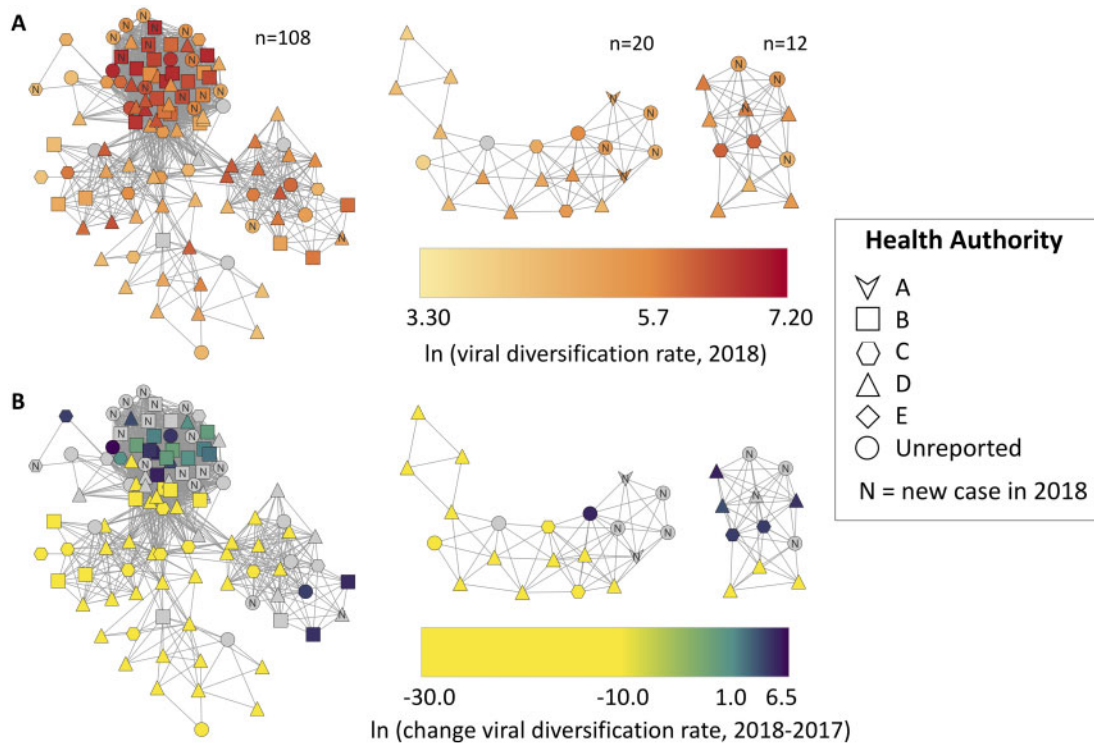


Figure 2. Representative phylogenetic clusters of different sizes include a range of viral diversification rates and changes in diversification rate. Node shapes represent individuals' health authority of residence and node colours represent either (A) log-transformed viral diversification rates in 2018 or (B) log-transformed changes in diversification rates between 2018 and 2017. Grey nodes without an N denote individuals who have passed away; grey nodes with an N newly diagnosed cases in 2018 that have an assigned change in diversification rate of 0

diversification rate analysis. Models were fully concordant with respect to age: individuals in the three younger age categories had significantly and progressively higher adjusted risks in all analyses. For example, the aRRs of elevated viral diversification rate among the clustered were 1.08 (1.01–1.15) for those aged 45–59, 1.35 (1.25–1.45) for those aged 30–44 and 1.59 (1.40–1.81) for those aged 29 and under, relative to those 60 and older.

The significance and effect direction of HA were not entirely concordant across all models. Individuals from health authorities B and C were at no elevated transmission risk compared to A in any of the models. Only the cluster membership analysis identified living in HA D relative to A as a significant risk factor (aOR = 1.36, 1.09–1.69), adjusting for other variables. Living within HA E was identified as a significant risk factor in the cluster membership analysis (aOR = 3.09, 2.19–4.38) and in the viral diversification rate (aRR = 1.36, 1.20–1.54), though not in the other two models. Participants who did not report a HA (due to recent migration, homelessness or other circumstances) had significantly lower adjusted odds of clustering (aOR = 0.45, 0.31–0.63) and a significantly lower adjusted risk of elevated viral diversification rate (aRR = 0.81, 0.72–0.92).

Models were not fully concordant in regard to the effect and significance of reported risk factors. While MSM had a significantly lower adjusted odds of clustering than non-MSM

(aOR = 0.65, 0.55–0.77), among the clustering population, MSM had 1.21 times (1.11–1.31) the adjusted risk of an elevated viral diversification rate than non-MSM. Heterosexual intercourse was associated with a lower adjusted odds of clustering (aOR = 0.83, 0.72–0.96) and a lower adjusted risk of elevated diversification rate (aRR = 0.94, 0.89–0.99). Previous HCV infection was associated with a significantly higher adjusted odds of clustering (aOR = 1.87, 1.60–2.18), higher adjusted risk of elevated diversification rate (aRR = 1.29, 1.22–1.37) and higher adjusted risk of elevated diversification rate among the clustered (aRR = 1.18, 1.09–1.28), but not associated with a higher risk of elevated change in diversification rate. Individuals with higher viral loads were at an elevated risk in all models.

Model sensitivity to subsampling

As expected, subsampling data resulted in left-shifted distributions of viral diversification rate (Supplementary Fig. S3) and cluster sizes (Supplementary Fig. S2), whereby the diversification rates were reduced and cluster sizes were smaller, with fewer clusters identified. Comparing across subsample datasets (25%, 50% and 75%) to the full dataset for each model strategy/outcome (Supplementary Figs S9–S11), variables with effect sizes further from the null in the original model tended to have their effect

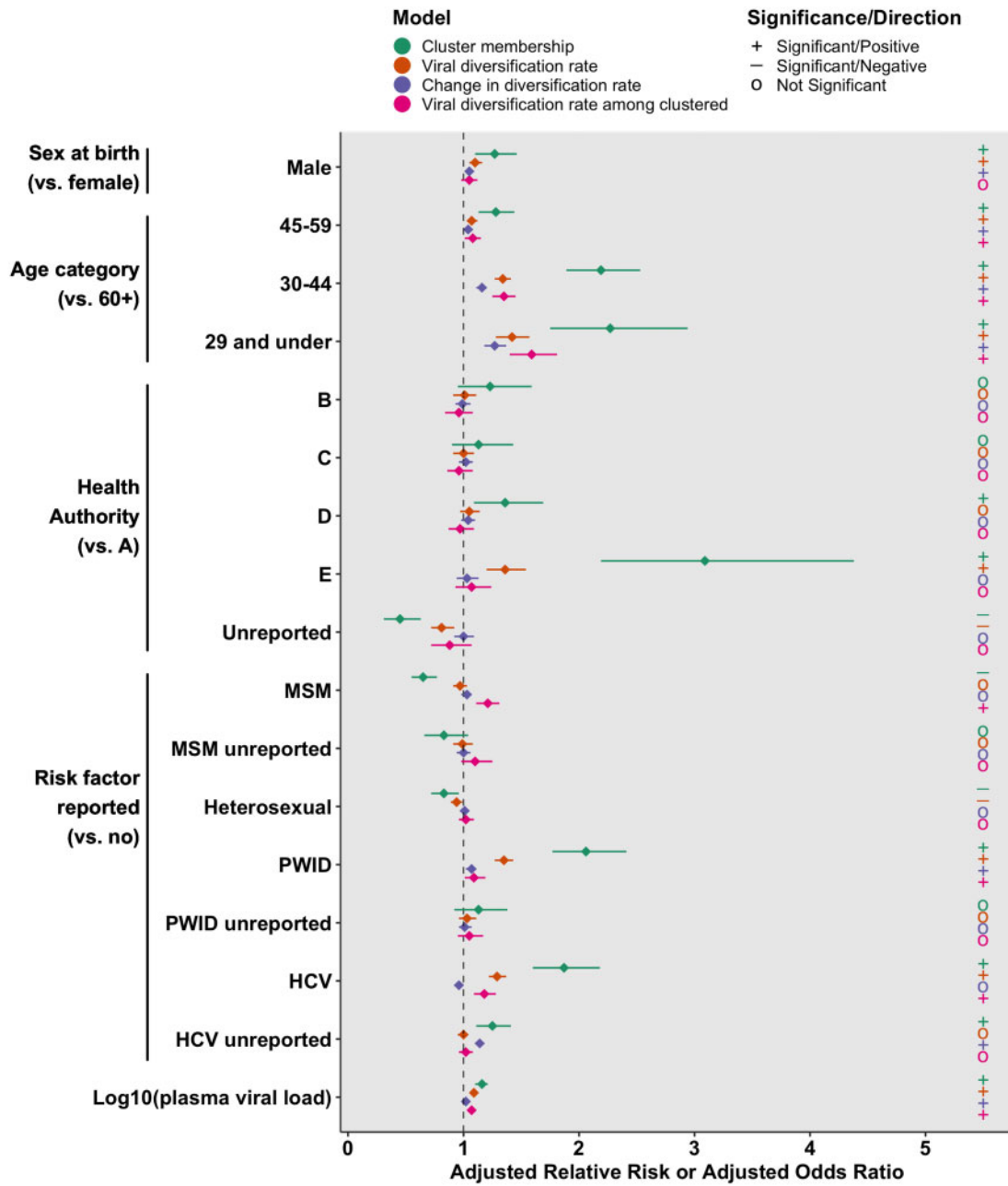


Figure 3. Adjusted relative risks and odds ratios for individual attributes associated with HIV cluster membership, viral diversification rate, annual changes in diversification rate and viral diversification rate among the clustered population. Relative risks were estimated using log linear regression models, while the odds ratios for cluster membership were estimated using a multiple logistic regression model with a logit linker equation. Clusters were defined using a pairwise patristic distance threshold of 0.02 substitutions/site, representing the 95th percentile of inpatient patristic distances, and contained a minimum of five individuals

direction and significance better recapitulated in the subsampled datasets. Viral diversification rate among clustered models most consistently generated the same significance and effect direction (Fig. 4A), meanwhile clustering had the least agreement of significance and effect direction for all subsample percentages (mean 0.59 agreement for 25% subsample). When we compared the difference in mean effect size estimates (Fig. 4B), the spread in differences was the widest for cluster membership for all

subsamples, although on average, difference in means remained within 0.2 of the original for all models. Notably, the fold increase in the CI width increased as data was removed, and this was particularly the case for the cluster membership models (for 25%: 2.7 mean fold change of CI width; Fig. 4C) and viral diversification rate among clustered models (for 25%: 3.4 mean fold change). Viral diversification rate and changes in diversification rate models had smaller increases in the CI width (for 25%: 1.5 and 1.7

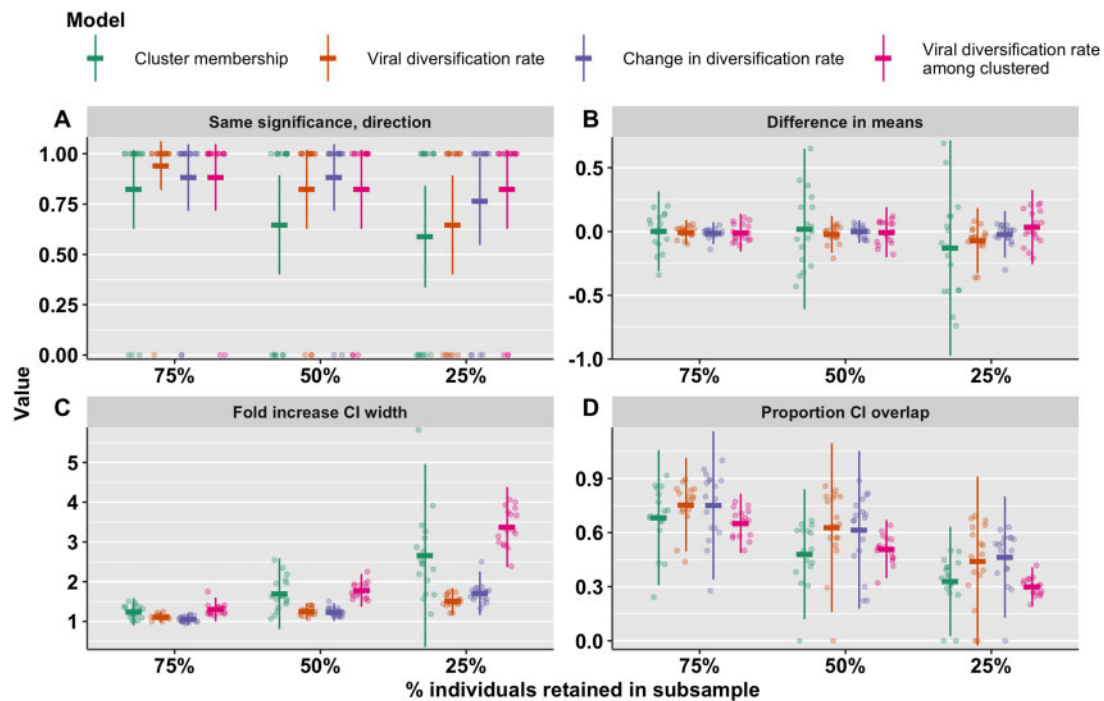


Figure 4. A summary of model robustness to subsampling. For datasets limited to 25%, 50% or 75% of the full dataset participants, resulting exponentiated model coefficient estimates were compared to those from the full dataset model in terms of (A) whether detected significance and effect direction were consistent; (B) the difference in mean effect size estimates; (C) the proportion of the subsampled confidence interval (CI) overlapping the full model CI; and (D) the fold increase in the width of the subsampled CI. Colours represent the various model outcomes/strategies, lightly coloured points represent model parameters, thick rectangle is the mean and confidence bars represent 95% confidence intervals

mean fold changes, respectively; Fig. 4C). Finally, we evaluated the percent of the subsampled CI that overlapped with the original CI—this decreased the more data was removed for all models (Fig. 4D). The greatest decreases in overlap were in the cluster membership and viral diversification rate among clustered models. Taken together, these sensitivity analyses suggest that cluster membership and viral diversification among clustered models were the least robust to recapitulating transmission risk estimates with less data available.

CONCLUSIONS AND IMPLICATIONS

Viral diversification rate is an informative metric to investigate risk factors associated with HIV transmission, which is generally concordant with clustering and robust to reduced sampling. Risk factors associated with both elevated viral diversification rates and cluster membership included being male, young, living in the HA E, previous injection drug use, previous HCV infection, not reporting an HCV test or a high most recent viral load.

Being young was a transmission risk factor in every model that we evaluated. Conceptually, it makes sense that younger people would be more active socially and sexually, and may be more likely to engage in risk-taking behaviour [29]. Participants previously infected with HCV were at a significant

risk of elevated diversification rate overall and among the clustered population, as well as at a higher odds of clustering, but not when evaluating the annual change in diversification rate. This is consistent with historical HIV transmission patterns in the province—in the 1990s, injection drug use was a primary risk factor for HIV and HCV transmission [30], but harm reduction measures such as InSite, North America's first safe injection site [31], have subsequently reduced new HIV cases in this population, resulting in fewer large changes in diversification rate.

Although odds ratios approximate relative risks for rare outcomes [32], 41.5% of our study population were in a cluster. Comparing odds ratios from the cluster membership analysis to relative risks from the viral diversification rate analyses is therefore not an equivalent comparison, as odds ratios tend to exaggerate the effect size [32]. This wider trend was consistent with our results—aORs for cluster membership were consistently further from 1 than the aRRs for viral diversification rate within the entire study population and among only the clustered subpopulation. This suggests that diversification rate offers modelling advantages over cluster membership as relative risks calculated from continuous outcomes more accurately estimate the effect size than odds ratios calculated from binary outcomes. Additionally, the CIs on the aORs for cluster membership were consistently wider than those for the aRRs

for viral diversification rate, especially in the subsampled datasets, where we detect the combined effects of clustering being a binary variable, as well as only accepting clusters with five or more members.

The interpretations of cluster membership and elevated viral diversification are similar; they identify viral lineages that branched rapidly, suggesting frequent between-host transmission. However, viral diversification rate gives more weight to recent branching events, does not rely on thresholding and has inherent modelling advantages as a continuous outcome. While it could be argued that the binary classification of cluster membership is more easily interpretable and actionable than a continuous metric with an abstract unit, dichotomizing the range of transmission activity obscures the detail necessary to prioritize treatment, particularly as clusters become larger and more numerous. Since clusters are easily interpretable to public health authorities and provide insights into how outbreaks span geographies and risk groups, we recommend viral diversification rates be used to complement phylogenetic clusters, resolving transmission dynamics within clusters and helping to both prioritize clusters of interest and make inferences about the mechanisms of cluster growth. Future work is needed to identify an optimal combined approach where viral diversification rate can be used to decipher differences within clusters, facilitating their prioritization. Our analyses showed that while larger clusters generally had higher diversification rates and lower changes in diversification rate, there is wide variation within clusters and cluster size alone does not predict individuals' viral diversification rates. Clustering was more sensitive to empirical subsampling than viral diversification rate, therefore viral diversification rate could be useful in geographies with limited sequencing capacity. Viral diversification rate may be partially subject to a similar challenge as clustering in that differences between subpopulations may reflect sampling differences more readily than transmission differences. While our empirical subsampling analysis showed that clustering was more sensitive to lower sampling rates than viral diversification rate, in future work, this could be explored further with simulations of structured populations with varying sampling and transmission rates.

Phylogenetic analyses of HIV sequences available through routine drug resistance genotyping highlight transmission risk factors that can help prioritize public health service allocation. We have shown that viral diversification rate, while generally concordant with phylogenetic clustering, was more robust to lower sampling representation. Viral diversification rate has untapped potential in the field of infectious disease epidemiology to decipher transmission dynamics.

SUPPLEMENTARY DATA

Supplementary data is available at *EMPH* online.

ACKNOWLEDGEMENTS

The authors thank their colleagues at the BC Centre for Excellence in HIV/AIDS, particularly those who contribute to the clinical laboratory, bioinformatics pipeline and Drug Treatment Programme. They also thank the members of the DTP cohort. All inferences, opinions and conclusions drawn in this paper are those of the authors, and do not reflect the opinions of the funders.

FUNDING

A.M. was supported by Canadian Institutes of Health Research (06556), a UBC Faculty of Medicine Dorothy Helmer award, and the Dr Ken Benson Memorial award through the Health Officers Council of BC. Z.L.B. was supported by a Scholar Award from the Michael Smith Foundation for Health Research. J.B.J. is funded by a Genome Canada and Genome BC Bioinformatics and Computational Biology grant (287PHY) and the Public Health Agency of Canada. J.S.G.M. is supported with grants paid to his institution by the British Columbia Ministry of Health.

Conflict of interest: None declared.

REFERENCES

1. Lima VD, Brumme ZL, Brumme C, STOP HIV/AIDS Study Group *et al.* The impact of Treatment as Prevention on the HIV epidemic in British Columbia, Canada. *Curr HIV/AIDS Rep* 2020; **17**:77–87.
2. British Columbia Centre for Disease Control. *HIV in British Columbia: Annual Surveillance Report* 2017. Retrieved from <http://www.bccdc.ca/health-professionals/data-reports/hiv-aids-reports> (1 May 2020, date last accessed).
3. Brown AJL, Lycett SJ, Weinert L, UK HIV Drug Resistance Collaboration *et al.* Transmission network parameters estimated from HIV sequences for a nationwide epidemic. *J Infect Dis* 2011; **204**:1463–9.
4. Leitner T, Escanilla D, Franzen C *et al.* Accurate reconstruction of a known HIV-1 transmission history by phylogenetic tree analysis. *Proc Natl Acad Sci U S A* 1996; **93**:10864–9.
5. Volz EM, Frost SDW. Inferring the source of transmission with phylogenetic data. *Plos Comput Biol* 2013; **9**:e1003397.
6. Poon A, Gustafson R, Daly P *et al.* Near real-time monitoring of HIV transmission hotspots from routine HIV genotyping: an implementation case study. *Lancet HIV* 2016; **3**:e231–8.
7. Ragonnet-Cronin M, Hodcroft E, Hue S *et al.*; UK HIV Drug Resistance Database. Automated analysis of phylogenetic clusters. *BMC Bioinformatics* 2013; **14**:317.
8. Pond SLK, Weaver S, Brown AJL *et al.* HIV-TRACE (TRANsmiSSion cluster engine): a tool for large scale molecular epidemiology of HIV-1 and other rapidly evolving pathogens. *Mol Bio Evol* 2018; **35**: 1812–9.
9. Wertheim JO, Murrell B, Mehta SR *et al.* Growth of HIV-1 molecular transmission clusters in New York City. *J Infect Dis* 2018; **218**:1943–53.
10. Poon AFY. Impacts and shortcomings of genetic clustering methods for infectious disease outbreaks. *Virus Evol* 2016; **2**:vew031.
11. Villandre L, Stephens DA, Labbe A *et al.*; Swiss HIV Cohort Study. Assessment of overlap of phylogenetic transmission clusters and

- communities in simple sexual contact networks: applications to HIV-1. *PLoS One* 2016;**11**:e0148459.
12. Chato C, Kalish ML, Poon AFY. Public health in genetic spaces: a statistical framework to optimize cluster-based outbreak detection. *Virus Evol* 2020;**6**:veaa011.
 13. Prosperi MCF, Ciccozzi M, Fanti I, ARCA collaborative group *et al.* A novel methodology for large-scale phylogeny partition. *Nat Commun* 2011;**2**:321.
 14. Brenner BG, Roger M, Routy J, Quebec Primary HIV Infection Study Group *et al.* High rates of forward transmission events after acute/early HIV-1 infection. *J Infect Dis* 2007;**195**:951–9.
 15. McCloskey RM, Poon AFY. A model-based clustering method to detect infectious disease transmission outbreaks from sequence variation. *PLoS Comput Biol* 2017;**13**:e1005868.
 16. Barido-Sottani J, Vaughan TG, Stadler T. Detection of HIV transmission clusters from phylogenetic trees using a multi-state birth–death model. *J R Soc Interface* 2018;**15**:20180512.
 17. Poon AFY, Joy JB, Woods CK *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**:926–35.
 18. McLaughlin A, Sereda P, Oliveira N *et al.* Detection of HIV transmission hotspots in British Columbia, Canada: a novel framework for the prioritization and allocation of treatment and prevention resources. *EBioMedicine* 2019;**48**:405–9.
 19. Joy JB, Liang R, McCloskey RM *et al.* Phylogenetically estimated HIV diversification rates reveal prevention of HIV-1 by antiretroviral therapy. In: *8th IAS Conference on HIV Pathogenesis, Treatment & Prevention 19–22 July 2015*, Vancouver, Canada. Abstract. *J Int AIDS Soc* 2015;**18**:20479.
 20. Hogg RS, Rhone SA, Yip B *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**:279–84.
 21. Katoh K, Toh H. Recent developments in the MAFFT multiple sequence alignment program. *Brief Bioinform* 2008;**9**:286–98.
 22. Larsson A. AliView: a fast and lightweight alignment viewer and editor for large datasets. *Bioinformatics* 2014;**30**:3276–8.
 23. Price MN, Dehal PS, Arkin AP. FastTree 2—approximately maximum-likelihood trees for large alignments. Poon AFY (ed.). *PLoS ONE* 2010; **5**:e9490.
 24. Shannon P, Markiel A, Ozier O *et al.* Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Research* 2003; **13**:2498–504.
 25. Paradis E, Claude J, Strimmer K. APE: analyses of Phylogenetics and Evolution in R language. *Bioinformatics* 2004; **20**:289–90.
 26. Paradis E, Schliep K. ape 5.0: an environment for modern phylogenetics and evolutionary analyses in R. Schwartz R (ed.). *Bioinformatics* 2019; **35**:526–8.
 27. Redding DW, Mooers AO. Incorporating evolutionary measures into conservation prioritization. *Conserv Biol* 2006;**20**:1670–8.
 28. Jetz W, Thomas GH, Joy JB *et al.* The global diversity of birds in space and time. *Nature* 2012;**491**:444–8.
 29. Landovitz RJ, Tseng C-H, Weissman M *et al.* Epidemiology, Sexual Risk Behavior, and HIV Prevention Practices of Men who Have Sex with Men Using GRINDR in Los Angeles, California. *Journal of Urban Health: bulletin of the New York Academy of Medicine* 2013; **90**: 729–39.
 30. Tyndall MW, Currie S, Spittal P *et al.* Intensive injection cocaine use as the primary risk factor in the Vancouver HIV-1 epidemic. *AIDS* 2003; **17**: 887–93.
 31. Pinkerton SD. How many HIV infections are prevented by Vancouver Canada's supervised injection facility? *International Journal of Drug Policy* 2011; **22**:179–83.
 32. Cummings P. The relative merits of risk ratios and odds ratios. *Archives of Pediatrics & Adolescent Medicine* 2009; **163**:438–45.