1

Phylodynamics analysis of the Human Immunodeficiency Virus type 1 (HIV-1) envelope gene in mother and child pairs.

Luciane Amorim Santos[1], Rebecca R Gray[3], Joana Paixão Monteiro-Cunha[1,2], Evandra Strazza[4], Simone Kashima[4], Edson de Souza[1], Thessika Hialla Almeida Araújo[1], Marilda de Souza Gonçalves[1], Marco Salemi[3], Luiz Carlos Junior Alcantara[1].

1- Laboratorio de Hematologia Genética e Biologia Computacional, Centro de Pesquisa Gonçalo Moniz, Fundação Oswaldo Cruz, Salvador, Bahia, Brasil. 2- Universidade Federal da Bahia. 3- Department of Pathology, Immunology, and Laboratory Medicine, University of Florida College of Medicine, Gainesville, FL, USA. 4- Hemocentro de Ribeirão Preto, Universidade de São Paulo-SP Brazil.

Running title: HIV-1 mother-to-child phylodynamics

Corresponding author:

Luiz Carlos Junior Alcantara, PhD

LASP/CPqGM/FIOCRUZ.

Rua Waldemar Falcão 121, Candeal, Salvador, Bahia, Brazil 40296-710.

Telephone # 55 71 3176 2255.

Fax # 55 71 3176 2300.

E-mail: lalcan@bahia.fiocruz.br

2

Abstract:

Characterizing the impact of HIV transmission routes on viral genetic diversity can improve the understanding of the mechanisms of virus evolution and adaptation. The HIV vertical transmission can occur *in utero*, during delivery or while breastfeeding. The present study investigated the phylodynamics of the HIV-1 *env* gene in mother-to-child transmission, by analyzing one chronically infected pair from Brazil and three acutely infected pairs from Zambia, with three to five time points. Sequences from 25 clones from each sample were obtained and aligned using Clustal X. ML trees were constructed in PhyML using the best evolutionary model. Bayesian analyses testing the relaxed and strict molecular clock were performed using BEAST and Bayesian Skyline Plot (BSP) was construed. The genetic variability of previously described epitopes were investigated and compared between each individual time points and between mother and child sequences. The relaxed molecular clock was the best-fitted model for all datasets. The tree topologies did not show differentiation in the evolutionary dynamics of the virus circulating in the mother from the child viral population. In the BSP, the effective population size was more constant along the time in the chronically infected patients while in the acute patients it was possible to detect bottlenecks. The genetic variability within viral epitopes recognized by the human immune system was considerably higher among the chronically infected pair in comparison with acutely infected pairs. These results contribute to the better understanding of the HIV-1 evolutionary dynamics in mother to child transmission.

In 2012, an estimated 2.5 million children under the age of 15 years old and approximately 15.5 million women were living with the Human Immunodeficiency Virus (HIV) worldwide[1]. Vertical transmission is the dominant mode of HIV infection among children and occurs when the infected mother transmits the virus to the child during pregnancy, at delivery or breastfeeding. Without prevention, the transmission rate can reach 30%, but adopting appropriate prophylaxis measures, which include the use of anti-retroviral drugs during pregnancy and when the baby is born, cesarean delivery and the substitution of breastfeeding for artificial milk[2], the vertical transmission rate can be reduced to less than 1%. It is known that the virus strain transmitted from the mother to the child is not, necessarily, the most prevalent strain found in the mother[3]. Besides, the HIV-1 strain found in the child may have a different evolutionary history when compared with the virus found in the mother. This dissimilarity may be due to differences in each individual environment as immune response, which differs from one individual to the other and depends on the phase of infection, and the use of therapy.

The HIV evolutionary dynamics studies can contribute to the understanding of the viral fitness mechanisms. Phylogenetic reconstructions can reveal the dynamics of HIV evolution intra and inter-host. The inter-host analyses allow inferring the movement of HIV lineages between locations, course of transmission, and estimate changes in viral effective population size over time. The selective pressure of the immune response, and viral replication rates influence the viral adaptation and dynamic, and can be evaluated by within host analyses[4, 5]. In this context, evolutionary pressure from the host immune response selects for HIV-1 mutations resulting in amino acid changes that allow the virus to evade cellular and humoral immune responses. Variable regions of the envelope (*env*)

4

gene are target of humoral and cellular immune responses and those patients with the strongest response will be able to control the virus infection for a longer period of time. The major consequence of this HIV escape mechanism is the great amount of nonsynonymous substitutions comparing to synonymous substitutions (positive selective pressure), in the *env* gene, contributing to the high diversity of this region in different time points within a host[6].

Considering that the time and mechanism of transmission are known, it is possible to infer the HIV-1 evolutionary dynamics. In this study, we evaluated the evolutionary dynamics differences of the HIV-1 *env* gene in vertical transmission, by a comparative analysis of HIV-1 subtypes B and C sequences isolated from mother and child pairs, obtained from three to five time points, in different phases of infection. The studied subjects included one mother and child pair assisted at the Sexually Transmitted Disease Reference Centre in Feira de Santana, Bahia, Brazil, whose samples were collected from three different time points, with an interval of, approximately, six months each, and three pairs from Zambia, Africa, whose *env* sequences were downloaded from the GenBank. Similarly as obtained for the first pair, each individual had viral sequences from three to five different time points with an interval of, approximately, six month from one sample to the other. In the case of the mother-child pair from Brazil, the samples were collected after signature of the informed consent letter. The Bahiana School of Medicine and Public Heath Ethics Committee approved this study.

Peripheral blood mononuclear cells (PBMCs) from the Brazilian pair were isolated from 10mL of total blood using the Ficoll-Hypaque method. DNA was extract from the PBMCs using QIAGEN kit (QIAamp® DNA Blood Kit). Nested PCR, using specific

5

primers, was performed in order to obtain an HIV-1 *env* (1476pb – from HXB2 position 6822 to 8298) gene fragment. The PCR product was purified using QIAquick PCR Purification kit (QIAGEN) and then cloned by TOPO TA Cloning kit. From each time point sample, approximately 25 clones were selected and sequenced at the Genome Sequence Service Laboratory, University of Florida, Gainesville, Florida, USA.

The sequences generated from all time points, for each patient, were aligned using Clustal X[7] and manually edited. For the Zambia pairs we also aligned the child sequences with the first time point sequences of the respective mother. Hudson test was performed to check if sequences from each different time point represent different subpopulations[8]. Considering that the presence of intra-patient recombinants can disturb the tree construction, the identification of these recombinants sequences was performed through the PHI (pair-wise homoplasy índex) test[9], implemented in the SplitsTree program[10]. Recombinant sequences were excluded until that the *p* value had not been significant ($p>0.05$)[11]. All recombinants detected were excluded from the dataset for posterior analyses.

Maximum Likelihood trees were generated using the PhyML[12] online tool applying the GTR evolution model to estimate the proportion of invariable sites and gamma shape parameter. The branch support was obtained by the bootstrap (1000 replicates). Bayesian analyses were performed with the BEAST v1.4.8 package[14] testing the strict molecular clock with constant population size prior and the relaxed molecular clock using the constant population size, Bayesian Skyline Plot (BSP) and exponential growth priors. The parameters for each model were estimated using the Monte Carlo Markov Chain (MCMC) method (50.000.000 generations with sampling every 5000 generations). The

6

analysis results were visualized using Tracer v1.4 software and the MCMC convergence was assessed calculating the effective sampling size (ESS) for each parameter, admitting that the sampling size were significant when ESS > 500[13]. The tested models were compared calculating the Bayes Factor (BF): the ratio of the marginal likelihood of the compared models. Evidence against the null model, the one with the lower marginal likelihood, is indicated by $2.\log_e.(BF) > 3$ is consider moderate evidence and > 10 strong evidence. The BF was calculated to compare the strict molecular clock with the relaxed molecular clock model, both using the constant population growth prior, and then, the relaxed molecular clock with the constant prior against the BSP prior and the exponential growth prior. The calculations were performed using BEAST v1.4.8 and Tracer v1.4 programs. Bayesian framework, using the relaxed molecular clock with the BSP prior, estimated the effective population size, an informative method of assessing the pathogen evolutionary history. Tracer v1.4 were used to perform the BSP reconstruction. Using TreeAnnutator v1.4.8 program, included in the BEAST package[13], the maximum clade credibility tree were selected from the posterior tree distribution after a 50% burn-in, for each dataset and all trees were visualized using FigTree v1.2.2.

Subtype B CTL and B-cell epitopes, previously described in Los Alamos database, were mapped[14] for the Brazilian pair sequences, which are subtype B, while previously described subtype C CTL and B-cell epitopes were mapped for the Zambia pairs, since their sequences clustered as subtype C. For this analysis we aligned the epitope amino acid sequence to the amino acid sequences from each study subject and detected the presence of the epitope and what mutations were found in the epitope region by visual analysis.

7

In this study one mother and child pair from Feira de Santana, FS17 and FS16, and three pairs from Zambia, MIP834, MIP2660 and MIP2953 were analyzed. The FS16 and FS17 isolates were HIV-1 subtype B and were chronically infected at the first time point. The pair was diagnosed with HIV-1 three years before the collection of the first time point sample. The child (FS16) was borne by natural birth and the mother (FS17) was under treatment in the first time point.

Different from the Brazil pair, the three isolates pairs from Zambia were subtype C and the first sample time point was the first HIV positive sample for the mother and child, showing that the transmission from the mother to the child and, then, the sampling, happened during the acute phase of infection of the mother. In addition, information on possible clinical symptoms of the mothers at sampling time were not available in the GenBank. These pairs dataset were selected for having three or more time points with multiple sequences from each time point and with approximately six month from one time point to the other.

The Hudson test was performed in all datasets to check if the different time points were different subpopulations. All datasets showed a value of $p < 0.01$, indicating that each time point of the mother and the child represents different subpopulations.

The PHI test did not show different patterns in the percentage of recombination along time points when comparing the mother to child or the acute infected with the chronically infected sequence datasets (Table 1). In most patients, a great number of recombinant sequences at the last time point were observed. The sequences from the mother MIP834_M showed the lower number of recombination events, presenting only one (3%)

8

recombinant sequence at the second time point, but some time points like FS-16 presented recombination as high as in 58% of the clone sequences (Table 1). The sequences from pair FS-16 and FS-17 were form clones of multiple PCR reactions. This could increase the variation among the clones and the number of recombination. The information if the Zambia pairs clones were from a single PCR is not available. The presence of recombinants sequences in the phylogenetic analyses can disrupt its inference given a less likely evolution results. Due to that, the sequence detected as recombinants by the PHI test were excluded from further analyses.

For sequence quality control the Quality Control tool, from Los Alamos HIV database was used. A maximum likelihood tree was also constructed using all the Zambia sequences, and showed no epidemiological link between the pairs.

Maximum likelihood trees inferred using the online tool PhyML did not showed a statistic branch support. The topology showed a higher relation between the child and the mother sequences at first time point, in the Zambia pairs, due to the proximity of the sampling time and the transmission event. However the chronically infected pair showed distant relation between the sequences from the mother and the child.

To study the HIV phylodynamics in these patients, Bayesian trees were constructed taking into account the sampling time. The relaxed molecular clock, that assigns different evolutionary rates for the different branches on the tree, was chosen the best fitting model for all datasets when compared to the strict clock, which assigns fix evolutionary rate for the analyzed gene.

9

The sequence datasets were then analyzed using the relaxed molecular clock with the BSP and exponential population growth. Each model was then compared with the constant size prior, calculating the Bayes Factor. For all datasets, the BSP was selected over the constant size, and the acutely infected pairs were strongly selected (> 10) while the chronically infected pair was moderately selected (>3). However, the exponential growth showed strong evidence over the constant model in the MIP834_I (child) dataset, and moderate evidence in the MIP2953_M (mother), MIP2953I+M (child + mother first time point) and FS17 (mother) datasets. The confidence intervals of the growth rate of MIP2662_I, MIP2953_I, FS16 and FS17 included the zero, indicating that there is no exponential growth in these viral populations.

Figure one shows the Bayesian tree constructed with the relaxed molecular clock model. We did not find any pattern or different behavior of the tree topologies between mother and child neither among subtypes or phase of infection. Some of them presented a perfect temporal structure like: MIP834_M and MIP2953_I (Figure 1).

The BSP graphics that estimates the viral effective population size showed different pattern for each pair, although the size had been around 2000. The pair from Feira de Santana showed a more constant growth over time, while in the Zambia pairs it was possible to observe the bottleneck between time points especially in the MIP2660 and MIP834 pairs (Figure 2) analysis.

Out of the 21 subtype C B-cell epitopes described in the Los Alamos database in this genomic fragment (*env*), four were wild type (for this analysis the Los Alamos subtype consensus sequence were considered as wild type) in the mother and the child of the

10

MIP834 and MIP2660 pairs and one in the MIP2953 pair. Concerning CTL epitopes, only two of the ten epitopes found in the subtype C consensus sequence were wild type in the children and three in the mothers of the MIP2660 and MIP2953 pairs. The MIP834 mother presented four wild type epitopes and the child presented five wild type epitopes. Most of the mutations found in the mother sequences were the same ones found in the child sequences. However, for the subtype B pair, the epitopes and mutations were different between mother and child sequences for both CTL and B-cell epitopes.

The genetic variability within B-cell and CTL epitopes was investigated and compared between different clones of each pair. Within the genomic fragment (*env*) analyzed in this study, there were 21 B-cell epitopes described in the Los Alamos database for the consensus subtype C sequence. Out of these, 17 (81%) epitopes were entirely conserved in the mother and child MIP834, 20 (95%) in the MIP2660 and six (29%) in the MIP2953 in all time points. The remaining four epitopes (19%) for the MIP834, one (5%) for the MIP2660 and 15 (71%) for the MIP2953 were entirely conserved in the different time points clones of the same individual, but viral strains with additional mutations were present in part of the viral population of either the mother or the child. Ten subtype C CTL epitopes were previously described in the Los Alamos database for the consensus subtype C sequence in the genomic fragment analyzed. Comparing the viral populations of mother and child, six (60%) epitopes were entirely conserved along the time in the MIP834 pair, three (30%) in the MIP2660 pair and seven (70%) in the MIP2953 pair. Mutations emerged or disappeared along the time within four (40%) epitopes in either the mother or the child of the MIP834, seven (70%) epitopes in either the mother or the child

11

of the MIP2660 and three (30%) epitopes in either the mother or the child of the MIP2953 pair.

For the chronically infected subtype B pair analysis, 32 B-cell epitopes were included in the analysis (previously described in the Los Alamos database for the subtype B consensus sequence). Of those, three (9%) epitopes were entirely conserved in the mother and child in all time points. Mother and child showed difference in 29 (91%) B-cell epitopes: 12 epitopes were conserved among the different time points clones of the same individual, but additional mutations were present in part of the viral population of either the mother or the child and 17 (53%) epitopes were completely different between mother and child and among the time points. Out of the 53 CTL epitopes analyzed within the FS16/FS17 pair sequences, only 2 (4%) were entirely conserved in the mother and the child along the time. Five (9%) epitopes were partially conserved along the time, since in the third time point (124 months) new viral strains emerged but disappeared in the fourth time point (129 months). Mother and child showed difference in 46 (87%) CTL cell epitopes: 4 epitopes were conserved among the different time points clones of the same individual, but additional mutations were present in part of the viral population of either the mother or the child and the remaining 42 (79%) epitopes were completely different between mother and child and among the time points.

In conclusion, the tree topologies generated in this study did not show differentiation in the evolutionary dynamics of the virus circulating in the mother from the child viral population. However, the trees and the BSP show that the strains from the mother are more closely related to the child strains in the acutely infected mothers when transmitted the virus to their child.

12

Studies have shown that during chronic HIV-1 infection only a few variants of quasispecies are transmitted [3, 15, 16, 17]. However, acutely infected mothers transmit multiple closely related variants to their child[18], which was possible to be identify in the trees constructed using the child sequences with the mother sequences obtained from the first time point (data not showed). In the acutely infected pairs, sequences from the first time point of the mother and their child are mixed, showing different variants giving rise to the next population, while in the chronically infected pair analysis, the mother and the child present two different viral populations. The transmission of lineages from the mother to the child without selection in the mother and the possible multiple transmissions during breastfeed are some of the possible explanations[18].

The effective population size reflects evolutionary relationship among strains and changes in the number of effectively infectious virus, rather than absolute number of circulating virions and viral load. Several studies have estimated the effective population size of HIV-1 quasispecies intra-patient[19, 20, 21, 22]. In this study, the calculus of the Bayes factor selected the BSP over the constant population growth in all dataset, while the exponential growth were only selected in a few datasets, showing that the BSP is a better fitting model to estimate the effective population size. The population growth estimative using different models indicate that the chronically infected pair has a more constant growth. Although the BSP model had been moderately selected over the constant, the BSP graphic shows a more constant effective population size over time, and the exponential growth rate did not exclude the zero, indicating constant growth. This could be related to a better adaptation of the virus in the chronic phase of infection and to a more compromised immune system in this pair.

13

The BSP of the acutely infected pairs shows a more dynamic growth so that the effective population size suffers a rapid decrease followed by growth, bottleneck event, due to the strong immune system pressure, from the beginning of the infection, selecting the more adapted strains overtime. A perfect temporal structure was found at MIP834_M (mother) dataset visualized in the tree and in the BSP, where the bottleneck events are clearly showed from one time point to the other.

The epitope mapping also evidences that the chronically infected mother presents a more distinct population from the child. Their different immune response and the long time of infection lead to this very different populations, while the acutely infected pairs present more similar sequences, with similar epitopes between mother and child.

These findings show how HIV-1 intra-host population dynamics can differ depending on the phase of infection and transmission, and contribute to a better understanding of the virus evolution mechanisms following vertical transmission.

Sequence Data

The sequences GenBank accession numbers for the F16-F17 pair are KF247318 – KF247430 and for the Zambia pairs are FJ854750 – FJ855125 and FJ859377 – FJ859679[18].
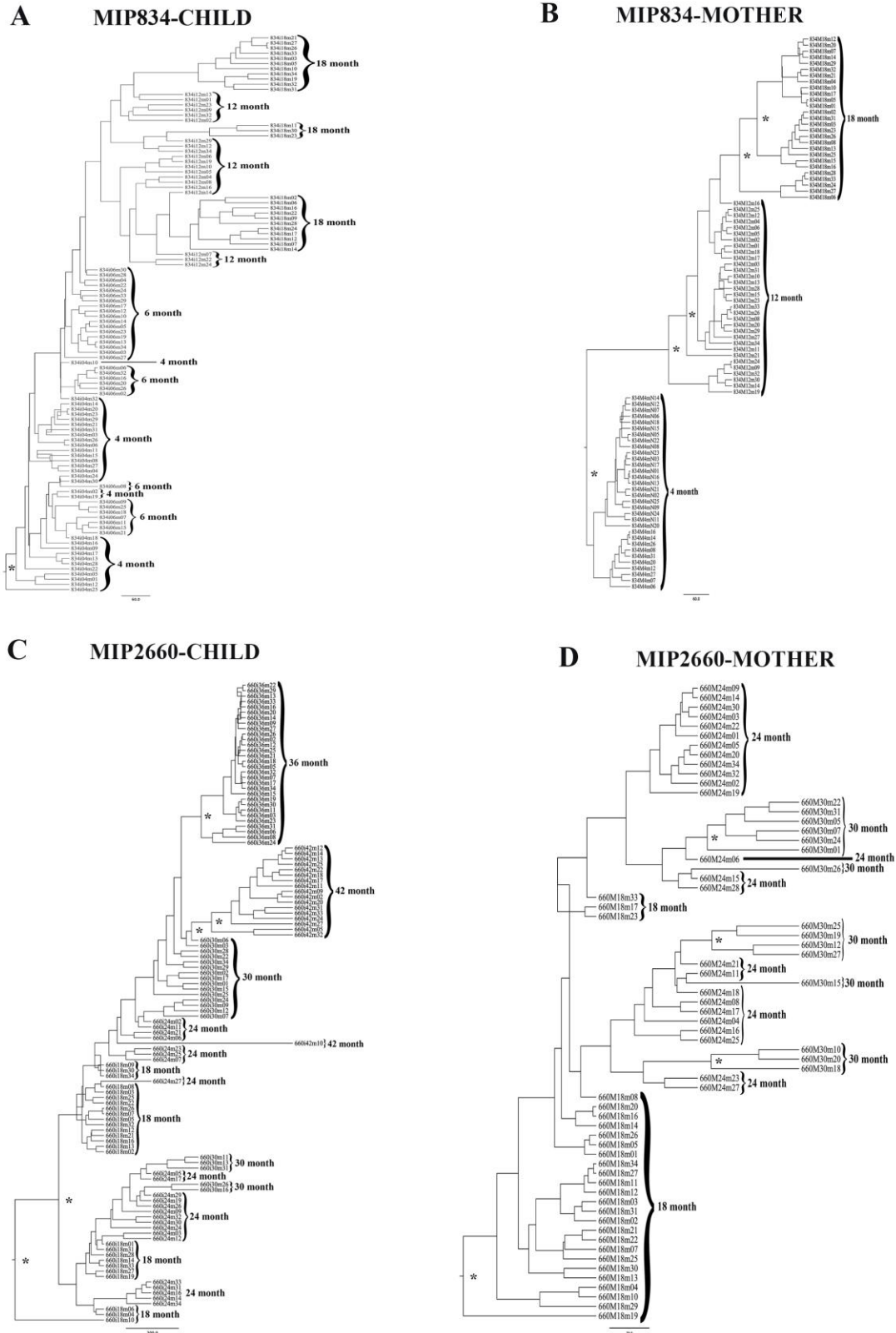
Acknowledgments

15

Reference

1. UNAIDS. AIDS epdemic update 2007. <http://www.unaids.org/en/KnowledgeCentre/HIVData/EpiUpdate/EpiUpdArchive/2007/default.asp.>

2. Ministério da Saúde. Epidemiologia HIV. <http://www.aids.gov.br/data/Pages/LUMIS72418C70PTBRIE.htm.>

3. Wolinsky SM, Wike CM, Korber BT, *et al.* Selective transmission of human immunodeficiency virus type-1 variants from mothers to infants. Science 1992; 255(5048): 1134-7.

4. Grenfell B, Pybus O, Gog J, et al. Unifying the epidemiological and evolutionary dynamics of pathogens. Science 2004; 303:327-32.

5. Lemey P, Rambaut A and Pybus O. HIV evolutionary dynamics within and among hosts. AIDS Reviews 2006;8:125-40.

6. Williamson S. Adaptation in the env gene of HIV-1 and evolutionary theories of disease progression. Mol Biol Evol 2003;20:1318-25.

16

7. Thompson JD, Gibson T J, Plewniak F, Jeanmougin F, Higgins D G. The ClustalX windows interface: flexible strategies for multiple sequence alignment aided by quality analysis tools. Nucleic Acids Research 1997;25: 4876-4882.

8. Hudson RR, Boos DD, Kaplan NL. A statistical test for detecting geographic subdivision. Molecular Biology Evolution 1992;9(1):138-51.

9. Bruen T, Philippe H, Bryant D. A simple and robust statistical test for detecting the presence of recombination. Genetics 2006; 72:2665-2681.

10. Huson D, Bryant D. Application of phylogenetic network in evolutionary studies. Molecular Biology Evolution 2006; 9:138-151.

11. Salemi M, Gray RR, Goodenow MM. An exploratory algoruthm to identify intra-host recombinat viral sequences. Molecular Phylogenetics and Evolution 2008;49(2):618-28.

12. Guindon S, Gascuel O. A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. Systematic Biology 2003;52(5):696-704.

13. Drummond AJ, Rambaut A. BEAST: Bayesian evolutionary analysis by sampling trees. BMC Evolutionary Biology 2007;**7**:214.

17

14. Bette, Korber T, Brander M, Haynes BF, Koup R, Moore JP, Walker BD, Watkins DI. HIV Molecular Immunology 2006/2007, Publisher: Los Alamos National Laboratory, Theoretical Biology and Biophysics, Los Alamos, New Mexico. LA-UR 07-4752.

15. Zhang H, Hoffmann F, He J, He X, Kankasa C, West JT, et al. Characterization of HIV-1 subtype C envelope glycoproteins from perinatally infected children with different courses of disease. Retrovirology 2006;3:73.

16. Zhang H, Hoffmann F, He J, He X, Kankasa C, Ruprecht R, et al. Evolution of subtype C HIV-1Env in a slowly progressing Zambian infant. Retrovirology 2005;2:67.

17. Derdeyn C A, Decker JM, Bibollet-Ruche F, Mokili JL, Muldoon M, Denham SA, et al. Envelope-constrained neutralization-sensitive HIV-1 after heterosexual transmission. Science 2004;303:2019–2022.

18. Hoffmann FG, He X, West JT, *et al*. Genetic variation in mother-child acute seroconverter pair from Zambia. AIDS 2008; 22:817-824.

19- Achaz G, Palmer S, Kearney M, Maldarelli F, Mellors JW, Coffin JM, Wakeley J. A robust measure of HIV-1 population turnover within chronically infected individuals. Mol Biol Evol 2004; 21**:**1902-12.

18

20. Seo TK, Thorne J L, Hasegawa M, Kishino H. Estimation of effective population size of HIV-1 within a host: a pseudomaximum-likelihood approach. Genetics 2002;160**,** 1283-93.

21. Brown A J. Analysis of HIV-1 env gene sequences reveals evidence for a low effective number in the viral population. Proc Natl Acad Sci USA 1997; 94:1862-5.

22. Shriner D, Shankarappa R, Jensen M A, Nickle D C, Mittler J E, Margolick J B, Mullins J I. Influence of random genetic drift on human immunodeficiency virus type 1 env evolution during chronic infection. Genetics 2004;166**,** 1155-64.

Figures Legends

A **MIP834-CHILD**

B **MIP834-MOTHER**

C **MIP2660-CHILD**

D **MIP2660-MOTHER**

E   MIP2953-CHILD

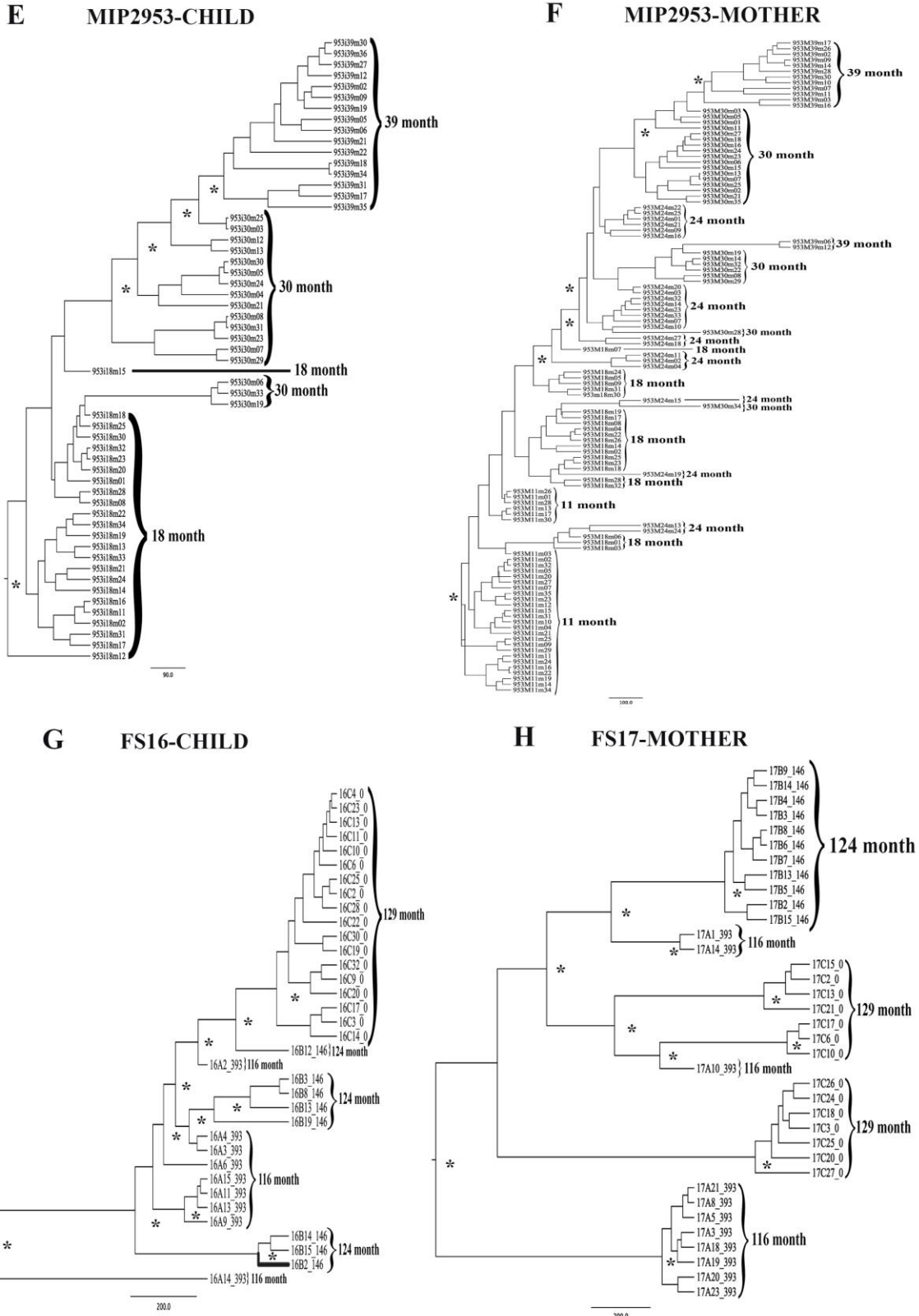F   MIP2953-MOTHER

G   FS16-CHILD

H   FS17-MOTHER

Figure 1: Bayesian maximum clade credibility phylogenetic trees. The trees were constructed using the relaxed molecular clock model and constant population size model. Tree roots were in the first time point considering that it is a temporal tree. The branch support are giving by the posterior probability > 0.80 and are indicated with an asterisk (*). The panels A and B are from the MIP834 pair; C and D are from the MIP2660 pair; E and F are from the MIP2953 pair and G and H are the FS16/FS17 pair. The number of months of each time point is related to the date of birth of each respective child.
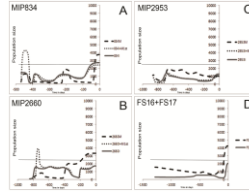
Figure 2. Bayesian Skyline Plots to estimate the effective population size. Each graphic shows the BSP mean value of the mother (M and FS17) and the child (I and FS16). For the Zambia pairs it is also found in the graphics the BSP of the child with the first time point of the mother (M+I). The confidence intervals were not plotted in the graphic and its approximately 10^3 higher and lower than the plotted values. The 2500 population size line was chosen based on the size of the different population along time for a better comparison between the different graphics. These analyses were performed in BEAST using all available time point for each patient.

Table 2: Recombinant sequence distribution

Sample	Time Point	# of Clones	#of Final Clones	% of recombinants
MIP834_I	4m	31	31	0%
	6m	32	32	0%
	12m	30	20	30%
	18m	32	25	21.88%
MIP834_M	4m	32	32	0%
	12m	33	32	3%
	18m	27	27	0%
MIP2660_I	18m	29	26	10.30%
	24m	32	24	25%
	30m	28	20	28.60%
	36m	33	30	9.10%
	42m	29	18	37.90%
MIP2660_M	18m	27	27	0%
	24m	28	25	10.70%
	30m	25	15	32%
MIP2953_I	18m	29	24	17.20%
	30m	25	17	28%
	39m	24	16	33%
MIP2953_M	11m	31	31	0%
	18m	27	22	18.50%
	24m	29	23	20.70%
	30m	33	25	24.30%
	39m	31	14	54.80%
FS16	116m	10	9	10%
	124m	19	8	57.90%
	129m	26	18	30.76%
FS17	116m	20	11	45%
	124m	14	11	21.42%
	129m	24	14	41.66%