Longitudinal Phylogenetic Surveillance Identifies Distinct Patterns of Cluster Dynamics

Manon Ragonnet-Cronin, BSc, Marianna Ofner-Agostini, MSc,† Harriet Merks, BSc,* Richard Pilon, BSc,* Michael Rekart, MD,‡ Chris P. Archibald, MD,† Paul A. Sandstrom, PhD,* and James I. Brooks, MD**

Objective: Through the application of simple, accessible, molecular epidemiology tools, we aimed to resolve the phylogenetic relationships that best predicted patterns of cluster growth using longitudinal population level drug resistance genotype data.

Methods: Analysis was performed on 971 specimens from drug naïve, first time HIV positive subjects collected in British Columbia between 2002 and 2005. A 1240bp fragment of the *pol* gene was amplified and sequenced with relationships among subtype B sequences inferred using Neighbour-Joining analysis. Apparent clusters of infections having both a mean within group distance <0.031 and bootstrap value >80% were systematically identified. The entire 2002–2005 dataset was then re-analyze to evaluate the relationship of subsequent infections to those identified in 2002. BED testing was used to identify recent infections (<156 days).

Results: Among the 2002 infections, 136 of 300 sequences sorted into 52 clusters ranging in size from 2 to 9 members. Aboriginal ethnicity and intravenous drug use were correlated, and both were linked to cluster membership in 2002. Although cluster growth between 2002 and 2005 was correlated with the size of the original cluster, more related infections were found in clusters seeded from nonclustered infections. Finally, all large growth clusters were seeded from infections that were much more likely to be recent.

Conclusions: This population level phylogenetic analysis suggests that a greater increase in cluster size is associated with recently infected individuals, which may represent the leading edge of the epidemic. The most impressive increase in cluster size is seen originating from initially nonclustered infections. In contrast, smaller existing clusters likely describe historical patterns of transmission and do not substantially contribute to the ongoing epidemic. Application of this method for cross-sectional analysis of existing sequences from defined geographic regions may be useful in predicting trends in HIV transmission.

Key Words: HIV, transmission dynamics, cluster dynamics, phylogenetic, public health

(*J Acquir Immune Defic Syndr* 2010;55:102–108)

BACKGROUND

The intimate modes of transmission and the clinical latency of HIV present impediments to accurate surveillance of HIV transmission. Analysis of characteristic patterns in HIV genetic sequences may serve as a proxy for epidemiological links among individuals. Although not providing information on routes or direction of transmission,[1,2] the phylogenetic relationships can be used to infer shared patterns of risk. Phylogenetic techniques have enhanced our understanding of HIV transmission by highlighting the role of primary HIV infections in driving the epidemic[3–7] and by establishing links in the spread of drug-resistant HIV.[8,9] HIV phylogenetic analysis shows great promise for public health surveillance through its ability to augment traditional epidemiologic surveillance.

Phylogenetic analysis of HIV *pol* sequences extant in laboratory databases may contribute significantly to the identification and hence mitigation of ongoing HIV transmission within a population.[10] Despite dedicated resources and public health programs, the estimated rates of HIV infection remain essentially unchanged in Canada.[11] The difficulty is that although public health intervention within a small group at high risk of transmission requires fewer resources, identifying those at risk using traditional surveillance may be onerous, expensive, and potentially unfeasible. New approaches may be found by using phylogenetic analysis of the HIV *pol* sequences produced in routine HIV drug resistance testing.[12,13] As genotyping is now recommended for all newly diagnosed individuals,[14–17] the increasingly comprehensive HIV sequence databases present an opportunity to gain a better understanding of population level HIV transmission patterns.

The aim of this study was to use a simple, cross-sectional phylogenetic analysis to divine predictive factors of HIV transmission at the population level. HIV *pol* sequences were generated from drug-naive newly diagnosed individuals in the province of British Columbia. The change in phylogenetic relationships among the specimens collected between 2002 and 2005 was used to determine if patterns of cluster dynamics were associated with specific risk groups.

Received for publication March 1, 2010; accepted May 19, 2010.

From the *National HIV and Retrovirology Laboratories, Public Health Agency of Canada, Ottawa, Canada; †Surveillance and Risk Assessment Division, Public Health Agency of Canada, Ottawa, Canada; and ‡British Columbia Centre For Disease Control, Vancouver, British Columbia.

Correspondence to: Dr. James I. Brooks, MD, Room 3172, Building #6, 100 Eglantine Driveway, Tunney's Pasture, AL0603A2, Ottawa, Ontario, Canada (e-mail: james_brooks@phac-aspc.gc.ca).

Supplemental digital content is available for this article. Direct URL citations appear in the printed text and are provided in the HTML and PDF versions of this article on the journal's Web site (www.jaids.com).

Copyright © 2010 by Lippincott Williams & Wilkins

Copyright © Lippincott Williams & Wilkins. Unauthorized reproduction of this article is prohibited.

This approach may assist in the prevention of new HIV infections by guiding public health interventions toward groups at higher risk of ongoing HIV transmission.

METHODS

Study Population

The Canadian HIV Strain and Drug Resistance Surveillance Program monitors drug resistance in specimens from newly HIV diagnosed, antiretroviral-naive individuals. Between 2002 and 2005, 971 remnant HIV-infected serum specimens from newly diagnosed HIV cases in British Columbia (BC) were sent to the National HIV and Retrovirology Laboratories in Ottawa for genotyping. For all individuals involved in the study, limited epidemiological data (age, sex, exposure category, and ethnicity) were collected during a short interview. Duplicates are removed from the dataset at various stages: upon data capture, at the time of data entry into the database, and when the data are cleaned before analysis. Epidemiological characteristics of the study population were compared to those of the total HIV diagnosed population in BC during the same time period using χ^2 tests.[18] The Canadian HIV Strain and Drug Resistance Surveillance Program has institutional REB approval for all participant sites.[19]

Laboratory Analysis

Nucleic acids from samples were extracted, amplified, and sequenced according to protocol (see **Table, Supplemental Digital Content 1**, http://links.lww.com/QAI/A80). Subtype determination was based upon submission of the 1240bp *pol* fragment to the REGA HIV-1 Subtyping Tool—Version 2.0 (http://dbpartners.stanford.edu/RegaSubtyping/).

Sequences were assembled, edited, and aligned to the National Centre for Biotechnology Information HIV-1 reference genome (Accession # NC_001802), trimmed to identical length (1240bp) and gap-stripped using BioEdit Software, version 7.0.9.[20] All sequences used for analysis were deposited in the GenBank database (Accession # HM468499-HM469374). The presence of drug resistance mutations was determined using the Stanford HIV Drug Resistance Database sequence analysis program accessed before January 1, 2008 (http://hivdb.stanford.edu/pages/asi/).[21]

Recent infections (<156 days) were identified using 10 μL of serum with the BED capture enzyme immunoassay (BED-CIA, Calypte).[22]

Phylogenetic Analysis

Cross-sectional phylogenetic analysis was performed on sequences from patients diagnosed in 2002. To determine relationships among infections over time, all sequences from infections diagnosed between 2002 and 2005 were then analyzed in aggregate. Phylogenetic interrelationships among viral sequences were estimated using neighbour-joining trees,[23] under the Kimura-2-Parameter (K2P) model,[24] as implemented in MEGA3.[25] Pair-wise Jukes Cantor distance was <1 indicating that the data were suitable for phylogenetic analysis using a distance method. Robustness of relationships among sequences was evaluated using bootstrap analysis with 100 replicates. In both phylogenetic trees, simple systematic identification of clusters was based upon bootstrap values ≥80% and short branch lengths with an intra cluster distance of ≤0.031 nucleotide substitutions/site. The existence of a cluster was interpreted as evidence of more closely related infections, suggesting a more recent common source of infection. Clusters from 2002 were re-analyzed based upon the intracluster genetic distances. As this project was designed to elucidate the predictive patterns of the infections diagnosed in 2002, only those clusters observed among the 2002–2005 sequences that included a member from the 2002 cohort were included.

Evaluation of Cluster Dynamics

For each cluster in the 2002–2005 aggregate cohort, cluster growth factor was defined as the number of new sequences per sequence in the 2002 baseline cluster. For example, in a cluster containing 3 infections in 2002 and 9 in 2005, the growth factor would be (9–3)/3 = 2, which represent the addition of 2 additional members for each original seed member. For initially nonclustered sequences in 2002 that became part of a cluster in the aggregate analysis, growth was equivalent to the number of new cluster members.

Statistical Analysis

Age, sex, hierarchical HIV risk factor,[26] drug resistance profiles, results of the BED assay, and cluster analysis data were entered into a database and analyzed using SAS version 4.1 (SAS Institute Inc, Cary, NC) and SPSS version 11.0 (Cary, NC). Aboriginal people constitute a major ethnic group in Canada, comprising First Nations, Inuit and Métis. Factors associated with cluster membership in 2002 were examined using likelihood ratio tests and Pearson χ^2 test, as appropriate. Groups of infections were stratified according to whether they existed as part of a cluster in 2002 and if they did, by the size of the cluster to which they belonged. An analysis of variance was performed to examine whether cluster size in 2002 was associated with cluster growth between 2002 and 2005.

Sequences from 2002 were then divided into groups based on cluster membership status in 2002 and their pattern of growth over the subsequent 3 years. A univariate analysis was performed, comparing the groups to determine whether epidemiological characteristics, transmitted drug resistance classification, and BED-CEIA results associated with the different rates of cluster growth using a χ^2 test with Yates correction, analysis of variance, and the Kruskal–Wallis test. Variables with $P < 0.20$ were further included in a multinomial regression analysis.

RESULTS

Study Population

More than 90% of the 971 HIV specimens were subtype B. Of the 876 subtype B HIV *pol* sequences, 300 specimens were from 2002, 184 from 2003, 305 from 2004, and 85 from 2005, representing 71%, 45%, 69%, and 21% of 1674 HIV diagnoses made in BC in the respective years. A portion of the 2003 specimens were destroyed in transit, and the 2005 specimens represent collection up to the date analyses began. Characteristics of the sampled population were not

© 2010 Lippincott Williams & Wilkins

Copyright © Lippincott Williams & Wilkins. Unauthorized reproduction of this article is prohibited.

significantly different from those of the total HIV diagnosed population in BC during the same period[18] other than for ethnicity, where smaller than expected numbers of whites and Aboriginals were observed (Table 1). Of the specimens tested, 247 (28.2%) were classified as recent (<156 days) by BED.

Phylogenetic Analysis

Only subtype B specimens were included in the analysis as they represented >90% of infections in BC and inclusion of the few non-Bs could result in artefactual clustering.

The baseline dataset consisted of sequences from 300 individuals diagnosed with HIV in 2002. In the phylogenetic tree, 136 (42.7%) individuals were identified as belonging to 52 separate clusters. Cluster size ranged between 2 and 9 individuals, and 32 of the clusters (61.5%) were made up of only 2 individuals. There was no difference in epidemiological variables, BED results or cluster growth factor when clusters were stratified according to intracluster distance (data not shown).

A second phylogenetic tree was built using sequences from all 876 B-subtype specimens obtained from 2002 to 2005. More than 93% of the clustered sequences identified in 2002 remained clustered in the aggregate analysis. Overall, 389 specimens (44.4% of the total) were found in 94 clusters that contained at least 1 sequence from 2002. The clusters were

TABLE 1. Epidemiological Characteristics and Laboratory Results of the 876 Individuals in the Cohort as Compared With the Total HIV Diagnosed Population in British Columbia During the Same Time Period

	Study Population		HIV diagnoses total population		
Characteristic	Number of Individuals	Percent	Number of Individuals	Percent	Pearson's χ^2
Sex					
Male	708	80.9	1332	79.6	$\chi^2 = 1.88$; $df = 1$; $P = 0.17$
Female	162	18.5	342	20.4	
Unknown	6	0.6	0	0	
Total	876	100	1674	100	
Age group					
14–19	8	0.9	19	1.1	$\chi^2 = 2.15$; $df = 4$; $P = 0.1$
20–29	131	14.9	276	16.5	
30–39	297	33.9	565	33.8	
40–49	281	32.1	520	31.1	
50+	156	17.8	285	17.0	
Unknown	3	0.3	9	0.6	
Exposure category					
MSM	337	38.5	631	37.7	$\chi^2 = 14.7$; $df = 9$; $P = 0.1$
MSM/IDU	31	3.5	55	3.3	
IDU	258	29.5	449	26.8	
Blood	7	0.8	26	1.6	
Heterosexual	154	17.6	361	21.6	
Perinatal	1	0.1	5	0.3	
Other	10	1.1	19	1.1	
NIR	2	0.2	3	0.2	
STW	3	0.3	6	0.4	
STW/IDU	43	4.9	98	5.9	
Unknown	30	3.4	21	1.3	
Ethnicity					
White	558	63.7	1078	64.4	$\chi^2 = 31.4$; $df = 2$; $P < 0.01$*
Aboriginal	105	12	258	15.5	
Other	208	23.7	279	16.7	
Unknown	5	0.5	59	3.6	
BED results (< > 156 days)					
Established	582	66.4	—	—	—
Recent	247	28.2	—	—	—
No result	47	5.4	—	—	—
Drug resistance					
WT	802	91.5	—	—	—
DR	74	8.5	—	—	—

Results that were not reported are indicated as "unknown". These were not included in the χ^2 tests. Results are shown for the entire 2002–2005 dataset; the same results were found when the data were broken down year by year (data not shown).

*Statistically significant.

MSM, men who have sex with men; NIR, no identifiable risk; STW, sex trade worker; WT, wild type; DR, drug resistant; df, degree of freedom.

© 2010 Lippincott Williams & Wilkins

Copyright © Lippincott Williams & Wilkins. Unauthorized reproduction of this article is prohibited.

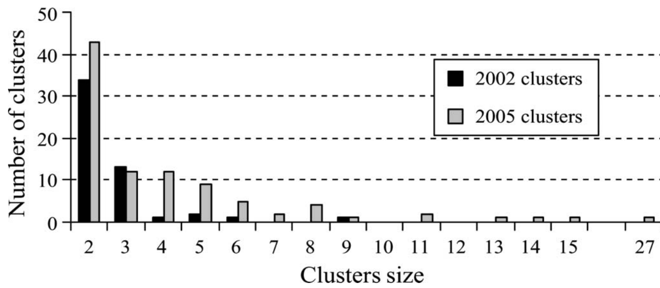

FIGURE 1. Distribution of clusters (2002 and 2005). The number of clusters was inversely related to cluster size, and there were a large number of 2-people clusters in 2002 and 2005. Results are shown for existing cluster in 2002 (black) and aggregate 2002–2005 clusters (grey).

evenly divided in number between pre-existing (47) and new (47) clusters. Forty-three sequences not in clusters in 2002 were identified as having formed clusters by 2005. Most clusters contained between 2 and 15 individuals, with a single cluster containing 27 individuals (Fig. 1). Thirteen sequences in 5 clusters were no longer associated with their previous clusters either because bootstrap values had fallen below 80% or intracluster distance had increased above 0.031 substitutions per site. The clusters from which they originated were not different from the persisting clusters with respect to intracluster distance (in 2002, data not shown) and were excluded from further analysis.

Cluster Dynamics

The relative changes in cluster size were examined to determine factors predictive of cluster growth that may be indicative of active subepidemics. Of the 47 clusters identified in 2002 and 2005, the composition of 21 clusters remained unchanged, whereas more than half (26 of 47) increased in size. The average growth factor for each 2002 cluster is shown in Figure 2, stratified by size of the seed cluster. As clusters with a growth factor of 0 did not, by definition, change in size, they were excluded from subsequent analysis. On average, the 26 clusters that increased in size grew by a factor of 0.99 (defined as new individuals/seed 2002 member). The size of the original cluster significantly influenced cluster dynamics with growth factors of 0.59 for cluster size ≤3, and 1.69 for clusters ≥ 4 ($P < 0.0005$). In contract, the majority of binary clusters, those with 2 individuals, remained stable (19 of 32, 60%) and did not increase in size.

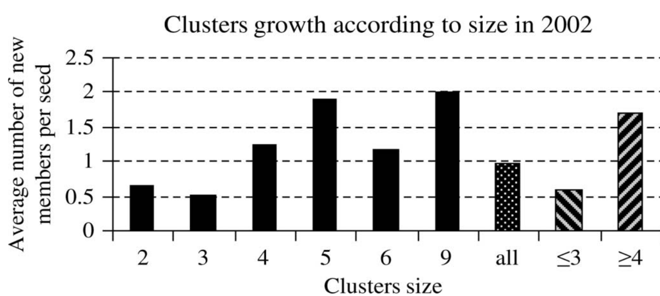

FIGURE 2. Average growth per cluster size. Average growth was calculated for each cluster as the number of new members per seed member. These were averaged for all clusters of the same size to trace a graph representing average growth per cluster size in 2002. Cluster growth increased with cluster size.

The 2002 cohort was stratified into sequences that were found as members of root clusters (RC, n = 136) or not clustered at all [NC, n = 164]. If these seed members were part of existing clusters, or went on to form new clusters in the aggregate analysis, they were categorized as having more than 1 new sequence per seed member (average growth, AG > 1) or less than 1 new sequence per seed member (AG ≤ 1). The reference group for subsequent analysis were sequences that were not clustered in 2002 and did not form clusters throughout the length of the study (NC-NC, n = 121). These groups were those compared in the logistic regression (Tables 2 and 3).

Of the 136 sequences found as members of root clusters, 42 of these sequences were part of clusters that continued to grow with AG > 1, with 81 RC sequences found to have an AG ≤1. Although 73.8% of the nonclustered sequences did not join clusters over time, 43 sequences formed the seed members of clusters comprising 111 new sequences. The average growth factor for this latter group was 2.4 which was greater than any other group ($P < 0.0005$). Approximately, half of this last group of newly formed clusters consisted of 2 sequences, whereas the remaining half of clusters comprised between 3 and 11 members.

Statistical Analysis

The baseline 2002 dataset was then interrogated in isolation to determine which factors were associated with cluster membership. A likelihood ratio test revealed an association between exposure category and cluster membership ($P < 0.001$, Table 4). IDUs were over represented within clusters making up 48.5% of this group (OR: 1.929, 95% Cl: 1.210 to 3.073) although only representing 38% of the specimens examined. Within the IDU exposure category, sequences identified as originating from Aboriginals were strongly associated with cluster membership (OR: 1.833, 95% Cl: 1.127 to 2.981) despite Aboriginals being under-represented in the dataset as compared with total diagnoses (Table 1). Aboriginal ethnicity and IDU exposure category were found to be correlated, with 33 of the 43 individuals with Aboriginal ethnicity exposed through IDU (76.7%), versus 91 of the 257 non-Aboriginals (35.4%). Analysis of the relationship between Aboriginal status and IDU as a risk factor revealed that IDU was a better predictor of cluster membership.

There was a slightly higher prevalence of recent infections among those sequences in clusters (31.6%) compared with those not (23.2%); however, overall differences were not statistically significant ($P = 0.10$). No association was found between cluster membership and sex or age.

In the univariate analysis, there were statistically significant differences between the 5 groups defined in both ethnicity and exposure category (Table 2). Variables with P values less than 0.20 included ethnicity, exposure category and STAHRS results, which were selected for further analysis using multinomial logistic regression. Individuals who never joined clusters (NC-NC) were used as a reference group for comparisons (Table 3). There was a significant presence of

© 2010 Lippincott Williams & Wilkins

Copyright © Lippincott Williams & Wilkins. Unauthorized reproduction of this article is prohibited.

TABLE 2. Characteristics of the 5 Groups Identified Amongst HIV First-Time Positive Treatment-Naive Cases in British Columbia and Results of Univariate Analysis

Variables	RC AG > 1, n = 42	RC AG ≤ 1, n = 81	NC AG > 1, n = 23	NC AG≤1, n = 24	NC-NC, n = 121	P
Ethnicity						
White	29 (69%)	57 (70.4%)	10 (43.5%)	13 (54.2%)	85 (70.2%)	0.0007*
Aboriginal	9 (21.4%)	14 (17.3%)	8 (34.8%)	1 (4.2%)	10 (8.3%)	
Other	4 (9.5%)	10 (12.3%)	5 (21.7%)	10 (41.7%)	26 (21.5%)	
Exposure						
Category	3 (7.1%)	14 (17.3%)	3 (13%)	5 (20.8%)	20 (16.5%)	0.02*
Heterosexual	20 (47.6%)	32 (39.5%)	10 (43.5%)	7 (29.2%)	27 (22.3%)	
IDU	10 (23.8%)	33 (40.7%)	5 (21.7%)	9 (37.5%)	54 (44.6%)	
MSM	5 (11.9%)	1 (1.2%)	2 (8.7%)	2 (8.3%)	9 (7.4%)	
STW/IDU	4 (9.2%)	1 (1.2%)	3 (13%)	1 (4.2%)	11 (9.1%)	
Other	3 (7.1%)	14 (17.3%)	3 (13%)	5 (20.8%)	20 (16.5%)	
BED result						
Recent	16 (38.1%)	22 (27.2%)	10 (43.5%)	6 (25%)	24 (19.8%)	0.06
Established	26 (61.9%)	59 (72.8%)	13 (56.5%)	18 (75%)	97 (80.2%)	

*Statistically significant.
RC, root cluster; NC, not clustered in 2002; NC-NC not clustered in 2002 or 2005; MSM, men who have sex with men; STW, sex trade worker.

sequences categorised as recent infections in the groups with the largest cluster growth factors. When compared with the reference group, clusters with large growth factors arising from an existing cluster or forming *denovo*, were 2.6 times (Cl: 1.184 to 5.861) or 3.2 times (Cl: 1.184 to 5.861) more likely to contain recent infections than those with low growth factors. Among those clustered infections with a large growth factor, sequences were 4.5 times (Cl: 1.131 to 18.001) more likely to have originated from IDU than the reference group.

DISCUSSION

Risk factors for HIV transmission are well understood; however, control of HIV transmission among diverse populations has proved difficult. Heterogeneity of infectiousness is likely to be an important determinant of the epidemiology of HIV, with a minority of infections resulting in onward transmissions.[27] Models have shown that the identification of predictive correlates of higher infectiousness can lead to dramatic improvements achieved through targeted prevention and control policies.[28] From the point of view of public health, these kinds of strategies are the most efficient, but they require a better knowledge of variability in individual infectiousness.[29]

This study sought to identify the epidemiological characteristics associated with differential cluster dynamics in BC. Cluster growth, as defined in this study, is not a measure of onward transmissions as our population is newly diagnosed and not newly infected. However, a cluster that forms around an originally isolated sequence still reflects related infections, be they new infections or newly diagnosed infections. When sampled over a period of time, clusters that have larger growth factors may represent subepidemics with greater relative transmission rates.

Large cluster growth factors were found to be associated with early infection, indicating that population phylogenetics may identify those groups at the leading edge of the epidemic. The estimated time of infection is unfortunately not entirely precise; the BED assay has been shown to have 86% specificity and 80% sensitivity.[30] However, the correlation that we found between recent infections and cluster growth is consistent with a number of other phylogenetic cohort studies in Switzerland,[7] United Kingdom,[4] and Quebec.[3] In contrast to other studies, our population level, longitudinal cluster identification strategy was not designed to specifically identify transmission chains. Instead, the goal was to identify the population level common risk behaviors, or social network factors, which drive the formation of these clusters. Within this framework, early infections associated with cluster growth over time and are interpreted as supporting the conclusion that

TABLE 3. Multinomial Regression Analysis for Different Groups of Newly Diagnosed Treatment-Naive HIV Cases in British Columbia

	RC AG>1, n = 42		RC AG≤1, n = 81		NC AG >1, n = 23		NC AG≤1, n = 24	
	OR (95% CI)	P	OR (95% CI)	P	OR (95% CI)	P	OR (95% CI)	P
Aboriginals	1.529 (0.455 to 5.139)	0.492	1.535 (0.664 to 3.546)	0.316	0.476 (0.133 to 1.712)	0.256	0.27 (0.009 to 0.798)	0.017*
BED result recent	2.634 (1.184 to 5.861)	0.018*	1.366 (0.691 to 2.702)	0.370	3.191 (1.184 to 5.861)	0.021*	1.504 (0.526 to 4.300)	0.447
IDU	4.512 (1.131 to 18.001)	0.033*	1.456 (0.600 to 3.532)	0.406	2.699 (0.595 to 12.250)	0.198	1.813 (0.449 to 7.325)	0.403

All groups in the table were compared with individuals who never joined clusters (NC-NC).
*Statistically significant.
Cl, confidence limits; NC, not clustered in 2002; NC-NC, not clustered in 2002 or 2005; OR, odds ratio; RC, root cluster.

© 2010 Lippincott Williams & Wilkins

Copyright © Lippincott Williams & Wilkins. Unauthorized reproduction of this article is prohibited.

TABLE 4. Cluster Membership According to Exposure Category

	Cluster Membership in 2002			
Exposure Category	**No**	**(%)**	**Yes**	**%**
Het	30	(18.3)	20	(14.7)
IDU	49	(29.9)	66	(48.5)
MSM	65	(39.6)	45	(33.1)
MSM/IDU	7	(4.3)	2	(1.5)
Other	5	(3.1)	3	(2.2)
Unknown	8	(4.9)	0	(0.0)
Total	164	(100)	136	(100)

MSM, men who have sex with men; Het, heterosexual.

clusters with large growth factors may represent the leading edge of the epidemic.

Conventional HIV phylogenetic analysis generates clusters that reflect proximal transmission events. However, these genetic associations are historical in nature and reflect epidemiologic associations between partners that existed at some time in the past. In contrast, this study used phylogenetic analysis of cross-sectional data to explore whether those associations, inferred by clustering, were still relevant and predictive of ongoing HIV transmission. The somewhat surprising result was that it was the nonclustered sequences that were associated with the largest growth factor of 2.4. This is in stark contrast to the clusters consisting of 2 or more related sequences which, when analyzed together, demonstrated an unsustainable growth rate of only 0.99. An interpretation is that analysis of existing clusters informs on historical transmission events, but it is the identification of unique features among those nonclustered infections with large growth factors that will predict future transmission points. The potential is that through orienting public health prevention toward features associated with certain nonclustered infections, a disproportionately greater effect on HIV prevention may be achieved.

Although identification of a cluster *per se* was not useful in determining which sequences were likely to be associated with ongoing transmission; when stratified by size, larger 2002 clusters were associated with larger cluster growth factors. Clusters comprising 4 or more individuals grew more than those of 3 and less members (1.69 vs. 0.59). Although this finding was driven by only 5 large clusters, it can be postulated that these large clusters may contain "super spreaders" or "network hubs" that are responsible for high transmission rates due to their increased connectivity.[29] In contrast, stable social situations or awareness of serostatus may limit the growth of smaller clusters, as has been shown elsewhere.[31] Providing further refinement to population level prevention strategies, the finding that nearly 2/3 of all 2002 clusters were binary in nature suggests that identification of epidemiologic features associated with these clusters may prevent nonproductive public health prevention efforts.

Although the findings are compelling, there are some limitations to our analysis. Specimens were passively collected and, therefore, not all of the new diagnoses in BC were available for the analysis despite possibly forming part of the infection clusters. Although this may result in selection bias, we were unable to demonstrate a difference in epidemiologic features in our sample set when compared with those reported through HIV case surveillance, other than for ethnicity where this did not bias subsequent results.[18] The large number of clusters containing only 2 individuals may reflect limited depth of sampling. However, the relative binary cluster dynamics likely reflect underlying differences in transmission characteristics among this group.

Non–subtype B sequences were deliberately excluded from the phylogenetic analysis, resulting in the exclusion of a small number of cases (<10%). As non–B subtypes are often associated with infections acquired abroad, this has the added limit of preventing any conclusion regarding the effect of migration and travel. Non–B sequences are being analyzed separately in a follow-up study, as has been done by Hughes et al[32] in the United Kingdom. A multi-layered approach to removing duplicate species from the dataset makes the possibility of a significant effect due to duplicate sampling remote. Finally, there is also the chance, as would exist for any study of this nature, that the results may be influenced by referral bias as the consequence of contact tracing for related infections. These are inherent limitations of the methodology involved in collecting these specimens, and efforts are being made to acquire information on whether the specimens were collected in the course of contact tracing. Conversely, this latter point hints at another application of population-level phylogenetics as a means of evaluating the success of contact tracing in reducing ongoing transmission.

The HIV *pol* gene was selected for phylogenetic analysis.[33] Although the selective pressure of antiretroviral treatment may theoretically drive convergent evolution, from a practical perspective, the response to drug selective pressure has not been shown to confound results.[12,34] The advantage of analysis on the *pol* gene is that one can ultimately leverage the vast database of HIV drug resistance genotypes for study. Consideration, however, must be given to all studies involving HIV phylogenetics, in that there is the potential for public health authorities to use existing genotyping databases to track HIV transmission events.[10]

The systematic cluster identification strategy allows for this analysis to be consistently implemented in any laboratory with a basic understanding of bioinformatics. We specifically defined clusters using less conservative criteria than in studies seeking to establish direct transmission links,[3,8,35] to enhance clustering of related infections although maintaining high support values. Our criteria were consistent with the bootstrap proportions ≥70% used for the identification of HIV transmission clusters in other work.[36,37] Our approach was to use the intracluster genetic distance to increase the resolution of cluster identification; however, when using this parameter, we found no clear correlation with cluster membership, size, or epidemiological characteristics. Regardless, our cluster definition remained robust as shown by the persistence of 93% of the 2002 clusters after the addition of nearly 500 more sequences.

In summary, a phylogenetic approach was developed that used pre-existing genotypes, and basic HIV surveillance data, to provide valuable insight into population-level HIV

© 2010 Lippincott Williams & Wilkins

Copyright © Lippincott Williams & Wilkins. Unauthorized reproduction of this article is prohibited.

transmission patterns. These data suggest that, analyzed in aggregate, existing clusters in a cross-sectional sample may not predict ongoing transmission routes. Specimens that were initially nonclustered ultimately became associated with a greater number of related infections. The greater accumulation of related sequences originating from unclustered, rather than existing clusters, suggests that, from a phylogenetic analysis perspective, it is the former group that represents the leading edge of the epidemic. Restricting public health inferences, based on phylogenetic analysis, to those factors only associated with the existence of a cluster may limit the predictive power of this type of analysis. The finding that at population-level fast growing clusters contain a disproportionate ratio of recent infections further supports the concept that cluster dynamics as opposed to cluster existence is the better predictor of population-level trends in HIV transmission. We are currently evaluating our model to determine if it is valid for predicting infections detected in the years 2006–2009.

ACKNOWLEDGMENTS

This work was supported by the Federal Initiative on HIV/AIDS, Canada. We are grateful to the National Laboratory for HIV Reference Services at the Public Health Agency of Canada for the BED testing; and to Ms. Elsie Wong, Mr. Darrel Cook, Dr. Mel Krajden, and Dr. Mark Gilbert at the BCCDC for supporting this ongoing collaboration.

REFERENCES

1. Pilcher CD, Wong JK, Pillai SK. Inferring HIV transmission dynamics from phylogenetic sequence relationships. *PLoS Med*. 2008;5:e69.
2. Pillay D, Rambaut A, Geretti AM, et al. HIV phylogenetics. *BMJ*. 2007; 335:460–461.
3. Brenner BG, Roger M, Routy JP, et al. High rates of forward transmission events after acute/early HIV-1 infection. *J Infect Dis*. 2007;195:951–959.
4. Pao D, Fisher M, Hue S, et al. Transmission of HIV-1 during primary infection: relationship to sexual risk and sexually transmitted infections. *AIDS*. 2005;19:85–90.
5. Pilcher CD, Tien HC, Eron JJ Jr, et al. Brief but efficient: acute HIV infection and the sexual transmission of HIV. *J Infect Dis*. 2004; 189:1785–1792.
6. Pinkerton SD. Probability of HIV transmission during acute infection in Rakai, Uganda. *AIDS Behav*. 2008;12:677–684.
7. Yerly S, von Wyl V, Ledergerber B, et al. Transmission of HIV-1 drug resistance in Switzerland: a 10-year molecular epidemiology survey. *AIDS*. 2007;21:2223–2229.
8. Kaye M, Chibo D, Birch C. Phylogenetic investigation of transmission pathways of drug-resistant HIV-1 utilizing pol sequences derived from resistance genotyping. *J Acquir Immune Defic Syndr*. 2008;49:9–16.
9. Pillay D, Bhaskaran K, Jurriaans S, et al. The impact of transmitted drug resistance on the natural history of HIV infection and response to first-line therapy. *AIDS*. 2006;20:21–28.
10. Torian LV, Blank S, Kellerman SE, et al. Investigation of a New Diagnosis of Multidrug-Resistant, Dual-Tropic HIV-1 Infection—New York City, 2005. *MMWR Morb Mortal Wkly Rep*. 2006;55:793–796.
11. Boulos D, Yan P, Schanzer D, et al. Estimates of HIV prevalence and incidence in Canada, 2005. *Can Commun Dis Rep*. 2006;32:165–174.
12. Hue S, Clewley JP, Cane PA, et al. HIV-1 pol gene variation is sufficient for reconstruction of transmissions in the era of antiretroviral therapy. *AIDS*. 2004;18:719–728.
13. Hue S, Clewley JP, Cane PA, et al. Investigation of HIV-1 transmission events by phylogenetic methods: requirement for scientific rigour. *AIDS*. 2005;19:449–450.
14. EuroGUidelines Group for HIV resistance. Clinical and laboratory guidelines for the use of HIV-1 drug resistance testing as part of treatment management: recommendations for the European setting. The EuroGUidelines Group for HIV resistance. *AIDS*. 2001;15:309–320.
15. Gazzard BG. British HIV Association guidelines for the treatment of HIV-1-infected adults with antiretroviral therapy 2008. *HIV Med*. 2008; 9:563–608.
16. Hirsch MS, Gunthard HF, Schapiro JM, et al. Antiretroviral drug resistance testing in adult HIV-1 infection: 2008 recommendations of an International AIDS Society-USA panel. *Clin Infect Dis*. 2008;47:266–285.
17. Panel on Antiretroviral Guidelines for Adults and Adolescents. Guidelines for the use of antiretroviral agents in HIV-1-infected adults and adolescents. Department of Health and Human Services. December 1, 2009; 1–161, p 9–15. Available at http://www.aidsinfo.nih.gov/ContentFiles/AdultandAdolescentGL.pdf. Accessed July 6, 2010.
18. Haag D, Kim P, Spencer D, et al. *HIV/AIDS Update Year End 2005*. Vancouver, BC: British Columbia Centre for Disease Control STI/HIV Prevention and Control; 2006.
19. Jayaraman GC, Archibald CP, Lior L, et al. Integrating laboratory and epidemiological techniques for population-based surveillance of HIV strains and drug resistance in Canada. *Can J Infect Dis*. 2000;11:74–80.
20. Hall TA. BioEdit: a user-friendly biological sequence alignment editor and analysis program for Windows 95/98/NT. *Nucleic Acids Symp Ser*. 1999;41:95–98.
21. Johnson VA, Brun-Vezinet F, Clotet B, et al. Update of the drug resistance mutations in HIV-1: Fall 2005. *Top HIV Med*. 2005;13:125–131.
22. Parekh BS, Kennedy MS, Dobbs T, et al. Quantitative detection of increasing HIV type 1 antibodies after seroconversion: a simple assay for detecting recent HIV infection and estimating incidence. *AIDS Res Hum Retroviruses*. 2002;18:295–307.
23. Saitou N, Nei M. The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Mol Biol Evol*. 1987;4:406–425.
24. Kimura M. A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. *J Mol Evol*. 1980;16:111–120.
25. Kumar S, Tamura K, Nei M. MEGA3: integrated software for molecular evolutionary genetics analysis and sequence alignment. *Brief Bioinform* 2004;5:150–163.
26. Public Health Agency of Canada. *HIV and AIDS in Canada: Selected Surveillance Tables to June 30, 2008. Surveillance and Risk Assessment Division, Centre for Communicable Diseases and Infection Control*. Ottawa, ON: Public Health Agency of Canada; 2009.
27. Woolhouse ME, Dye C, Etard JF, et al. Heterogeneities in the transmission of infectious agents: implications for the design of control programs. *Proc Natl Acad Sci U S A*. 1997;94:338–342.
28. Lloyd-Smith JO, Schreiber SJ, Kopp PE, et al. Superspreading and the effect of individual variation on disease emergence. *Nature*. 2005; 438:355–359.
29. Galvani AP, May RM. Epidemiology: dimensions of superspreading. *Nature*. 2005;438:293–295.
30. Loschen S, Batzing-Feigenbaum J, Poggensee G, et al. Comparison of the human immunodeficiency virus (HIV) type 1-specific immunoglobulin G capture enzyme-linked immunosorbent assay and the avidity index method for identification of recent HIV infections. *J Clin Microbiol*. 2008; 46:341–345.
31. Lewis F, Hughes GJ, Rambaut A, et al. Episodic sexual transmission of HIV revealed by molecular phylodynamics. *PLoS Med*. 2008;5:e50.
32. Hughes GJ, Fearnhill E, Dunn D, et al. Molecular phylodynamics of the heterosexual HIV epidemic in the United Kingdom. *PLoS Pathog*. 2009; 5:e1000590.
33. Sturmer M, Preiser W, Gute P, et al. Phylogenetic analysis of HIV-1 transmission: pol gene sequences are insufficient to clarify true relationships between patient isolates. *AIDS*. 2004;18:2109–2113.
34. 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;93:10864–10869.
35. Smith DM, May SJ, Tweeten S, et al. A public health model for the molecular surveillance of HIV transmission in San Diego, California. *AIDS*. 2009;23:225–232.
36. Cuevas M, Fernandez-Garcia A, Sanchez-Garcia A, et al. Incidence of non-B subtypes of HIV-1 in Galicia, Spain: high frequency and diversity of HIV-1 among men who have sex with men. *Euro Surveill*. 2009;14(47): Article 3.
37. Nguyen L, Hu DJ, Choopanya K, et al. Genetic analysis of incident HIV-1 strains among injection drug users in Bangkok: evidence for multiple transmission clusters during a period of high incidence. *J Acquir Immune Defic Syndr*. 2002;30:248–256.

© 2010 Lippincott Williams & Wilkins

Copyright © Lippincott Williams & Wilkins. Unauthorized reproduction of this article is prohibited.