VIRUS EVOLUTION

Virus Evolution, 2017, 3(1): vex013

doi: 10.1093/ve/vex013
Research article

Rising prevalence of non-B HIV-1 subtypes in North Carolina and evidence for local onward transmission

Ann M. Dennis,[1,*] Stephane Hué,[2,†] Emily Learner,[3] Joseph Sebastian,[4] William C. Miller,[5] and Joseph J. Eron[1,3]

[1]Division of Infectious Diseases, University of North Carolina, Chapel Hill, NC, USA, [2]London School of Hygiene & Tropical Medicine, London, UK, [3]Department of Epidemiology, University of North Carolina, Chapel Hill, NC, USA, [4]Laboratory Corporation of America, Research Triangle Park, NC, USA and [5]Department of Epidemiology, Ohio State University, Columbus, OH, USA

*Corresponding author: E-mail: adennis@med.unc.edu
†http://orcid.org/0000-0002-8580-6905

Abstract

HIV-1 diversity is increasing in North American and European cohorts which may have public health implications. However, little is known about non-B subtype diversity in the southern United States, despite the region being the epicenter of the nation's epidemic. We characterized HIV-1 diversity and transmission clusters to identify the extent to which non-B strains are transmitted locally. We conducted cross-sectional analyses of HIV-1 partial *pol* sequences collected from 1997 to 2014 from adults accessing routine clinical care in North Carolina (NC). Subtypes were evaluated using COMET and phylogenetic analysis. Putative transmission clusters were identified using maximum-likelihood trees. Clusters involving non-B strains were confirmed and their dates of origin were estimated using Bayesian phylogenetics. Data were combined with demographic information collected at the time of sample collection and country of origin for a subset of patients. Among 24,972 sequences from 15,246 persons, the non-B subtype prevalence increased from 0% to 3.46% over the study period. Of 325 persons with non-B subtypes, diversity was high with over 15 pure subtypes and recombinants; subtype C (28.9%) and CRF02_AG (24.0%) were most common. While identification of transmission clusters was lower for persons with non-B versus B subtypes, several local transmission clusters (≥3 persons) involving non-B subtypes were identified and all were presumably due to heterosexual transmission. Prevalence of non-B subtype diversity remains low in NC but a statistically significant rise was identified over time which likely reflects multiple importation. However, the combined phylogenetic clustering analysis reveals evidence for local onward transmission. Detection of these non-B clusters suggests heterosexual transmission and may guide diagnostic and prevention interventions.

Key words: molecular epidemiology; HIV-1; transmission; southeastern United States.

Introduction

Surveillance of HIV diversity in the United States (US) and worldwide remains important as HIV diversity can impact diagnosis, pathogenesis, transmission, and vaccine development (Hemelaar et al. 2011). The global spread of HIV is defined by the virus' high genetic variability and rapid evolution giving rise to distinct subtypes, circulating recombinant forms (CRFs), and unique recombinants. Subtype distribution also has discrete geographic patterns with subtype-B dominating in US and other industrialized countries. The etiology of such distinct patterns is likely multifactorial including founder effects and social/migration factors and it remains unclear whether biologic differences in subtypes contribute to global spread (Hemelaar et al. 2011). The prevalence of non-B subtypes in Europe is increasing

(Neogi et al. 2014; U.K. Collaborative Group on HIV Drug Resistance 2014) which has historically been attributed to heterosexual transmission among immigrants. However, there is increasing evidence for domestic transmission of non-B subtypes, including among men who have sex with men (MSM) (Fox et al. 2010; Brand et al. 2014; Ragonnet-Cronin et al. 2016).

Little is known about HIV diversity in the southern US despite the region being the epicenter (Centers for Disease Control Prevention 2016) of the US HIV epidemic. While other US areas, such as Massachusetts and other Northeastern states, have reported increased HIV diversity (Pyne et al. 2013), southeastern US states are underrepresented in these analyses. Southern states experience the greatest burden of HIV infection compared to other US regions. The region has higher HIV diagnosis rates in non-urban areas (Centers for Disease Control Prevention 2016), increasing incidence among black and Latino MSM, and continued heterosexual transmission. In North Carolina (NC), over 28,000 individuals are living with HIV and ~1,500 new HIV diagnoses are reported annually (North Carolina HIV/STD Surveillance Unit 2015). The state has also experienced a rise in immigration; the proportion of the population that is foreign born rose from 1.7% in 1990 (Malone et al. 2003) to 7.6% in 2013 (United States Census Bureau 2013). HIV diagnoses among foreign-born persons in NC have also increased reflecting these migration patterns (North Carolina HIV/STD Surveillance Unit 2012). Among foreign-born persons with non-B subtypes in the US, heterosexual contact is the major risk factor (Prosser et al. 2012).

We characterized the HIV-1 subtype diversity in NC to evaluate trends in non-B subtype prevalence and identify transmission clusters involving non-B subtypes. A more detailed understanding of diversity trends and local, ongoing transmission of non-B subtypes will help identify populations that will benefit from tailored HIV prevention interventions.

Methods

Study population

We conducted a cross-sectional analysis of all HIV-1 *pol* sequences derived from drug resistance sequences performed by the largest reference laboratory in NC (Laboratory Corporation of America®). The majority of the sequences were generated with the GenoSure® MG assay, sampled between 1997 and mid-2014, and spanned protease nucleotide positions 1–297 and reverse transcriptase positions 1–1200. All sequences were collected from adult patients (≥18 years old) accessing clinical care in NC. Demographic variables collected at the time of sample collection included age, sex, and geographic location of clinic. Chart review of medical records was performed for the subset of individuals with non-B subtypes who received care at the University of North Carolina (UNC) Infectious Disease Clinic to collect risk behaviors, diagnosis year, race/ethnicity, and originating country. Following final data collection, all phylogenetic and statistical analyses were performed on de-identified datasets to protect subjects' anonymity. The study was approved by the University of North Carolina Intuitional Review Board.

HIV-1 subtype and sequence analyses

Sequences were aligned using MUSCLE (Edgar 2004) and edited manually in Bioedit (Hall 1999). Gapped positions were stripped and the final sequence length was 1497 bases. Drug resistance mutations (DRM) and initial subtype assignments were identified using The Stanford HIV Web Service (Sierra v.1.1) to query the Stanford HIVdb Program (Liu and Shafer 2006). Major DRMs were selected using the 2009 standardized surveillance list from the World Health Organization (Bennett et al. 2009). Non-B subtype sequences were initially identified by the Stanford HIVdb Program and then further confirmed and analyzed with the Context-based Modeling for Expeditious Typing (COMET HIV-1) tool (Pineda-Peña et al. 2013; Struck et al. 2014) and by phylogenetic reconstruction with reference sequences. HIV-1 pure subtypes and CRF references were obtained from the Los Alamos National Laboratory (LANL) HIV database (http://www.hiv.lanl.gov; 2010 reference alignment, $n = 159$ sequences). Further sensitivity analysis was performed among the B subtypes in COMET to determine the degree of potential misclassification by Stanford.

Phylogenetic analyses and clusters

A maximum-likelihood (ML) phylogenetic tree was constructed in FastTree (Price et al. 2009) v.2.1.4 with the general time reversible model of nucleotide substitution using the earliest available sequence from each individual including all B and non-B subtypes. Statistical support of clades was assessed with local support values (Shimodaira–Hasegawa-like test [SH-test]) in FastTree. Putative transmission clusters were identified using the automated tool ClusterPicker v1.3 (Ragonnet-Cronin et al. 2013). We defined clusters as clades with high branch support (probability ≥0.90; SH-test) and a maximum pairwise genetic distance <3.5% difference between all sequences. For every non-B study sequence identified in a putative cluster, we conducted a BLAST search to identify the 10 most closely related sequences available in the HIV LANL database. All the non-B subtype sequences were evaluated further with the LANL and reference sequences in a ML tree under the same model conditions.

Clusters involving non-B subtypes were then confirmed through Bayesian Markov Chain Monte Carlo (MCMC) inference in BEAST v.1.8.2 (Drummond and Rambaut 2007). Temporal signal was low when initially evaluated across the entire dataset using TempEst v.1.5 (Andrew Rambaut et al. 2016), likely due the mixture of different subtypes. However, when broken into subtypes we did find evidence for significant clock-like behavior of varying intensity. In the BEAST analysis, a conjunction of models including the SRD06 nucleotide substitution model, a lognormal relaxed molecular clock model and the Bayesian Skyline model as coalescent tree prior were used. An informative prior was imposed on the tree root height, as a normal distribution of mean 94 and standard deviation of 5 years before the most recent tip (Faria et al. 2014). The MCMC chain was run for 50 million generations, sampling every 10,000 generations. Convergence of the estimates was considered satisfactory when the effective sample size (ESS) calculated in Tracer v1.6.0 (A Rambaut and Drummond 2007) was >200 in all parameters; 10% of the generations were discarded as burn-in. The maximum clade credibility tree (MCCT) was summarized using TreeAnnotator v1.8.2 keeping the median height over the posterior distribution of trees. Clades with high posterior probability (i.e. ≥0.95) were considered highly supported.

Statistical analyses

Descriptive statistics and bivariate associations were used to assess differences between non-B and B subtypes using chi-squared test for categorical variables and Kruskal–Wallis for continuous variables. The first available sequence for each individual was used to assess the overall distribution of non-B

subtypes and to assess prevalence trends over time. The trend in non-B subtype prevalence over time was assessed using the chi-squared trend test and linear regression with calendar year, sex, and sampling region as explanatory variables. Analyses were performed using Stata 13.0 (StataCorp, College Station, TX).

Results

Study population

A total of 24,972 HIV-1 sequences were available from 15,246 individuals aged ≥18 years at time of sample collection. On initial analysis, 461 (1.9%) sequences were non-B subtypes from 325 (2.1%) individuals and subjected to further subtyping and phylogenetic analysis. Four sequences were possible non-B subtypes by Stanford HIV-dB but subtype B in COMET and therefore classified as subtype B. Among all persons ($n = 15,246$), most were men ($n = 10,680$; 70.1%), the median age was 40 (IQR 32–48) years, and most had an initial sample before 2009 ($n = 8,193$; 53.7%). As a sensitivity analysis, the 14,921 sequences (first available per individual) determined as subtype B in Stanford, were also evaluated in COMET. Of these, 99.7% were determined to be pure B subtype in COMET indicating very little misclassification. The remaining 0.3% were reported as unassigned B subtype recombinants in COMET.

Several demographic features differed between persons with non-B subtype HIV compared to those with subtype B (Table 1). Persons with non-B subtypes were more likely to be women (58.8% vs. 28.0%), have a more recently obtained initial sample (median year 2009 [IQR 2007–12] vs. 2008 [IQR 2005–11]), and have a sample sent from a clinic in the urban regions of Charlotte or Raleigh (85.9% vs. 68.0%) compared to persons with subtype B (all $P < 0.01$). Prevalence of major DRMs also differed significantly as 34.8% of persons with B versus 19% of non-B subtypes ($P < 0.001$) had at least one DRM. Prevalence was lower in all three major drug classes among non-B subtypes (Table 1). Among subtype B sequences with at least one major DRM, the most common DRMS were: K103N (45%), M184V (42%), M41L (17%), D67N (14%). Among the 63 sequences with one DRM from non-B subtypes, most common mutations were: M184V (52%), K103N (32%), Y181C (15% vs. 10% among subtype B), D67N (11%), M41L (11%). The distribution was only significantly different between prevalence of K103N ($P = 0.049$). Of the 325 persons with non-B subtypes, 54 (17%) received care at UNC. These persons were similar to those who received care at other clinics by age and sex. Among chart review of these 54 patients, most had an originating country documented ($n = 50$), with African countries predominating ($n = 29$; 58%), followed by US (32%), and Asia (8%).

Trends over time showed that the proportion of non-B subtypes among all sequences (B and non-B subtypes) increased significantly by calendar year (Fig. 1). The non-B subtype prevalence, based on the first available sequence, ranged from 0% to 1.63% in 1997–2001, to 0.92–2.53% in 2002–6, to 1.75–3.12% in 2007–11, and 2.69–3.46% by 2012–4 ($P < 0.0001$ for chi-squared trend). In linear regression, the trend in non-B subtype prevalence remain significantly associated with calendar year after adjusting for male sex and sampling from the urban regions of Charlotte or Raleigh ($R^2 = 0.77$, $P = 0.002$) (Fig. 1A).

Viral diversity among non-B subtypes

Substantial diversity was found among 325 persons with non-B subtypes based on the earliest sequence available in the COMET

Table 1. Characteristics of 15,246 HIV-1 infected men and women with first available HIV-1 *pol* sequence sampled in North Carolina from 1997 to 2014 by HIV-1 subtype.

Characteristic	**Overall** (n=15,246)		**Subtype B** (n=14,921)		**Non-B subtypes** (n=325)		P value[a]
	n	%	n	%	n	%	
Age							
≥35 years	10,327	67.7	10,107	67.7	220	67.7	0.99
<35 years	4,919	32.3	4,814	32.3	105	32.3	
Sex							
Male	10,680	70.1	10,550	70.7	130	40.0	<0.001
Female	4,364	28.6	4,173	28.0	191	58.8	
Sample year							
2009–14	7,053	46.3	6,874	46.1	179	55.1	0.001
2003–8	6,113	40.1	5,993	40.2	120	36.9	
1997–2002	2,080	13.6	2,054	13.8	26	8.0	
Clinic location by region							
Charlotte or Raleigh	10,422	68.4	10,143	68.0	279	85.9	<0.001
Other	4,824	31.6	4,778	32.0	46	14.1	
In phylogenetic cluster[b]							
No	7,603	49.9	7,361	49.3	242	74.5	<0.001
Yes	7,643	50.1	7,560	50.7	83	25.5	
Major DRM[c]							
Any	5,248	34.4	5,186	34.8	62	19.1	<0.001
NRTI	3,568	23.4	3,525	23.6	43	13.2	<0.001
NNRTI	3,324	21.8	3,287	22.0	37	11.4	<0.001
PI	1,436	9.4	1,422	9.5	14	4.3	0.001

DRM, drug resistance mutation; NRTI, nucleoside reverse transcriptase inhibitor; NNRTI, non-nucleoside reverse transcriptase inhibitor; PI, protease inhibitor.
[a]P value based on comparisons between subtype B and non-B subtypes with the chi-squared test.
[b]Identified in a putative transmission clusters, based on the maximum-likelihood tree in FastTree.
[c]Major drug resistance mutation from 2009 WHO list.

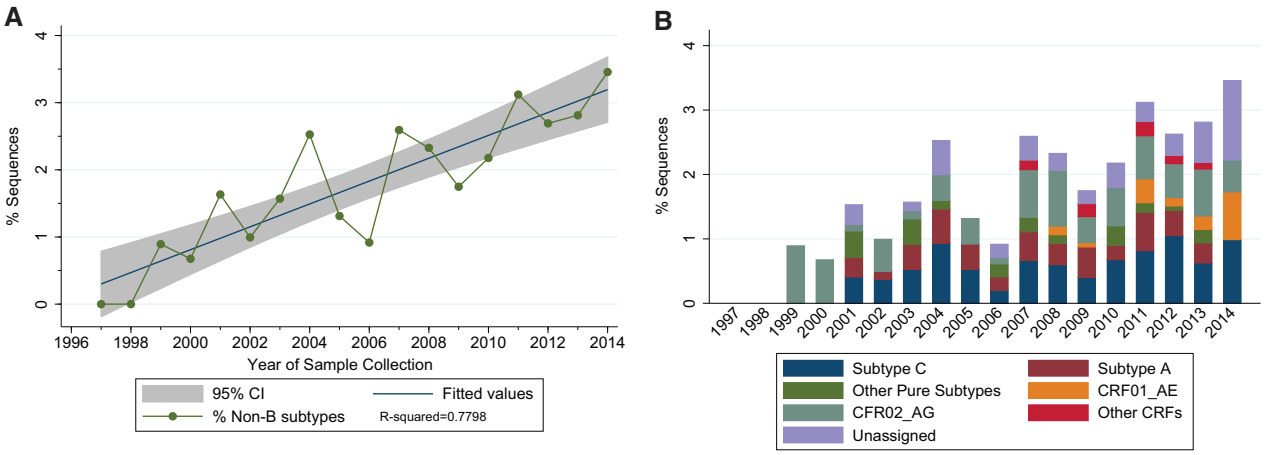

Figure 1. (A) Change in the proportion of non-B subtypes over time among 15,246 HIV-1 pol sequences collected in North Carolina from 1997 to 2014. Proportions are based on sampling date for the first sequence available per individual by calendar year. (B) Percentage of non-B subtypes by total number of non-B subtype sequences by calendar year of sample acquisition.

analysis (Table 2 and Fig. 1B). The most common subtypes/CRFs were C (28.9%), CRF02_AG (24.2%), and A (16.2%). Four other pure subtypes (D, G, H, J; $n = 24$), eight other CRFs ($n = 27$), and numerous unique/other recombinants (unassigned; $n = 47$) were identified. Seventy-five (23.1%) persons had more than one sequence (median 2, range 2–10), totaling 211 sequences. COMET assignments for intra-individual sequences were identical for 94.7% (71/75) of these persons. Of the four persons with sequences with differing subtype assignments, all involved recombinants. In the phylogenetic tree constructed with reference alignments and all non-B study sequences, all sequences stemming from the same individual cluster together (Fig. 2).

Phylogenetic clusters

To identify potential transmission clusters, a ML tree was initially constructed with the earliest sequence available from all individuals ($n = 15,246$). A total of 2,316 putative clusters were identified; 33 (1.5%) were among non-B subtypes. Persons with non-B subtypes were less likely to be identified in a putative transmission cluster compared to persons with subtype B (25.5% vs. 50.7% B-subtype sequences were in a clusters; $P < 0.001$) (Table 1). Cluster sizes were also significantly smaller among non-B subtypes: non-B clusters had mean size of 2.5 (range 2–7; SD 1.3) compared to 3.3 (range 2–36; SD 2.7) members in B-subtype clusters ($P = 0.004$). Non-B clusters also had smaller proportion of men; on an average, non-B clusters were 38.7% male versus 68.9% male in B-subtype clusters ($P < 0.001$). The 33 non-B clusters involved 83/325 (25.5%) individuals. While most clusters (27/33) were pairs ($n = 2$ individuals), six clusters involved 3–7 individuals. To further characterize the non-B clusters, we constructed a ML tree including LANL references identified through BLAST ($n = 269$ unique references) and all non-B sequences (Fig. 2). Only two clusters (both pairs) included LANL references; one of these had a low SH-test at the clade (0.82) and was not considered robust (Cluster# 2043 in Fig. 2 among CRF02_AG strains). Subtype diversity was high among the remaining 32 clusters: CRF02_AG (11), C (4), A (4), D (1), G (1), H (1), CRF01_AE (1), CRF18_cxp (1), and ORF/URF (8) (Fig. 2).

In the BEAST analysis, all 32 clusters were highly supported with posterior probabilities = 1 (Fig. 3). All 32 robust clusters included women. Of the 26 pairs, most were male/female (22/26; 84%) and the four other pairs included only women. No clusters contained sequences only from men, suggesting that MSM were not likely the dominant mode of transmission in these clusters. Overall 20 of the 81 clustered individuals had risk behavior information and all reported heterosexual risk and no intravenous drug use. The six larger clusters involved 29 individuals and ranged from 29% to 100% women (Cluster #1915, 7 members ORF/URF strain and Cluster#2010, 4 members, subtype G, respectively) (Fig. 3). The originating country was available for 11/29 individuals in these clusters, and of these 35% (10/29) were US-born. All these large clusters had estimated origins (tMRCA, time of the most common ancestor) from 1995 to 2005 (Fig. 3). Additionally, 42% of clustered individuals had CRF02_AG strains forming 11 clusters. Three of these CRF02_AG clusters stemmed from a larger highly supported clade (14 individuals) suggesting an earlier importation, tMRCA= 1994.1 (95% HPD, highest posterior density 1989.6–1997.8) and local propagation of subclusters. This larger cluster had an overall timespan of ~18 years (most recent sequence was sampled in 2012) translating into an average of 0.78 transmissions/year. The complete consensus BEAST tree is available as Supplementary Fig. S1.

Table 2. Distribution of non-B HIV-1 subtypes among 325 HIV-infected men and women based on the first available sequence.

Subtype	*n*	%
Pure Subtype		
C	94	28.9
A	54	16.2
D	9	2.8
G	9	2.8
H	5	1.5
J	1	<1
CRFs		
CRF02_AG	78	24.0
CRF01_AE	15	4.6
Others[a]	12	3.7
Unassigned		
URF/ORF	47	14.5

CRF, circulating recombinant form; URF unique recombinant form; ORF, other recombinant form.
Subtype assignments based on COMET.
[a]Other CRFs (number reported): CRF06_cpx (5), CRF18_cpx (2), CRF09_cpx (1), CRF13_cpx (1), CRF25_cpx (1), CRF29_BF (1), CRF45_cxp (1).

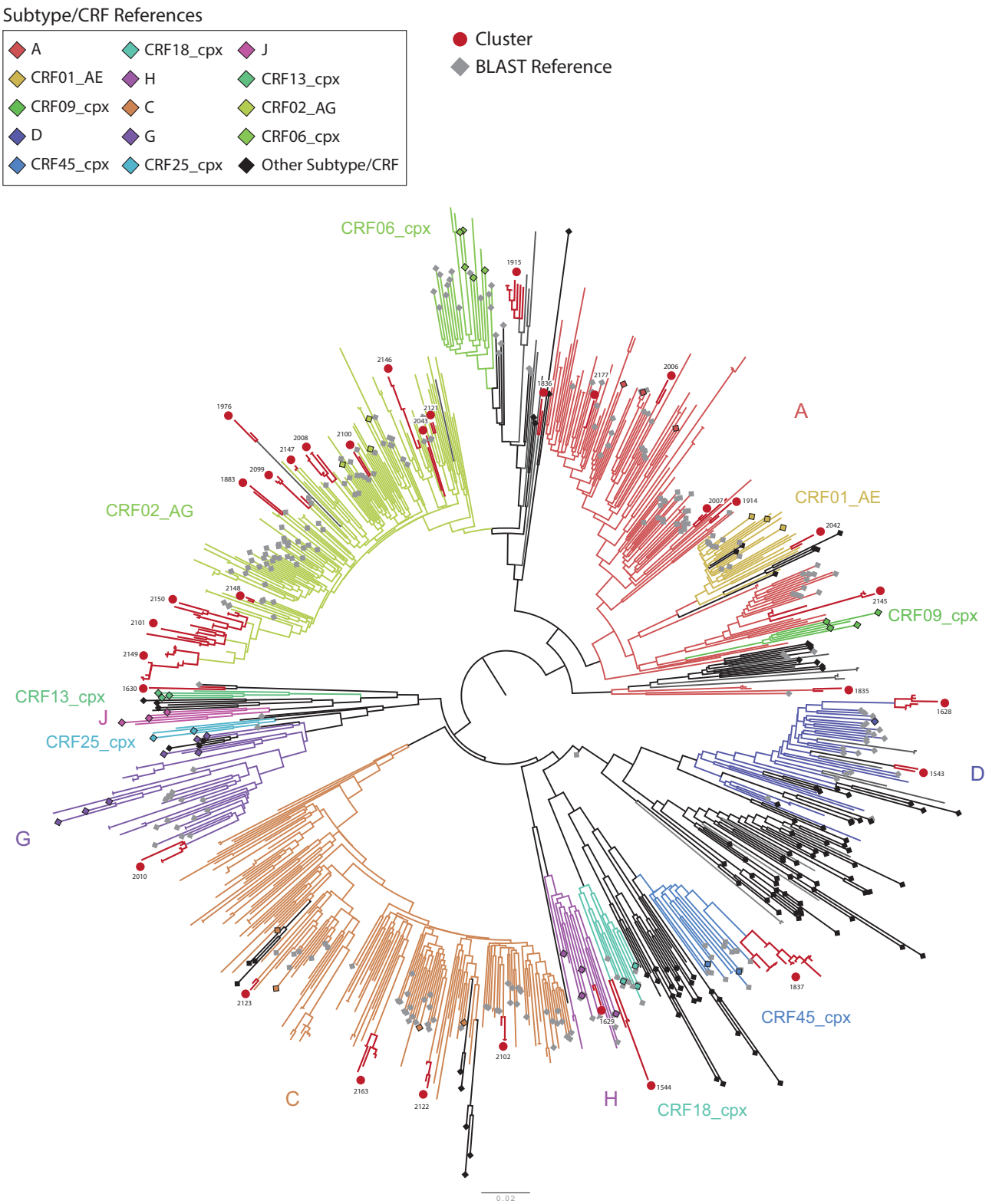

Figure 2. Maximum Likelihood phylogenetic tree of HIV-1 non-subtype B pol sequences (*n*=461) from 325 HIV-infected men and women in North Carolina and 428 references from the Los Alamos National Laboratory (LANL) HIV database (http://www.hiv.lanl.gov). Subtype and Circulating Recombinant Forms (CRF) references are indicated by diamonds with black outline (*n*=159). Color coding indicates the common subtypes and CRFs among study sequences. LANL references identified through BLAST search of study sequences are shown with grey diamonds (*n*=269). Putative transmission clusters are highlighted in red and indicated with red dots. Number label at tips are cluster identifiers. Branch lengths represent nucleotide substitutions per site.

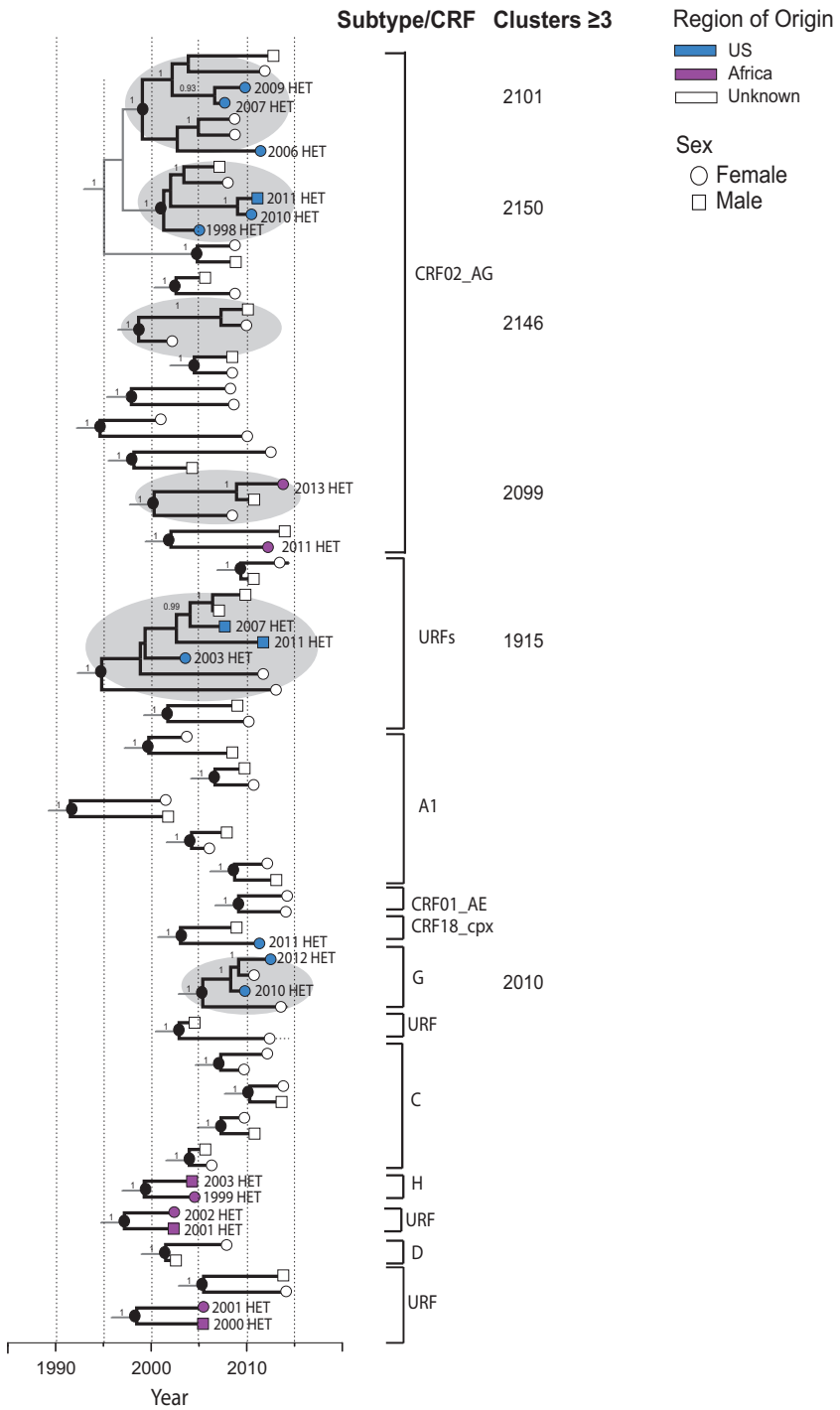

Figure 3. Non-B subtype transmission clusters (n=32) involving 81 HIV-1 pol sequences from individuals sampled in North Carolina. Clusters were confirmed through Bayesian Markov Chain Monte Carlo inference in BEAST. The scale bar (x-axis) is calendar years. Black circles at nodes indicate cluster origin and estimated time of the most recent common ancestor (tMRCA) and tips correspond to sampling date. Nodes with posterior probabilities >0.90 are labeled. Region of origin (if known) is indicated by tip colors. Numbers at tips indicate year of diagnosis (if available). HET indicates reported heterosexual transmission.

Discussion

We conducted a combined phylogenetic and demographic analysis of 24,972 HIV-1 *pol* sequences derived from resistance tests collected during clinical care. The dataset represents one of the largest samples to date stemming from a single US state to investigate subtype diversity and local HIV transmission clusters. We found that the overall prevalence of non-B subtypes remains low at 2.1% overall, but has increased significantly from 0% in 1997 to 3.46% by 2014. Among non-B subtypes, a high degree of viral diversity was seen with over 15 pure subtypes and CRFs, which is indicative of multiple viral introductions in the region. Despite this, we found evidence for several local transmission clusters involving non-B strains of which all were presumptively due to heterosexual transmission.

While HIV-1 subtype-B remains dominant in North America, an increase in non-B subtypes has been reported (Pyne et al. 2013; Siemieniuk et al. 2013). We found a slow but steady increase in HIV-1 diversity in NC. Consistent with our results, among 24,286 *pol* sequences from 46 US states, the non-B subtype prevalence was 3.27% with an increasing trend of 0–4.12% from 2004 to 2011 (Pyne et al. 2013). However, the geographical distribution of non-B subtypes were largely widespread and sampling between states varied greatly—the number of samples included from NC was relatively small and only subtype C was identified among non-B strains. In our study, while subtype C was the most common non-B subtype (28.9%), five other subtypes and numerous recombinant strains were identified. Diversity also remained high among non-B subtypes over time.

In contrast, much higher HIV-1 diversity has been reported in other immigrant-rich North American cohorts. In the US, non-B subtype prevalence was reported as 5.4% in US military (Brodine et al. 2003), 8.3% in Rhode Island (Chan et al. 2014) cohorts, and up to 43.4% among immigrants in New York City (Lin et al. 2006). The strains in these studies largely reflected countries of travel or origin suggesting HIV acquisition abroad. In Canada, the non-B prevalence rose from 9.6% to 32.4% in one cohort from 1994 to 2010 (Siemieniuk et al. 2013). Such increases may be partially explained by increased immigration from endemic countries with non-B subtypes; however, there is mounting evidence for local transmission of non-B strains (Luft et al. 2011). Despite an increase in HIV diagnoses among foreign-born persons in NC, we found a much lower non-B subtype prevalence. North Carolina has experienced one of the largest increases in immigration since 1990; however, most of the increase is due to immigration from Latin America (Migration Policy Institute 2016) where subtype-B predominates. Immigration from Asia and Africa have also increased in NC from 1990 to 2014, but to a lesser extent (Migration Policy Institute 2016). In our study, we found that two-thirds of persons with non-B subtypes who had an originating country documented were born in African or Asian countries which supports the ideas that most non-B diversity reflects immigration patterns. Persons with non-B subtypes were more likely to have sequences sent from the urban areas of Charlotte and Raleigh which are the areas with the largest immigrant populations in NC. The finding that persons with non-B subtypes were less likely to have major DRM and have a more recent first sequence suggests that a higher proportion could be antiretroviral naïve with more recent HIV diagnoses compared to those with subtype B. This may reflect immigration patterns; less transmitted or acquired drug resistance among this group is possible but cannot be discerned due to lack of drug exposure data.

The high density of viral sampling in our study allows the investigation of phylogenetic clustering to detect local transmission of non-B strains which been done by few US studies (Chan et al. 2014; Wertheim et al. 2016). Subjects with non-B subtypes in our study were more likely to be female which suggests predominately heterosexual transmission as reported in prior studies (Pyne et al. 2013). In fact, we found no evidence of MSM transmission among non-B strains. All the non-B clusters in our study included women and were smaller compared to HIV-1B clusters. The smaller size suggests importation (with the remainder of the transmission chain in country of origin) or that a fraction of these clusters have not been diagnosed or sampled (such as the case with female predominate clusters). In Rhode Island, 22% of non-B samples formed clusters and there was evidence for ongoing transmission in the region among African immigrants though at a low level (Chan et al. 2014). Among 41,539 HIV-1 *pol* sequences reported to the US National HIV surveillance system, evidence for importation and domestic transmission of non-B subtypes were also found (Wertheim et al. 2016). While sequences linked to non-US sampled sequences from the HIV LANL database was overall relatively rare in that study, most internationally linked US sequences were from persons born outside the US. However, 13 non-B clusters were identified that involved at least two US and two non-US sequences suggesting importation and propagation in the US. In contrast to the non-B clusters identified in our study, most included MSM and none were sampled in the southern US.

While a wide diversity among non-B subtypes was identified in our study, 42% of individuals in non-B transmission clusters had CRF02_AG strains. The CRF02_AG strain is the most prevalent CRF worldwide and the predominant HIV-1 strain in West and West Central Africa, accounting for 50–70% of the circulating strains in the region (Hemelaar et al. 2011). However, the CRF02_AG strain has also increased in prevalence from 5.4% to 7.7% globally and from 2.9% to 4.5% in Europe from 2001 to 2007 (Hemelaar et al. 2011). In Europe, CRF02_AG has previously been more commonly identified in African immigrants. However, the strain has been recently described among multicultural MSM transmission networks in Spain (Bracho et al. 2014) and in France (Tamalet et al. 2015). In our study, while risk behavior information was only available for a subset of individuals, we found no evidence for MSM transmission among CRF02_AG or other non-B subtypes. Our findings, however, do suggest that the CRF02_AG is circulating in local transmission chains among US-born heterosexuals.

Although our study is limited by the retrospective nature and reference laboratory bias, the results nonetheless offer important insights into HIV diversity trends into the region. Subjects would only be included here if they had a sequence performed by LabCorp; undiagnosed patients, those not linked to care or if sequencing was performed by another laboratory are obviously not included. Since HIV-1 genotypes were not recommended at entry to care until 2007 (Hirsch et al. 2008), most sequences in our dataset prior to this time were likely sent to evaluate treatment failure and all sampling times do not reflect diagnosis or infection dates. We also have clinical, demographic, and migration information on only a minority of subjects. The temporal analysis of the non-B subtype clusters in the study also has limitations that should be mentioned. The non-B sequences were analyzed together in BEAST due the limited number of sequences per subtype. Analyzing diverse linages in a single phylogeny can lead to inconsistencies in the estimated the age of HIV-1 subtypes (Wertheim et al. 2011). However, this mainly affects aging estimations at the oldest, deep nodes in the tree rather than the origins of the transmission clusters which was of interest in our study. Despite these limitations, the large sequence dataset offers a high sampling density which likely improves the ability to detect local transmission clusters. Our study included over 15,000 individuals which is a large fraction of the state's HIV population. Of the 42,889 HIV cases reported in NC from 1983 to 2013, an estimated 28,101 are currently alive and residing in the state (North Carolina HIV/STD Surveillance Unit 2015).

Continued surveillance of HIV-1 viral diversity and transmission remains important in southern US states which are now central to the national HIV epidemic. Phylogenetic analyses combined with traditional epidemiological monitoring can be used to uncover HIV transmission clusters and better understand the dynamics of viral spread. Such tools can be conducted using large sequence datasets which are routinely collected for

clinical care and allow monitoring and characterization of subtype-B and non-B strains. Detection of non-B clusters strongly suggests heterosexual transmission in our southern US cohort and may help guide diagnostic and prevention interventions. Detecting and characterizing clustered transmissions allow more specific interventions to be designed based on geographic, transmission risk behavior, and demographic features of clusters.

Acknowledgements

The project was supported by the National Insitute of Allergy and Infectious Diseases (NIAID), NIH, through Grant Award Number K08AI112432-01, and the UNC Center for AIDS Research through Grant Award Number P30AI50410. The content is soley the responsibility of the authors and does not necessarily represent the official views of the NIH. We thank patient and staff of the UNC Center for AIDS Reasarch Clinical Cohort.

Author contributions

HIV pol nucleotide sequence data (J.S.), Planning of analyses (A.D., S.H., W.M. and J.E.), data assembly and review (A.D. and E.L.), phylogenetic analyses (A.D. and S.H.), statistical analyses (A.D.), drafting manuscript (A.D.), manuscript review and edit (all authors).

Conflict of interest: None declared.

Data availability

The GenBank acession numbers for 325 non-B subtype pol sequences and 100 randomly selected subtype B pol sequences are: KY579388-KY579812. A subset of sequences is made available due to the potential for identification of direct transmission among individuals in such large datasests which could have near complete sampling of local HIV cases.

Supplementary data

Supplementary data are available at *Virus Evolution* online.

References

Bennett, D. E. et al. (2009) 'Drug Resistance Mutations for Surveillance of Transmitted HIV-1 Drug-Resistance: 2009 Update', *PLoS One*, 4/3: e4724

Bracho, M. A. et al. (2014) 'Emerging Trends in CRF02_AG Variants Transmission Among Men Who Have Sex With Men in Spain', *Journal of Acquired Immune Deficiency Syndromes*, 65/3: e130–3

Brand, D. et al. (2014) 'Characteristics of Patients Recently Infected with HIV-1 Non-B Subtypes in France: A Nested Study Within the Mandatory Notification System for New HIV Diagnoses', *Journal of Clinical Microbiology*, 52/11: 4010–6

Brodine, S. K. et al. (2003) 'Diverse HIV-1 Subtypes and Clinical, Laboratory and Behavioral Factors in a Recently Infected US Military Cohort', *Aids*, 17/17: 2521–7

Centers for Disease Control Prevention. *HIV in the Southern United States. CDC Issue Brief*. <https://www.cdc.gov/hiv/pdf/policies/cdc-hiv-in-the-south-issue-brief.pdf> accessed 12 Jul 2016.

Chan, P. A. et al. (2014) 'Phylogenetic and Geospatial Evaluation of HIV-1 Subtype Diversity at the Largest HIV Center in Rhode Island', *Infection, Genetics and Evolution*, 28: 358–66

Drummond, A. J., and Rambaut, A. (2007) 'BEAST: Bayesian Evolutionary Analysis by Sampling Trees', *BMC Evolutionary Biology*, 7: 214.

Edgar, R. C. (2004) 'MUSCLE: Multiple Sequence Alignment with High Accuracy and High Throughput', *Nucleic Acids Research*, 32/5: 1792–7

Faria, N. R. et al. (2014) 'The Early Spread and Epidemic Ignition of HIV-1 in Human Populations', *Science*, 346/6205: 56.

Fox, J. et al. (2010) 'Epidemiology of Non-B Clade Forms of HIV-1 in Men Who Have Sex with Men in the UK', *Aids*, 24/15: 2397–401.

Hall, T. A. (1999) 'BioEdit: A User-Friendly Biological Sequence Alignment Editor and Analysis Program for Windows 98/98/NT', *Nucleic Acids Symposium Series*, 41: 95–8.

Hemelaar, J. et al. (2011) 'Global Trends in Molecular Epidemiology of HIV-1 During 2000-2007', *Aids*, 25/5: 679–89.

Hirsch, M. S. et al. (2008) 'Antiretroviral Drug Resistance Testing in Adult HIV-1 Infection: 2008 Recommendations of an International AIDS Society-USA panel', *Clinical Infectious Diseases*, 47/2: 266–85.

Lin, H. H. et al. (2006) 'Genetic Characterization of Diverse HIV-1 Strains in An Immigrant Population Living in New York City', *Journal of Acquired Immune Deficiency Syndromes*, 41/4: 399–404.

Liu, T. F., and Shafer, R. W. (2006) 'Web Resources for HIV Type 1 Genotypic-Resistance Test Interpretation', *Clinical Infectious Diseases*, 42/11: 1608–18.

Luft, L. M., Beckthold, B., and Gill, M. J. (2011) 'Increasing HIV Subtype Diversity in Canadian-Born Patients Living in Southern Alberta, Canada', *Journal of Acquired Immune Deficiency Syndromes*, 57/2: e27–9.

Malone, N. et al. (2003) *The Foreign-Born Population: 2000. Census 2000 Brief*. Issued Dec. 2003. <https://www.census.gov/prod/2003pubs/c2kbr-34.pdf>.

Migration Policy Institute. State Immigration Data Profiles. <http://www.migrationpolicy.org/data/state-profiles/state/demographics/NC> accessed 20 Oct 2016.

Neogi, U. et al. (2014) 'Temporal Trends in the Swedish HIV-1 Epidemic: Increase in Non-B Subtypes and Recombinant Forms Over Three Decades', *PLoS One*, 9/6: e99390.

North Carolina HIV/STD Surveillance Unit. *North Carolina HIV/STD Epidemiologic Profile for HIV/STD Prevention & Care Planning*. Dec 2012. North Carolina Department of Health and Human Services, Raleigh, North Carolina. <http://epi.publichealth.nc.gov/cd/stds/figures/Epi_Profile_2012.pdf> accessed 12 Jul 2016.

North Carolina HIV/STD Epidemiologic Profile (2013) Issued March 2015. *North Carolina Department of Health and Human Services, Raleigh, North Carolina*. <http://epi.publichealth.nc.gov/cd/stds/figures/Epi_Profile_2013.pdf> accessed 12 Jul 2016.

Pineda-Peña, A.-C. et al. (2013) 'Automated Subtyping of HIV-1 Genetic Sequences for Clinical and Surveillance Purposes: Performance Evaluation of the New REGA Version 3 and Seven Other Tools', *Infection, Genetics and Evolution*, 19: 337–48.

Price, M. N., Dehal, P. S., and Arkin, A. P. (2009) 'FastTree: Computing Large Minimum Evolution Trees with Profiles Instead of a Distance Matrix', *Molecular Biology and Evolution*, 26/7: 1641–50.

Prosser, A. T., Tang, T., and Hall, H. I. (2012) 'HIV In Persons Born Outside the United States, 2007–2010', *Jama*, 308/6: 601–7.

Pyne, M. T. et al. (2013) 'Large-Scale Analysis of the Prevalence and Geographic Distribution of HIV-1 Non-B Variants in the United States', *Journal of Clinical Microbiology*, 51/8: 2662–9.

Ragonnet-Cronin, M. et al. (2013) 'Automated Analysis of Phylogenetic Clusters', *BMC Bioinformatics*, 14: 317.

—— (2016) 'Transmission of Non-B HIV Subtypes in the United Kingdom Is Increasingly Driven by Large Non-Heterosexual Transmission Clusters', *Journal of Infectious Diseases*, 213/9: 1410–8.

Rambaut, A., and Drummond, A. J. (2007) *Tracer v1.4*. <http://beast.bio.ed.ac.uk/Tracer>.

——, and Lam, T. T. et al. (2016) 'Exploring the Temporal Structure of Heterochronous Sequences Using TempEst (formerly Path-O-Gen)', *Virus Evolution*, 2/1: vew007.

Siemieniuk, R. A., Beckthold, B., and Gill, M. J. (2013) 'Increasing HIV Subtype Diversity and Its Clinical Implications in a Sentinel North American Population', *Canadian Journal of Infectious Diseases & Medical Microbiology*, 24/2: 69–73.

Struck, D. et al. (2014) 'COMET: Adaptive Context-Based Modeling for Ultrafast HIV-1 Subtype Identification', *Nucleic Acids Research*, 42/18: e144.

Tamalet, C. et al. (2015) 'Emergence of Clusters of CRF02_AG and B Human Immunodeficiency Viral Strains Among Men Having Sex with Men Exhibiting HIV Primary Infection in Southeastern France', *Journal of Medical Virology*, 87/8: 1327–33.

U.K. Collaborative Group on HIV Drug Resistance (2014) 'The Increasing Genetic Diversity of HIV-1 in the UK, 2002-2010', *Aids*, 28/5: 773–80.

United States Census Bureau (2013) *Selected Social Characteristics in the United States. 2013 American Community Survey 1-Year Estimates*. <https://factfinder.census.gov> accessed 25 Jul 2016.

Wertheim, J. O., Fourment, M., and Kosakovsky Pond, S. L. (2011) 'Inconsistencies in Estimating the Age of HIV-1 Subtypes Due to Heterotachy', *Molecular Biology and Evolution*, 29/2: 451–6.

——, Oster, A. M., and Hernandex, N. et al. (2016) 'The International Dimension of the U.S. HIV Transmission Network and Onward Transmission of HIV Recently Imported into the United States', *AIDS Research and Human Retroviruses*, 32/10–11: 1046–53.