# Supplementary Text S1 - Additional information on Methods

**Patients, samples and new sequences:**

From October 2009 to February 2014 HIV-1 seropositive patients from 13 cities in Santa Catarina and Rio Grande do Sul were invited to enroll in this study which was approved by the ethics committees of the Federal University of Santa Catarina (UFSC) and the Foundation of Research and Production in Health of the Rio Grande do Sul state (FEPPS-RS). Informed consent was obtained from all patients enrolled.

Briefly, 360 blood samples were collected in FTA cards (Whatman Inc.) and purified to downstream molecular procedures following the manufacturer's protocol. HIV-1 envelope (HXB2 6846-7360 bp) and polymerase (HXB2 2274-3545 bp) fragments were amplified by nested-PCR and sequenced as described elsewhere (1). Sequences were assembled and visually inspected using the SeqMan software in the LaserGene package (DNASTAR).

**Sequence dataset compilation:**

We constructed an as complete as possible dataset of Brazilian *pol* and *env* subtype C sequences (HIV-1C_BR) with sampling time and location annotation by complementing our new sequence data with publically available HIV-1C sequences. To this end, all *pol* and *env* HIV-1C_BR sequences were downloaded from the Los Alamos HIV database ($n$ = 385). When sampling location and/or time was not available, we attempted to retrieve this from the study reporting the sequences. We added the 50 best hits per query in a BLAST search of HIV1C_BR against all non-Brazilian HIV-1 subtype C sequences (2), but kept only those assigned as subtype C by REGA, RIP and SCUEAL (3-5) for further analyses. To avoid erroneous clustering due to convergent evolution as a consequence of drug selective pressure (6) we checked all *pol* sequences - including the newly generated ones- for the presence of Surveillance Drug Resistance Mutations (SDRM) by submitting them to the CPR tool available in the Stanford HIV Database (http://hivdb.stanford.edu) (7). Low-quality sequences, sequences with >= 5 SDRMs and sites with a SDRM detected in at least two different sequences were removed.

The sequences were aligned using Muscle (8) and the alignment was manually edited using AliView (9). Phylogenetic trees were reconstructed using RAxML (10) under a GTR+I+Γ4 nucleotide substitution model and Brazilian sequences that clustered outside the general HIV-1C_BR clade were removed. Sequences outliers in a regression of root-to-tip divergence versus sampling time identified using Path-O-Gen software (http://tree.bio.ed.ac.uk/software/pathogen/) were also removed. This procedure resulted in datasets with 1522 *pol* and 621 *env* sequences, which were down-sampled to around 500 sequences each. For this we kept all Brazilian subtype C sequences and selected 120 *pol* and 170 *env* international sequences using CD-HIT (11). Six additional sub-samplings containing only Brazilian sequences were made for *pol* and *env* (see the main text).

**Adjusting predictors for phylogeographic inference**

All epidemiological information was collected from the DATASUS system, a databank hosted by the Brazilian Ministry of Health (http://www2.datasus.gov.br/DATASUS/index.php?area=0203). To extract the GLM predictors, we focused on a 10 years period (2004-2013) as a representative time frame for the current infected population. Population data and geographic coordinates were obtained from the Brazilian Institute of Geography and Statistics (IBGE) (http://cidades.ibge.gov.br/xtras/home.php) and air traffic information was collected as previously described (12). For location-specific measures (3 to 7), we considered separate origin and destination predictors for the rates in the GLM design matrix, which implies that, for origin predictors, the measures for the origin locations are used to predict the pairwise transition rates between the origin and destination location and vice versa.

**References:**

1. **Librelotto CS, Gräf T, Simon D, Almeida SEM, Lunge VR**. 2015. HIV-1 epidemiology and circulating subtypes in the countryside of South Brazil. Rev Soc Bras Med Trop **48**:249–257.

2. **Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, Madden TL**. 2009. BLAST+: architecture and applications. BMC Bioinformatics **10**:421.

3. **de Oliveira T, Deforche K, Cassol S, Salminen M, Paraskevis D, Seebregts C, Snoeck J, van Rensburg EJ, Wensing AMJ, van de Vijver DA, Boucher CA, Camacho R, Vandamme A-M**. 2005. An automated genotyping system for analysis of HIV-1 and other microbial sequences. Bioinformatics **21**:3797–3800.

4. **Pond SLK, Posada D, Stawiski E, Chappey C, Poon AFY, Hughes G, Fearnhill E, Gravenor MB, Brown AJL, Frost SDW**. 2009. An evolutionary model-based algorithm for accurate phylogenetic breakpoint mapping and subtype prediction in HIV-1. PLoS Comput Biol **5**:e1000581.

5. **Siepel AC, Halpern AL, Macken C, Korber BTM**. 1995. A Computer Program Designed to Screen Rapidly for HIV Type 1 Intersubtype Recombinant Sequences. AIDS Res Hum Retroviruses **11**:1413–1416.

6. **Lemey P, Derdelinckx I, Rambaut A, Van Laethem K, Dumont S, Vermeulen S, Van Wijngaerden E, Vandamme A-M**. 2005. Molecular Footprint of Drug-Selective Pressure in a Human Immunodeficiency Virus Transmission Chain. J Virol **79**:11981–11989.

7. **Gifford RJ, Liu TF, Rhee SY, Kiuchi M, Hue S, Pillay D, Shafer RW**. 2009. The calibrated population resistance tool: Standardized genotypic estimation of transmitted HIV-1 drug resistance. Bioinformatics **25**:1197–1198.

8. **Edgar RC**. 2004. MUSCLE: a multiple sequence alignment method with reduced time and space complexity. BMC Bioinformatics **5**:113.

9. **Larsson a.** 2014. AliView: a fast and lightweight alignment viewer and editor for large data sets. Bioinformatics **30**:btu531–.

10. **Stamatakis A**. 2014. RAxML version 8: A tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics **30**:1312–1313.

11. **Huang Y, Niu B, Gao Y, Fu L, Li W**. 2010. CD-HIT Suite: A web server for clustering and comparing biological sequences. Bioinformatics **26**:680–682.

12. **Lemey P, Rambaut A, Bedford T, Faria N, Bielejec F, Baele G, Russell CA, Smith DJ, Pybus OG, Brockmann D, Suchard MA**. 2014. Unifying Viral Genetics and Human Transportation Data to Predict the Global Transmission Dynamics of Human Influenza H3N2. PLoS Pathog **10**:e1003932.

**Table S1**. Number of Brazilian sequences and sampling time span according to state and city in the complete dataset and subsamplings for *pol* and *env* genes.

| State | City | Sampling time span | *pol* datasets | | |
|---|---|---|---|---|---|
| | | | Complete | Rand10 | Rand20 |
| RS | CRA | 2011-2012 | 17 | 10 | 17 |
| | CXS | 2013-2014 | 11 | 10 | 11 |
| | LAJ | 2013 | 2 | 6 | 6 |
| | POA | 2002-2011 | 50 | 10 | 20 |
| | RIG | 2002-2007 | 20 | 10 | 20 |
| | STI | 2013 | 2 | 2 | 2 |
| | STL | 2013-2014 | 4 | 4 | 4 |
| | URU | 2009-2010 | 6 | 6 | 6 |
| SC | BLU | 2012 | 15 | 10 | 15 |
| | CHA | 2012 | 10 | 10 | 10 |
| | CRI | 2007-2013 | 30 | 10 | 20 |
| | FLP | 2008-2011 | 63 | 10 | 20 |
| | ITA | 2004-2014 | 35 | 10 | 20 |
| | JOI | 2012-2013 | 25 | 10 | 20 |
| | LGE | 2013 | 3 | 3 | 3 |
| GO | GOI | 2003-2011 | 17 | 10 | 17 |
| MS | CPG | 2008-2010 | 4 | 4 | 4 |
| PR | CTB | 2004-2006 | 15 | 10 | 15 |
| RJ | RJN | 2004-2011 | 19 | 10 | 19 |
| SP | SPL | 2003-2010 | 21 | 10 | 20 |
| TO | PAL | 2008 | 2 | 2 | 2 |
| **TOTAL** | | | 371 | 167 | 271 |

| State | City | Sampling time span | *env d*atasets | | |
|---|---|---|---|---|---|
| | | | Complete | Rand10 | Rand20 |
| RS | CRA | 2011-2012 | 19 | 10 | 19 |
| | CXS | 2013-2014 | 18 | 10 | 18 |
| | LAJ | 2013 | 15 | 10 | 15 |
| | POA | 2003-2011 | 18 | 10 | 18 |
| | STI | 2013 | 11 | 10 | 11 |
| | STL | 2013-2014 | 8 | 8 | 8 |
| | URU | 2009-2010 | 9 | 9 | 9 |
| SC | BLU | 2012 | 19 | 10 | 19 |
| | CHA | 2012 | 16 | 10 | 16 |
| | CRI | 2007-2013 | 20 | 10 | 20 |
| | FLP | 2008-2011 | 64 | 10 | 20 |
| | ITA | 2004-2014 | 16 | 10 | 16 |
| | JOI | 2012-2013 | 34 | 10 | 20 |
| | LGE | 2013 | 5 | 5 | 5 |
| AM | MAN | 2006 | 2 | 2 | 2 |
| RJ | RJN | 2008-2010 | 3 | 3 | 3 |
| SP | SPL | 2003-2010 | 12 | 10 | 12 |
| **TOTAL** | | | 289 | 147 | 231 |

**Acronyms for states:** AM: Amazonas; GO: Goiás; MS: Mato Grosso do Sul; PR: Paraná; RJ: Rio de Janeiro; SC: Santa Catarina; SP: São Paulo; TO: Tocantins; RS: Rio Grande do Sul. **Acronyms for cities:** BLU: Blumenau; CHA: Chapecó; CPG: Campo Grande; CRA: Cruz Alta; CRI: Criciúma; CTB: Curitiba; CXS: Caxias do Sul; FLP: Florianópolis; GOI: Goiania; ITA: Itajaí; JOI: Joinville; LAJ: Lajeado; LGE: Lages; MAN: Manaus; PAL: Palmas; POA: Porto Alegre; RIG: Rio Grande; RJN: Rio de Janeiro; SPL: São Paulo; STI: Santiago; STL: Santana do Livramento; URU: Uruguaiana.

## Table S2. Monophyletic Clade analysis for each location in *pol* and *env* datasets

***pol* - Monophyletic Clade (MC)**

| City | Sample Number | observed mean | lower 95% CI | upper 95% CI | null mean | lower 95% CI | upper 95% CI | *p* value |
|---|---|---|---|---|---|---|---|---|
| FLP | 63 | 2.89 | 2 | 4 | 2.30 | 1.92 | 3.03 | 0.091 |
| BLU | 15 | 2.00 | 2 | 2 | 1.17 | 1.00 | 2.00 | 0.048* |
| CHA | 10 | 2.00 | 2 | 2 | 1.07 | 1.00 | 1.38 | 0.02* |
| CRA | 17 | 3.03 | 3 | 3 | 1.23 | 1.00 | 2.00 | 0.002* |
| CRI | 30 | 2.55 | 2 | 3 | 1.59 | 1.06 | 2.09 | 0.008* |
| CXS | 11 | 1.01 | 1 | 1 | 1.10 | 1.00 | 1.85 | 1.000 |
| ITA | 35 | 2.54 | 2 | 3 | 1.74 | 1.14 | 2.18 | 0.341 |
| JOI | 25 | 2.00 | 2 | 2 | 1.45 | 1.02 | 2.02 | 0.171 |
| LAJ | 2 | 1.00 | 1 | 1 | 1.03 | 1.00 | 1.12 | 1.000 |
| LGE | 3 | 1.00 | 1 | 1 | 1.01 | 1.00 | 1.00 | 1.000 |
| POA | 50 | 6.10 | 4 | 10 | 2.05 | 1.54 | 2.89 | 0.001* |
| STI | 2 | 1.09 | 1 | 2 | 1.00 | 1.00 | 1.00 | 1.000 |
| STL | 4 | 2.00 | 2 | 2 | 1.01 | 1.00 | 1.01 | 0.002* |
| URU | 6 | 5.90 | 6 | 6 | 1.03 | 1.00 | 1.12 | 0.001* |
| GOI | 17 | 2.02 | 2 | 2 | 1.24 | 1.00 | 2.00 | 0.080 |
| CPG | 4 | 1.00 | 1 | 1 | 1.01 | 1.00 | 1.01 | 1.000 |
| PAL | 2 | 1.00 | 1 | 1 | 1.00 | 1.00 | 1.00 | 1.000 |
| SPL | 21 | 3.05 | 3 | 3 | 1.34 | 1.01 | 2.00 | 0.003* |
| RJN | 19 | 4.00 | 4 | 4 | 1.26 | 1.00 | 2.00 | 0.001* |
| RIG | 20 | 2.05 | 2 | 2 | 1.31 | 1.00 | 2.00 | 0.118 |
| CTB | 15 | 2.00 | 2 | 2 | 1.18 | 1.00 | 2.00 | 0.057 |

***env* - Monophyletic Clade (MC)**

| City | Sample Number | observed mean | lower 95% CI | upper 95% CI | null mean | lower 95% CI | upper 95% CI | *p* value |
|---|---|---|---|---|---|---|---|---|
| BLU | 19 | 2.00 | 2 | 2 | 1.36 | 1.00 | 2.01 | 0.137 |
| CHA | 16 | 2.00 | 2 | 2 | 1.26 | 1.00 | 2.00 | 0.093 |
| CRA | 19 | 2.04 | 2 | 2 | 1.36 | 1.00 | 2.01 | 0.144 |
| CRI | 20 | 2.00 | 2 | 2 | 1.39 | 1.01 | 2.01 | 0.159 |
| CXS | 18 | 3.99 | 4 | 4 | 1.33 | 1.00 | 2.00 | 0.001* |
| FLP | 64 | 3.16 | 2 | 4 | 2.45 | 2.04 | 3.21 | 0.111 |
| ITA | 16 | 2.01 | 2 | 2 | 1.25 | 1.00 | 2.00 | 0.084 |
| JOI | 34 | 2.06 | 2 | 3 | 1.85 | 1.17 | 2.33 | 0.483 |
| LAJ | 15 | 2.00 | 2 | 2 | 1.22 | 1.00 | 2.00 | 0.072 |
| LGE | 5 | 1.00 | 1 | 1 | 1.02 | 1.00 | 1.05 | 1.000 |
| POA | 18 | 1.60 | 1 | 2 | 1.33 | 1.00 | 2.00 | 0.129 |
| STI | 11 | 2.01 | 2 | 2 | 1.13 | 1.00 | 1.95 | 0.038* |
| STL | 8 | 2.00 | 2 | 2 | 1.06 | 1.00 | 1.46 | 0.02* |
| URU | 9 | 4.99 | 2 | 6 | 1.08 | 1.00 | 1.63 | 0.001* |
| SPL | 12 | 2.00 | 2 | 2 | 1.15 | 1.00 | 2.00 | 0.054 |
| RJN | 3 | 2.00 | 2 | 2 | 1.01 | 1.00 | 1.01 | 0.005* |
| MAN | 2 | 1.00 | 1 | 1 | 1.00 | 1.00 | 1.00 | 1.000 |

***** Significant *p* values ($p<0.05$); **CI:** confidential interval. **Acronyms for cities:** BLU: Blumenau; CHA: Chapecó; CPG: Campo Grande; CRA: Cruz Alta; CRI: Criciúma; CTB: Curitiba; CXS: Caxias do Sul; FLP: Florianópolis; GOI: Goiania; ITA: Itajaí; JOI: Joinville; LAJ: Lajeado; LGE: Lages; MAN: Manaus; PAL: Palmas; POA: Porto Alegre; RIG: Rio Grande; RJN: Rio de Janeiro; SPL: São Paulo; STI: Santiago; STL: Santana do Livramento; URU: Uruguaiana.

**Table S3: Time of first transition and 95%HPD from Porto Alegre to all Brazilian sampled cities as estimated for *pol* and *env* Rand20 datasets***

| City | *pol* | *env* |
|---|---|---|
| | Year (95% HPD) | Year (95% HPD) |
| RIG | 1979 (1970-1986) | NA |
| FLP | 1980 (1972-1987) | 1979 (1969-1987) |
| URU | 1980 (1972-1987) | 1979 (1971-1986) |
| CRI | 1981 (1973-1987) | 1980 (1971-1988) |
| CTB | 1981 (1973-1987) | NA |
| ITA | 1981 (1973-1987) | 1983 (1974-1991) |
| RJN | 1981 (1973-1987) | 1993 (1980-2005) |
| GOI | 1981 (1973-1988) | NA |
| SPL | 1981 (1973-1988) | 1983 (1973-1991) |
| BLU | 1982 (1974-1989) | 1982 (1972-1989) |
| CRA | 1982 (1974-1989) | 1980 (1971-1989) |
| JOI | 1982 (1974-1989) | 1980 (1972-1989) |
| CXS | 1982 (1974-1990) | 1982 (1972-1989) |
| LAJ | 1985 (1975-1995) | 1982 (1972-1990) |
| CHA | 1985 (1976-1994) | 1982 (1973-1990) |
| CPG | 1986 (1977-1997) | NA |
| STL | 1987 (1977-1999) | 1985 (1976-1995) |
| LGE | 1988 (1976-2003) | 1989 (1979-1999) |
| STI | 1988 (1977-2003) | 1983 (1973-1991) |
| MAN | NA | 1989 (1978-2001) |
| PAL | 1990 (1977-2003) | NA |

*mean from the three Rand20 sub-samplings. **HPD:** highest probability density; **NA**: not available; **Acronyms for cities:** BLU: Blumenau; CHA: Chapecó; CPG: Campo Grande; CRA: Cruz Alta; CRI: Criciúma; CTB: Curitiba; CXS: Caxias do Sul; FLP: Florianópolis; GOI: Goiania; ITA: Itajaí; JOI: Joinville; LAJ: Lajeado; LGE: Lages; MAN: Manaus; PAL: Palmas; POA: Porto Alegre; RIG: Rio Grande; RJN: Rio de Janeiro; SPL: São Paulo; STI: Santiago; STL: Santana do Livramento; URU: Uruguaiana.

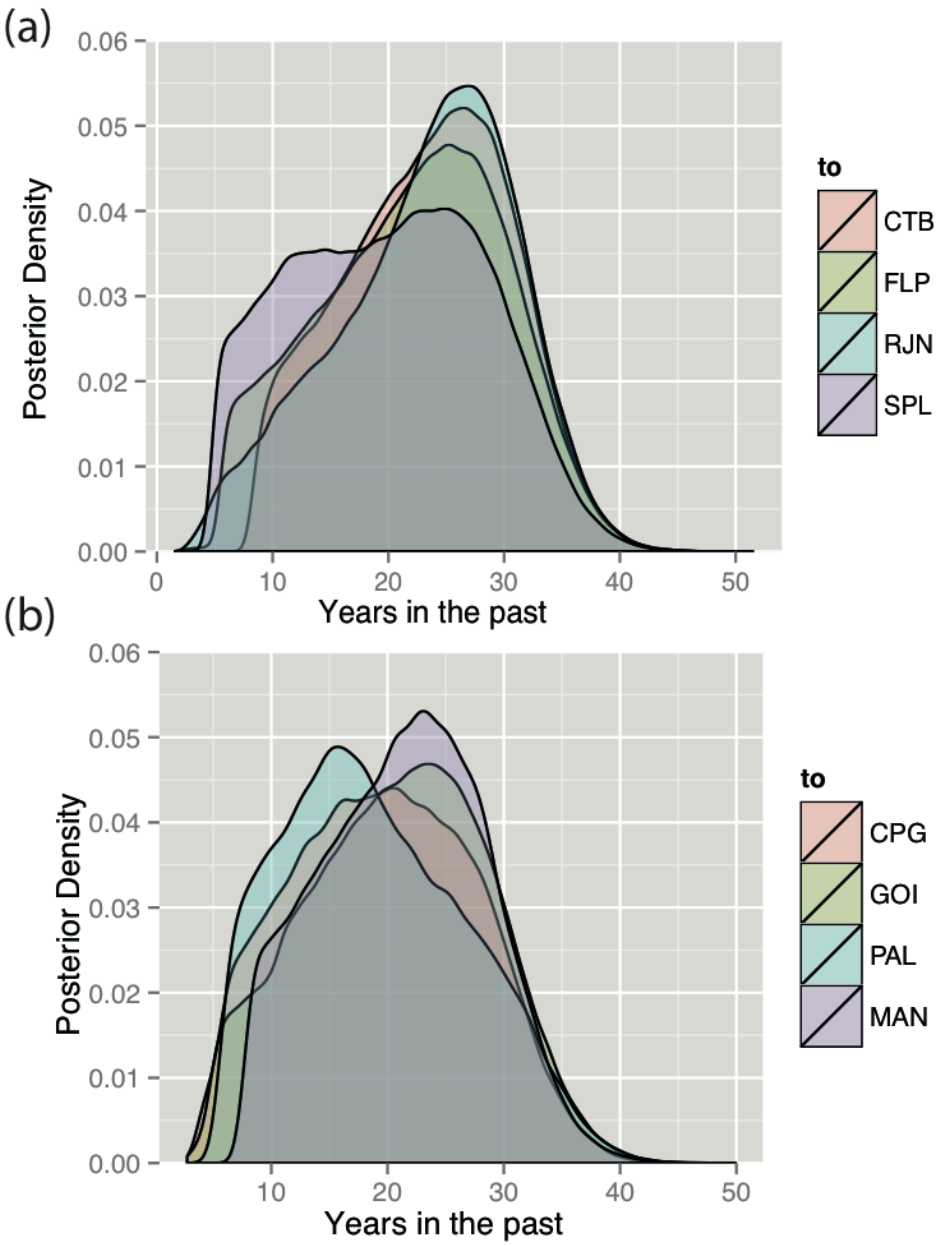


**Figure S1. Estimated density of location transitions by time from Porto Alegre to other Brazilian capitals.** Smooth density estimates of all transitions counted in *pol* and *env* Rand20 datasets analysis by time, with origin in Porto Alegre. Time zero in Y-axis is equal to 2014. Colors are according to the legend on the right. a) Transitions to South (FLP and CTB) and Southeast (RJN and SPL) regions; b) Transitions to Central-West (CPG and GOI) and to North (MAN and PAL) regions. CTB: Curitiba; FLP: Florianópolis; RJN: Rio de Janeiro; SPL: São Paulo; CPG: Campo Grande; GOI: Goiânia; PAL: Palmas; MAN: Manaus.