Epidemiological Evaluation of Phylogenetic Clustering and Meeting Sexual Partners at Social Venues Among Men Who Have Sex With Men and Transgender Women in Lima, Peru

Audrey V.M. Brezak

A thesis

submitted in partial fulfillment of the

requirements for the degree of

Master of Public Health

University of Washington

2018

Committee:

Ann C. Duerr

Joshua T. Herbeck

Angela K. Ulrich

Jennifer E. Balkus

Program Authorized to Offer Degree:

Epidemiology

University of Washington

Abstract

Epidemiological Evaluation of Phylogenetic Clustering and Meeting Sexual Partners at Social Venues Among Men Who Have Sex With Men and Transgender Women in Lima, Peru

Audrey V.M. Brezak

Chair of the Supervisory Committee:

Ann C. Duerr

Department of Epidemiology

New HIV infections in the Americas are predominately occurring among men who have sex with men (MSM) and transgender women (TW).[1] In Peru, there is a concentrated epidemic among MSM and TW, for whom the HIV prevalence exceeds 20% while the general HIV prevalence is less than one percent.[2] This study sought to identify geographic or behavioral traits associated with HIV infection among MSM and TW in Lima, Peru, using spatial clustering and phylogenetic clustering analyses.

The spatial clustering analysis evaluated the association of residential location with HIV infections using kernel density estimations using a case-cohort design and included 375 HIV cases and 322 in the sub-cohort as controls. This analysis did not find significant evidence of clustering of HIV infections by residential location. The phylogenetic clustering analysis assessed the association between belonging to a phylogenetic cluster and sexual partnering at specific venues within 60 days of incident HIV infection for 202 participants. We found an association between belonging to a phylogenetic cluster and reporting meeting a sexual partner at any venue or a specific venue (OR for any venue compared to reporting meeting a sexual partner at no venues 60 days before diagnosis= 2.33, $p<0.01$; OR for Vale Todo compared to none= 2.02, $p=0.03$). These findings provide support for more targeted HIV prevention interventions in social venues rather than traditional HIV outreach activities in residential neighborhoods.

Introduction:

In the Americas, men who have sex with men (MSM) represent the largest proportion of new HIV infections each year.[3] In Peru, there is a concentrated HIV epidemic among MSM and transgender women (TW), populations who have limited access to HIV prevention, testing, and care services. In Lima, the HIV prevalence among MSM and TW exceeds 20% while the general HIV prevalence is less than 1%. [2,3] A better understanding of drivers of HIV transmission among MSM and TW in Lima is needed to develop more targeted prevention activities to serve these populations.

Modeling studies of North American MSM estimate early infection is linked to one-third to two-thirds of onward transmissions.[4] Studies of HIV transmission among MSM populations incorporating phylogenetic methods have revealed that sequences from between 40-60% of MSM within their study sample belong to clusters of highly related infections, indicating rapid onward HIV transmission shortly after HIV acquisition and before sequences diverge.[5–9] Investigation of these rapid transmission events using molecular epidemiologic techniques could inform how to potentially disrupt transmission cycles in these populations.The combination of phylogenetic clustering analysis with epidemiologic methods can provide insight into transmission-network level phenomena, such as rates or factors associated with transmissions.[5–8,10,11] Transmission clusters may be characterized by geographic, demographic, or behavioral characteristics using molecular epidemiologic approaches to identify factors associated with rapid, ongoing transmission.[9,12–14] Logistic regression has been applied widely in molecular epidemiology to identify factors associated with ongoing transmission in pathogens such as HIV, hepatitis C virus, and *Mycobacterium tuberculosis*.[15–19]

To better understand potential drivers of rapid, onward transmission of the MSM/TW HIV epidemic in Lima, this study integrates spatial, phylogenetic, and behavioral data of high-risk men who have sex with men and transgender women who participated in the *Sabes* study. In this analysis, we link phylogenetic cluster analysis with individual's sexual partnership data to identify potential 'hotspots' for disease transmission, hotspots being defined as social venues frequented by individuals with increased

likelihood of being in a high-risk network (defined by membership within a phylogenetic cluster). This research seeks to evaluate the association between: 1) residential location and incident HIV infection; and 2) phylogenetic clustering and meeting sexual partners at social venues in the 60 days prior to HIV diagnosis. Findings of this study have the potential to guide the targeting of HIV intervention efforts, depending on whether incident HIV infections are associated with residential location or clusters of infection can be linked to venues in which sexual partners are established.

Methods:

Study Design, Data Sources, and Population:

This research was based on data collected from the parent study, *Sabes*, which means "do you know?" in Spanish.[20] The *Sabes* study sought to evaluate an expanded treatment as prevention intervention in Lima, Peru among MSM and TW. In this study, 3,336 high-risk men who have sex with men and transgender women were screened for HIV-infection. Those who tested negative at baseline were eligible for enrollment into a 24-month cohort study to investigate factors related to HIV acquisition. There were 2,084 HIV-uninfected MSM and TW that enrolled in the study and were tested each month for incident HIV. Participants completed monthly questionnaires detailing HIV risk behaviors including drug and alcohol use, venue attendance, and numbers of sexual partners met at venues. Those who acquired HIV during follow-up were enrolled into a separate study on the early treatment of HIV. The cohort HIV incidence was 10.4 per 100 person-years (n=256 infections). Over the course of the study, individuals who acquired HIV had their virus sequenced by the UW Mullins Virology Lab. Residential addresses were also collected from all cohort members.

In the first analysis, a case-cohort design was used to examine geographic variation in HIV acquisition risk. The cohort for analysis included 375 HIV-positive cases and 322 individuals who were randomly selected from the at-risk population of the cohort of ~3,000 high-risk MSM and TW (see Table 1). The second analysis is a cross-sectional study that uses molecular epidemiological methods to identify

genetically related clusters of HIV infections among the cohort and then, tests the association of membership in a network with venues at which individuals reported meeting a sexual partner in the 60 days prior to HIV diagnosis. The second analysis will include *Sabes* participants who acquired incident HIV infections which were sequenced (n= 201, see Table 1) and contextual sequences (see Figure 5 legend), which are used to provide sufficient variation and definition in the phylogenetic tree to allow for characterization of local transmission networks.

Outcomes:

For the geographic analysis, our outcome was significant spatial clustering of residential locations of individuals with HIV infections, when mapped over the sub-cohort residential distribution. Areas with significant spatial clusters of infection may indicate the presence of residential areas with higher HIV risk in Lima.

For the phylogenetic clustering analysis, the outcome was belonging to a transmission cluster. Clusters were defined by patristic distance threshold of 0.02 and a minimum cluster size greater than 2 *Sabes* individuals. Inferences from clustering analysis were supported by sensitivity analyses with variations in patristic-distance defined cluster thresholds (0.01, 0.02, 0.045, 0.06) of cluster size minimums of $\geq$ 2 *Sabes* individuals as well as 3 individuals with at least of $\geq$ 2 *Sabes* individuals (see Table 4). Belonging to a cluster (containing at least 2 *Sabes* cohort members) of related HIV infections, was used as the outcome as an indication of membership within a related sexual network within the *Sabes* cohort.

Exposures:

For the geographic analysis, the primary exposure of interest was residential location. The phylogenetic clustering analysis evaluated the exposure of reporting meeting a sexual partner at one of 16 listed social venues prior to two months of HIV diagnosis. Participants in the longitudinal study came to one of three clinics for HIV testing every month and answered questions about where they met sex

partners in the past 30 days using Computer Assisted Self-Interviewer (CASI). The CASI is designed to ask participants about sensitive information using a private, self-directed approach and was used in the study to understand HIV risk behaviors including drug use, sexuality, and sexual habits, and their associations with incident HIV infection. The CASI questionnaire used in the *Sabes* study asks about 16 social venues in Lima that were identified in focus group studies with individuals in this community as places that were frequented by MSM and TW and used for meeting sexual partners.

We focused our analysis on the period 60 days prior to HIV diagnosis (if individuals were missing 60-day data, the 30 days prior to HIV diagnosis was used as the relevant exposure period). This allows the assessment of the odds of membership within a HIV transmission network associated with meetung partners at high-risk venues during a time when HIV was likely acquired or when there was a high risk of onward HIV transmission. The exposures evaluated included the venues where individuals with incident HIV infection reported meeting sexual partners during the 60 days before diagnosis. Meeting a sexual partner at a venue was considered as a binary response (e.g. participant met a partner at venue "X" 60 days prior to HIV diagnosis vs. did not report meeting a partner at venue "X" 60 days prior to HIV diagnosis). Exposures were considered in combination, for example, if an individual reported meeting a sexual partner at *any* of the 16 venues in the 60 days prior to HIV diagnosis the participant was considered exposed, as well as individually (i.e. participant was considered exposed for venue "X" if they reported meeting a sexual partner at venue "X"). Analyses did not adjust for individuals meeting sexual partners at multiple venues during the exposure period.. Future analyses will evaluate evaluating behavioral confounders, such as reporting meeting partners at multiple venues, or the number of partners per venue,

Geographic analysis

To assess the impact of residential location on HIV acquisition, 322 individuals were randomly selected from the at-risk cohort of ~3,000 high-risk MSM and TW and their residential addresses were mapped to provide the underlying residential distribution of the cohort. Next, the residential locations of 375 incident HIV infections identified during the study were mapped on top of the sub-cohort

distribution. Kernel density estimates (smoothing technique used in spatial statistics) of the densities of cases and sub-cohort were calculated and evaluated on a grid. The bandwidth was chosen to be .01 for the kernel density estimates through trial and error until desired degree of discrimination between regions was achieved.

In comparing these maps, a spatial risk surface was estimated, and the odds of incident HIV infection was estimated across the spatial surface. The null hypothesis represents a constant odds surface and is evaluated via a Monte Carlo test in which the points are randomly re-labelled and the integral is evaluated on a grid. We used the overall p-value for the Monte Carlo test to determine if there is evidence of non-constant odds over the map as a whole. This indicates if there are geographic areas of higher HIV risk in Lima, which may identify residential neighborhoods or specific areas with higher HIVprevalence.

Phylogenetics

For the phylogenetic analysis, the *Sabes* study provided sequence data on ~700 nucleotides of the *pol* (partial reverse-transcriptase and partial protease) gene from 342 participants, 201 of these have exposure information and were included in the logistic regression analyses.

To provide definition to the local epidemic, we added to this tree approximately 700 contextual sequences to root the tree (see Figure 5 legend). We used Geneious to align the sequences using MUSCLE and reconstructed maximum-likelihood phylogenies under the General Time Reversible model of nucleotide substitution with substitution rate heterogeneity, and then imported the phylogeny into R software package *Seattlepolphylogenetics* to conduct patristic distance-based clustering analyses. The use of patristic distance-based clustering methods was shown to have the highest performance in identifying majority and minority populations modeled under a number of scenarios among six popular clustering methods.[21] Patristic distance, which is defined as the sum of branch lengths on the path from one tip to another in the tree,[22] is used similarly as an extension of clustering by pairwise genetic distance but is reputed to be less sensitive to population coverage.[19,21]

Using this method, we identified phylogenetic clusters with a range of conservative and strict patristic distance clustering thresholds (0.01, 0.02, 0.045, 0.06) and varying minimum cluster sizes (clusters containing at least 2 individuals that are both *Sabes* cohort members and clusters containing at least 3 individuals with 2 that are both *Sabes* cohort members).[5,6,8,] Identifying individuals whose sequences cluster by a defined genetic or patristic distance-based threshold is used as a proxy for characterizing individuals within a related transmission network. [16] Analyses were conducted to explore the effect of increasing the patristic-distance clustering thresholds and minimum cluster sizes on the proportion of the cohort that is identified as members in clusters (see Table 3).[2,9,21] We calculated the proportion of the cohort that belonged to a phylogenetic cluster for the set of cluster analysis thresholds (see Table 3), as is often done in studies of HIV transmission to evaluate the appropriate set of thresholds for the data and epidemic structure.[13,16,21]

Statistical Analyses

Descriptive analyses were performed to characterize the study populations according to being in a cluster or not being in a cluster defined by the cluster thresholds. Participants in these categories were compared using Fisher's exact and Kruskal-Wallis tests (as appropriate).

Logistic regression analyses were used to identify factors associated with being in a cluster. All analyses were univariate and evaluated the odds of cluster membership associated with either reporting meeting a sexual partner at any venue or an individual venue of the 16 high-risk venues listed in the *Sabes* monthly CASI questionnaire. This analysis intended to establish whether each participant reported meeting a sexual partner at a pre-identified list of 16 venues in the 60 days prior to HIV diagnosis (when 60-day data was unavailable, the 30 days prior to HIV diagnosis was used) was associated with phylogenetic cluster membership.

Both the geographic and phylogenetic clustering analyses were conducted using Geneious (version 10.2.3) and R (version 3.4.3) through the RStudio interface (version 1.1.419).

Results:

Participant characteristics of geographic analysis

In total, 375 HIV-positive and 322 participants selected to be the sub-cohort were included in the case-cohort analysis (Table 1). The HIV-positive participants including 37 (10%) participants who were diagnosed with acute infection, 71 (19%) participants who were diagnosed with recent infection, 267 (71%) number of participants who were diagnosed with chronic infection. Those in the sub-cohort were selected from the cohort at risk of HIV at baseline and included 37 (11%) individuals who were diagnosed with HIV by the end of the 2-year follow-up study, 28 (9%) individuals diagnosed between 3-6 months after infection, 67 (21%) individuals diagnosed with recent HIV infection, and 190 (59%) individuals who remained HIV-negative.

Overall, the median age of the study population was 29 years (IQR: 22 – 34 years), most were well-educated, and generally had moderate incomes (68% of study population had a monthly income between $400-799). Most individuals in the population identified as bisexual (31%) or homosexual (62%), and 18% of individuals identified as transgender in the sub-cohort. Incident HIV seroconversion was observed in 31% of participants in the sub-cohort.

Mapping variation in HIV risk

The geographic analysis found that the distribution of the residential addresses of the cases and sub-cohort were not highly variable (see Figure 1). The kernel density estimates of the case (see Figure 2) and sub-cohort distribution (see Figure 3) did not differ much by observation, though visually, the HIV cases are more densely packed around the city center than sub-cohort, which is also where the social venues that were reported as frequent sources of sexual partners. In comparing these maps, a spatial risk surface was estimated, and the odds of incident HIV infection was estimated across the spatial surface (see Figure 4). The log odds ratio surface does have some variability, and there are higher odds of incident HIV as demonstrated by presence of yellow. A Monte Carlo test was used to evaluate if there is a constant odds surface over the map as a whole. The findings of the Monte Carlo test were that there was no evidence of non-constant odds over the map as a whole (p-value =0.27), indicating that there were no

areas identified with significantly increased odds of HIV infection based on residential addresses within the study population.

Population sources of sequences for phylogenetic tree reconstruction

In total, there were 342 HIV-positive individuals from the *Sabes* cohort that were sequenced and included in the phylogenetic tree. HIV infections identified either at baseline screening or during follow-up were eligible for sequencing, and a combination of both was used to reconstruct the phylogenetic tree. In baseline HIV screening for the *Sabes* study, 19.6% (654 of 3,336) were HIV positive. Among the individuals who were HIV-positive at baseline, 93 individuals' viruses were sequenced to include in the phylogenetic tree reconstruction: 14 individuals diagnosed with acute infection (acute HIV infection was operationally defined as a positive plasma HIV RNA test in a person with a negative third-generation HIV antibody test), 40 individuals diagnosed with recent infection (recent HIV infection was diagnosed by a positive third-generation rapid HIV test with a documented negative third-generation HIV-antibody or HIV-RNA test in the previous 90 days), and 39 individuals diagnosed with chronic infection (chronic HIV infection diagnosed after at least 3 months of infection). The phylogenetic tree was populated with sequences from these 93 baseline diagnoses within the cohort to provide greater population coverage and improve our ability to identify linked transmissions within this population. Among participants who were HIV negative at baseline screening (n=2,682), 249 participants acquired HIV during follow-up and were included in the phylogenetic tree. Of the 249 participants, 73 were diagnosed during acute infection, 127 were diagnosed with recent infection, and 49 were diagnosed with chronic infection.

To provide context to the tree networks, we added sequences from heterosexual populations in Peru (n=21) as well as publicly available sequences from LANL from South America from the past 10 years (n=572).[23,24] The addition of these outgroups to the tree allows us to better distinguish between local transmission clusters (see Figure 5 legend).

Participant characteristics in clustering analysis

Among the 249 participants diagnosed with incident HIV infection during the *Sabes* study period of follow-up, 207 had complete questionnaire data as the questionnaire was changed to include venue data

a few months into the study. Of the 207 with questionnaire data, 201 also had HIV sequence data and were included in the phylogeny and clustering analyses.

Among the 201 *Sabes* participants in the tree with venue attendance and partnering data covering the time of infection, 30% were acute diagnoses observed in the study, 46% were recent diagnoses, and 24% were considered 'chronic' diagnoses as their diagnosis was recognized between 3-6 months post-infection.

Comparing participants who did not cluster (n = 87) versus those were considered members of a phylogenetic cluster (n = 114) under the default analysis thresholds (see Table 1, 2), there were approximately equal distribution of acute (29% of those not in clusters; 31% of those in clusters), recent (46% versus 46%), and chronic infections (26% versus 24%). Individuals that were identified as members in a cluster of related infections were generally younger (51% participants of those not in a cluster were 18-24 years versus 59% participants who clustered were 18-24 years), more likely to have attended some post-secondary education (75% of those in clusters attended some post-secondary education, compared to 68%), and less likely to identify as a transwoman (18% of those not in clusters were transwomen compared to 6% of those in clusters were transwomen); however, none of these differences were statistically significant.

In evaluating the risk of meeting a sexual partner at any venue within 60 days of diagnosis, there were significant differences by participants who clustered and who did not. Among participants who clustered, 58% (66 of 114) reported meeting a sexual partner within 60 days of diagnosis at any venue, compared to 38% (33 of 87) participants not in clusters (p=0.004). In terms of evaluating an individual venue, among participants in clusters, 27% (31 of 114) reported meeting a sexual partner within 60 days of diagnosis at Vale Todo, compared to 16% (14 of 87) participants not in clusters (p=0.057).

HIV phylogenetic cluster composition

A range of patristic-distance based clustering thresholds and minimum cluster sizes identified from those commonly used for studies of HIV transmission among MSM were applied to the phylogeny to explore tendencies in proportion of the data clustering and size of clusters.[9,19]

The selection of a default cluster threshold was guided by estimating that individuals who are diagnosed with acute/recent infections are likely to cluster, particularly those diagnosed within the same 2-year time interval and geographic area. In the sample that was sequenced, 60% of individuals were diagnosed within 3 months of infection, consequently we should expect a similar proportion of the sequences to cluster if we use the findings of other HIV transmission dynamic models. Therefore, the thresholds that produced 55% of the study population belonging to a cluster became the default analysis threshold for reporting odds ratios. Our clustering analysis thresholds were also consistent with other studies of HIV microepidemics in MSM that use patristic-distance based clustering methods to investigate HIV epidemics.[9,19,21]

Factors associated with membership in a pair/cluster

In analyses to estimate the odds of membership in a phylogenetic cluster associated with meeting partners at specific venues in the 60 days prior to incident HIV diagnosis or 30 days prior to chronic HIV diagnosis, estimates were unstable in most cases due to generally small samples sizes (9 – 45 individuals exposed per venue (exposed being considered as reported meeting a partner in the 60 days before diagnosis at that venue). In logistic regression analyses evaluating the odds of membership in a cluster associated with meeting a sexual partner at *any* venue in the 60 days prior to infection, we observed estimates of the odds of membership of a phylogenetic cluster from 1.89 to 6.10 with varying clustering analysis thresholds (see Table 4).

Discussion:

In this study of incident HIV infections among MSM and TW in Lima, Peru, there was not a significant elevation of HIV risk associated with residential location, but there was an association between meeting a sexual partner at high-risk venues in the 60 days prior to HIV diagnosis and membership in a phylogenetic cluster.

The case-cohort geographic analysis indicates that there is some evidence of difference in risk of incident HIV infection, but not significant variation in the log odds over the entire region of Lima, indicating that areas of higher HIV risk are not localized in one spatial cluster, but more uniformly

distributed in the region. This study may provide support to implement HIV prevention and treatment outreach services in communities which were seen to have higher log odds of infection (see Figure 4), such as closer to the social venues and in certain outer areas of town.

The phylogenetic clustering analysis found an increased risk of belonging to a phylogenetic cluster associated with meeting sexual partners from various venues with 60 days of HIV diagnosis compared to those who did not report meeting sexual partners at social venues in the 60 days prior to diagnosis. This risk varies by venue; individuals who reported meeting sexual partners at certain venues (particularly the saunas and clubs) had higher risks of being a member of a phylogenetic cluster than at other types of venues (movie theaters and street corners). Consistent with findings from other studies of sexual partner meeting places among MSM, individuals who reported meeting sexual partners from common sources were more likely to be found in phylogenetic clusters, indicating membership within a common sexual network.[6,19,25]

In terms of our ability of identify clusters, we estimate our sampling density was robust enough to provide sufficient population coverage. This is a challenge we addressed by incorporating both incident and chronic infections into the study, as well as individuals for whom do not have exposure data in our phylogeny. Our confidence in population coverage is supported by other studies of HIV transmission among MSM, in which similar proportions of MSM fall out in clusters using the same thresholds used in our study (40-60% sequences cluster vs 55% in our study).[1,6,19,26]

Limitations of the geographic analysis are that we were not able to use time of diagnosis data to incorporate into an infectious disease model nor were we able to incorporate the reported sourcing of sexual partners at specific social venues (though they were mapped over the case-cohort residential data). Next steps in this analysis would be to use spatial regression to examine the association of living close to one of these commonly reported venues for sourcing sexual partners and incident HIV infection.

Limitations of our phylogenetic analyses include that the venue partnering–clustering logistic analysis did not consider a dose-response effect in the number of partners met per venue. We considered

the exposure binary in these analyses (e.g. did they report meeting sexual partners at venue X during the 60 days prior to HIV diagnosis or not?). However, we recognize that this is valuable information and future analyses of this data could incorporate social network analysis to account for the range in number of partners met at each venue during the 60-day exposure period prior to HIV diagnosis.

Our study made contributions in adding 342 new HIV sequences from MSM in Peru to the HIV genomics community as well as documenting the observed associations between meeting sexual partners at social venues in the 60 days prior to HIV infection and being a member of a phylogenetic cluster of related HIV infections. This analysis identified microepidemics associated with sexual partnership data at social venues, however it does not indicate whether transmission itself is associated with the sites identified. We identified an association between membership to a network of related transmissions and sexual partnership at any (any one of the 16 venues listed on the questionnaire) or sexual partnership at a few specific venues in Lima within 60 days of incident HIV infection. The implications of these findings are the venues identified associated with higher odds of phylogenetic clustering may be places where sexual networks are established. We recognize that venues are not high-risk themselves as they are likely not where transmissions occur, but rather the environment and high-HIV prevalence clientele that intersect at these venues that may drive HIV transmission. These findings demonstrate that social venues may be a potential place for HIV prevention interventions in this microepidemic among MSM and TW, where we may be able reach individuals who are at higher risk for HIV acquisition and onward transmission. These higher-risk venues could be points of intervention for HIV prevention services, testing, and linkage-to-care activities to augment traditional HIV prevention outreach activities conducted in neighborhoods.

References:

1. Beyrer, C. *et al.* The Expanding Epidemics of HIV Type 1 Among Men Who Have Sex With Men in Low- and Middle-Income Countries: Diversity and Consistency. *Epidemiol. Rev.* **32,** 137–151 (2010).
2. Baral, S., Sifakis, F., Cleghorn, F. & Beyrer, C. Elevated risk for HIV infection among men who have sex with men in low- and middle-income countries 2000-2006: a systematic review. *PLoS Med.* **4,** e339 (2007).
3. Beyrer, C. *et al.* The Expanding Epidemics of HIV Type 1 Among Men Who Have Sex With Men in Low- and Middle-Income Countries: Diversity and Consistency. *Epidemiol. Rev.* **32,** 137–151 (2010).
4. Cohen, M. S., Shaw, G. M., McMichael, A. J. & Haynes, B. F. Acute HIV-1 Infection. *N. Engl. J. Med.* **364,** 1943–54 (2011).
5. Brenner, B. G. *et al.* High Rates of Forward Transmission Events after Acute/Early HIV-1 Infection. *J. Infect. Dis.* **195,** 951–959 (2007).
6. Brenner, B. G., Ibanescu, R.-I., Hardy, I. & Roger, M. Genotypic and Phylogenetic Insights on Prevention of the Spread of HIV-1 and Drug Resistance in "Real-World" Settings. *Viruses* **10,** (2017).
7. Brenner, B. G. & Wainberg, M. A. Future of phylogeny in HIV prevention. *J. Acquir. Immune Defic. Syndr.* **63 Suppl 2,** S248-54 (2013).
8. Pao, D. *et al.* Transmission of HIV-1 during primary infection: relationship to sexual risk and sexually transmitted infections. *AIDS* **19,** 85–90 (2005).
9. Poon, A. F. Y. *et al.* Near real-time monitoring of HIV transmission hotspots from routine HIV genotyping: an implementation case study. *lancet. HIV* **3,** e231-8 (2016).
10. Lam, T. T.-Y., Hon, C.-C. & Tang, J. W. Use of phylogenetics in the molecular epidemiology and evolutionary studies of viral infections. *Crit. Rev. Clin. Lab. Sci.* **47,** 5–49 (2010).
11. Leigh Brown, A. J. *et al.* Transmission Network Parameters Estimated From HIV Sequences for a Nationwide Epidemic. *J. Infect. Dis.* **204,** 1463–1469 (2011).
12. Chalmet, K. *et al.* Epidemiological study of phylogenetic transmission clusters in a local HIV-1 epidemic reveals distinct differences between subtype B and non-B infections. *BMC Infect. Dis.* **10,** 262 (2010).
13. Hué, S., Clewley, J. P., Cane, P. A. & Pillay, D. HIV-1 pol gene variation is sufficient for reconstruction of transmissions in the era of antiretroviral therapy. *AIDS* **18,** 719–28 (2004).
14. Hall, M., Woolhouse, M. & Rambaut, A. Epidemic Reconstruction in a Phylogenetics Framework: Transmission Trees as Partitions of the Node Set. (2015). doi:10.1371/journal.pcbi.1004613
15. Jacka, B. *et al.* Drug use and phylogenetic clustering of hepatitis C virus infection among people who use drugs in Vancouver, Canada: A latent class analysis approach. *J. Viral Hepat.* **25,** 28–36 (2018).
16. Jacka, B. *et al.* Phylogenetic clustering of hepatitis C virus among people who inject drugs in Vancouver, Canada. *Hepatology* **60,** 1571–1580 (2014).
17. Slattery, M. L. The science and art of molecular epidemiology. *J. Epidemiol. Community Health* **56,** 728–9 (2002).
18. Tuite, A. R. *et al.* Epidemiological evaluation of spatiotemporal and genotypic clustering of <I>Mycobacterium tuberculosis</I> in Ontario, Canada. *Int. J. Tuberc. Lung Dis.* **17,** 1322–1327 (2013).
19. Poon, A. F. Y. *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.* **211,** 926–35 (2015).
20. Lama, J. R. *et al.* Design Strategy of the Sabes Study: Diagnosis and Treatment of Early HIV Infection Among Men Who Have Sex With Men and Transgender Women in Lima, Peru, 2013–2017. *Am. J. Epidemiol.* (2018). doi:10.1093/aje/kwy030

21. Poon, A. F. Y. Impacts and shortcomings of genetic clustering methods for infectious disease outbreaks. *Virus Evol.* **2,** vew031 (2016).
22. Farris J. The Meaning of Relationship and Taxonomic Procedure. *Syst. Zool.* **16,** 44–51 (1967).
23. Stekler, J. D. *et al.* Prevalence and impact of minority variant drug resistance mutations in primary HIV-1 infection. *PLoS One* **6,** e28952 (2011).
24. Soria, J. *et al.* Transmitted HIV resistance to first-line antiretroviral therapy in Lima, Peru. *AIDS Res. Hum. Retroviruses* **28,** 333–8 (2012).
25. Lubelchek, R. J. *et al.* Transmission clustering among newly diagnosed HIV patients in Chicago, 2008 to 2011: using phylogenetics to expand knowledge of regional HIV transmission patterns. *J. Acquir. Immune Defic. Syndr.* **68,** 46–54 (2015).
26. Bezemer, D. *et al.* Transmission networks of HIV-1 among men having sex with men in the Netherlands. *AIDS* **24,** 271–282 (2010).

Tables and Figures

Table 1. Descriptive statistics by study population

	HIV-positive **Analysis 1**		Sub-cohort **Analysis 1**		HIV-positive **Analysis 2**		Member of cluster **Analysis 2**	
	N	%	N	%	N	%	N	%
Total Participants	375	100	322	100	201	100	114	57
Age								
18-24	171	45	130	40	111	55	67	59
25-34	167	44	138	43	75	37	42	37
35+	37	10	54	17	15	8	5	4
Education								
No secondary	24	6	23	7	7	4	3	3
Some secondary	98	26	110	34	49	24	25	22
Post-secondary	253	67	189	59	145	72	86	75
Monthly Income								
<$400	41	16	36	16	36	18	19	16
$400-$799	214	64	202	68	134	67	79	70
≥$800	110	20	112	16	28	14	15	13
No response	0	0	0	0	3	1	1	1
Sexuality and Gender Identity								
Bisexual	91	24	99	31	62	35	37	32
Homosexual	233	62	155	48	112	63	67	59
Transwoman	44	12	58	18	23	11	7	6
Heterosexual	3	1	5	2	4	2	3	3
Sex Work Status								
Yes	46	12	62	19	21	10	9	8
No	327	88	258	81	179	89	104	91
No response	0	0	0	0	1	1	1	1
HIV Status								
Acute	37	10	37	11	60	30	35	31
Chronic	267	71	28	9	49	24	27	24
Recent	71	19	67	21	93	46	52	46
Negative	-	-	190	59	-	-	-	-

Table 2. Logistic regression analysis of reported meeting of sexual partners at select high-risk venues within 60 days of HIV diagnosis and being a in phylogenetic cluster

	Cohort participants N	%	Odds ratio of clustering**	95% CI	p-value
Total	201	100	-	-	-
Did not meet sex partner at venues (ref)	102	51	1.0	-	-
Met a sex partner* at:					
any venue	99	49	2.33	(1.44, 3.77)	<0.01
Vale Todo	45	22	2.02	(1.04, 3.97)	0.03
Sagitario	30	15	1.52	(0.73, 3.16)	0.26
Sauna 240	10	5	1.85	(0.46, 7.37)	0.37
Legendarias	9	4	1.57	(0.38, 6.46)	0.37
Sauna 69	9	4	2.81	(0.56, 13.86)	0.17
Discoteca80divas	9	4	0.96	(0.25, 3.68)	0.95
PlazaSanMartin	6	3	N/A***	-	-
Hostal Paraiso	6	3	N/A	-	-
LaCueva	5	2	N/A	-	-
Kapital	4	2	N/A	-	-
LaJarrita	4	2	N/A	-	-
Parque Central	3	1	N/A	-	-
WashingtonConQuila	3	1	N/A	-	-
VideoKaleta	2	1	N/A	-	-
Sauna Open	1	1	N/A	-	-
VideoClubMinotauro	0	0	N/A	-	-
Ellolas	0	0	N/A	-	-

*Refers to participants who reported meeting sexual partner(s) at venue(s) within 60 days of HIV diagnosis
** Cluster thresholds used were*:* patristic distance 0.02, cluster size of $\geq$ 2 *Sabes* participants; odds ratio compares the odds of clustering amongst individuals who reported sourcing sexual partner from any/specific venue in the 60 days prior to HIV diagnosis compared to individuals who did not report sourcing sexual partner from any/specific venue in the 60 days prior to HIV diagnosis
*** N/A due to small sample size

Table 3. Application of patristic-distance based clustering analyses to *Sabes* phylogeny

Analysis Threshold	Patristic Distance Threshold	Minimum Cluster Size	Number of Clusters	% Sabes Cohort in Clusters
1	0.01	2	95	27%
*2**	*0.02*	*2*	*179*	*50%*
3	0.045	2	282	79%
4	0.06	2	315	89%
5	0.01	3	43	12%
6	0.02	3	129	36%
7	0.045	3	265	75%
8	0.06	3	313	88%

*Analysis threshold 2 was selected as the default cluster analysis threshold

Table 4. Range in Odds of Clustering Associated with Meeting Partners at any Social Venue during 60 Days Before Diagnosis with Varying Cluster Thresholds
*Exposed in this analysis considers all participants who reported sourcing partners at *any venue* in the 60 days before diagnosis

Analysis Threshold	Number exposed* classified in clusters (n=99)	Odds Ratio of Clustering associated with identifying partners at *any* venue	95% CI	p-value
1	37	1.89	(1.15, 3.12)	0.01
2	66	2.33	(1.44, 3.77)	<0.01
3	92	3.41	(1.62, 7.16)	<0.01
4	98	5.73	(1.72, 19.0)	<0.01
5	21	2.7	(1.40, 5.16)	<0.01
6	49	1.98	(1.23, 3.17)	<0.01
7	88	2.87	(1.51, 5.47)	<0.01
8	98	6.1	(1.84, 20.21)	<0.01

Figure 1. Map of residential addresses of *Sabes* HIV cases, sub-cohort, and reported social venues used for meeting sexual partners, Lima,

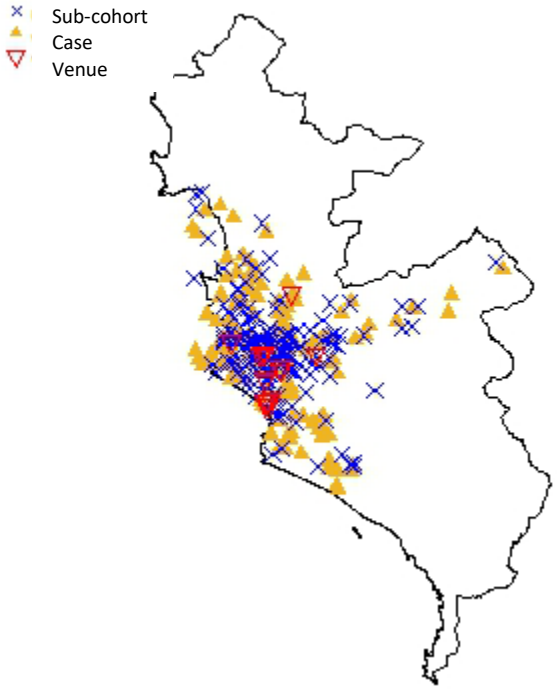


Figure 2. Map of Kernel Density Estimates of *Sabes* incident HIV case distribution, Lima, Peru 2013-2015

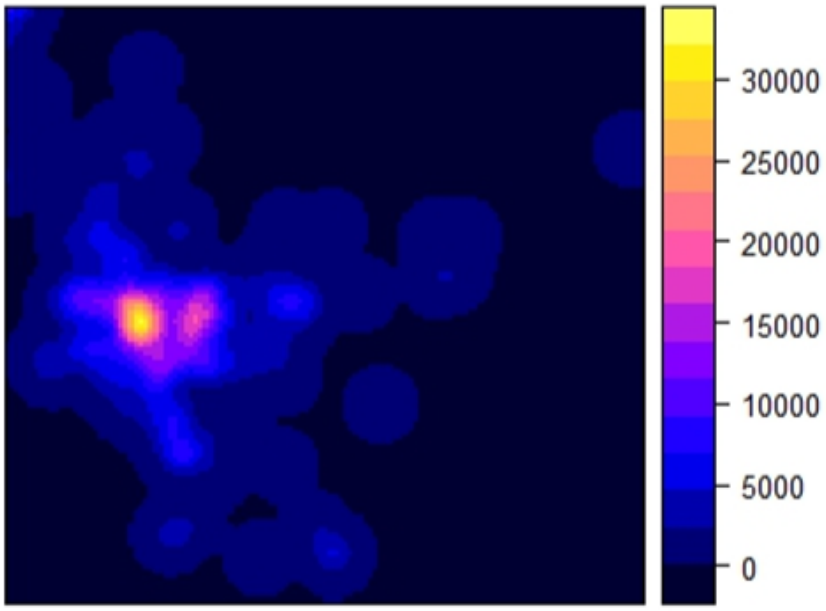


Figure 3. Map of Kernel Density Estimates of *Sabes* sub-cohort distribution, Lima, Peru 2013-2015

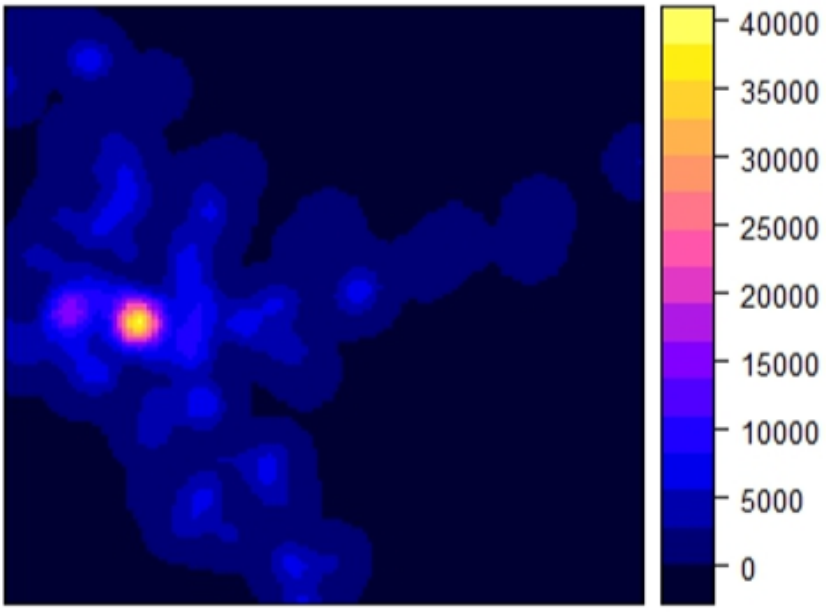


Figure 4. Map of the log odds ratio surface, Lima, Peru 2013-2015

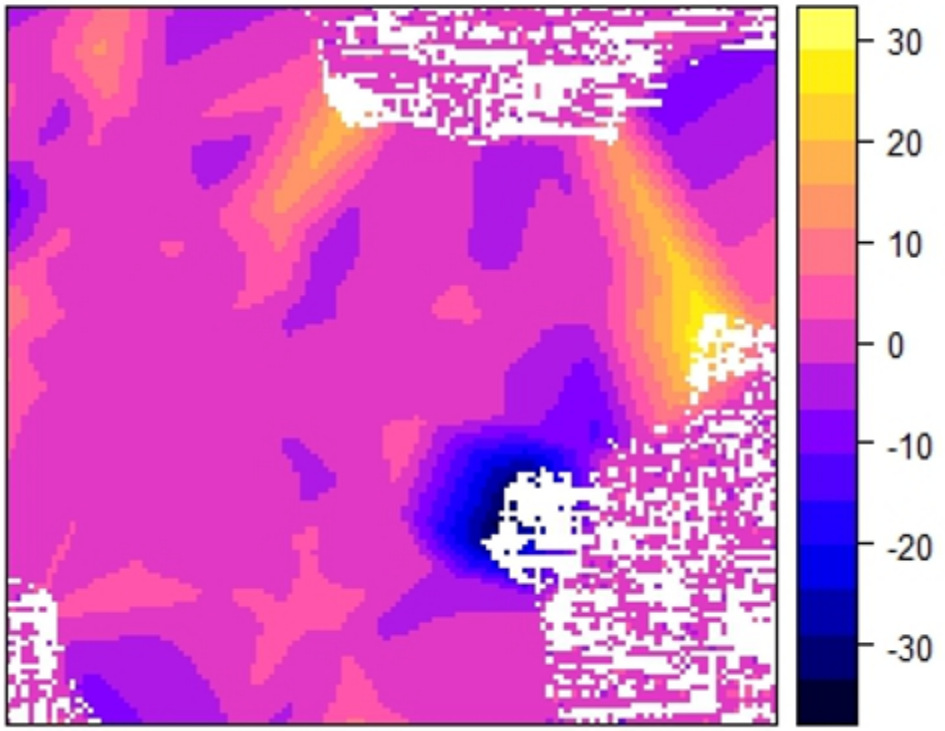
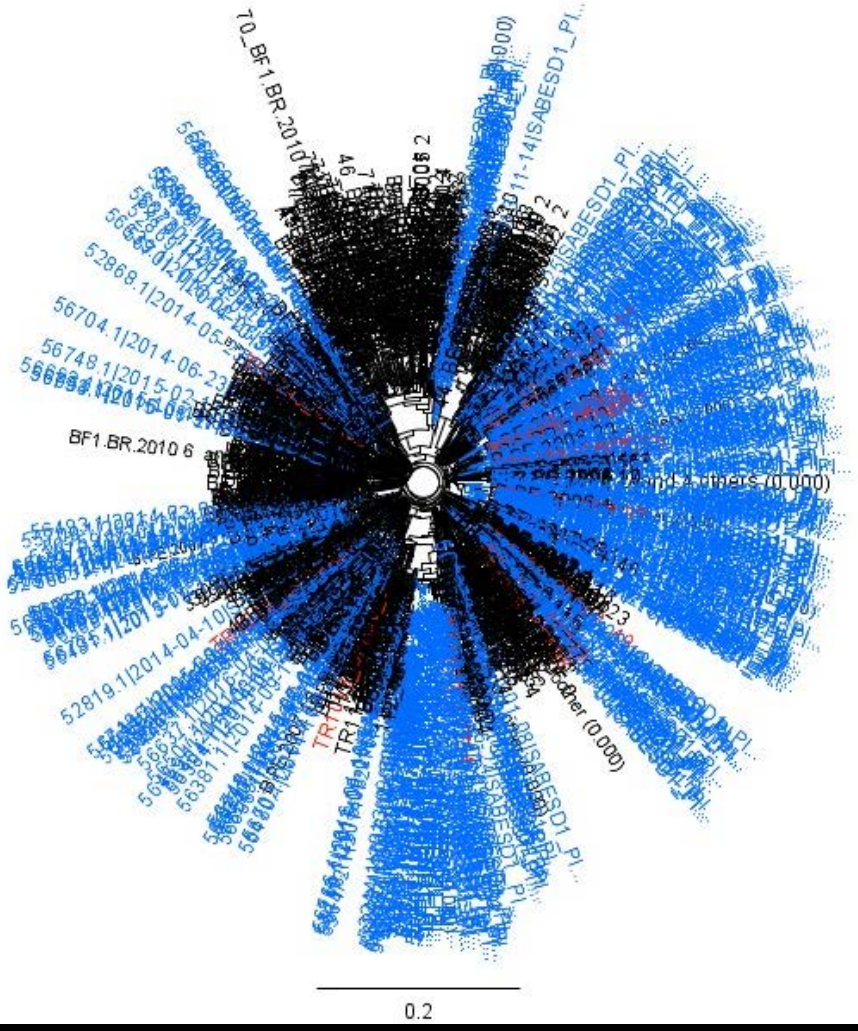


Figure 5. Phylogenetic reconstruction of HIV infections within the *Sabes* cohort and South America



Source	Number n=935	Infection Duration	Sampling Date	Population	Legend
Sabes	342	Incident	2013-2015	MSM	
LANL	572	Unknown	2006-2015	Unknown, MSM, Heterosexuals	
Frenkel	21	Unknown	2010	Pregnant women	