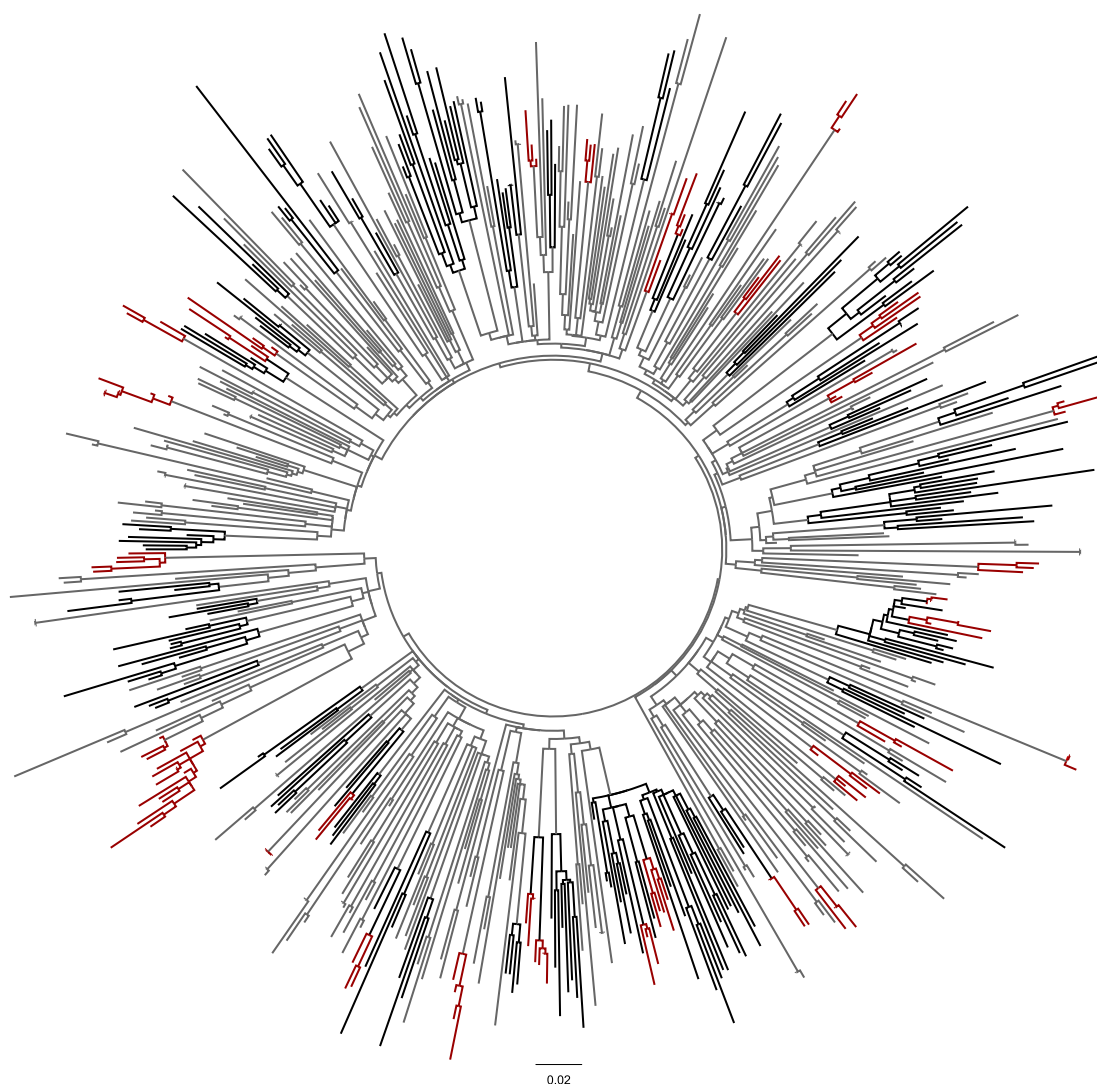


**Supplemental Figure 1.** Racial/ethnic distribution of sequences obtained from 1997-2009. Sampling density among ethnic group appeared to be proportional to composition of clinic cohort (and of HIV cases reported to the state among patients diagnosed from 2000-2009).



**Supplemental Figure 2.** Maximum likelihood phylogenetic tree of 646 sequences that were initially identified via  $<4.5\%$  difference in pairwise genetic distance with at least one other sequence among 1671 B-subtype sequences, each from a unique patient. The 67 identified transmission clusters ( $\geq 3$  sequences) supported by Bayesian Posterior Probabilities = 1 are highlighted in black. The 33 transmission clusters with mean intra-cluster pairwise genetic distance difference  $\leq 1.5\%$  are shown in red.

**Supplemental Table.** Comparison of factors associated with membership in phylogenetic transmission clusters defined by mean intra-cluster genetic distance difference  $\leq 0.015$  and Bayesian Posterior Probability=1, among patients diagnosed 2000-2009 (n=889). Logistic regression models show associations for membership in transmission clusters compared to those not in clusters or pairs.

Characteristic	Not in Cluster	In Pair n=2	In Cluster n≥3	Membership in Cluster n≥3			
				Bivariable		Multivariable	
				OR (95% CI)	P	OR (95% CI)	P
Total n (%)	631 (71)	156 (18)	102 (11)				
Age ≤30 years	239 (66)	62 (17)	61 (17) <sup>ab</sup>	2.29 (1.50-3.51)	<0.001	2.26 (1.41-3.62)	0.001
Year of Diagnosis >2004	252 (69)	66 (18)	49 (13)	1.36 (0.90-2.06)	0.14	NA	
Race/Ethnicity							
Black	372 (71)	83 (16)	66 (13) <sup>ab</sup>	Ref		Ref	
White	163 (70)	39 (17)	31 (13)	1.06 (0.67-1.67)	0.81	1.11 (0.68-1.81)	0.69
Latino	60 (68)	25 (28)	3 (3)	0.24 (0.07-0.79)	0.02	0.24 (0.07-0.79)	0.02
Other/Unknown	36 (77)	9 (19)	2 (4)	0.31 (0.07-1.29)	0.11	0.45 (0.06-3.68)	0.46
Sex and Transmission Risk							
Women	164 (73)	39 (17)	23 (10)	Ref		NA	
Men – non-MSM	174 (71)	47 (19)	25 (10)	1.00 (0.55-1.81)	1.00		
MSM	265 (69)	66 (17)	53 (14)	1.41 (0.84-2.38)	0.19		
Duration of Infection							
Chronic	433 (72)	108 (18)	62 (10)	Ref		Ref	
Recent	90 (72)	19 (15)	16 (13)	1.28 (0.71-2.30)	0.41	1.81 (0.89-3.67)	0.10
Acute	108 (67)	29 (18)	24 (15)	1.53 (0.92-2.54)	0.10	1.55 (0.87-2.78)	0.14
ART-naïve	541 (70)	140 (18)	94 (12)	1.83 (0.86-3.88)	0.12	NA	
Genotype <1 year of diagnosis	499 (70)	129 (18)	84 (12)	1.18 (0.69-2.02)	0.54	NA	
CD4 cell count >200 cells/mm <sup>3cd</sup>	225 (67)	70 (21)	42 (12) <sup>abf</sup>	1.75 (1.00-3.06)	0.05	NA	
HIV RNA >5.0 log copies/ml <sup>d</sup>	245 (71)	62 (18)	39 (11)	0.97 (0.63-1.49)	0.89	NA	
TDR	55 (60)	21 (23)	16 (17)	1.63 (0.91-2.94)	0.10	1.60 (0.86-2.97)	0.14
Residence in primary catchment area <sup>e</sup>	432 (69)	117 (19)	77 (12)	1.26 (0.78-2.03)	0.34	2.24 (1.25-4.02)	0.007

OR, odds ratio; NA not applicable; MSM, men who have sex with men; ART, antiretroviral therapy; TDR, transmitted drug resistance

<sup>a</sup>  $P < 0.05$  based on Pearson's  $\chi^2$  for group trend

<sup>b</sup>  $P < 0.05$  based on Pearson's  $\chi^2$  for membership in cluster  $n \geq 3$  versus not in a cluster

<sup>c</sup> CD4 cell counts only available for patients with chronic duration of infection ( $n=603$ )

<sup>d</sup> Values recorded at date closest to diagnosis

<sup>e</sup> Primary catchment area includes 16 out of 100 NC counties that account for approximately 75% of the UNC clinic population

<sup>f</sup>  $P < 0.05$  based on Pearson's  $\chi^2$  for membership in pair  $n=2$  versus not in a cluster