FIOCRUZ

FUNDAÇÃO OSWALDO CRUZ
CENTRO DE PESQUISAS GONÇALO MONIZ

Curso de Pós-Graduação em Biotecnologia em Saúde e Medicina Investigativa

DISSERTAÇÃO DE MESTRADO

USO DE FERRAMENTAS DE BIOINFORMÁTICA PARA ESTUDOS DE EPIDEMIOLOGIA MOLECULAR, FILOGEOGRAFIA E FILODINÂMICA VIRAL.

Luciane Amorim Santos

Salvador - Brasil
2010

FUNDAÇÃO OSWALDO CRUZ
CENTRO DE PESQUISAS GONÇALO MONIZ

FIOCRUZ

Curso de Pós-Graduação em Biotecnologia em Saúde e Medicina Investigativa

USO DE FERRAMENTAS DE BIOINFORMÁTICA PARA ESTUDOS DE EPIDEMIOLOGIA MOLECULAR, FILOGEOGRAFIA E FILODINÂMICA VIRAL.

Luciane Amorim Santos

Orientador: Prof. Dr. Luiz Carlos Júnior Alcântara
Co-orientador: Prof. Dr. Marco Salemi

Dissertação apresentada ao Curso de Pós-Graduação em Biotecnologia em Saúde e Medicina Investigativa para a obtenção do grau de Mestre.

Salvador - Bahia – Brasil
2010

“E ele muda os
tempos e as estações;
ele remove os reis e
estabelece os reis; ele
dá sabedoria aos sábios
e conhecimento aos
entendidos.”
Daniel 2:21

AGRADECIMENTOS

A Deus, “ao único Deus, nosso Salvador, mediante Jesus Cristo, Senhor nosso, glória, majestade, império e soberania, antes de todas as eras, e agora, e por todos os séculos. Amém!” (Jd 1:25).

Ao meu orientador Dr Luiz Carlos Junior Alcântara pela grande contribuição na minha formação científica, pela paciência e incentivo, dando-me a oportunidade de desenvolver este trabalho, e me ensinando em todo o tempo.

Ao Dr Marco Salemi, por me receber em seu laboratório, me ensinando e me orientando sempre com toda a sua alegria e entusiasmo italiano.

Ao Dr Bernardo Galvão Castro Filho, por me receber e dar a oportunidade de realizar este trabalho no Laboratório Avançado de Saúde Pública e no Centro de HTLV da Escola Bahiana de Medicina e Saúde Pública, e com seus sábios conselhos.

Ao meu “papai godinho”, Roberto Santos, pelo amor, apoio e incentivo, sendo um exemplo de vida pra mim.

Á minha mãe e melhor amiga, Ana Lúcia, pelo amor, amizade e pelas palavras de sabedoria que sempre vem na hora certa.

Aos meus irmãos, Juliana e Gustavo, e cunhados, Anderson e Ana Carolina, pelos momentos de “lezeiras” e por sempre acreditarem em mim.

Á minha avó, Celina, pelo seu grande amor por mim, pelos agrados e paparicações.

Á amiga e “roomate” Nazle, por ser minha família durante o ano longe de casa, com seu sorriso contagiante todos os dias, além dos ensinamentos científicos.

Á Rebecca Gray, por tudo que me ensinou sempre com muita disposição, paciência e carinho.

A todos do laboratório da Dra Goodenow na Universidade da Flórida, por me receberem de uma forma tão acolhedora, e por todos os ensinamentos.

Ás amigas Aline e Hialla, pela ajuda, paciência e amizade. Vocês são um presente de Deus na minha vida.

Á Joana, pela grande ajuda e colaboração neste trabalho.

Aos Amigos do LASP e do Centro de HTLV. É muito bom conviver todos os dias com vocês.

Aos meus amigos baianos, cariocas, pernambucanos e mineiros, pelas orações, incentivo, e grande amizade em todo o tempo. Vocês fazem a minha vida muito mais divertida.

A todos os professores do Curso de Pós-Graduação em Biotecnologia em Saúde e Medicina Investigativa com os quais tive a oportunidade e o prazer de aprender.

A todos os co-autores dos trabalhos pela confiança e colaboração.

Aos colegas da pós-graduação pela convivência e amizade durante todo o curso.

Ao CNPq pelo apoio financeiro.

RESUMO

As ferramentas de bioinformática tem sido amplamente utilizada para o melhor entendimento de diversos microorganismos. Neste trabalho foram realizados três estudos utilizando estas ferramentas para avaliar diferentes questões biológicas. No primeiro estudo realizou-se uma caracterização molecular de 57 sequências do gene *pol,* provenientes de pacientes infectados pelo HIV-1 de Salvador, Bahia, Brasil. Para identificar os subtipos e formas recombinantes do HIV-1 circulante na cidade de Salvador foi realizado análises filogenéticas, e através do algoritmo do banco de dados *Stanford HIV resistance* as mutações associadas à resistência aos ARVs foram detectadas. Entre as 57 sequências analisadas foram identificados neste estudo 45 (77,2%) pertencem ao subtipo B, 11 (21,0%) recombinantes BF e uma (1,8%) do subtipo F1. Além disto, uma alta frequência de eventos de recombinação entre os subtipos B e F foram detectados com 5 padrões de recombinação, duas intergênicas e três intragênicas, mostrando uma alta diversidade. As mutações encontradas com uma maior prevalência foram: I54V (PI) em 7,0%; M184V (NRTI) em 14,0% e K103N (NNRTI) em 10,5% das sequências analisadas. Estes resultados contribuem para traçar o perfil da epidemiologia molecular e diversidade do HIV-1 em Salvador. O segundo estudo avaliou a filodinâmica do HIV-1 em pares de mãe e filho infectados, e em diferentes fases da infecção, três pares na fase aguda e um na fase crônica, e que apresentavam sequências de diferentes tempos. Para este fim foi realizado inferências filogenéticas bayesianas, onde a hipótese do relógio molecular e de diferentes crescimentos populacional foram testadas. Não foi possível observar uma diferença entre a dinâmica da população viral da mãe e a encontrada no filho. Porém, quando observamos o crescimento populacional e o tamanho da população efetiva, ao longo do tempo, sequências provenientes de pares em fase crônica da infecção tem um crescimento mais constante, enquanto as sequências dos pares na fase aguda da infecção se observa uma dinâmica das populações virais, provavelmente devido à pressão do sistema imune e a não adaptação destes vírus. No terceiro estudo, 104 sequências do genoma completo do WNV, disponíveis no Genebank, foram estudadas para identificar a região genômica que apresenta máximo poder interpretativo para inferir relações temporais e geográficas entre as cepas do vírus. Alinhamentos de cada gene foram submetidos à avaliação do sinal filogenético através do programa TREEPUZZEL. As regiões NS3 e NS4 apresentaram um sinal filogenético acima de 70%, sendo as regiões mais indicadas para construção filogenética. Além disto, árvores bayesianas foram inferidas utilizando as regiões NS3, NS5 e E, onde os clados das árvores NS3 e NS5 apresentaram um maior suporte e estrutural temporal geográfica, diferente da região E. Estes achados mostram que os genes NS3 e NS5 são os mais indicados para análises filogenéticas. Neste trabalho foi demonstrando o uso de ferramentas de bioinformática para a melhor caracterização da diversidade, epidemiologia molecular, dinâmica populacional e determinação das relações temporal e geográfica dos vírus.

Palavras-chave: HIV, WNV, Bioinformática, Filodinâmica, Filogeografia.

ABSTRACT

The bioinformatics tools have been widely used for better understanding of several microorganisms. Here three studies were performed using these tools to answer different biological questions. In the first study, it was conducted the molecular characterization of 57 HIV-1 *pol* gene sequences from infected patients from Salvador, Bahia, Brazil. To identify the HIV-1 subtypes and recombinants forms, phylogenetics analyses were performed and the Stanford HIV resistance Database were used to analyze the antiretroviral susceptibility. Among all analyzed sequences, 45 of them were (77.2%) subtype B, 11 (21.0%) were BF recombinant and one sequence was (1.8%) subtype F1. Furthermore, a high frequency of recombination events between subtypes B and F was detected with five different patterns: two intergenic and three intragenic. The mutations found with higher prevalence were: I54V (PI) in 7.0%; M184V (NRTI) in 14.0% and K103N (NNRTI) in 10.5% of the analyzed sequences. These results contribute for the knowledge of the molecular epidemiology and diversity of HIV-1 in Salvador. The second study have evaluated the HIV-1 phylodynamics in mother and child infected pairs in different stages of infections: three pairs acutely infected and one chronically infected. Phylogenetic inference was performed using the Bayesian framework were the molecular clock and different population growth models hypothesis were tested. We did not find any difference of the population dynamics between mother and child. However, when observing the population growth and the effective population size through time, the chronically infected pair sequences showed a constant growth, while the acutely infected pair sequences showed a more dynamic population growth, probably due to the immune system selective pressure. In the third study, 104 WNV full genome sequences were selected from Genbank, to identify the best genomic region, which could provide the maximal interpretative power to infer temporal and geographic relationships among the virus strains. The phylogenetic signal was evaluated using the TREEPUZZEL program. The results showed that the NS3 and NS5 regions are the best ones to infer phylogeny since their phylogenetic signal was higher than 70%. Furthermore, Bayesian trees were constructed using the NS3, NS5 and E regions, and the NS3 and NS5 tree clades showed a higher support and a temporal geographic structure, different from the E region. These findings show that the NS3 and NS5 genes are the most informative genes for phylogenetic analyses. These studies demonstrated the use of bioinformatics tools for the better characterization of the virus diversity, molecular epidemiology, and population dynamics.

Key words: HIV, WNV, Bioinformatics, Phylodynamics, Phylogeography.

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE ABREVIATURAS E SIGLAS

3'	Região carboxi-terminal do ácido nucléico
3TC	Lamivudina
5'	Região amino-terminal do ácido nucléico
ABC	Abacavir
AIDS	Síndrome da Imunodeficiência Adquirida (*Acquired Immune Deficiency Syndrome*)
ATV	Atazanavir
ARV	Antiretrovirais (*Antiretroviral)*
ARVT	Tarapia antiretroviral (A*nti-retroviral Therapy*)
AZT	Azidotimidina
BEAST	*Bayesian Evolutionary Analysis Sampling Trees*
BF	Fator de Bayes (*Bayes Factor*)
BSP	*Bayesian skyline plot*
CA	Capsídio
CCR5	Receptor Quimiocina da Família Cisteína-Cisteína 5 (*Cysteine-Cysteine-Chemokine Receptor-5*)
CDC	*Centers for Disease Control*
cDNA	DNA complementar
CDST	Centro de Referência em Doenças Sexualmente Transmissíveis
CPqGM	Centro de Pesquisa Gonçalo Moniz
CRF	Forma recombinante circulante (C*irculant Reconbinant Form*)
CTL	Linfócito T citotoxico
CXCR4	Receptor Quimiocina da Família Cisteína-X-Cisteína 4 (*Cysteine-X-Cysteine Chemokine Receptor-4*)
D4T	Estavudine
DDI	Didanosine
DLV	Delavirdine
DNA	Ácido desoxirribonucléico (*Desoxyribonucleic Acid*)
dNTP	Deoxinucleosídeo-trifosfato

DRV	Darunavir
E	Envelope
EFV	Efavirenz
env	Gene do envelope do HIV
ETR	Etravirina
FIOCRUZ	Fundação Oswaldo Cruz
FPV	Fosamprenavir
FTC	Emtricitabine
G	Distribuição gama
gag	Gene antigeno de grupo do HIV
gp 41	Glicoproteína do envelope do HIV com peso de 41kd
gp120	Glicoproteína do envelope do HIV com peso de 120kd
gp160	Glicoproteína precursora do envelope do HIV com peso de 160kd
GTR	*General time reversible*
HAART	Terapia anti-retroviral de alta potência (*Highly Active Anti-retroviral Therapy*)
HIV	Vírus da imunodeficiência humana (*Human Immunodeficiency Virus*)
HIV-1	Vírus da imunodeficiência humana tipo 1 (*Human Immunodeficiency Virus-1*)
HIV-2	Vírus da imunodeficiência humana tipo 2 (*Human Immunodeficiency Virus-2*)
HLA	Antígenos Leucocitários Humanos (*Human Leukocyte Antigen***)**
HUPES	Hospital Professor Edgar Santos
I	Proporção de sítios invariáveis
IDV	Indinavir
IN	Enzima integrase
kb	Kilobase
LASP	Laboratório Avançado de Saúde Pública
LTR	Longas regiões terminais repetidas (*Long Terminal Repeat's*)
LPV	lopinavir
M	Membrana
MA	Matriz
MCMC	Cadeias markovianas de Monte Carlo (*Markov chain Monte Carlo)*

min	Minutos
mL	Mililitros
µl	microlitros
ML	Máxima Verossimilhança (*Maximum Likelihood*)
mM	Milimolar
mRNA	Ácido ribonucléico mensageiro (*Messenger Ribonucleic Acid*)
NC	Nucleocapsídio
Nef	Gene do fator negativo (*Negative Factor Gene*)
NFV	Nelfinavir
ng	Nanogramas
NJ	Agrupamento de vizinhos (*Neighbor-Joining*)
nm	nanometros
NNRTI	Inibidores da transcriptase reversa não nucleosidico (*Non-nucleoside reverse transcriptase inhibitors*)
NRTI	Inibidores da transcriptase reversa nucleosidico (*nucleoside reverse transcriptase inhibitors*)
NS	não-estruturais
NVP	Nevirapine
pb	Pares de base
PBMC	células mononucleares de sangue periférico (*Peripheral blood mononuclear cells)*
PCR	Reação em Cadeia da Polimerase (*Polymerase Chain Reaction*)
pH	Potencial de hidrogênio
PHI	Índice de homoplasia par-a-par (pair-wise homoplasy índex)
PI	Inibidores da protease
pmoles	Pico Moles
pol	Gene da polimerase
PR	Enzima protease
prM	pré-membrana
Rev	(*Anti-repression transactivator protein*)
RNA	Ácido ribonucléico (*Ribonucleic Acid*)

s	Segundos
SCUEAL	*Subtype Classification using Evolutionary Algorithms*
SQV	Saquinavir
Taq	Enzima DNA polimerase codificada pela bactéria termófilo *Thermus aquaticus*
Tat	Gene ativador da transcrição (*Trans-acting transcription transactivator*)
TDF	Tenofovir
TPV	Tipranavir
TR	Enzima transcriptase reversa
UPGMA	Agrupamento de pares não ponderados baseado na média aritmética (*Unweighted Pair Group Method with Arithmetic Mean)*
Vif	Gene da maturação e infectividade *(Virion Infectivity Factor)*
Vpr	Gene promotor do complexo pré-integração (*Viral Protein R*)
Vpu	Gene promotor da liberação de vírus do HIV-1 (*Viral Protein U*)
WNV	Vírus do Oeste do Nilo (*West Nile Vírus*)

SUMÁRIO

1. INTRODUÇÃO

1.1 Vírus da Imunodeficiência Humana

1.1.1 Descoberta do HIV

No início dos anos 80, foram relatados os primeiros casos da Síndrome da Imunodeficiência Adquirida (AIDS- *Acquired Immune Deficiency Syndrome*) nos Estados Unidos. Nesses casos, observou-se o aparecimento de Sarcoma de Kaposi, pneumonia por *Pneumocystis carinii* e comprometimento do sistema imune em indivíduos adultos, do sexo masculino e que apresentavam comportamento homossexual. Em 1983 foi então identificado e isolado o vírus da imunodeficiência humana (HIV- *Human Immunodeficiency Virus*) como o agente etiológico da AIDS (BARRE-SINOUSI *et al.*, 1983).

O HIV-1 (vírus da imunodeficiência humana do tipo 1) é um retrovírus que infecta principalmente linfócitos T CD4+. A principal manifestação característica desta infecção é a queda no número desses linfócitos, levando a uma deficiência múltipla do sistema imune, deixando assim o organismo susceptível a infecções oportunistas por outros patógenos.

1.1.2 Epidemiologia do HIV

Atualmente, existe aproximadamente 33,3 milhões de pessoas vivendo com o HIV no mundo, nas quais dois terços (22,5 milhões) vivem na África sub-Saariana. Dos 30,8 milhões de adultos vivendo com o vírus, 15,9 milhões são mulheres, e entre as crianças e jovens infectados, 2,5 milhões estão abaixo de 15 anos de idade. Dados mais recentes demonstram que apenas no ano de 2009 ocorreram 2,6 milhões de novas infecções e 2,8 milhões de mortes pelo vírus (UNAIDS, 2010). No Brasil, estima-se que 630 mil pessoas vivam com o vírus (MINISTÉRIO

DA SAÚDE, 2010), sendo que a maior parte concentra-se nas regiões sul e sudeste do país. Apesar destes dados alarmantes, e segundo a Organização Mundial da Saúde, o Brasil mantém a sua posição de país com a epidemia controlada.

A infecção pelo HIV-1 no Brasil iniciou-se entre homens homossexuais, depois passando para usuários de drogas endovenosas, e então para a população geral, aumentando significativamente o número de mulheres infectadas no país (DOURADO *et al*., 2007). Este aumento da proporção da infecção entre as mulheres provavelmente está atribuída ao comportamento dos seus parceiros sexuais masculinos (SILVA; BARONE, 2006).

1.1.3 Transmissão vertical do HIV-1

A transmissão vertical do HIV-1 ocorre quando a mãe infectada pelo vírus transmite para seu filho durante a gestação, no parto ou através da amamentação. A taxa de transmissão vertical pode chegar a 20%, porém com medidas de prevenção ela pode chegar menos de 1%. Essa medidas inclui o diagnóstico precoce da gestante, uso de drogas antirretrovirais durante a gravidez e no recém-nascido, parto cesariano programado e substituição da amamentação da criança com o leite materno por leite artificial (MINISTÉRIO DA SAÚDE, 2010).

Uma mulher grávida infectada pelo HIV-1 tem de 5 a 10% de chance de infectar o seu bebe *in útero*, de 10 a 20% durante o parto e 10 a 20% na amamentação (LEHMAN *et al*. 2007). Sabe-se que a cepa do vírus transmitido da mãe para o filho não é necessariamente o mais abundante na mãe (WOLINSKY *et al.*, 1992). Desta forma, o isolado do HIV-1 prevalente no filho pode ter uma evolução diferente daquele encontrado na mãe. Essa diferença também se dá pelas diferentes condições de cada indivíduo, como a resposta imune de cada um. O vírus pode ser transmitido na forma de vírus livre ou associado a uma célula e, dependendo do tipo de transmissão vertical, uma dessas formas terá uma participação maior, podendo assim, também variar a dinâmica da evolução do vírus no indivíduo. Na transmissão *in útero* e pela amamentação, o vírus transmitido, em sua maior parte, é associado a uma célula e, durante o parto, é vírus livre (LEHMAN *et al*., 2007).

1.1.4 Epidemiologia Molecular do HIV

O HIV é um Lentivírus pertencente à família *Retroviridae* e em sua classificação filogenética apresenta-se dividido em dois tipos, o HIV-1 e o HIV-2. O HIV-1 é dividido em 3 grupos: M ("major", principal), O ("outlier", atípico) e N ("new", novo, não-M e não-O). O grupo M é responsável pela pandemia mundial e apresenta 9 subtipos (subtipos A, B, C, D, F, G, H, J e K (MCCUTCHAN, 2000)), além das suas, aproximadamente, 49 formas recombinantes circulantes (CRF) descritas (LOS ALAMOS), mostrando, assim, a alta diversidade genética do vírus.

Os subtipos do HIV-1 apresentam uma distribuição diferente pelas diversas regiões mundiais. O subtipo C, que é o mais encontrado na África Sub-Saariana, é responsável pelo maior número de infecções no mundo seguido pelos subtipos A e B. O subtipo B, mesmo não sendo o responsável pelo maior número das infecções, atualmente, é o mais estudado no mundo, pois é o mais encontrado na América do Norte, Europa e Austrália (Figura 1.4.1).

No Brasil, o subtipo predominante é o B, seguido dos subtipos F (MORGADO *et al*, 1994) e C (SOARES *et al*, 2003), e, em menor freqüência, os subtipos D (MORGADO *et al*, 1998) e A (CARIDE *et al.*, 2001), além de formas recombinantes B/F e B/C (SABINO *et al*., 1994; GUIMARÃES *et al*., 2001) (Figura 1.4.1). Apesar da grande parte das cepas virais encontradas no Brasil hoje ainda serem do subtipo B, tem se observado um grande crescimento do subtipo C, mostrando a importância de se ter um controle epidemiológico.

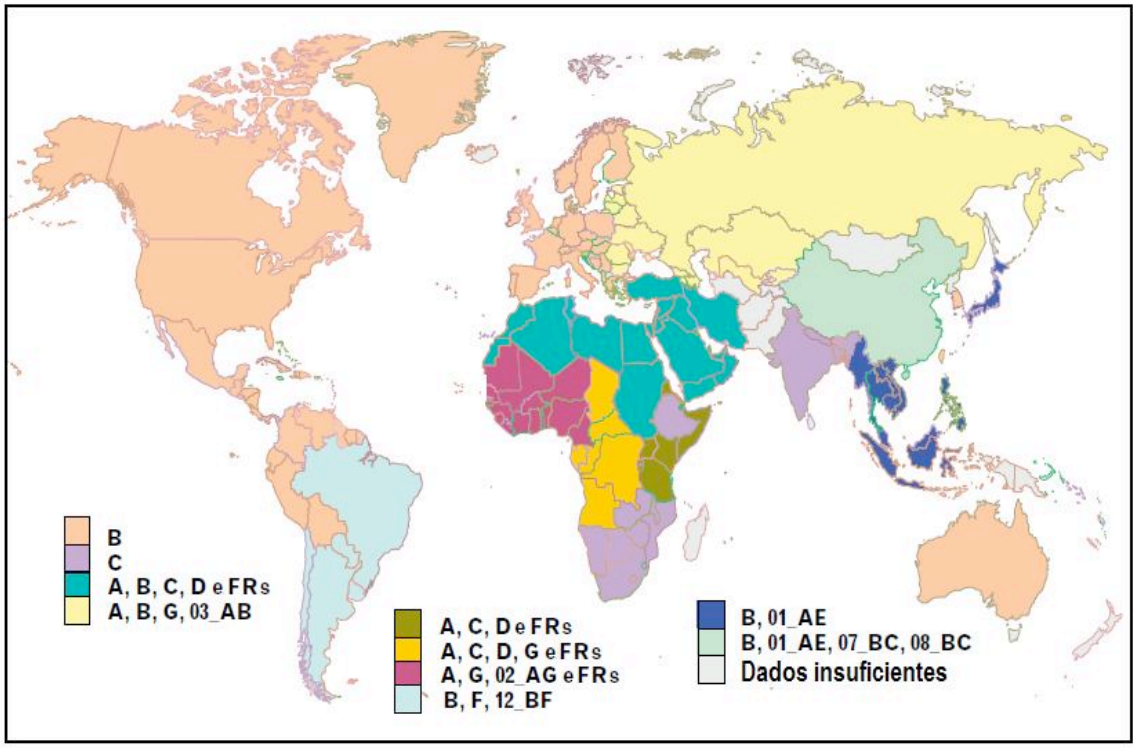


Figura 1.1.4.1: Distribuição dos subtipos do HIV-1 no mundo. Adaptado de Woodman e Williamson, 2009.

1.1.5 Estrutura do HIV

A estrutura do HIV-1 possui um formato icosaédrico com 110nm de diâmetro. Seu capsídio viral é coberto por uma camada lipoprotéica (envelope) derivada da membrana do hospedeiro. Este envelope apresenta duas glicoproteínas: a glicoproteína 120 (gp120), que é de superfície e a glicoproteína 41(gp41), que é transmembranar (Figura 1.5.1) (CHAN, 1998).

A matriz viral localizada abaixo do envelope, é formada pela proteína p17 (MA). No centro está o capsídio em forma de cone que é formado pela proteína p24 (CA). Já o nucleocapsídio é formado pelas proteínas p7 e p9 (NC) onde se localizam as duas fitas simples de RNA, material genético do vírus (Figura 1.5.1).

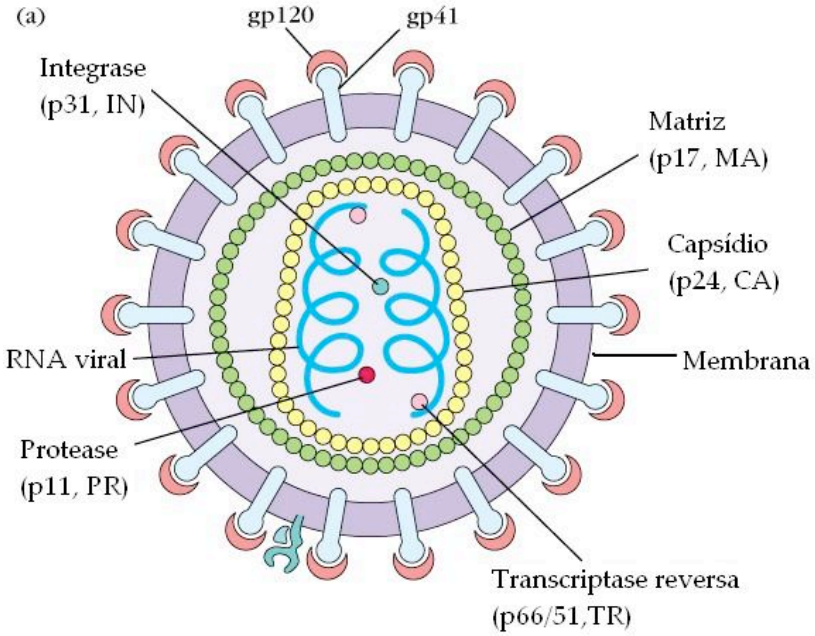


Figura 1.1.5.1: Desenho esquemático da estrutura morfológica do HIV-1. Adaptado de http://tutor.lscf.ucsb.edu/instdev/sears/immunology/images/figure19-08a.jpg

No interior do capsídio do HIV-1 também estão localizadas as enzimas transcriptase reversa (p66/51, TR), integrase (p31, IN) e protease (p11, PR) que estão envolvidas no processo de replicação, integração do genoma viral e maturação respectivamente (Figura 4.1.1). As chamadas proteínas acessórias VIF, VPR e NEF também são encontradas no capsídio.

1.1.6 Estrutura genética do HIV-1

O genoma do HIV-1 é formado por duas fitas simples de RNA com um comprimento de aproximadamente 9,5 kb. Este genoma é composto por 9 genes que são flanqueados por duas regiões com sequências repetitivas denominadas de longas regiões terminais repetidas (LTR) (Figura 1.6.1). Estas regiões LTR são necessárias para a integração do genoma viral ao do

hospedeiro e é onde estão localizados os sítios responsáveis pela transcrição gênica das proteínas virais (GREENE, 2002).

O HIV-1 apresenta 3 genes estruturais: *gag* (gene antígeno de grupo), *pol* (polimerase) e *env* (envelope). O gene *env* codifica a glicoproteína 160 (gp160) que sofre o processo de clivagem gerando a gp120 e a gp41. Essas proteínas estão localizadas no envelope viral e são responsáveis pelo reconhecimento e fusão da célula alvo (CHAN, 1998). O gene *pol* irá codificar as enzimas transcriptase reversa (TR), integrase (IN) e protease (PR). A TR é responsável pela transcrição reversa do RNA viral em partícula de cDNA, a IN irá integrar o cDNA ao genoma da célula hospedeira e a PR irá participar do processo de maturação da partícula viral. O gene *gag* codifica a proteína precursora p55 que sofrerá processo de clivagem dando origem às proteínas da matriz (p24), do capsídio (p17) e do nucleocapsídio viral (p9 e p7).

Os genes reguladores são *tat* e *rev*. A proteína TAT é transativadora da transcrição viral tendo um papel importante na replicação do HIV. Já a proteína REV é responsável pela exportação de mRNAs viral para o citoplasma da célula (HOPE, 1997).

Os genes acessórios do HIV-1 são *vif, vrp*, *vpu* e *nef*. A proteína VIF está relacionada com o transporte de componentes virais. A VRP participa do processo de transporte do cDNA para ser integrado ao genoma da célula alvo. A VPU atua no brotamento da partícula viral. Já a proteína NEF está relacionada com processos de escape viral, pois reduz a expressão de CD4 e das moléculas de HLA de classe I e II.

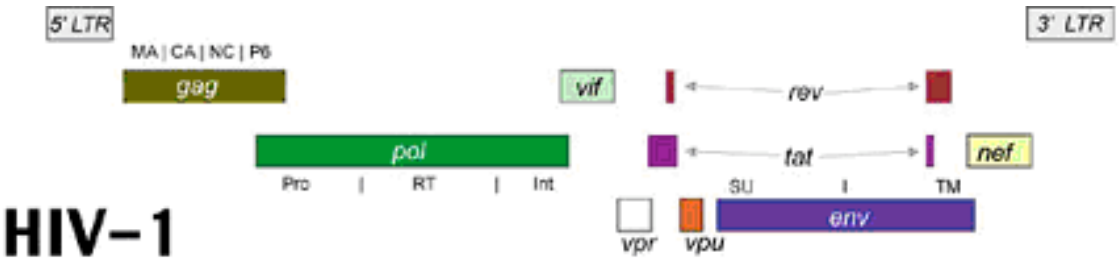


Figura 1.1.6.1: Desenho representando o genoma do HIV-1. Adaptado de: www.aids.harvard.edu/research/discoveries.html.

1.1.7 Ciclo de Replicação do HIV-1

Para o HIV-1 entrar na célula-alvo ele precisa reconhecer o receptor CD4 e os co-receptores CCR5 e CXCR4 (UGOLINI *et al*., 1999). A proteína de superfície do envelope viral gp120 reconhece e se liga ao receptor CD4 (KEDZIERSKA *et al*., 2003). Esta ligação provoca uma mudança conformacional permitindo a ligação da gp120 com o co-receptor CCR5 ou CXCR4 resultando em uma segunda mudança conformacional (UGOLINI *et al*., 1999). Esta segunda mudança na conformação irá expor a proteína transmembranar do envelope viral, gp41, permitindo que esta se ligue à membrana plasmática da célula-alvo, promovendo a fusão da membrana plasmática com o envelope viral e permitindo a entrada do capsídio viral.

Após a entrada do capsídio viral, ocorrerá decapsidação liberando o material genético e as enzimas virais no citoplasma da célula. A enzima transcriptase reversa começa a atuar sintetizando uma molécula de cDNA dupla fita a partir do RNA viral. Juntamente com a enzima integrase, o cDNA viral será transportado para o núcleo da célula hospedeira, onde a integrase irá atuar integrando o cDNA ao DNA da celular.

Utilizando o maquinário da célula-alvo, o cDNA do vírus será transcrito, levado para o citoplasma da célula, onde serão então sintetizadas as proteínas virais. A enzima protease irá clivar as proteínas precursoras virais iniciando o processo de maturação. Essas proteínas clivadas migrarão para os sítios de maturação próximos à membrana plasmática, juntamente com o RNA genômico, onde as novas partículas virais sairão por brotamento. A figura 1.7.1 ilustra todo o ciclo de replicação.

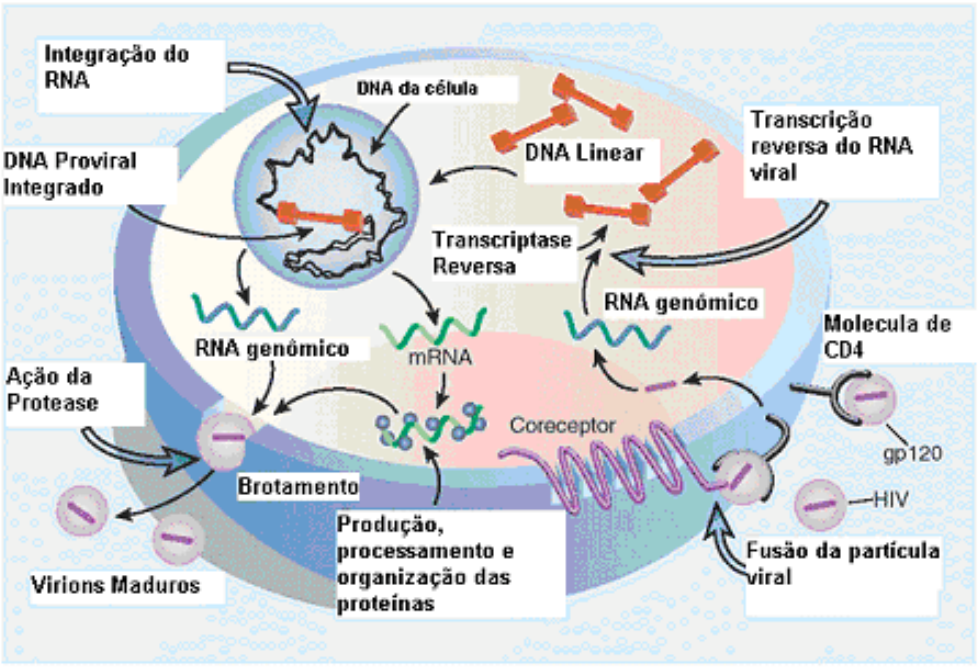


Figura 1.1.7.1: Figura esquemática resumida do ciclo de replicação do HIV-1. Adaptado de: http://www.biology.arizona.edu/immunology/tutorials/AIDS/graphics/hiv_biology.gif

1.1.8 Aspectos clínicos e imunológicas do HIV-1

A transmissão ocorre a partir do contato com o sangue, sêmen, leite materno e outras secreções de indivíduos HIV infectados. Durante o início da infecção (2 a 6 semanas), ocorre um pico da replicação viral, com consequente aumento da carga viral, um aumento de linfócitos T CD8+ no sangue periférico (FAUCI, 1993) e uma pequena queda de linfócitos T CD4+ (PANTALEO; FAUCI, 1996). Durante esta infecção primária, cerca de 50% dos pacientes infectados apresentam sintomas semelhantes aos de uma gripe que dura uma ou mais semanas (COHEN *et al*., 1997). Durante esta fase não é possível diagnosticar a infecção com o uso de testes para detecção de anticorpos, pois os seus níveis são muito baixos e não detectados nos testes de triagem atuais (FAUCI, 1993).

Após a infecção primária, inicia-se a fase de latência clínica. Nesta fase, o número de células T CD4+ volta ao normal e vai diminuindo gradativamente ao longo deste período. Esta

fase também se caracteriza pela alta de anticorpos específicos contra o HIV, grande aumento de células T CD8+ (CTL) contra o vírus, o que irá controlar a replicação viral e, consequentemente, há uma queda da carga viral. Esta fase dura entre 2 a 10 anos dependendo do paciente (COHEN *et al*.,1997).

Quando os níveis de células T CD4+ chegam a valores inferiores a 200 células/μL, inicia-se a fase sintomática, caracterizada pela Síndrome da Imunodeficiência Adquirida (AIDS) (COHEN *et al*., 1999). Durante este período, além da queda de células T CD4+, também ocorre a queda de células T CD8+ e anticorpos neutralizantes, e um aumento da carga viral do indivíduo. Nesta fase, ocorre o aparecimento de infecções oportunistas, neoplasias secundárias e manifestações neurológicas (COFFIN *et al*., 1995; KAHN *et al*., 1998). A figura 1.8.1 representa dinâmica imunológica da infecção.

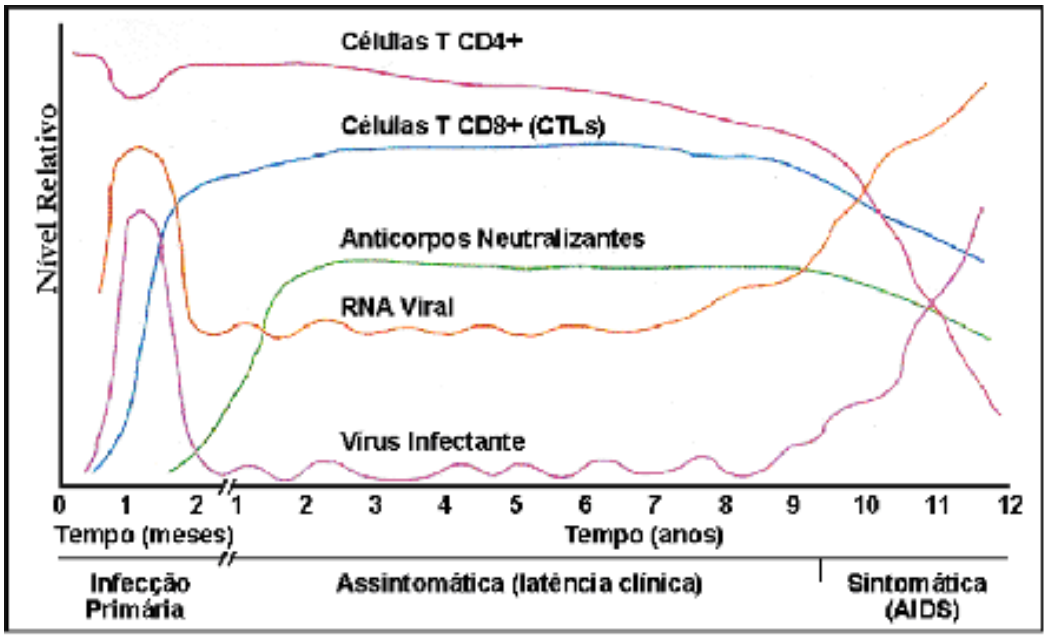


Figura 1.1.8.1: História natural da Infecção pelo HIV. Adaptado de Poignard *et al*., 1996.

1.1.9 Terapia com antirretrovirais

A primeira droga terapêutica contra o HIV surgiu em 1987. Conhecida como zidovudina, esta droga é um análogo de nucleosídio que age inibindo a atividade da enzima transcriptase reversa impedindo a replicação viral (YARCHOAN *et al.,* 1986). Subsequentemente, muitas outras drogas que agem sobre a transcriptase reversa, protease e integrase começaram a surgir no intuito de controlar a carga viral e aumentar a sobrevida do paciente. Em 1996, surgiu a chamada terapia antirretroviral de alta potência ("Highly Active Anti-Retroviral Therapy"- HAART) ou "Coquetel Anti-AIDS", que se baseia na combinação de três drogas de classes diferentes. (PERELSON *et al*., 1996; SEPKOWITZ, 2001; PALELLA *et al*, 1998).

Com o uso da HAART não foi possível eliminar completamente o vírus. Porém, foi possível reduzir a carga viral, reconstruir o sistema imune do paciente, retardando assim, a progressão para AIDS, além de melhorar a qualidade de vida dos portadores do vírus (PERELSON *et al*., 1996; SEPKOWITZ, 2001; PALELLA *et al*, 1998).

Existem três tipos mais importantes de antirretrovirais, que compõem o coquetel de tratamento: os inibidores de protease (PI), os inibidores da transcriptase reversa nucleosídico (NRTI) e os inibidores da transcriptase reversa não nucleosídico (NNRTI). Os inibidores da protease são drogas que se ligam ao sítio ativo da enzima, ou mimetizam o estado de transição durante a clivagem do peptídeo, ou, ainda, agem como um complemento simétrico ao sítio ativo da enzima, inibindo a ação da protease, impedindo a maturação da partícula viral. Os PIs em uso hoje são atazanavir (ATV), darunavir (DRV), fosamprenavir (FPV), indinavir (IDV), lopinavir (LPV), nelfinavir (NFV), saquinavir (SQV) e tipranavir (TPV). Os inibidores da transcriptase reversa análogos de nucleosídios são substâncias análogas de nucleosídios que não apresentam o grupo hidroxila no carbono 3'. Devido a esta modificação, durante a transcrição reversa o nucleosídio modificado irá se ligar à nova fita que está sendo sintetizada, impedindo que a enzima transcriptase reversa continue a síntese do cDNA e inibindo, assim, a replicação viral. Os NRTIs em uso clínico hoje são lamivudina (3TC), abacavir (ABC), zidovudina (AZT), estavudina (D4T), didanosina (DDI), emtricitabina (FTC) e tenofovir (TDF). Já os inibidores da transcriptase reversa não nucleosídico são drogas que agem se ligando reversivelmente a um sítio próximo ao sitio ativo da enzima transcriptase reversa, inibindo a sua ação, impedindo, assim, a

replicação do HIV-1. Essas drogas tem ação específica na transcriptase reversa do HIV-1. As drogas da classe dos NNTRIs em uso clínico são: delavirdina (DLV), efavirenz (EFV), etravirina (ETR) e nevirapina (NVP).

Apesar desta terapia estar contribuindo para que os pacientes mantenham uma baixa carga viral e números mais altos de linfócitos T CD4+, retardando a progressão para AIDS, muitos pacientes apresentam falha terapêutica. Esta falha pode ser em decorrência da não adesão do paciente a terapia, a dificuldade do paciente a ter acesso ao medicamento e ao acompanhamento médico. No entanto, a falha terapêutica ocorre principalmente devido a mutações no gene *pol* nas regiões que codificam as enzimas-alvo dos antirretrovirais (protease e transcriptase reversa) que levam à resistência (COFFIN, 1995). A identificação destas mutações pode ajudar a definir qual o melhor e mais eficaz tratamento para o paciente.

A falha na terapia pode ser detectada pelo aumento da carga viral, pela queda da contagem de linfócitos T CD4+ ou pela progressão clínica. No Brasil, ao ser detectado esta falha na terapia a múltiplos esquemas terapêuticos é realizada a detecção da resistência pela genotipagem por sequenciamento, onde é possível detectar as mutações que conferem resistência tanto na enzima transcriptase reversa quanto na protease.

1.2 Vírus do Oeste do Nilo

O vírus do Oeste do Nilo (*West Nile Vírus* – WNV) pertence a família *Flaviviridae* do gênero *Flavivirus*. Apresenta um tamanho de aproximadamente 50nm e seu material genético é composto de RNA de fita simples sentido positivo de 11 a 12 mil pares de base. Seu genoma é composto por quatro gene estruturais: nucleocapsideo (NC), pré-membrana (prM), membrana (M) e envelope (E) e sete não-estruturais (NS1, NS2A, NS2B, NS3, NS4A, NS4B e NS5).

O WNV foi primeiramente isolado em 1937 na Uganda, região do oeste do rio Nilo (SMITHBURN *et al*., 1940). Após isto foi encontrado em diversas regiões da África. Em 1957 ocorreram um surto em Israel. No início dos anos 60 foi encontrado em cavalos na França, se espalhando assim em seguida pela Europa (MURGUE *et al*., 2002). Somente em 1999 o WNV foi encontrado no continente americano primeiramente em um surto na cidade de Nova Iorque e

região (LANCIOTTI *et al*., 1999). A partir dai o vírus se espalhou por todo os Estados Unidos, Canadá, México, America Central e Caribe.

Sua transmissão ocorre principalmente pela picada de mosquitos do gênero *Culex* infectados. Estes se infectam ao picar pássaros infectados, que são os reservatórios naturais do vírus, transmitindo assim para o homem e outros animais. O vírus é transmitido em períodos quentes em lugares de clima tropical e temperado (CDC). A transmissão também pode ocorrer, em poucos casos, através de transfusão sanguínea, transplantes de órgãos e de mãe para filho (CDC).

1.2.1 Epidemiologia Molecular do WNV

A reconstrução da história evolutiva do WNV utilizando análises filogenéticas foram feitas em diversos estudos sendo estes agrupados em dois principais linhagens genéticas. A linhagem 1 contém isolados da Europa, Estados Unidos, Israel, Índia, Rússia e Austrália. Já a linhagem 2 contém isolados da África subsaariana e Madagascar (BERTHET *et al*., 1997; LANCIOTTI *et al*., 2002).

A linhagem 1 é subdividida em três sub-clados: 1a que inclui isolados da África, Europa, Estados Unidos, Oriente Médio e Rússia; 1b que consiste de isolados Kunjin da Austrália; e 1c com isolados da Índia (LANCIOTTI *et al*., 2002; SCHERRET *et al*., 2001).

Dentro dos Estados Unidos, dois clados tem sido descritos: um contendo isolados do surto de 1999 a 2000 ocorrido no nordeste do país, e o segundo com isolados de 2002 até o presente que estão distribuídos por todo o país (LANCIOTTI *et al*., 2002).

1.2.2 Sintomatologia e Tratamento do WNV

Após serem picadas pelo mosquito infectado pelo WNV os indivíduos levam de três a 14 dias para desenvolverem alguma sintomatologia. Cerca de 80% dos indivíduos não desenvolvem

sintomatologia. Porém, aproximadamente 20% apresenta sintomatologia moderada que inclui febre, dor de cabeça, dores no corpo, náusea, vômitos, e às vezes aumento dos gânglios linfáticos ou erupção cutânea no tórax, barriga e dorso; e alguns caso pode desenvolver sintomas graves apresentando febre alta, dor de cabeça, rigidez do pescoço, torpor, desorientação, coma, tremores, convulsões, fraqueza muscular, perda de visão, entorpecimento e paralisia (CDC).

Não existe tratamento especifico contra o WNV. Nos casos moderados a infecção cura por si só, porém é necessário um tratamento de suporte com fluidos intravenosos e auxílio na respiração para os casos mais graves de doença (CDC).

1.3 Bioinformática

A bioinformática é a ciência que utiliza a informática, a estatística e a matemática na biologia molecular. O termo "Bioinformática" foi primeiramente usado por Pauline Hogeweg em 1979 para estudos de processos de informática em estudos de biologia sistemacional. Desde então o seu principal uso tem sido nos ramos da genética e da genômica em especial para auxiliar no manejo da grande quantidade de dados gerado no sequenciamento de DNA, RNA e aminoácidos.

Hoje, a bioinformática tem sido utilizada em diversas áreas como a construção de banco de dados e a mineração de dados; análises de sequências; para identificar gene, predizer sua funções e demonstrar relações entre genes e proteínas; prever a conformação tridimensional das proteínas; construir árvores filogenéticas e modelos evolutivos; construir bibliotecas genômicas; estudar as funções biológicas; *design* de drogas entre muitas outras.

1.3.1 Análise filogenética

A história evolutiva entre espécies e gene pode ser representada por uma filogenia. A árvore filogenética é um diagrama que representa essas relações e é assim chamado pela sua similaridade com a estrutura de uma árvore. Os nós externos representam os táxons existentes e

os nós internos o hipotético ancestral comum de um conjunto de táxons. A extensão dos ramos representa o número de substituições por sítio ou uma estimativa de tempo de divergência de um táxon para o outro. Os táxons com o mesmo ancestral são chamados de grupo monofilético. O padrão de ramos de uma árvore filogenética representa a relação evolutiva entre os táxons e esse padrão é chamado de topologia. A topologia será diferente para cada conjunto de dados analisados dependendo do seu grau de similaridade (VANDAMME, 2009).

Obter uma reconstrução filogenética que represente a relação verdadeira entre os táxons estudados é algo muito difícil de se atingir. Existem muito modelos evolutivos e métodos de construção da filogenia disponível e a escolha do melhor para os seu dados não é uma garantia de que obteve a árvore verdadeira. Diante disto, o que se busca é chegar o mais perto possível do que seria a verdadeira construção filogenética. Os métodos mais utilizados hoje são os que se baseiam em distancia: Agrupamento de pares não ponderados baseado na média aritmética (UPGMA, do inglês: *Unweighted Pair Group Method with Arithmetic Mean*) e agrupamento de vizinhos (NJ, tradução de *neighbor-joining)*; análises de estado de caráter: máxima parcimônia, máxima verossimilhança (ML, do inglês: *Maximal likelihood*); e inferência Bayesiana com análises de cadeias markovianas de Monte Carlo (MCMC do inglês, Markov chain Monte Carlo). Além disto todos estes métodos dependem de um modelo matemático que representa como a evolução ocorreu a partir do alinhamento de sequências de nucleotídeo ou aminoácido.

1.3.2 Filodinâmica

A filodinâmica é o uso de análises filogenéticas em estudos da dinâmica de microorganismos na combinação de processos evolutivos e ambientais. Estudos da dinâmica de populações irão observar as mudanças que ocorrem na população ao longo do tempo e a fixação dessas mutações (PYBUS e RAMBAUT, 2009).

Quando mutações que ocorrem em um gene são passados para a geração seguinte coexistindo com o seu gene original tem como resultado um polimorfismo. Em um sítio polimórfico, dois ou mais variantes de um gene circulam na população simultaneamente. A dinâmica da frequência dessas mutações na população pode mudar ao longo do tempo. Quando

uma mutação surge, e sua frequência aumenta até 100% essa mutação se torna fixa na população. A taxa de fixação é o numero de mutações novas pelo tempo que se torna fixa. Já a taxa de mutação é o número de mudança de nucleotídeo, ou aminoácido, por sítio por ciclo de replicação. A razão com que uma mutação se torna fixa em uma população depende do tamanho da população efetiva, além de eventos ambientais como migração ou catástrofes que pode gerar o efeito do gargalo de garrafa (VANDAMME, 2009).

Os vírus são microorganismos ideais para estudos de filodinâmica por sofrerem mutações e se adaptarem ao seu ambiente em um período de tempo mais rápido. Nestes estudos a relação do surgimento de novas mutação e sua fixação poderá fornecer informação sobre o vírus e o ambiente dessas populações virais. Este ambiente envolve genética, sistema imune, forma de transmissão, tratamento, dentre outros fatores que pode contribuir para a adaptação ou não do vírus mutado (PYBUS e RAMBAUT, 2009).

1.3.3 Filogeografia

Filogeografia é o estudo da relação entre as populações e sua localização na Terra. Estes estudos auxiliam a investigações de eventos históricos e geográficos e seus efeitos na distribuição de genes e microorganismos, descrevendo sinais genéticos geograficamente estruturados dentro e entre espécies (AVISE, 2000).

O uso de ferramentas filogenéticas tem sido amplamente utilizada para traçar a rota e história de doenças virais como dengue, raiva, *influenza* e HIV (HOLMES, 2004). Os vírus, principalmente de RNA, apresentam uma alta taxa de mutação em um período de tempo relativamente curto sendo assim ideais para estes estudos (HOLMES, 2004). Através destas análises é possível correlacionar os processos históricos e geográficos com migrações, guerras, e catástrofes, entendendo melhor à migração do vírus entre populações e prever possíveis rotas de vírus, contribuindo assim para a prevenção e controle de doenças virais.

2. JUSTIFICATIVA

A bioinformática é uma ciência que disponhe de diversas ferramentas que podem auxiliar em diferentes estudos de microorganismos como os vírus. Diversos estudos tem utilizado essas ferramentas para um melhore entendimento do genoma, evolução, migração entre regions, além de outras fatores virais importantes. Devido a isto, essas ferramentas são de grande auxilio em estudos para a melhor caracterização da diversidade, epidemiologia molecular, dinâmica populacional e determinação das relações temporal e geográfica dos vírus.

Inúmeros trabalhos realizados em todo mundo, buscam uma vacina e terapias mais eficazes contra HIV-1. Baseando-se na literatura, sabe-se que este vírus é muito variável e com uma distribuição mundial diferente e específica para cada região geográfica. Devido a isto, são de extrema importância estudos para mapear as características moleculares dos isolados do HIV-1 que circulam em cada região. A caracterização das cepas virais circulantes no Brasil poderá contribuir para o desenho de uma vacina, bem como o entendimento do comportamento das cepas de nosso país. Informações de grande valor a respeito da diversidade e variabilidade viral das sequências brasileiras podem ser obtidas através de análises utilizando as ferramentas da bioinformática. Isto, em conjunto com o cruzamento de sequências consensos do genoma viral e com as caracterizações biológicas e fenotípicas das variantes do HIV-1 (subtipos B, C, F e formas recombinantes) facilitarão, significantemente, o desenho de uma vacina.

A avaliação da dinâmica evolucionária dos isolados do HIV-1 poderá contribuir para o conhecimento do *fitness* viral. Reconstruções filogenéticas podem revelar a dinâmica da evolução viral intra e inter hospedeiros. Estas investigações inter-hospedeiros permitem avaliar o movimento das cepas do HIV-1 entre diferentes locais e o curso da transmissão, além de estimar mudanças no tamanho da população viral efetiva, ao longo do tempo. Os estudos da resposta imune, pressão seletiva e taxa de replicação viral contribuirá para o conhecimento da dinâmica e adaptação viral (GRENFELL *et al*., 2004; LEMEY *et al*., 2006). As regiões variáveis do envelope viral são alvos para a resposta imune celular e humoral, e àqueles indivíduos infectados com a mais alta resposta serão capazes de controlar o vírus por um maior tempo. O principal mecanismo de escape viral e a grande quantidade de substituições sinônimas, quando comparadas com as substituições não-sinônimas (pressão seletiva positiva), no gene *env* do HIV-1, contribuindo para a alta diversidade desta região em diferentes tempos num hospedeiro

(WILLIAMSON *et al*., 2003). Considerando que o tempo e mecanismo de infecção são conhecidos num estudo de transmissão vertical, utilizando as análises e ferramentas ideais é possível inferir sobre a dinâmica evolucionária do HIV-1 intra e inter pacientes, em tempos diferentes.

Inúmeros estudos reconstruíram a história evolutiva do Vírus do Oeste do Nilo (WNV) utilizando filogenia (BRETHET *et al*., 1997; LANCIOTTI *et al*., 2002). Porém, tem se encontrado dificuldades para determinar essa história nas amostras mais recentes encontradas em diversas regiões dos Estados Unidos (BERTOLOTTI *et al*., 2007; TANG *et al*., 2008). Observa-se também que não há estudos que indique a melhor região a ser utilizada em análises filogenéticas e para conseguir estudos melhores tem sido necessário o sequenciamento do genoma total (DAVIS *et al*., 2005; PARREIRA *et al*., 2007). Para isso, é necessário um estudo que determine qual a região do genoma mais indicada para se analisar reduzindo assim o custo e tempo desses estudos. Além disto, com o uso da região e ferramentas corretas será possível conhecer mais sobre as relações temporais e geográficas deste vírus.

3. OBJETIVOS

3.1. Objetivo Geral

Demonstrar a aplicação de ferramentas de bioinformática em estudos de epidemiologia molecular do HIV-1, melhor entendimento das dinâmica das populações virais ao longo do tempo e das suas relações temporais e geográficas.

3.2. Objetivos Específicos

3.2.1. Caracterização Molecular do Gene *pol* do HIV-1 de Indivíduos Infectados de Salvador, Bahia, Brasil.

- Subtipar os isolados de HIV-1 em pacientes do Hospital Professor Edgar Santos, Salvador, Bahia, através da análise do gene *pol;*
- Verificar a presença de recombinações intersubtipos e os padrões dessas recombinações;
- Verificar a presença de mutações, no gene *pol* do HIV-1, associadas à resistência aos antirretrovirais.

3.2.2. Avaliação filodinâmica de isolados do HIV-1 na transmissão materno-fetal.

- Avaliar as diferenças evolutivas (estudo filodinâmico) no gene *env* de diversos clones do HIV-1, em três tempos diferentes, provenientes do PBMCs, de mães e filhos infectados;
- Avaliar essas diferenças evolutivas nos subtipos B, provenientes de Feria de Santana, e C

da Zâmbia, como também em mães com infecção aguda e crônica.

3.2.3. Caracterização evolutiva do genoma total do Vírus do Oeste do Nilo.

- Avaliar os padrões evolutivos de cada gene a fim de determinar qual região apresenta máximo poder interpretativo para inferir relações temporais e geográficas entre as cepas do WNV.

4. MATERIAL E MÉTODOS

4.1 Caracterização Molecular do Gene *pol* do HIV-1 de Indivíduos Infectados de Salvador, Bahia, Brasil

4.1.1 População de Estudo

Neste trabalho foram analisadas 57 amostras de pacientes infectados pelo HIV-1 atendidos no ambulatório de retrovírus do Hospital Universitário Professor Edgard Santos (HUPES) da cidade de Salvador, Bahia. Foram coletados 10 mL de sangue e realizada uma entrevista para a obtenção de dados étnicos, socioeconômicos e da história da infecção destes indivíduos. Os prontuários também foram analisados para a obtenção dos dados clínicos, laboratoriais e de tratamento dos pacientes. A realização da coleta de dados e da amostra do paciente só foi executada após a assinatura do Termo de Consentimento Livre e Esclarecido.

4.1.2 Aspectos Éticos

Essas amostras são provenientes do projeto "Caracterização étnica/geográfica da população de Salvador e de portadores do HIV-1 e a correlação entre o índice de ancestralidade africana e vulnerabilidade a HIV/AIDS", que tem a aprovação do Comitê de Ética em Pesquisa do Centro de Pesquisas Gonçalo Moniz da Fundação Oswaldo Cruz, de parecer Nº 84/2006 (Anexo A).

4.1.3 Procedimentos Experimentais

Extração de DNA

Com a obtenção das amostras, o DNA foi extraído a partir do sangue total utilizando o kit QIAamp DNA (QIAamp DNA minikit, Hilden, Düsseldorf, Alemanha). Este método utiliza primeiramente a proteinase K que irá realizar a digestão das proteínas, eliminando completamente a atividade de enzimas como RNAse e DNAse. Depois, a amostra é incubada com um tampão de lise e centrifugada em uma coluna contendo uma membrana sílica. As condições do tampão de lise permitem que ocorra adsorção da molécula de DNA à membrana sílica. Então são utilizados dois tampões de lavagem que removem qualquer resíduo contaminante, sem afetar a ligação do DNA a membrana. As condições de salinidade e pH do lisado asseguram que proteínas e outros contaminantes que podem inibir a reação da PCR não fiquem retidos na membrana. Para finalizar o método é utilizada água bidestilada a 37°C para que o DNA aderido na coluna passe para o tubo onde será armazenado o DNA extraído.

Reação em Cadeia de Polimerase (PCR)

O fragmento do gene *pol* estudado apresenta aproximadamente 1000 pb, correspondendo aos nucleotídeos de 2253 a 3260 do isolado referência HXB2. Esta região foi amplificada pela utilização da *nested* PCR (PCR aninhado) com *primers* específicos (Tabela 4.1.4). A *nested* PCR ocorre realizando duas reações sucessivas de PCR com dois pares de *primers* diferentes. A primeira reação ou primeiro *round* ocorre utilizando o par de *primers* mais externo. Já a segunda reação ou segundo *round* ocorre utilizando o produto do primeiro *round* e o par de *primers* mais internos.

O protocolo utilizado para os primeiro e segundo *rounds* foi: 5,0μl de tampão de reação 10X; 2,5 μl de $MgCl_2$ 50 mM; 12 μl de dNTP 1,25 mM; 0,5 μl de cada um dos *primers* direto e reverso à 20,0 pmoles/μl; 0,3 μl de *Taq polimerase* (LGC); 5 μl (100 ng/μl) de DNA e H_2O bidestilada estéril na quantidade suficiente para 50 μl final. As condições de amplificação utilizadas no termociclador (*Applied Biosystens*) foram: 3 ciclos de 95°C por 3min, 55°C por 1min, 27°C por 1min; 35 ciclos de 95°C por 1min, 55°C por 45s e 72°C por 1min; 72°C por 10min.

Os produtos da PCR foram analisados em gel de agarose a 1%, com brometo de etídeo e visualizado em luz ultravioleta.

Purificação e Sequenciamento

O produto da PCR foi purificado utilizando o kit QIAquick PCR (QIAquick Gel Extraction Kit, Hilden, Düsseldorf, Alemanha), conforme instruções do fabricante. Os produtos amplificados nas PCR e purificados foram sequenciados no sequenciador automático ABI3100 utilizando o kit *Taq* FS *Dye terminator cycle sequencing kit* (APPLIED BIOSYSTEMS). Nas reações de sequenciamento utilizamos os *primers* internos da PCR e *primers* específicos do sequenciamento (Tabela 4.1.3).

As sequências geradas foram analisadas no programa *SeqScape* (APPLIED BIOSYSTEMS) para definição das sequências-consenso de cada amostra. Devido à extrema sensibilidade das técnicas de amplificação de sequências nucleotídicas, as sequências geradas passaram por um rigoroso controle de qualidade para evitar a contaminação das amostras deste estudo entre si, e com amostras de outros estudos previamente conduzidos no mesmo laboratório, ou com cepas usadas em procedimentos laboratoriais.

Tabela 4.1.3.1: *Primers* utilizados na PCR e no sequenciamento.

Primer	**Região**	**Sequência**
ED5F (1º round)	*env*	5’ ATGGGATCAAGCCTAAAGCCATGTG 3’
MM1R (1º round)	*env*	5’GGTGAATATCCCTGCCTAA 3’
ED31F (2º round e sequenciamento)	*env*	5’ CCTCAGCCATTACACAGGCCTGTCCAAAG 3’
MM4R (2º round e sequenciamento)	*env*	5’CCTCCTACTATCATTATGAA 3’
ES7 (sequenciamento)	*env*	5’ CTGTTAAATGGCAGTCTAGC 3’
ED14 (sequenciamento)	*env*	5’ TCTTGCCTGGAGCTGTTTGATGCCCCAGAC 3’
DP10 (1º *round*)	*pol*	5’ TAACTCCCTCTCAGAAGCAGGAGCCG 3’
LR54 (1º *round*)	*pol*	5’ TAGGCTGTACTGTCCATTTAT 3’
DP16 (2º *round* e sequenciamento)	*pol*	5’ CCTCAAATCACTCTTTGGCAAC 3’
RT12 (2º *round* e sequenciamento)	*pol*	5’ ATCAGGATGGAGTTCATAACCCATCC 3’
LR51(sequenciamento)	*pol*	5’TGTGG TATTCCTAATTGAACTTCCC 3’
LR49(sequenciamento)	*pol*	5’ CAATGGCCATTGACAGAAGA 3’

4.1.4 Análises das Sequências

Após a obtenção das sequências foram realizadas as análises utilizando programas de bioinformática disponibilizados na unidade de bioinformática do LASP/CPqGM/Fiocruz (http://lasp.cpqgm.fiocruz.br).

Alinhamento Múltiplo

As sequências consenso geradas foram alinhadas utilizando-se o programa CLUSTAL-X (THOMPSON *et al*., 1997), juntamente com amostras de referência de todos os subtipos e principais formas recombinantes do grupo M do HIV-1 obtidas no banco de dados do Laboratório Nacional de Los Alamos, Estados Unidos (http://www.hiv.lanl.gov/). O alinhamento gerado foi então manualmente editado utilizando o programa GENEDOC (NICHOLAS *et al*., 1997).

Análises Filogenéticas

Para a determinação dos subtipos do HIV-1 as sequências geradas foram submetidas à análise filogenética. Como grupo externo foi utilizada a sequência genômica do vírus da imunodeficiência humana do tipo 1 do grupo O de número de acesso MVP5180. As inferências filogenéticas foram realizadas pelos métodos "neighbor-joining" (NJ) e máxima verossimilhança ("Maximum Likelihood"-ML), utilizando o modelo de substituição de nucleotídeos GTR (que assume frequências de base diferentes, bem como um viés transição-transversão, com taxas diferentes para cada uma das quatro transversões) com taxa de variação ao longo dos sítios obedecendo a uma distribuição gama (α=0,82), além de uma fração dos sítios tida como invariável (I=0,31) (GTR+I+G), selecionado pelo ModelTest implementado no programa PAUP* versão 4.02 (SWOFFORD, 2002). O cálculo de *bootstrap*, baseado em 1000 reamostragens, foi utilizado para determinação da consistência dos ramos e as árvores foram visualizadas utilizando-se o programa FigTree. A ferramenta de subtipagem REGA (DE OLIVEIRA *et al*., 2005) também foi utilizada para confirmação dos subtipos.

As sequências referências utilizadas para a análise filogenética e seus números de acesso são: AF004885 (A); AF069670 (A); AF286237 (A); AF286238 (A); AY173951 (B); AY331295 (B); AY423387 (B); K03455 (B); AF067155 (C); AY772699 (C); U46016 (C); AY253311 (D); AY371157 (D); K03454 (D); U88824 (D); AF005494 (F1); AF075703 (F1); AF077336 (F1); AJ249238 (F1); AF377956 (F2); AJ249236 (F2); AJ249237 (F2); AY371158 (F2); AF061641

(G); AF061642 (G); AF084936 (G); U88826 (G); AF005496 (H); AF190127 (H); AF190128 (H); AF082394 (J); AF082395 (J); AJ249235 (K); AJ249239 (K); AF385934 (CRF12_BF); AF385935 (CRF12_BF); DQ085873 (CRF28_BF); DQ085872 (CRF28_BF); DQ085876 (CRF29_BF); DQ085871 (CRF29_BF); EU735534 (CRF39_BF); EU735535 (CRF39_BF); EU735538 (CRF40_BF); EU735539 (CRF40_BF).

Identificação de Recombinantes

Para o estudo de presença de recombinação intersubtipos, todas as sequências geradas foram, uma a uma, avaliadas através do método de *bootscanning* implementado no programa SIMPLOT, versão 2.5 (SALMINEN, 1995). Este programa compara a nova sequência com um conjunto de sequências referência de cada subtipo do grupo M e CRFs BF, para avaliar a similaridade de cada região da nova sequência com as referências. Desta forma, o programa gera um gráfico onde é possível observar se há presença de recombinação intersubtipos na sequência analisada e o ponto onde ocorre esta recombinação. Neste trabalho, esta análise foi conduzida em uma janela deslizante de 200 nucleotídeos de extensão da sequência, em um estudo movendo-se, através de incrementos de 20 bases, frente a um alinhamento de sequências de referência, representativas dos diversos subtipos do grupo M. Para cada janela, 100 ciclos de replicação de *bootstrap* foram conduzidos através do algoritmo de *neighbor-joining* usando os programas SEQBOOT, DNADIST (usando o modelo de dois parâmetros de Kimura (KIMURA, 1980)), NEIGHBOR, e CONSENSE contidos no pacote PHYLIP (FELSENSTEIN, 1989), e os valores de *bootstrap* foram visualizados frente às posições nucleotídicas do alinhamento de sequências referência. As sequências referencia utilizadas nesta análise foram as mesma da construção da árvore filogenética. O programa GENEDOC (NICHOLAS *et al*., 1997) foi utilizado para analisar visualmente o ponto de mudança de subtipo nos recombinantes encontrados. Para cada parte da estrutura dos mosaicos foi construída uma árvore NJ para confirmar o perfil de cada recombinante. Além disto, as sequências também foram submetidas a ferramenta online SCUEAL (*Subtype Classification using Evolutionary Algorithms*) (POND, 2009).

Mapeamento de mutações que levam a resistência aos antirretrovirais

O nível de susceptibilidade das sequências foi inferido através do algoritmo do banco de dados *Stanford HIV resistance* (SHAFER, 2000). Foi submetido o arquivo contendo as sequências em formato *fasta* ao sítio eletrônico http://hivdb.stanford.edu/. O algoritmo de Stanford atribuiu para cada mutação um escore de resistência. O conjunto dos escores é traduzido em 5 patamares de susceptibilidade inferida a cada fármaco: susceptível, susceptibilidade potencialmente diminuída em baixo grau, baixo grau de resistência, resistência intermediária e alta resistência. A partir das informações geradas foi possível observar quais as mutações mais encontradas nas sequências estudadas e a qual medicamento essas mutações conferem maior resistência. Estes dados também foram comparados com o tratamento que cada paciente está submetido. A frequência de mutações dita secundárias para resistência aos inibidores da protease foi observada em pacientes virgens de tratamento. Estas mutações, que não devem ser confundidas com resistência secundária ao tratamento, por si só não provocam queda da sensibilidade e geralmente emergem no gene da protease após o surgimento das mutações primárias, uma vez que tendem a melhorar o *fitness* da replicação viral. As mutações secundárias na protease também são comumente encontradas como polimorfismos em variantes nunca antes expostas aos antirretrovirais. Consideraremos como mutações secundárias na protease aquelas descritas na mais recente revisão da *International AIDS Society* (JOHNSON, 2006).

4.2 Avaliação filodinâmica de isolados do HIV-1 na transmissão materno-fetal.

População de estudo

4.2.1 População de Estudo

Para este estudo foram incluídos oito pares de mãe e filho infectados pelo HIV-1 provenientes do Centro de Referência em Doenças Sexualmente Transmissíveis

(CDST/HIV/AIDS) da Secretaria Municipal de Saúde, na cidade de Feira de Santana, Bahia, Brasil. Dos oito pares, apenas foi possível coletar amostras de três tempos diferentes, com intervalo médio de cinco meses de uma coleta para a outra, de dois pares (Tabela 4.2.1). Os demais pares foram excluídos do estudo pois, em dois, apenas dois tempos foram coletados, e em quatro, apenas um tempo. A análise dos prontuários para a coleta dos dados clínicos, laboratoriais e de terapia dos pacientes foi realizado. A coleta foi realizada após a assinatura de um termo de consentimento informado. As crianças foram incluídas no estudo mediante consentimento dos responsáveis maiores de idade.

Dos dois pares em que foi possível obter os três tempos, o par 131+88 foi excluído do estudo por problemas no sequenciamento dos clones.

Foram também incluído nesse estudo três pares de mãe e filho infectados pelo HIV-1 da Zâmbia (HOFFMANN, 2008). Essas sequências são provenientes do GenBank (BENSON, 2008) e cada paciente apresenta de três a cinco tempos diferentes com intervalo médio de seis meses entre cada tempo (Tabela 4.2.1).

Tabela 4.2.1.1: Tempos das coletas em meses a partir do nascimento da criança

Nome		Coletas*	Local
MIP834_M	Mãe	4, 12 e 18 meses	Zâmbia
MIP834_I	Filho	4, 6, 12 e 18 meses	Zâmbia
MIP2660_M	Mãe	18, 24 e 30 meses	Zâmbia
MIP2660_I	Filho	18, 24, 30, 36 e 42 meses	Zâmbia
MIP2953_M	Mãe	11, 18, 24, 30 e 39 meses	Zâmbia
MIP2953_I	Filho	18, 30 e 39 meses	Zâmbia
FS17	Mãe	116, 124 e 129 meses	Feira de Santana
FS16	Filho	116, 124 e 129 meses	Feira de Santana
FS88	Mãe	44, 50 e 54 meses	Feira de Santana
FS131	Filho	45, 50 e 60 meses	Feira de Santana

* em meses após nascimento da criança

4.2.2 Aspectos Éticos

Este projeto foi aprovado pelo Comitê de Ética e Pesquisa da Escola Bahiana de Medicina e Saúde Pública com protocolo de número 86/2007 intitulado "Variabilidade genética dos isolados do HIV-1 em mulheres e crianças infectadas de Feira de Santana" e posterior adendo datado de 19/08/2008 com objetivo de título "Estudar a dinâmica evolutiva do gene *env* do HIV-1". O termo de consentimento livre e esclarecido foi assinado por todos os indivíduos participantes da pesquisa (Anexo B).

4.2.3 Procedimentos Experimentais

Estes procedimentos foram realizados apenas nas amostras de Feira de Santana.

Separação de PBMCs

As células mononucleares de sangue periférico (PBMC) dos pares para a análise da filodinâmica foram isoladas a partir de 10 mL de sangue total por meio do método de centrifugação em gradiente de densidade Ficoll-Hypaque (Histopaque Sigma, EUA). As células foram recuperadas e lavadas, e em seguida realizada a contagem do número total de células viáveis, armazenando em alíquotas de 10^6 células a 4°C.

Extração de DNA

O DNA foi extraído a partir de PBMCs utilizando o kit QIAamp DNA (QIAamp DNA minikit, Hilden, Düsseldorf, Alemanha).

Reação em Cadeia de Polimerase (PCR)

O fragmento do gene *env* estudado apresenta aproximadamente 1480pb. Esta região foi amplificada pela utilização da *nested* PCR (PCR aninhado) com *primers* específicos (Tabela 4.1.3).

O protocolo utilizado para os primeiro e segundo *rounds* foi: 5,0 uL de tampão de reação 10X; 2,5 uL de $MgCl_2$ 50 mM; 12 uL de dNTP 1,25 mM; 0,5 uL de cada um dos *primers* direto e

reverso à 5 pmoles/uL; 0,3 uL de *Taq polimerase* (LGC); 5 uL (100 ng/uL) de DNA (primeiro *round)* ou produto do primeiro *round* (segundo *round)* e H_2O bidestilada estéril na quantidade suficiente para 50 uL final. As condições de amplificação utilizadas no termociclador (Applied Biosystens) foram: 1 ciclo à 97^o C por um minuto; 35 ciclos à 94^o C por 1 minuto, 55^o C por 45 segundos e 72^o C por 2 minutos; e um ciclo final à 72^o C por 10 minutos.

Os produtos da PCR foram analisados em gel de agarose a 1%, com brometo de etídeo e visualizado em luz ultravioleta.

Clonagem

A clonagem foi realizada a partir de produto fresco de PCR utilizando o vetor de clonagem TOPO TA 2.1 (Invitrogen) e posterior transformação utilizando células competentes DH5α (Invitrogen). Após verificar a presença do fragmento nos clones, aproximadamente vinte clones, de cada amostra contendo o fragmento clonado, foram selecionados e transferido para crescimento em meio de cultura liquido. A partir da cultura com o clone, foi realizada a extração do DNA plasmidial utilizando o kit PureLink Quick Plasmid Miniprep (Invitrogen). Estas técnicas foram realizadas na Fundação Hemocentro de Ribeirão Preto com a supervisão da Dra Simone Kashima.

Sequenciamento dos Clones

O sequenciamento dos clones contendo o fragmento de 1480 pb do gene *env* foi realizado Laboratório do Serviço de Sequenciamento Genômico da Universidade da Florida, Gainesville, Florida, EUA (Genome Sequence Service Laboratory, University of Florida) sob supervisão do Dr. Marco Salemi.

4.2.5 Análises das Sequências

As análises das sequências foram realizadas no Departamento de Patologia e Imunologia da Universidade da Florida (Department of Pathology and Immunology, University of Florida), Gainesville, Florida , EUA, sob supervisão do Dr. Marco Salemi.

Alinhamento múltiplo

As sequências dos pares de Feira de Santana e da Zâmbia foram então alinhados cada par de mãe com seu respectivo filho e cada paciente separadamente utilizando o programa ClustalX (THOMPSON *et al*., 1997). O alinhamento foi então editado utilizando o software BioEdit (http://www.mbio.ncsu.edu/BioEdit/bioedit.html). Após a edição o tamanho final da sequência de cada dataset foi: de 1350 pb para FS17, 1434 pb para FS16, 987 pb para o 834, 912 pb para o 2660 e 906 pb para o 2953.

Teste de Hudson

Para verificar se o vírus encontrado em cada tempo dentro do mesmo paciente e o encontrado na mãe e filho eram de subpopulações diferentes foi realizado o teste de Hudson (HUDSON, 1992). Esse teste é realizado online (**http://wwwabi.snv.jussieu.fr/achaz/hudsontest.html**) no qual inseri-se o alinhamento das duas populações que se deseja comparar e a ferramenta fará o teste estatístico baseado nas diferenças par a par entre os conjuntos de dados.

Teste para verificar presença de recombinantes intra-paciente

A presença de recombinantes intra-paciente cria o chamado genoma mosaico, interferindo assim na correta representação da evolução por uma árvore. Devido a isto, para verificar a presença de recombinantes intra-paciente foi realizado o teste estatístico PHI (pair-wise homoplasy índex) (BRUEN *et. al*., 2006), implementado no programa SplitsTree (HUSON e BRYANT, 2006). Dado um alinhamento, o programa cria uma rede de ligações entre as sequências, permitindo assim a identificação de incertezas filogenéticas, e o teste PHI calcula o índice de homoplasia par a par. As sequências são então progressivamente retiradas até que o valor do teste PHI não seja mais significativo ($p>0,05$) Os recombinantes encontrados foram retirados do conjunto de dados para as análises posteriores (SALEMI, 2008).

Analises filogenéticas

Para verificar a relação entre as sequências de cada tempo de cada paciente e a relação entre as sequências presentes na mãe com as presentes no filho, primeiramente foi inferida um árvore filogenética utilizando o método de máxima verossimilhança. As árvores foram inferidas utilizando a ferramenta online PhyML (GUINDON e GASCUEL, 2003) onde foi inserido o alinhamento e a ferramenta estima o modelo a ser utilizado, o valor da distribuição gama e dos sítios invariáveis. Além disto o PhyML também realiza o cálculo do *bootstrap*, com 1000 de reamostragem para dar suporte estatístico aos ramos. As árvores foram visualizadas utilizando o programa FigTree v.1.2.2.

Árvores Bayesianas onde as datas de coletas são levadas em consideração na sua construção foram inferidas utilizando o programa BEAST (*Bayesian Evolutionary Analysis Sampling Trees*) versão 1.4.8. (DRUMMOND AND RAMBAUT 2007, DRUMMOND *et al*. 2005). Foram testados os modelos de relógio molecular rígido e relaxado com o conhecimento a

priori da taxa de crescimento demográfico constante, além do relógio molecular relaxado com conhecimento a priori da taxa de crescimento demográfico exponencial e *Bayesian skyline plot* (BSP), para estimar o crescimento da população efetiva. Os parâmetros para cada modelo foi estimando usando o Método de Monte Carlo com Cadeias de Markov (MCMC) rodando 50.000.000 de gerações, com amostragem a cada 5.000 gerações. Os resultados das análises foram visualizados no software Tracer v.1.4 e a convergência da Cadeia de Markov foi acessada a partir do cálculo do tamanho efetivo de amostragem para cada parâmetro, onde o valor ideal é > 500, indicando amostragem suficiente (DRUMMOND e RAMBAUT, 2007). Para cada conjunto de dados a árvore com máxima credibilidade do clado, que é a árvore com o maior produto da probabilidade posterior do clado, foi selecionada a partir da distribuição das árvores posteriores após um *burnin* de 50% utilizando TreeAnnotator v 1.4.8. As árvores foram visualizadas e manipuladas no programa FigTree v.1.2.2.

Os diferentes modelos utilizados foram comparados para selecionar o modelo mais adequado para cada conjunto de dados, calculando o Fator de Bayes (Bayes Factor-BF), que é a razão de verossimilhança marginal (*marginal likelihood*) dos dois modelos comparados (KASS e RAFTEY, 1995; SUCHARD, WEISS e SINCHEIMER, 2001). Para cada modelo coalescente foi calculado o valor aproximado de verossimilhança marginal através de amostragem (1000 *bootstraps*) usando a média da amostragem de verossimilhança. A diferença em log dos valores de verossimilhança marginal entre os dois modelos é o log do Fator de Bayes. Evidências contra o modelo nulo (o que apresentar o menor valor de verossimilhança marginal) é indicado quando o valor de 2 log(BF) for > 3 (moderado) e >10 (forte). Esses cálculos foram realizados no software BEAST v 1.4.8 e Tracer v.1.4.

A estimativa da taxa de crescimento populacional foi obtida utilizando as análises bayesianas implementadas no BEAST, que utilizaram o modelo do relógio molecular relaxado e crescimento exponencial. Se o intervalo de confiança de 95% da taxa de crescimento contiver zero, o modelo constante não pode ser rejeitado.

Para estimar o tamanho da população efetiva foi realizado a análise Bayesiana com o modelo do relógio relaxado utilizando o conhecimento a priori BSP. A reconstrução do BSP foi então realizado no programa Tracer v.1.4.

Mapeamento de epítopos

Nas sequências dos pacientes de Feira de Santana, que são do subtipo B, foram mapeados os epítopos de CTL e células B pertencentes ao subtipo B descritos em Los Alamos (BETTE, 2006/2007). Para as sequências provenientes dos pares da Zâmbia, que são do subtipo C, foram mapeados apenas os epítopos de células B do subtipo C descritos em Los Alamos, e não os de CTL pois esse pacientes encontravam-se na fase aguda da infecção não apresentando resposta CTL.

4.3 Caracterização evolutiva do genoma total do Vírus do Oeste do Nilo.

4.3.1 Conjunto de dados utilizados

Todas as sequências (104) de genoma completo do WNV disponíveis no Genbank até o dia 15 de outubro de 2008 que apresentavam local e ano da coleta foram baixadas. Esse conjunto de dados foi então dividido em quatro grupos não exclusivos: completo, contendo todas as 104 sequências do estudo; America do Norte, com 76 sequências; Linhagem 1, com 95 sequências; Linhagem 2, com 9 sequências. Cada um destes conjuntos de dados foi então dividido em alinhamentos separados por gene: estruturais (M, prM e E) e não-estruturais (NC, NS1, NS2a, NS2b, NS3, NS4a, NS4b e NS5). Dois conjuntos de dados também foram criados com genes concatenados apresentando mais de 2000 pb de tamanho para comparação: NC, prM, M e E (denominado 5'); E + NS1. Todos os alinhamentos foram realizados utilizando o programa ClustalX (THOMPSON *et al*., 1997) e editado manualmente.

4.3.2 Analises das sequências

Determinação do modelo evolutivo

O melhor modelo de substituição de nucleotídeos foi testado para cada gene dos diferentes conjunto de dados. Foi realizado um teste de razão de verossimilhança hierárquica onde a taxa da matriz, o parâmetro de sítios invariáveis e a distribuição das razões foram estimados em um árvore-base Neigbor-joining com correção de distancias LogDet. Estas análises foram realizadas no software PAUP* versão 4.02 (SWOFFORD, 2002).

Mapeamento de verossimilhança

Para investigar o sinal filogenético de cada gene nos diferentes conjunto de dados foi realizado o mapeamento de verossimilhança analisando 10.000 quartetos randomizados utilizando o programa TREE-PUZZLE (SCHMIDT *et al*., 2002). Esta análise avalia grupos de quatro sequências randomicamente selecionadas utilizando máxima verossimilhança. Para cada quarteto as três possíveis árvores não enraizadas são pesadas. Esses pesos a posteriore são então plotados em uma superfície triangular. As topologias das árvores altamente resolvidas são plotadas nas pontas do triângulo, que indica presença de um sinal filogenético tipo árvore, e os quartetos não resolvidos por uma árvore filogenética são mostrados no centro do triângulo. Um sinal filogenética tipo estrela ocorre quando mais de 30% das medidas dos quartetos ficam no centro do triangulo, indicando assim um baixo sinal filogenético (STRIMMER e HAESELER, 1997).

Analises Filogenéticas Bayesianas

Para estimar a genealogia e a escala evolutiva temporal dos genes E, NS3 e NS5 do conjunto de dados da America do Norte análises Bayesianas foram realizadas através do pacote do software BEAST versão 1.4.8 (DRUMMOND e RAMBAUT, 2007) utilizando o modelo do relógio molecular relaxado *log-normal,* com o modelo de substituição nucleotídica GTR+I e crescimento populacional constante. As análises foram realizadas como descrito anteriormente (4.2.5, análises filogenéticas). As árvores foram visualizadas e manipuladas no FigTree v.1.1.2.

Avaliação da Hipótese do Relógio Molecular

Análises do relógio molecular foram realizadas nos genes E, NS3 e NS5 dos conjuntos de dados Mundo e America do Norte. O relógio molecular estrito assume que todos os ramos na árvore tem a mesma taxa evolutiva, enquanto o relógio molecular relaxado permite taxas diferentes construída com uma distribuição específica para todos os ramos da árvore. Esses dois modelos foram comparados calculando o Fator de Bayes (BF) como explicado anteriormente (4.2.5, análises filogenéticas).

5. RESULTADOS

Os resultados estão descritos em três artigos com os resultados de cada estudo realizado.

5.1 Caracterização Molecular do Gene *pol* do HIV-1 de Indivíduos Infectados de Salvador, Bahia, Brasil.

Manuscrito em preparação: SANTOS, L.A.; MONTEIRO, J.P.; ARAUJO, A.F.; BRITES, C.; GALVÃO-CASTRO, B.; ALCANTARA, L.C.J. Detection of distinct Human Immunodeficiency Virus type 1 (HIV-1) Cirulating Recombinant Forms (CRFs) in Northeast Brazil.

Figuras que não foram incluídas neste artigo, mas que melhor detalham o perfil dos recombinantes, encontram-se no Apêndice A.

Detection of distinct Human Immunodeficiency Virus type 1 (HIV-1) Circulating Recombinant Forms (CRFs) in northeast Brazil.

Running head: HIV-1 molecular characterization in Salvador.

Manuscrito em preparação

Santos, L.A.[1]*; Monteiro-Cunha, J.P.*[1]*; Araujo, A.F.*[1]*; Brites, C.*[2]*; Galvão-Castro, B.*[1,3] *and Alcantara, L.C.J.*[1,3,4]

1– Advanced Public Health Laboratory, Gonçalo Moniz Research Center, Oswaldo Cruz Foundation, Salvador, Bahia, Brazil; 2– Federal University of Bahia, Salvador, BA, Brazil; 3– HTLV Center/ Bahia School of Medicine and Public Health/Bahia Foundation for Science Development, Salvador, Bahia, Brazil; 4– National Cancer Institute, National Institutes of Health, Bethesda, MD, USA.

Corresponding author:

Luiz Carlos Junior Alcantara, PhD

NIH-NCI, Vaccine Branch

Building 41, Room C303, 41 Library Drive, MSC 5055, Bethesda, MD 20892, USA

Telephone: (301) 402-6158

Fax: (301) 402-0055

email: alcantaralc@mail.nih.gov

Abstract:

The HIV-1 is characterized by great genetic diversity, therefore it is important to characterize the circulating strains in different geographic regions. The enzymes encoded by the *pol* gene are the major target for antiretroviral therapy that can lead to drug resistance associated mutations. This study performs the molecular characterization of the *pol* gene of 57 HIV-1 infected individuals from Salvador, Bahia, Brazil. DNA sequences were obtained by PCR, followed by sequencing. The subtypes were determined by phylogenetic analyses and intersubtype recombination was investigated using the *bootscanning* method. The *pol* subtypes were compared with *gag* and *env* subtypes. Antiretroviral susceptibility was evaluated through the Stanford HIV resistance Database. The subtypes frequencies were: 77.2% of subtype B, 1.8% of subtype F and 21.0% of BF recombinant. Comparing with *gag* and *env* regions, two intergenic and three intragenic recombinant patterns were observed. Among the nine BF recombinants detected in the *pol* fragment six (10.5%) viruses were related to CRF28/CRF29, two were related to CRF12 (3.5%) and one (1.8%) virus was related to CRF39. The antiretroviral resistance analyses showed that 24.6% of the strains had resistance to nucleoside reverse transcriptase inhibitors (NRTIs), 21.0% to Non-nucleoside reverse transcriptase inhibitors (NNRTIs) and 7% to the Protease Inhibitors (PI). This HIV-1 characterization could contribute to the better understanding of the circulate strains in Salvador. They could also support the search for better treatment monitoring.

Introduction

Today approximately 33 million people are infected with Human Immunodeficiency Virus (HIV) worldwide [UNAIDS, 2009]. The characteristics of the viruses found in each geographic region are different, mainly due to the HIV high diversity, host genetic factors and the selective pressure of antiretroviral therapy (ARV). The HIV is classified in two types: HIV-1 and HIV-2. The HIV-1 is highly spread through the world and is responsible for the virus pandemic. The type 1 HIV is divided in three groups: group M, N and O. The group M is then divided in 9 subtypes (A, B, C, D, F, G, H, J and K) [Mccutchan, 2000] and 49 Circulating Recombinant Forms (CRF). Besides, there are many Unique Recombinant Forms (URF) with different patterns of recombination, increasing the HIV diversity. The distribution of these subtypes is different around the world. In Brazil, the subtype B is the most frequent, followed by the BF recombinants and subtypes F and C, with remarkable differences among the geographic regions. It is important to monitor the subtype frequencies distribution and the emergence of new CRF in the country for the better understanding of the HIV behave and history in Brazil.

There is no vaccine today against the HIV. Its diversity is one of the major difficulties for the development of an effective vaccine. Although the ARVs are able to reduce the viral load, increase the TCD4 cell count, slowing the progression to AIDS and improving the patients quality of life, many patients develop treatment failure [Perelson et al., 1996; Sepkowitz, 2001; Palella et al, 1998]. This failure is caused mainly by mutations at the *pol* gene, which encodes the ARV target enzymes (protease and reverse transcriptase), leading to resistance [Coffin, 1995]. The resistance mutation monitoring is important for the better understanding of the HIV diversity and epidemic, controlling the transmission of strains with anti-retroviral resistance mutation and finding better treatments.

In this study we performed the molecular characterization of the HIV-1 *pol* gene in infected patients from Salvador, Bahia, Brazil. We subtyped the *pol* region checking for inter-subtype recombinants and the presence of anti-retroviral resistance mutations.

Methods

Population

Fifty-seven HIV-1 infected patients followed at the Hospital Professor Edgar Santos in Salvador, Bahia, Brazil were included in this study. Ten mL of whole blood samples were collected during the year of 2006. Samples were stored at -20^{o}C at the Advanced Laboratory of Public Health (LASP)/CPqGM/FIOCRUZ in Salvador, Bahia, Brazil. All patients included in this study signed the letter of informed consent. The clinical data available for each patient were collected from medical records. The CPqGM/FIOCRUZ Ethics Committee approved this study.

DNA extraction and PCR

The genomic DNA of all samples was extract using Qiagen extraction kit (QIAGEN, Valencia, CA). The *pol* fragment was amplified using a nested polymerase chain reaction (PCR). The primers used were DP10 and LR54 for the first round and DP16 and RT12 for the second round. The PCR cycling conditions were as follow: three cycles of denaturing at 95^{o}C for 3 min, annealing at 55^{o}C for 1 min and primer extension at 72^{o}C for 1 min; another 35 cycles of denaturing at 95^{o}C for 1 min, annealing at 55^{o}C for 45 sec and primer extension at 72^{o}C for 1 min and a final extension at 72^{o}C for 10 min. The PCR products were then purified using the Qiagen columns (QIAGEN, Valencia, CA) and sequenced in an ABI 3100 Genetic Analyzer (Applied Biosystems, Foster City, CA) using a Big Dye Terminator kit (Applied Biosystems, Foster City, CA). The primers used for sequencing were the DP16 and RT12 with the inner primers LR49 and LR51.

Sequence analyses

The sequences were analyzed using the SeqScape v2.1.1 software (Applied Biosystems, Foster City, CA) to define the consensus sequence for each sample. We used BLAST search against the HIV-1 sequences database (http://blast.ncbi.nlm.nih.gov) to check for contamination.

Phylogenic analyses were performed to subtype each sequence. Reference sequences form all HIV-1 subtypes were downloaded from Los Alamos database, aligned with the new sequences using CLUSTAL-X [Thompson et al., 1997] and manually edited using GENDOC [Nicholas et al., 1997]. One sequence of the HIV-1 group O (MVP5180) was used as outgroup. Phylogenic reconstruction was performed using the PAUP* software v4.02[a] [Swofford, 1997] to generate neighbor-joining (NJ) and maximum likelihood (ML) trees applying the GTR nucleotide substitution model with invariable sites and gamma distribution. The bootstrap value (1000 replicates) was calculated to check the node reliability and was considered significant when above 70%. To calculate de statistical support for the tree branches the likelihood-ratio test was used. All trees were visualized using FigTree v1.2.2. Intersubtype recombination was identified using Bootscanning implemented in SIMPLOT software, version 2.5 [Salminen, 1995] and the GENEDOC software was used to analyze the crossover points by visual inspection of the alignment. Each part of the mosaic structure was confirmed by the reconstruction of bootstrapped NJ trees as described above. The BF recombinant forms determined were aligned with reference sequences of CRFs from the Los Alamos database and analyzed by phylogenetic reconstruction as described above. All sequences were submitted to the REGA subtyping tool [De Oliveira et. al., 2005].

Drug resistance analyzes

All sequence was submitted to the Stanford HIV resistance database (http://hivdb.stanford.edu/.) to check for the presence of antiretroviral resistance mutation.

Results

The HIV-1 *pol* sequences of 57 HIV-1 infected patients were included in this study. The mean age was 42.41 ranging from 12 to 64 year old. Thirty-six individuals (63.2%) were male and 21 (36.8%) were female. The major risk behavior identified was heterosexual transmission with 59.65% (34), followed by men who have sex with men (MSM) with 15.79% (9), vertical transmission with 5.26% (3), intravenous drug user (IDU) with 5.26% (3), bisexual with 3.51% (2) and blood transfusion with 1.75% (1). The risk behavior was unknown in 8.77% (5) of the patients (Table 1). The CD4 cell count was bellow 200 cells/ul in 12 (21.1%) patients and the viral load was higher than 10,000 copies/ml in 20 (35.1%) patients. Eleven (19.3%) of the patients were treatment naïve and 44 (77.2%) were using antiretroviral treatment, but the drug regimen information was only available for 14 patients. Two patients (3.5%) had no information regarding ARV therapy.

The protease and reverse transcriptase regions of the *pol* gene were used to subtype the sequences of 57 individuals included in this study. Fourty-five (78.9%) of the sequences were subtype B, three (5.3%) were subtype F1 and nine (15.8%) were BF recombinant in *pol* (Table 1). The NJ tree constructed had similar topology with the ML tree. Each subtype formed a monophyletic clad that were well supported with a bootstrap support higher than 70% with the exception of the subtype B+BF clad (Figure 1). The subtype B+BF clad with the study isolates had a bootstrap support of 57% probable due to the CRF sequences. When this analysis was performed only with the pure subtypes the clad bootstrap support were 95% (data not showed). The bootstrap support of subtype F clad were 70% and 100% in the pure subtype tree. The B clad had a high significant ML value ($p<0.001$) and the F clad ML value of 0.002.

The *pol* subtypes were compared with the subtypes found for the same samples in a previous study [Araujo et al, 2010] using *gag* and *env* regions. Three sequences were intergenic BF recombinants. One subtype B sequence in *pol* was F1 in *env* and two subtype F sequences in *pol* were BF in *gag* and B in *env.* Therefore the subtype prevalence in this population was 77.2% (44) of B, 1.8% (1) of F and 21.0% (12) of BF recombinants.

Among the nine (15.8%) BF recombinants detected in the *pol* fragment, three different mosaic patterns were observed. Six viruses (BR08BA, BR15BA, BR23BA, BR46BA, BR69BA and BR79BA) showed the same recombination pattern and breakpoints in the SIMPLOT analysis and clustered together with CRF28_BF and CRF29_BF sequences. Two viruses (BR75BA and BR81BA) were closely related to each other and to CRF12_BF in all analysis, presenting the same recombination profile and breakpoints and one sequence (BR42BA) was classified as CRF39_BF. The sequence alignment was divided based on the related CRF breakpoints and each subregion was analyzed. NJ and ML analyses of the individual fragments confirmed the Bootscanning recombination results (Figure 1).

The sequences subtypes were compared with the gender of the patients. The ratio of male-to-female was 0.71:1 (5 males and 7 females) for the BF infected individuals and 2.14:1 (30 males and 14 females) for subtype B infected individuals. The subtype F infected patient was a male.

The antiretroviral resistance analyses showed that 43 (75.4%) isolates were susceptible to NRTIs, 45 (79.0%) to NNRTIs and 53 (93.0%) to PIs. The NRTIs drugs lamivudine (3TC) (10, 17.5%) and zidouvudine (AZT) (11, 19.3%) and the NNRTI drugs delavirdine (DLV) (12, 21.1%), nevirapine (NVP) (12, 21.1%) and efavirenz (EFV) (12, 21.1%) presented the highest numbers of sequences with associated-drug resistance mutations. All PI drugs had the same number of

sequences, four (7%) sequences, with mutations that confer resistance. All patients with resistance mutations were under ARV therapy, with the exception of patient BR28BA that were not under treatment at sampling time, but used ARVs approximately five years before. This patient strain presents the V108I mutation that causes low-level resistance to each of the NNRTIs except etravirine (ETR). The most frequent mutations found that confer resistance to each class of drugs was: I54V for the PI with 4 sequences (7.0%), M184V for the NRTI with 8 sequences (14%) and K103N for the NNRTI with 6 sequences (10.5%). Minor PI mutations were also found in a few sequences (Table 2).

Discussion

The study of the HIV-1 molecular epidemiology and diversity is of great importance for the better understanding of its epidemic in Brazil. The HIV-1 molecular profile in Brazil is of a majority of the subtype B and the presence of subtypes F1, C, D, A and recombinants forms BF and BC [MORGADO et al, 2002]. However, in the south of the country the subtype C is found in a much higher prevalence than in other regions of the country [SOARES et al, 2003]. Due to this, the identification of the HIV-1 circulating forms in Salvador city is necessary for the monitoring of the emergence of new recombinant forms, the migration of subtype C from the south of the country and the introduction of new subtypes from different geographic regions. In this study we found 77.2% of subtype B, 21.0% of BF recombinant and 1.8% of subtype F. In agreement with previous studies of the Northeast Brazil epidemiology, no subtype C was found. A study with 83 patients from Santos, São Paulo, Sá-Filho and collaborators identified a subtype frequency of 65.1% of subtype B, 28.9% of BF recombinants, 4.8% of subtype F and 1.2% of subtype C, using the *pol* gene. However, in a different study with 93 HIV infected blood donors from Salvador, Fortaleza and Goiânia from 2001 to 2003, using three different genes (*gag*, *pol* and *env*) the subtypes frequencies were: 75,9% of subtype B, 22.7% BF recombinants, 2.3% of subtype and 1.2% of subtype C [Brennan et al., 2007]. These subtype frequencies were more similar to the frequencies found in this study probably due to the geographic location and the use of different genome regions, giving a more consistent result.

Five different recombination structures were found in this population. Two of them were intergenic recombinant patterns: one represented by BR01BA ($B_{pol}/F1_{env}$) and the other represented by BR26BA and BR96BA ($BF_{gag}/F1_{pol}/B_{env,}$). These two sequences presented the same genotype in all three genes, and therefore could be representatives of the same recombinant

ancestral. However, they showed distinct intra-gag breakpoints [Araujo et al., 2010], which points to different recombination events or, less likely, to a common ancestral that went through subsequent recombination processes. In fact, BR26BA and BR96BA did not show a direct phylogenetic relationship in the trees based on *gag, pol* or *env* fragments. The other three recombination patterns were represented by intragenic recombinant viruses. Analyzing their *pol* sequences, six (10.5%) were related to CRF28/CRF29, two (3.5%) were related to CRF12 and one (1.8%) was related to CRF39. Out of the six CRF28/29 representatives, five were previously characterized in *env* (positions 6945 to 8183 relative to HXB2 genome) as subtype B, the same genotype found in CRF28/29 sequences [Araujo et al., 2010]. However, the *gag* region (positions 898 to 1968 relative to HXB2 genome) of the two viruses with available sequences (BR08BA and BR79BA) were also classified as pure subtype B, while CRF28/29 sequences present a crossover point (1323) in this fragment. Similarly, samples BR75BA and BR81BA showed the same recombination pattern as CRF12 in *pol*, but had subtype B in *gag*, while CRF12 have a breakpoint at position 952 and are mostly subtype F in this fragment. Moreover, BR81BA was B in *env*, while CRF12 is F in the same fragment. BR75BA was not amplified in *gag*. On the contrary, sample BR42BA showed the same recombination pattern in *pol* and the same *env* genotype (subtype F) as CRF39. The above observations suggest that different BF CRFs are co-circulating in Northeast Brazil and that they are undergoing further recombination processes. In addition, all the intragenic recombinants identified in this study were previously classified as pure subtypes in the study based on *gag* and *env* regions [Araujo et al., 2010]. This reinforces the importance of multiple genomic region characterization in order to obtain a reliable profile of the HIV-1 molecular epidemiology, especially in regions with high diversity of genotypes.

The transmission of anti-retrovirus resistance mutations can be a public heath problem. Therefore

it is important to monitor the rising of these mutations on the HIV infected population. The frequency of strains with resistance to ARVs may vary among populations. In this study, 24.6% of the strains showed some level of resistance to the NRTI drugs, 21.0% to the NNRTI drugs and 7.0% to the PI. The mutations I54V (PI), M184V (NRTI) and K103N (NNRTI) were the most frequent mutations found in 7.0%, 14.0% and 10.5% of the patients, respectively. When compared with mutations found in other studies these mutations are also found but with different prevalence. [SA-FILHO et al., 2008; LLOYD et al., 2008 and MARCONI et al., 2008]. In studies with patient that use different treatment regimes or with treatment failure these mutations were found in a higher prevalence: I54V in 24% and M184V in 88% of the sequences; and M184V/I in 64.3% and K103N in 51.3% of patients, respectively [SA-FILHO et al., 2008; MARCONI et al., 2008]. However, a study with treatment naïve patients presented a much lower prevalence, M184V in 6.0% and K103N in 3.0% of the patients [LLOYD et al., 2008], showing that these mutations are being transmitted, limiting the possible treatment regimes. The mutation M184V confers high levels of resistance to the NRTIs drugs lamivudine, emtricitabine and low level of resistance to the abacavir, that are important drugs used in treatment regimes, being an important mutation to monitor. Most of the mutations that lead to NNRTIs resistance will confer reduce of susceptibility of at least 3 of the 4 NNRTIs. The K103N mutation reduces susceptibility to all NNRTIs and confers high levels of resistance to 3 of the NNRTIs drugs, delavirdine, efavirenz, nevirapine. Mutations with this type of characteristic can make the therapy harder because patients with one single mutation probably will not be able to use any of the NNRTIs drugs in their treatment regimen.

These finding can contribute for the better understanding of the HIV-1 molecular epidemiology of Salvador. The identification and characterization of subtypes, recombinant forms and ARV

resistance mutation found in Salvador contributes for the knowledge of HIV-1 molecular diversity in Brazil and better treatment strategies.

Acknowledgments

Bioinformatics analysis was performed at the LASP/CPqGM/FIOCRUZ Bioinformatics Unit, supported by FAPESB (grants 303/03) and the Brazilian Ministry of Health (306/04 and 307/04). The authors thank CNPq for the financial support.

References

Araujo FA, Brietes C, Monteiro JP, Santos LA, Galvão-Castro B and Alcantara LCJ. 2010. Lower Prevalence of Human Immunodeficiency Virus Type 1 Brazilian Subtype B Found in Northeastern Brazil with Slower Progression to AIDS. AIDS Research and Human Retroviruses, 26(11):1249-54, 2010.

Brennan, C. A.; Brite, C.; Bodelle, P.; Golden, A.; Hackett JR, J. *et al*.. HIV-1 strains identified in Brazilian blood donors: significant prevalence of B/F1 recombinantes. AIDS Res Hum Retroviruses, 23:1434-1441, 2007.

Coffin JM. HIV population dynamics *in vivo*: Implictions for genetics variation, pathogenesis, and therapy. Science, 267:483-489, 1995.

De Oliveira T, Deforche K, Cassol S, Salminen M, Paraskevis D, Seebregt C et al. 2005. An automated genotyping system for analysis of HIV-1 and other microbial sequences. Bioinformatics, 21(19):3797-800.

Lloyd, B.; O'Connell, R. J.; Michael, N. L.; Aviles, R.; Palou, E. *et al*.. Prevalence of resistence mutations in HIV-1-infected Hondurans at tha biginning of the National Antiretroviral Therapy Program. AIDS Res Hum Retroviruses, 24:529-535, 2008.

Marconi, V. C.; Sunpath, H.; Zhigang LU; Gordon, M.; Koranteng-Apeagyei, F.; HAMPTON, J. *et al*.. Prevalence of HIV-1 Drug Resistance after Failure of a First Highly Active Antiretroviral Therapy Regimen in KwaZulu Natal, South Africa. Clin Infect Dis, 46(10):1589-97, 2008.

Mccutchan FE. Understanding the genetic diversity of HIV-1. AIDS 14, [s.l.], p. S31-S44, 2000. Supplement 3.

Morgado, M. G.; Guimaraes, M.L; Galvao-Castro, B. HIV-1 Polymorphism: a challenge for Vaccine Develoment. Mem Inst Oswaldo Cruz, 97:143-150, 2002.

Nicholas KB, Nicholas HBJ, Deerfield DW. 1997.GeneDoc: Analysis and visualization of genetic variation. EMB News, 14:30.

Palella FJ Jr, Delaney KM, Moorman AC, Loveless MO, Fuhrer J, Satten GA, Aschman DJ, Holmberg SD. 1998. Declining morbidity and mortality among patients with advanced human immunodeficiency virus infection. HIV Outpatient Study Investigators. N Engl J Med. 26;338(13):853-60

Perelson AS, Neumann AU, Markowitz M, Leonard JM, Ho DD. HIV-1 dynamics in vivo: virion clearance rate, infected cell life-span, and viral generation time. Science. 1996 Mar 15;271(5255):1582-6.

Sa-Filho, D. J.; Soares, M. S.; Candido, V.; Gagliani, L. H.; Cavaliere, E. *et al*.. HIV type 1 pol

gene diversity and antiretroviral drug resistance mutations in Santos, Brazil. AIDS Res Hum Retroviruses, 24:347-353, 2008.

Salminen MO, Carr JK, Burke DS, Mccutchan FE. 1995. Identification of breakpoints in intergenotypic recombinants of HIV type 1 by bootscanning. AIDS Res Hum Retroviruses, 11:1423-5.

Sepkowitz KA. 2001. AIDS--the first 20 years. N Engl J Med. 7;344(23):1764-72.

Soares, M.A.; De Oliveira, T.; Brindeiro, M.; Diaz, R.S.; Sabino, E.C. *et al*.. A specific subtype C of human immunodeficiency virus type 1 circulates in Brazil. AIDS, [s.l.], v. 17, p. 11-21, 2003.

Swofford D. 1997. PAUP*: Phylogenetic analysis using parsimony. Version 4.0b10. Smithsonian Institution, Washington, D.C.

Thompson JD, Gibson TJ, Plewniak F, Jeanmougin F, e Higgins DG. 1997. The ClustalX windows interface: flexible strategies for multiple sequence alignment aided by quality analysis tools. Nucleic Acids Research, 25: 4876-4882.

UNAIDS. Global Report. AIDS epdemic update 2010. Disponível em: <http://www.unaids.org/GlobalReport/Global_report.htm >

Table 1: Epidemiologic and clinical characteristics of the HIV-1-infected individuals

Parameters	Number of patients (n=57)	Percentage (%)
Age (years)*	42.41	
Gender		
Male	36	63.2
Female	21	36.8
Transmission route		
Heterosexual	34	59.65
MSM[a]	9	15.79
Bisexual	2	3.51
Vertical	3	5.26
Blood transfusion	1	1.75
IDU[b]	3	5.26
Unknown	5	8.77
Subtype (*pol*)		
B	45	78.9
BF	9	15.8
F1	3	5.3

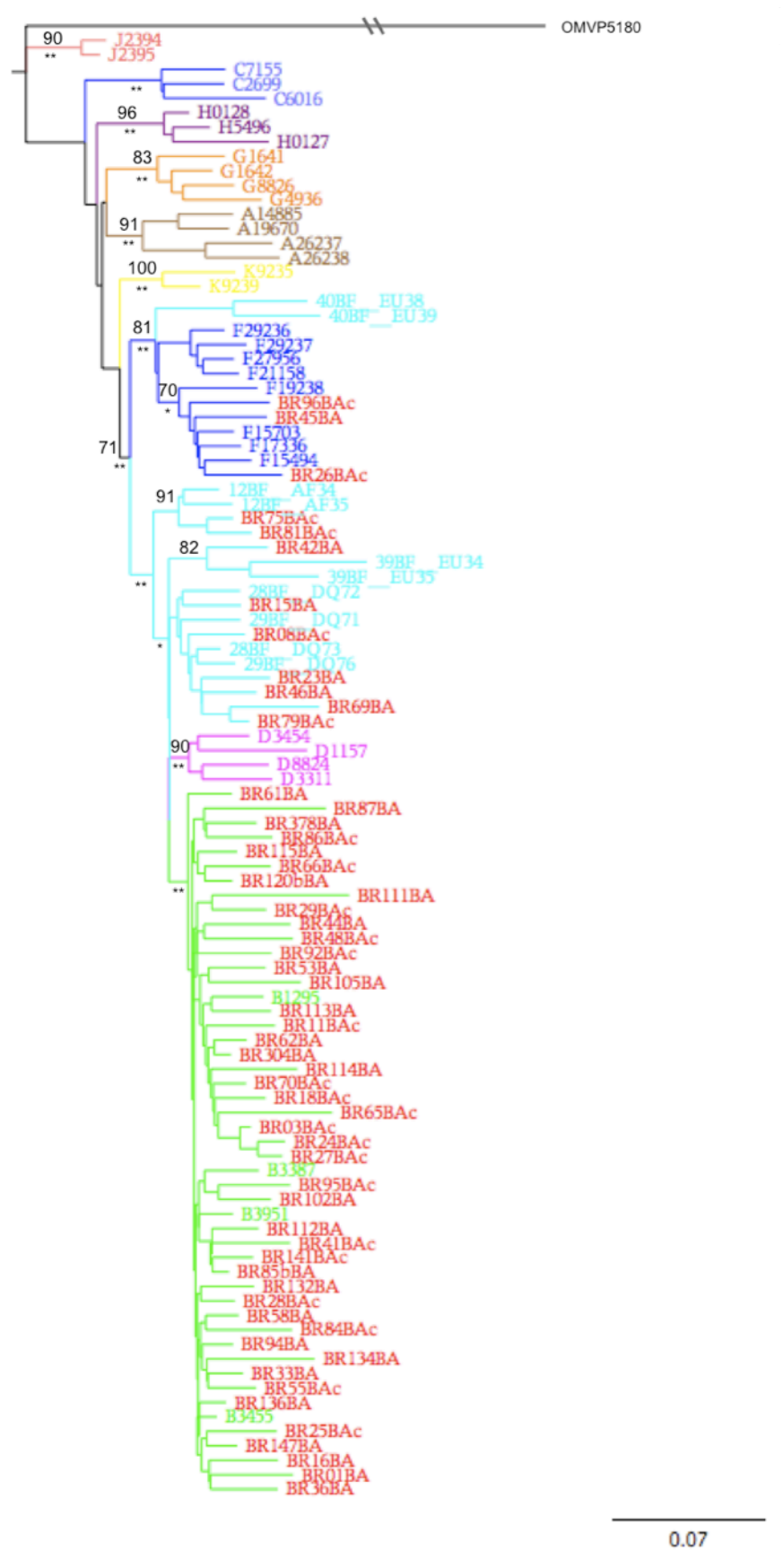
a Men who have sex with men; b Intravenous drug user; *mean

Table 2. Number of mutations/polymorphisms at anti-retroviral HIV-1 resistance-associated positions, in 57 samples from Salvador, Bahia, Brazil.

	PI	N (%)	NRTI	N (%)	NNRTI	N (%)
Mutation	I54V/L	4 (7.0)	M184V	8 (14.0)	K103N/E/S	6 (10.5)
	V82A	3 (5.3)	V118I	5 (8.8)	V90I	3 (5.3)
	M46I/L	2 (3.5)	K70R/I	5 (8.8)	Y188L	2 (3.5)
	L90M	2 (3.5)	M41L	5 (8.8)	V108I	2 (3.5)
	L33F	2 (3.5)	L210W/E	5 (8.8)	V181C	2 (3.5)
	L24I	1 (1.8)	D67N	4 (7.0)	L100I	2 (3.5)
	F53L	1 (1.8)	T215Y/F	4 (7.0)	V106I	2 (3.5)
	L23I	1 (1.8)	K219N/Q	3 (5.3)	G190A	1 (1.8)
	I50V	1 (1.8)	T69N/A	2 (3.5)	V179D	1 (1.8)
	N88D	1 (1.8)	E44D	2 (3.5)		
			D177E	1 (1.8)		
	A71T/V*	5 (8.8)	T200A	1 (1.8)		
	F53V*	1 (1.8)	R211K	1 (1.8)		
	V32M*	1 (1.8)	K49R	1 (1.8)		
	G73C*	1 (1.8)	I135T	1 (1.8)		
	T74S*	1 (1.8)	S162C	1 (1.8)		
	K43T*	1 (1.8)	A62V	1 (1.8)		
	N88H*	1 (1.8)				

n Number of sequences with anti-retroviral resistance mutation; (%) percentage of sequences analyzed with the mutation. *Minor mutations.

FIG 1: Neighbor-joining (NJ) phylogenetic tree constructed using the GTR+I+G model of nucleotides substitution, showing the phylogenic relationship among the HIV-1 *pol* strains from this study (red tips) and HIV-1 group M reference sequences from different subtypes (subtypes: A in brown; B in green; C in light blue; D in magenta; F in blue; G in orange; H in purple; K in yellow; J in salmon) and CRFs BF (CRF12_BF; CRF28_BF; CRF29_BF; CRF39_BF and CRF40_BF, all in cyan). The sequence OMVP5180 from the HIV-1 group O was used as outgroup. The 1000 replicate bootstrap values that were >70% are indicated in the branches. The ML method branches statistic support are indicated with one asterisk (*) when significant ($p<0.05$), and with two asterisks (**) when highly significant ($p<0.001$).


OMVP5180
90
J2394
J2395
C7155
C2699
C6016
96
H0128
H5496
H0127
83
G1641
G1642
G8826
G4936
91
A14885
A19670
A26237
A26238
100
K9235
K9239
40BF_EU38
40BF_EU39
81
F29236
F29237
F27956
F21158
F19238
70
BR96BAc
BR45BA
F15703
F17336
71
F15494
BR26BAc
12BF_AF34
12BF_AF35
91
BR75BAc
BR81BAc
82
BR42BA
39BF_EU34
39BF_EU35
28BF_DQ72
BR15BA
29BF_DQ71
BR08BAc
28BF_DQ73
29BF_DQ76
BR23BA
BR46BA
BR69BA
BR79BAc
D3454
90
D1157
D8824
D3311
BR61BA
BR87BA
BR378BA
BR86BAc
BR115BA
BR66BAc
BR120bBA
BR111BA
BR29BAc
BR44BA
BR48BAc
BR92BAc
BR53BA
BR105BA
B1295
BR113BA
BR11BAc
BR62BA
BR304BA
BR114BA
BR70BAc
BR18BAc
BR65BAc
BR03BAc
BR24BAc
BR27BAc
B3387
BR95BAc
BR102BA
B3951
BR112BA
BR41BAc
BR141BAc
BR85bBA
BR132BA
BR28BAc
BR58BA
BR84BAc
BR94BA
BR134BA
BR33BA
BR55BAc
BR136BA
B3455
BR25BAc
BR147BA
BR16BA
BR01BA
BR36BA
0.07

5.2 Avaliação filodinâmica de isolados do HIV-1 na transmissão materno-fetal.

Manuscrito em preparação: SANTOS LA; GRAY, R.R.; STRAZZA, E.; KASHIMA, S.; SOUZA, E.; GALVAO-CASTRO, B.; SALEMI, M.; ALCÂNTARA, L.C. Phylodynamics analysis of the Human Immunodeficiency Virus type 1 (HIV-1) *env* gene in Mother and Child Pairs.

Phylodynamics analysis of the Human Immunodeficiency Virus type 1 (HIV-1) *env* gene in Mother and Child Pairs.

Manuscrito em preparação

Santos LA [1], *Gray RR*[2], *Strazza E*[3], *Kashima S*[3], *Souza E*[4], *Galvao-Castro B*[1], *Salemi M*[2], *Alcantara LC*[1].

1- Laboratorio Avancado de Saude Publica, Centro de Pesquisa Goncalo Moniz, Fundacao Oswaldo Cruz, Salvador, Bahia, Brasil. 2- Department of Pathology, Immunology, and Laboratory Medicine, University of Florida College of Medicine, Gainesville, FL, USA. 3- Regional Blood Center of Ribeirão Preto, University of São Paulo, Brazil. 4- Centro de Referência em DST/HIV/Aids da Secretaria de Saúde de Feira de Santana, Bahia, Brazil.

Corresponding author:

Luiz Carlos Junior Alcantara, PhD

NIH-NCI, Vaccine Branch

Building 41, Room C303, 41 Library Drive, MSC 5055, Bethesda, MD 20892, USA

Telephone: (301) 402-6158

Fax: (301) 402-0055

email: alcantaralc@mail.nih.gov

Abstract:

The HIV vertical transmission can occur *in utero*, during delivery or while breastfeeding. Besides this, the immune system response and the use of anti-retroviral drugs can influence the evolutionary dynamics of the virus in each patient. In this report, the phylodynamics of the HIV-1 *env* gene in mother-to-child transmission was investigated. One chronically infected pair from Brazil and three acute infected pairs from Zambia, available at Genbank, with three to five time points, were studied. Sequences from 25 clones from each sample were obtained, aligned using Clustal X and manually edited. Maximum likelihood trees were constructed in PhyML using the best evolution model for each dataset. Bayesian analyses testing the relaxed and strict molecular clock, to check which is the best-fitted one, were performed using BEAST. Bayesian Skyline Plot (BSP) was construed for each patient. We also searched for previously described epitopes, from Los Alamos, and compared the epitope sequence and found mutations, in each time point, and between mother and child. We have found that the relaxed molecular clock was the best-fitted model for all datasets. Comparing the tree topology from each analysis, we did not find any difference in the evolutionary dynamics that could differ the mother from the child. In the BSP, the effective population size is more constant along the time in the chronically infected patients while in the acute patients it is possible to detect bottlenecks from one time point to the other. The epitopes and mutations found in the epitope region were different between mother and child chronically infected pair, while the acute infected pairs showed similar epitope profile. These results contribute to the better understanding of the HIV-1 evolutionary dynamics in mother to child transmission.

Introduction

To date, there are 2.5 million children under the age of 15 years old living with the Human immunodeficiency virus (HIV) in the world and approximately 15.5 million of women. The majority of these women and children live in the Sub-Saharan Africa (UNAIDS, 2007). In Brazil the vertical transmission rates is decreasing due to the country Ministry of Heath efforts to prevent the mother to child transmission.

The HIV vertical transmission occurs when the infected mother transmit the virus to the child during pregnancy, at delivery or breastfeeding. Without prevention the transmission rate can be of 20%, however using the right measures it can be less than 1%. These measures include the use of anti-retroviral drugs during pregnancy and when the baby is born, choose for cesarean delivery and substitution of breastfeeding for artificial milk (MINISTÉRIO DA SAÚDE, 2008).

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 (WOLINSKY *et al.*, 1992). The HIV-1 strain found in the child may have a different evolutionary history when compared with the virus found in the mother. This difference is also due to each individual different environment like immune response, that can differ from one individual to the other and depending on the phase of infection, and the use of therapy. The virus can be transmitted as cell-free or cell-associated virus. The majority of the virus transmitted *in útero* and by breastfeeding is cell-associated virus, while during delivery is cell-free virus, which indicates that depending on how the virus was transmitted the HIV-1 intra-host evolution may change (LEHMAN *et al*., 2007).

The HIV evolutionary dynamics study can contribute to the understanding of virus fitness. Phylogenetic reconstructions can reveal the dynamics of HIV evolution intra-host and inter-host.

These inter-host evaluations are able to investigate the movement of HIV lineages between locations, course of transmission, and estimate changes in viral effective population size over time. The immune response, selective pressure and viral replication rates will contribute to understand the viral adaptation and dynamic, when performing within host analyses (GRENFELL et al., 2004; LEMEY et al., 2006). Variable regions of the envelope (*env*) 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 synonymous substitutions comparing to nonsynonymous substitutions (positive selective pressure), into the *env* gene, contributing to the high diversity of this region in different time points within a host (WILLIAMSON, 2003).

Considering that the time and mechanism of transmission is known when studying vertical transmission it is possible to infer about the HIV-1 evolutionary dynamics. In this study, we have evaluated the evolutionary dynamics differences of the HIV-1 *env* gene in vertical transmission. We have studied sequences from different time points of mother and child belonging to subtypes B and C, and in different phase of infection.

Materials and Methods

Study population

In this study we have included one mother and child pair from Feira de Santana, Bahia, Brazil and 3 pairs from Zambia, Africa. It were collected samples from three different time points from the mother and child Brazilian pair, that were assisted at the Sexually Transmitted Disease Reference Centre in Feira de Santana. Each patient had sequences from three different time points with the interval of, approximately, six month from one sample to the other. The samples were collected after signature of letter of informed consent.

We also studied three pairs from Zambia. The *env* sequences from these pairs (mother and child) were downloaded from the GenBank (BENSON, 2008). Each individual had viral sequences from three to five different time points with the interval of, approximately, six month from one sample to the other.

The Bahiana School of Medicine and Public Heath Ethics Committee approved this study.

Laboratory experiments

For the Brazilian pair samples PBMCs were isolated from 10mL of total blood using the Ficoll-Hypaque method. The DNA was extract from the PBMCs using QIAGEN kit (QIAamp® DNA Blood Kit). Nested PCR, using specific primers, was performed in order to obtain an HIV-1 *env* (1200pb) gene fragment. The PCR product was purified using QIAquick PCR Purification kit (QIAGEN) and then cloned by TOPO TA Cloning kit. From each sample, approximately 25 clones were selected and sequenced at the Genome Sequence Service Laboratory, University of Florida, Gainesville, Florida, USA.

Sequence analyses

The sequences generated from all time points, for each patient, were aligned using Clustal X (THOMPSON *et al*., 1997) and manually edited. For the Zambia pairs we also aligned the child sequences with the respective mother sequences from the first time point.

Hudson test was performed to check if sequences from each different time point represent different subpopulations (HUDSON, 1992).

The presence of intra-patient recombinants can disturb the tree construction. To identify the recombinants sequences, the PHI (pair-wise homoplasy índex) test (BRUEN *et al*., 2006), implemented in SplitsTree program (HUSON & BRYANT, 2006), was performed. Recombinant sequences were excluded until that the *p* value had not been significant ($p>0.05$) (SALEMI, 2008). All the recombinants detect were excluded from the dataset for posterior analyses.

Maximum Likelihood trees were generated using the online tool PhyML (GUINDON & GASCUEL, 2003) using the GTR evolution model to estimating 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 (DRUMMOND & RAMBAUT, 2007) 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 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 (DRUMMOND & RAMBAUT 2007). The

models tested 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 than, 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 (DRUMMOND & RAMBAUT 2007), the maximum clade credibility tree were selected from the posterior tree distribution after a 50% burnin, for each dataset and all trees were visualized using FigTree v1.2.2.

For the Brazilian pair sequences, which are subtype B, CTL and B-cell subtype B epitopes, previously described in Los Alamos database, were mapped (BETTE, 2006; 2007). For the Zambia pairs, subtype C, B-cell previously described subtype C epitopes were mapped. However, due to the fact that these pair are in the acute phase of infection, CTL epitopes were not checked.

Results

In this study one mother and child pair from Feira de Santana, FS16 FS17, 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. However, the three isolates pairs from Zambia were subtype C and the first sample time point were 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.

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 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 sequences datasets. In most patients it was found a great number of recombinant sequences at the last time point. The sequences from the mother MIP834_M was the one with the lower number of recombinant sequences presenting only one (3%) recombinant sequence ate the second time point.

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 mother at first time point, in the Zambia pairs, due to these sequences time points being close to the time of transmission. However the chronically infected pair showed distant relation among the mother and child sequences.

To study the HIV phylodynamics in these patients Bayesian trees were constructed taking into

account the sampling time. First, for each dataset, we used the strict molecular clock model, which fix evolutionary rate for the analyzed gene during time, and the relaxed molecular clock, that assigns different evolutionary rates for the different branches on the tree, both with the constant size prior. The Bayes Factor was calculated comparing the strict clock analyses with the relaxed clock and the relaxed molecular clock were chosen the best fitting model for all datasets.

The 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 were selected over the constant size, and the acute pairs were strongly selected while the chronic pair was moderate selected. However, the exponential growth showed strongly evidence over the constant model for the MIP834_I (child), and moderate evidence in the MIP2953_M (mother), MIP2953I+M (child + mother first time point) and FS17 (child). 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 (Table 1).

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 at the MIP2660 and MIP834 pairs (Figure 2).

Out of the 22 subtype C B-cell epitopes searched at the Zambia pairs sequences, only three of them were present in the mother and child of the MIP834 and MIP2660 pairs and one epitope at the MIP2953 pair. Most of the mutations found in the mother sequences epitope region were the same ones found in the child sequences. However, for the subtype B pair, the epitopes and mutations found were different between mother and child sequences for the CTL epitopes and B-cell epitopes.

Discussion

The results found in this study did not show any pattern that differs the mother to the child at HIV-1 dynamics. However, the trees and the BSP show that the strains from the mother are more closely related to the child strains in the acute infected mothers when transmitted the virus to their child.

The macrophages have an important role in the recombination process (LAMERS *et al.,* 2009). Since the HIV infects the macrophages during a more posterior phase of infection, it was expected to find a larger percentage of recombinants intra-patient in the chronically infected patient. However these differences were not found in these analyzes. Even though the MIP834 and MIP2660 showed a lower number of recombinants the MIP2953 pair presented similar percentages with the chronically infected pair.

Studies have shown that during chronic HIV-1 infection only a few variants of quasispecies are transmitted (WOLINSKY, 1992; ZHANG, 2006; ZHANG, 2005; DERDEYN, 2004). However, acute infected mothers transmit multiple close related variants to their child (HOFFMANN et al., 2008), which is possible to visualize in the trees constructed using the child sequences with the mother first time point. In the acute infected pairs, the first time point of the mother and their child are mixed, showing different variants giving rise to the next population, while at the chronic infected pair the mother and the child presents two different virus 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 (HOFFMANN et al., 2008).

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 (ACHAZ et al., 2004; SEO et al., 2002; BROWN, 1997; SHRINER et al., 2004). 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 moderate 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 is probably due to that in the chronic phase of infection the virus is more adapted and in this pair the immune system is more compromised.

The BSP of the acute 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 is also evidence that the chronically infected mother presents a different population from the child. Their different immune response and the long time of infection leads to this very different population, while the acute pairs presents a more similar sequences with similar epitopes between the mother and the child.

These findings show how HIV-1 intra-host population dynamics can differ depending on the phase of infection and transmission, contributing for the better understanding of the virus dynamics over time.

Acknowledgments

The authors thank the CNPq (Conselho Nacional de Desenvolvimento Científico e Tecnológico) for the financial support.

Reference

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

Benson DA, Karsch-Mizrachi I, Lipman DJ, Ostel J, Wheeler DL. 2008. GenBank. Nucleic Acids Res. 36:D25-30.

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.

Brown A J. 1997. Analysis of HIV-1 env gene sequences reveals evidence for a low effective number in the viral population. Proc Natl Acad Sci USA, 94, 1862-5.
Bruen T, Philippe H, Bryant D. 2006.A simple and robust statistical test for detecting the presence of recombination. Genetics, 172:2665-2681.

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. 2004. Science, 303:2019–2022.

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

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

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

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

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

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

Lamers SL, Salemi M, Galligan DC, et al.. 2009. Extensive HIV-1 intra-host recombination is common in tissues with abnormal histopathology. PLoS One, 4(3):e5065.

Lehman DA, Farguhar C. 2007. Biological mechanisms of vertical human immunodeficiency virus (HIV-1) transmission. Rev. Med. Virol, 17:381-403.

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

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

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

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

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

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

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

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

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

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

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

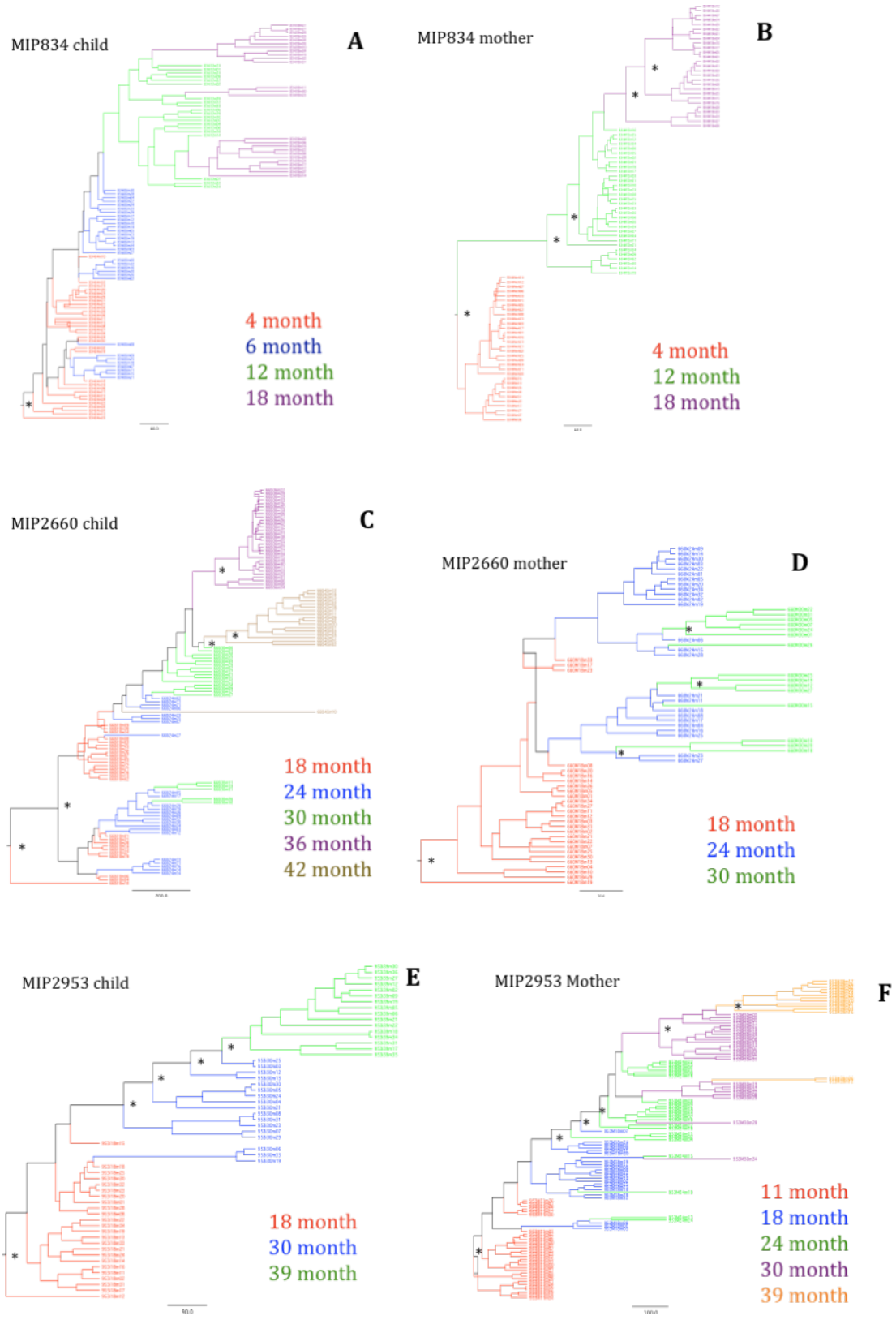
Table 1: Relaxed molecular clock comparison using Bayes Factor of the constant population size

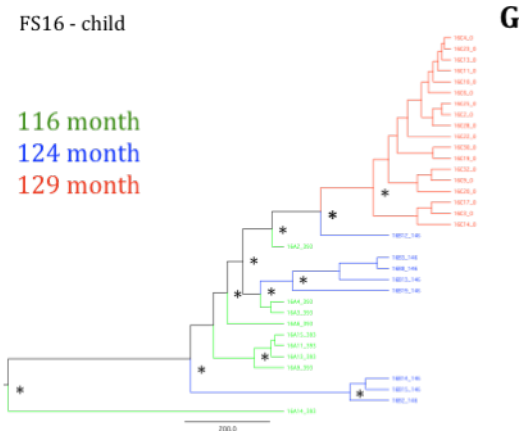
Sample	Clock	Model	Marginal Lik*	Bayes Factor	Root Height	Root Height**	Growth Rate	Growth Rate**
MIP834_I	RC	CONT	3826.945		554.55	490.13 - 648.06		
MIP834_I	RC	BSP	3817.284	19.322	468.63	449.66 - 497.08		
MIP834_I	RC	EXP	3818.659	16.574	560.88	494.78 - 655.11	4.89E-03	2.879E-3 - 7.25E-3
MIP834_M	RC	CONT	2859.562		525.64	448.40 - 634.77		
MIP834_M	RC	BSP	2841.519	36.086	481.35	449.06 - 530.86		
MIP834_M	RC	EXP	2859.037	1.05	542.41	462.54 - 659.80	2.17E-03	2.113E-4 - 4.14E-3
MIP834_I+M	RC	CONT	4269.614		550.13	491.40 - 634.32		
MIP834_I+M	RC	BSP	4252.346	34.54	474.27	452.10 - 503.95		
MIP2660_I	RC	CONT	4354.726		1053.53	837.92 - 1408.02		
MIP2660_I	RC	BSP	4345.188	19.076	832.49	766.76 - 1061.32		
MIP2660_I	RC	EXP	4354.002	1.448	1055.383	836.24 - 1357.98	9.34E-04	-3.057E-5 - 2.035E-3
MIP2660_M	RC	CONT	2738.816		594.57	461.16 - 833.29		
MIP2660_M	RC	BSP	2744.967	12.3	468.77	414.66-570.39		
MIP2660_M	RC	EXP	2737.876	1.88	583.24	470.58 - 715.23	7.90E-03	4.085E-3 - 1.253E-2
MIP2660_I+M	RC	CONT	4689.241		995.78	844.78 - 1230.25		
MIP2660_I+M	RC	BSP	4680.955	16.572	857.033	772.43 - 981.95		
MIP2953_I	RC	CONT	3275.053		844.28	726.36 - 1038.67		
MIP2953_I	RC	BSP	3119.374	311.356	763.25	698.66 - 858.17		
MIP2953_I	RC	EXP	3274.22	1.664	844.02	729.81 - 1038.23	7.22E-04	-5.48E-4 - 2.006E-3
MIP2953_M	RC	CONT	4417.711		979.87	910.18 - 1101.23		
MIP2953_M	RC	BSP	4409.993	15.436	877.09	860.29 - 900.21		
MIP2953_M	RC	EXP	4415.083	5.256	958.844	898.29 - 1062.54	2.30E-03	1.194E-3 - 3.509E-3
MIP2953_I+M	RC	CONT	3493.517		940.46	888.45 - 1026.05		
MIP2953_I+M	RC	BSP	3485.655	16.536	871.85	856.52 - 891.62		
MIP2953_I+M	RC	EXP	3491.391	5.064	940.15	882.07 - 1029.57	1.62E-03	5.378E-4 - 2.734E-3
FS16	RC	CONT	3962.271		1049.218	601.255-1662.772		
FS16	RC	BSP	3963.902	3.262	940.166	501.642-1459.25		
FS16	RC	EXP	3963.07	1.598	953.383	539.301-1416.94	1.19E-03	-7.382E-4 - 3.391E-3
FS17	RC	CONT	3592.213		1271.931	696.586-2248.081		
FS17	RC	BSP	3589.511	5.406	1297.749	603.121-2730.772		
FS17	RC	EXP	3594.902	5.378	1187.25	692.726-2032.158	6.87E-04	-9.871E-4 - 2.582E-3

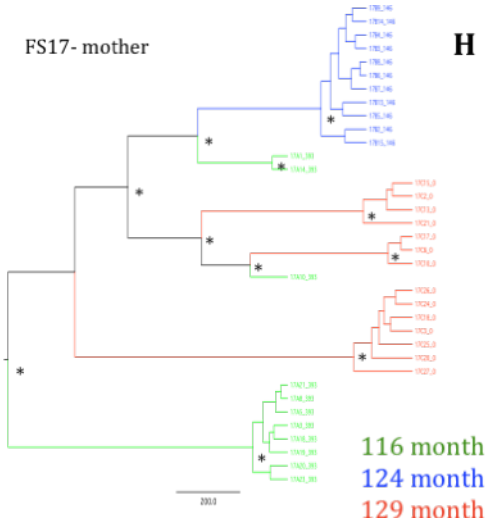
The selected models are highlighted. Marginal Likelihood (*); confidence intervals (**); RC = Relaxed Molecular Clock; CONT = Constant Growth; EXP = Exponential growth; BSP = Bayesian skyline plot; I = child; M = mother

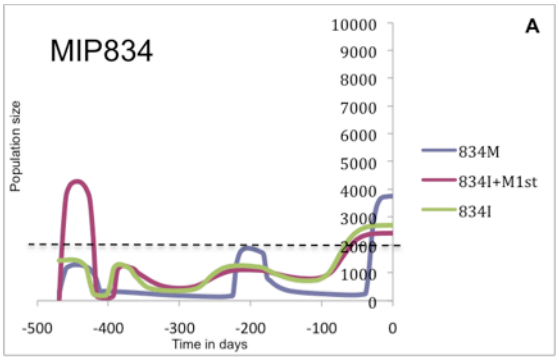
Figure 1: Bayesian maximum clade credibility phylogenetic trees. The trees were constructed using the relaxed molecular clock model and constant population size model. The branch support are giving by the posterior probability > 0.80 and are indicated with an asterisk (*). The subtitles for the meaning of each branch color of the trees are indicated in each panel. The panels in the left side are the child trees and on the right side the panels with each respective mothers tree.

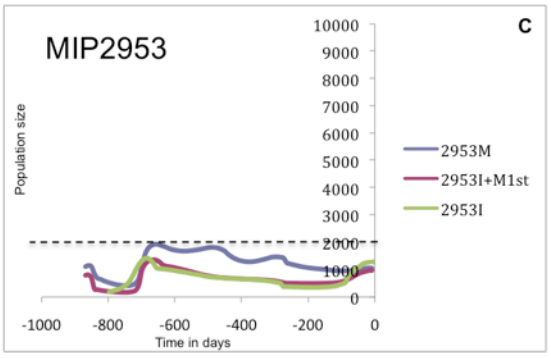
Figure 2. Bayesian Skyline Plots to estimate the effective population size. Each graphic shows the BSP mean value of the mother (M and FS17) and the child (I and FS16). For the Zambia pairs it is also found in the graphics the BSP of the child with the first time point of the mother (M+I). These analyses were performed in BEAST using all available time point for each patient.

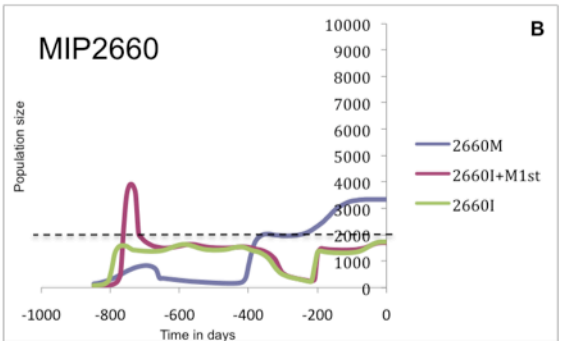

MIP834 child
A
4 month
6 month
12 month
18 month
MIP834 mother
B
4 month
12 month
18 month
MIP2660 child
C
18 month
24 month
30 month
36 month
42 month
MIP2660 mother
D
18 month
24 month
30 month
MIP2953 child
E
18 month
30 month
39 month
MIP2953 Mother
F
11 month
18 month
24 month
30 month
39 month

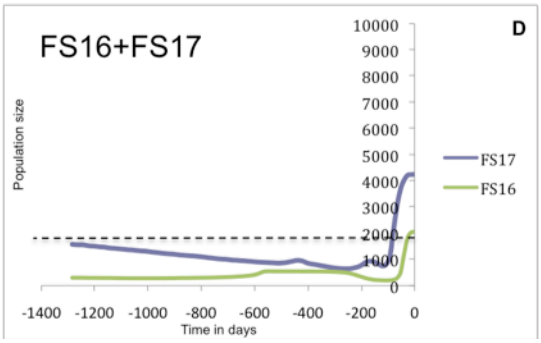

FS16 - child
G
116 month
124 month
129 month


FS17- mother
H
116 month
124 month
129 month


MIP834
A
Population size
Time in days
834M
834I+M1st
834I


MIP2953
C
Population size
Time in days
2953M
2953I+M1st
2953I


MIP2660
B
Population size
Time in days
2660M
2660I+M1st
2660I


FS16+FS17
D
Population size
Time in days
FS17
FS16

5.3 Caracterização evolutiva do genoma total do Vírus do Oeste do Nilo.

Artigo publicado: GRAY, R.R.; VERAS, N.M.C.; SANTOS, L.A.; SALEMI, M. Evolutionary characterization of the West Nile Vírus complete genome. **Molecular Phylogenetics and Evolution**, 56(1):195-200. 2010.

Molecular Phylogenetics and Evolution xxx (2010) xxx–xxx

Contents lists available at ScienceDirect

Molecular Phylogenetics and Evolution

journal homepage: www.elsevier.com/locate/ympev

Evolutionary characterization of the West Nile Virus complete genome

R.R. Gray [a,*], N.M.C. Veras [a,b], L.A. Santos [a,c], M. Salemi [a,*]

[a] Department of Pathology, Immunology, and Laboratory Medicine, University of Florida College of Medicine, Gainesville, FL 32610, USA
[b] Instituto de Biologie, Universidade de Brasilia, Brasilia, DF, Brazil
[c] Oswaldo Cruz Foundation, Salvador, BA, Brazil

ARTICLE INFO

Article history:
Received 7 October 2009
Revised 14 January 2010
Accepted 19 January 2010
Available online xxxx

Keywords:
West Nile Virus
Phylogeny
Geography
Evolution

ABSTRACT

The spatial dynamics of the West Nile Virus epidemic in North America are largely unknown. Previous studies that investigated the evolutionary history of the virus used sequence data from the structural genes (prM and E); however, these regions may lack phylogenetic information and obscure true evolutionary relationships. This study systematically evaluated the evolutionary patterns in the eleven genes of the WNV genome in order to determine which region(s) were most phylogenetically informative. We found that while the E region lacks resolution and can potentially result in misleading conclusions, the full NS3 or NS5 regions have strong phylogenetic signal. Furthermore, we show that geographic structure of WNV infection within the US is more pronounced than previously reported in studies that used the structural genes. We conclude that future evolutionary studies should focus on NS3 and NS5 in order to maximize the available sequences while retaining maximal interpretative power to infer temporal and geographic trends among WNV strains.

1. Introduction

West Nile Virus (WNV) is a mosquito-borne *Flavivirus* with a natural reservoir in birds, and incidental infection in humans and horses. The WNV genome is comprised of one long open reading frame that includes four 5′ structural genes (nucleocapsid [NC], pre-membrane [prM], membrane [M] and envelope [E]) and seven 3′ non-structural [NS] genes (NS1, NS2A, NS2B, NS3, NS4A, NS4B, and NS5). The virus was first isolated in Uganda in 1937 (Smithburn et al., 1940). During the remainder of the century the virus remained confined to the African and Asian continent (Murgue et al., 2002), and was only detected in North America in 1999 (Lanciotti et al., 1999b). Numerous studies have reconstructed the evolutionary history of the virus outside and within North America using phylogenetic analyses. Based on these studies, WNV has been grouped into two major genetic lineages (Berthet et al., 1997; Lanciotti et al., 2002). Lineage 1 contains strains from United States, Europe, Israel, India, Russia, and Australia, while the Lineage 2 contains strains from sub-Saharan Africa and Madagascar (Lanciotti et al., 2002). Lineage 1 is further divided into 3 sub-clades: 1a (including strains from Africa, Europe, US, Middle East, and Russia), 1b (Kunjin strain from Australia), and 1c (India) (Lanciotti et al., 2002; Scherret et al., 2001). Within North America, two major clades have been described: one that contains the strains isolated from the northeastern US during 1999–2000 (Lanciotti et al., 2002), and those sampled from 2002-present. The 2002-present clade, including sequences from geographically diverse areas in the US replaced the 1999–2000 one (Davis et al., 2005; Snapinn et al., 2007).

Although temporal structure is clearly evident in comprehensively sampled phylogenetic reconstructions, several studies reported a lack of geographic structure in the US based on phylogenetic analysis (Bertolotti et al., 2007, 2008; Tang et al., 2008). These studies used sequences from the structural genes (prM and E) to infer evolutionary relationships, as have many other phylogenetic studies (Anderson et al., 2001; Banet-Noach et al., 2003; Blitvich et al., 2004; Briese et al., 2002; Davis et al., 2003, 2007; Ebel et al., 2004; Lanciotti et al., 1999a; Savini et al., 2008). However, a comprehensive assessment of the reliability of each gene has not yet been reported. In fact, a lack of phylogenetically informative sites in the prM and E regions has been noted (Davis et al., 2005) and comprehensive studies that used full genome sequences from the US noted the improved resolution (Davis et al., 2005; Grinev et al., 2008; Herring et al., 2007). Interestingly, full genome sequencing has revealed geographic structure in the Mediterranean (Parreira et al., 2007) and Africa (Botha et al., 2008). However, sequencing the full genome is expensive, both from a computational and a laboratory perspective. Furthermore, for many computationally intensive analyses, utilizing the full genome is unfeasible. It would be, therefore, beneficial to use only those genomic regions that contain the highest phylogenetic signal to reduce cost without losing valuable information. This study

* Corresponding authors. Fax: +1 352 273 8284.
E-mail addresses: grayr@pathology.ufl.edu (R.R. Gray), salemi@pathology.ufl.edu (M. Salemi).

1055-7903/$ - see front matter © 2010 Elsevier Inc. All rights reserved.
doi:10.1016/j.ympev.2010.01.019

systematically evaluated the evolutionary patterns in each gene in order to determine which region(s) provided the maximal interpretative power to infer temporal and geographic relationships among WNV strains.

2. Methods

2.1. Sequence data

We downloaded all full genome sequences (104) available at Genbank as of 10/15/08 (http://www.ncbi.nlm.nih.gov/) for which the sampling year and country of collection was recorded. The dataset was divided into four non-mutually exclusive groups: full (104 sequences), North America (76 sequences), Lineage 1 (95 sequences) and Lineage 2 (9 sequences). Each of those datasets was then divided in different alignments by gene: structural (premembrane, membrane, E) and non-structural (nucleocapsid, NS1, NS2a, NS2b, NS3, NS4a, NS4b ad NS5). We created two additional concatenated datasets >2000 nucleotides for comparison: (1) NC, PM, MEM, and ENV (denoted as 5′); (2) ENV + NS1. Sequences were aligned using Clustal (Thompson et al., 1997) and manually edited.

2.2. Maximum likelihood determination of the model of evolution

The best fitting nucleotide substitution model was tested for each gene and each of the four datasets. A hierarchical likelihood ratio test in which the rate matrix, a parameter of invariable sites, and the distribution of rates were estimated on a neighbor-joining (NJ) base-tree with LogDet corrected distances (Swofford and Sullivan, 2009). The analyses were conduced with PAUP* 4.02 version (Swofford, 2002).

2.3. Likelihood mapping

To investigate the phylogenetic signal of each dataset, likelihood mapping was performed using TREE-PUZZLE program by analyzing 10,000 random quartets (Schmidt et al., 2002). According to this strategy, groups of four sequences (quartets), randomly chosen, are evaluated using maximum likelihood. For each quartet, the three possible unrooted tree topologies are weighted. The posterior weights are then plotted into a triangular surface. The fully resolved tree topologies are plotted in the three corners, which indicate the presence of tree-like phylogenetic signal, and the unresolved quartets are shown in the central region of the triangle. A star-like signal occurs when more then 30% of the dots fall within the central area (Strimmer and von Haeseler, 1997).

2.4. Bayesian MCMC phylogenetic analyses

To estimate the genealogy and the evolutionary timescale of E, NS3 and NS5 alignments using the North American dataset, we used the Bayesian framework implemented in BEAST software package version 1.4.8 (Drummond and Rambaut, 2007) under an uncorrelated log-normal relaxed clock model, the GTR + Γ model of nucleotide substitution, and a constant population size. The MCMC analysis was run until convergence with sampling every 10,000th generation. The results were visualized in Tracer v.1.4, and convergence of the Markov chain was assessed by calculating the effective sampling size (ESS) for each parameter (Drummond and Rambaut, 2007). All ESS values were >500 indicating sufficient sampling. For each dataset, the maximum clade credibility (MCC) tree, which is the tree with the largest product of posterior clade probabilities, was selected from the posterior tree distribution after 50% burnin using the program TreeAnnotator version 1.4.8. Final trees were manipulated in FigTree v.1.1.2 for display.

2.5. Evaluation of the molecular clock hypothesis

Molecular clock analyses were performed based on the E, NS3 and NS5 worldwide and North American datasets. The strict molecular clock assumes the same evolutionary rates along all branches in the tree, while the relaxed molecular clock allows different rates drawn from a specified distribution for all tree branches. These two models were compared by calculating the Bayes Factor (BF), which is the ratio of the marginal likelihoods (marginal with respect to the prior) of the two models being compared (Kass and Raftey, 1995; Suchard et al., 2001). We calculated approximate marginal likelihoods for each coalescent model via importance sampling (1000 bootstraps) using the harmonic mean of the sampled likelihoods (with the posterior as the importance distribution). The difference (in $\log_e$ space) of marginal likelihood between any two models is the $\log_e$ of the Bayes Factor, $\log_e$(BF). Evidence against the null model (i.e. the one with lower marginal likelihood) is indicated by $2 \cdot \log_e(BF) > 3$ (moderate) and >10 (strong). The calculations were performed with BEAST version 1.4.8 and Tracer v.1.4.

3. Results

3.1. Phylogenetic signal using likelihood mapping

For each gene in each of the four datasets, the phylogenetic noise was calculated using likelihood mapping analysis (Schmidt et al., 2002). Simulation studies have shown that datasets with less than 30% noise are usually reliable for phylogenetic inference (Strimmer and von Haeseler, 1997). In the analysis including all worldwide sequences (Table 1), the phylogenetic noise ranged from 21.4% (NS3) to 76% (NC). Only two individual genes (NS3 and NS5) contained <30% noise. These were also the longest genes (1857 and 2715 nt, respectively). E, the third longest gene (1503 nt), contained 32.2% noise, while prM (276 nt) contained 45.8% noise. The percentage of constant sites (54.4–61.1%), as well as the alpha value of the gamma-distribution (0.14–0.26), was similar among different genes, suggesting that the length of the region was the most important factor contributing to lower levels of phylogenetic noise. To confirm this observation, we tested the concatenated 5′ and E + NS1 regions of length 2373 and 2559 nt, respectively. 5′ also showed <30% noise (25.4%), while E + NS1 was slightly higher (30.6%).

For the Lineage 1 and North American dataset, the noise for each gene was higher, yet a similar pattern emerged as in the full dataset. In Lineage 1, NS3 and NS5 showed the lowest noise (27.5% for both), while E and prM were quite high (52.7% and 59.8%). The 5′ region was also low (24.2%), while the E + NS1 region was higher (37.6%). In the North American dataset, NS3 and NS5 again were lowest (26% and 26.7%, respectively), and E and prM much higher (40.6% and 50.9%, respectively). Interestingly, both 5′ and E + NS1 were above 30% (36.7 and 52.2%, respectively), even while being of a longer length than NS3. In Lineage 2, all genes contained <30% noise, likely due to the high diversity among these variants. The relationship between nucleotide length and phylogenetic signal was plotted for the worldwide and North American datasets (Fig. 1). The linear correlation was significant in both cases (worldwide: $R^2 = 0.64$, $p = 0.001$; North American: $R^2 = 0.58$, $p = 0.003$). These results suggest that while length is significantly correlated with phylogenetic signal, NS3 and NS5 contain the greatest phylogenetic signal for geographically reduced datasets. For the North American dataset in particular, E alone

Please cite this article in press as: Gray, R.R., et al. Evolutionary characterization of the West Nile Virus complete genome. Mol. Phylogenet. Evol. (2010), doi:10.1016/j.ympev.2010.01.019

Table 1
Determination of phylogenetic signal/noise by Likelihood Mapping analysis.

Dataset	Gene	Length	% Noise	Constant sites	Alpha
WORLDWIDE	NC	369	76.0	60.7	0.18
WORLDWIDE	PM	276	45.8	53.3	0.21
WORLDWIDE	MEM	225	70.0	55.6	0.29
WORLDWIDE	ENV	1503	32.2	54.8	0.22
WORLDWIDE	NS1	1056	49.8	55.2	0.23
WORLDWIDE	NS2a	693	44.5	54.4	0.26
WORLDWIDE	NS2b	393	52.8	61.1	0.14
WORLDWIDE	NS3	1857	21.4*	59.0	0.19
WORLDWIDE	NS4a	447	46.2	56.8	0.21
WORLDWIDE	NS4b	768	35.9	55.2	0.22
WORLDWIDE	NS5	2715	22.4*	54.4	0.24
WORLDWIDE	ENV + NS1	2559	30.6	54.9	0.22
WORLDWIDE	5′	2373	25.4*	55.6	0.20
LINEAGE 1	NC	369	85.0	73.2	0.29
LINEAGE 1	PM	276	50.9	67.0	0.24
LINEAGE 1	MEM	225	80.9	67.6	0.37
LINEAGE 1	ENV	1503	40.2	69.3	0.17
LINEAGE 1	NS1	1056	66.5	66.6	0.17
LINEAGE 1	NS2a	693	57.0	69.1	0.16
LINEAGE 1	NS2b	393	62.1	65.6	0.19
LINEAGE 1	NS3	1857	26.0*	69.4	0.14
LINEAGE 1	NS4a	447	57.9	68.2	0.17
LINEAGE 1	NS4b	768	45.8	67.4	0.23
LINEAGE 1	NS5	2715	26.7*	70.4	0.14
LINEAGE 1	ENV + NS1	2559	37.6	63.2	0.19
LINEAGE 1	5′	2373	24.2*	65.1	0.23
NORTH AMERICA	NC	369	99.8	94.9	97.9
NORTH AMERICA	PM	276	59.8	93.8	0.14
NORTH AMERICA	MEM	225	99.3	92.0	0.69
NORTH AMERICA	ENV	1503	52.7	91.0	0.70
NORTH AMERICA	NS1	1056	88.8	93.0	8.46
NORTH AMERICA	NS2a	693	75.7	93.8	0.10
NORTH AMERICA	NS2b	393	83.1	92.4	0.53
NORTH AMERICA	NS3	1857	27.5*	93.6	0.02
NORTH AMERICA	NS4a	447	62.1	93.7	0.08
NORTH AMERICA	NS4b	768	51.5	91.3	0.30
NORTH AMERICA	NS5	2715	27.5*	93.5	0.03
NORTH AMERICA	ENV + NS1	2559	52.2	93.0	0.51
NORTH AMERICA	5′	2373	36.7	93.3	0.33
LINEAGE 2	NC	369	24.6*	67.5	0.23
LINEAGE 2	PM	276	20.6*	59.4	0.25
LINEAGE 2	MEM	225	13.5*	61.3	0.19
LINEAGE 2	ENV	1503	5.6*	61.2	0.23
LINEAGE 2	NS1	1056	8.7*	62.2	0.28
LINEAGE 2	NS2a	693	17.5*	61.3	0.21
LINEAGE 2	NS2b	393	23*	68.4	0.16
LINEAGE 2	NS3	1857	7.9*	65.3	0.14
LINEAGE 2	NS4a	447	19.0*	63.1	0.22
LINEAGE 2	NS4b	768	10.3*	60.2	0.18
LINEAGE 2	NS5	2715	6.3*	64.2	0.27
LINEAGE 2	ENV + NS1	2559	2.4*	59.1	0.23
LINEAGE 2	5′	2373	4.8*	59.5	0.22

* Strong phylogenetic signal (<30% noise).

does not contain enough phylogenetic signal to infer reliable evolutionary relationships.

3.2. Model of evolution

The rate matrix, the alpha parameter in the gamma-distribution, and the proportion of invariable site (*Pinv*) were calculated for each gene (NC, prM, M, E, NS1, NS2a, NS2b, NS3, NS4a, NS4b, NS5) in the four datasets (Table S2). The estimated rate matrix was similar between genes within each dataset, particularly within the full and Lineage 1 datasets. For all datasets, the best model of evolution for the E, NS3 and NS5 genes were essentially the same (ABACDA, $+\Gamma$, suggesting that differential evolutionary patterns are not responsible for the increased signal in NS3 and NS5. Furthermore, these results indicate that a concatenated alignment of multiple genes analyzed under the same nucleotide substitution model is appropriate.

3.3. Evolutionary rate and TMRCA of the North American clade

To compare the effect of the reduced phylogenetic signal on the estimation of the evolutionary rate for the E, NS3, and NS5 genes in the North American and worldwide datasets, the strict clock (SC) and relaxed molecular clock (RC) hypotheses were compared using a Bayesian framework implemented in the program BEAST (Drummond and Rambaut, 2007) that incorporated the year of sampling for each sequence (Drummond et al., 2002). Analyses were conducted using a constant coalescent prior and the GTR + G model of evolution. For all three genes in the worldwide dataset, the SC hypothesis was strongly rejected (Bayes Factors >70) when compared to the RC model (Table 2) suggesting rate heterogeneity among branches. The mean evolutionary rate estimated under the relaxed clock assumption was one order of magnitude faster than under the strict clock assumption, and was slightly faster for E (2.01×10^{-3}) than for NS3 (1.5×10^{-3}) and NS5 (1.18×10^{-3}), although the 95% high posterior density intervals (95% HPDs) were overlapping. The TMRCA of the North American dataset was estimated for each gene/clock model (two additional sequences from Hungary [DQ118127] and Israel [AF481864] were included because for some analyses they clustered in the North American clade). Under the relaxed clock assumption, the TMRCA was similar for all genes (E: 12.8 years before present [YBP], NS3: 12.7 YBP, NS5: 10.2 YBP). The apparent inconsistency in the E gene (slightly older TMRCA and faster evolutionary rate) is likely due to the larger 95% HPDs for the TMRCA. Under the strict clock assumption, the estimated TMRCAs vary widely, likely due to the poor fit of this model.

For the North American dataset, the relaxed clock was again favored over the strict clock for all three genes (BF > 10). The mean evolutionary rate was again slightly higher for E (8.18×10^{-4}) than NS3 (7.67×10^{-4}) and NS5 (7.88×10^{-4}), but again the HPDs were overlapping. The mean rate was ~0.5 order of magnitude lower than in the worldwide dataset; however, the 95% HPDs were much larger and overlapped with those of the worldwide dataset. The estimates of the TMRCA of the North American clade (in this case, the root of the entire tree) were similar to those in the worldwide dataset. The estimate for the E gene under the relaxed clock model (8.7 YBP) was lower than NS3 (12.1 YBP) and NS5 (13.0 YBP), consistent with the faster evolutionary rate.

For both the worldwide and the North American dataset, the NS3 and NS5 genes were concatenated to determine whether HPDs would be reduced with longer sequences. However, multiple analysis with the chain run for 3×10^8 generations failed to converge (data not shown), suggesting that current computational resources are insufficient for large (in this case, a dataset with 104 taxa and >4500 nt) alignments. This result highlights the need to carefully select phylogenetically informative regions rather than using the full genome.

3.4. Phylogenies for the North American datasets

In order to determine how the topology and support for the genealogies differed between genes with strong phylogenetic signal (NS3 and NS5) and weak signal (E), we inferred the maximum clade credibility (MCC) tree for the E, NS3, and NS5 genes for the North American dataset under the relaxed molecular clock and the constant size coalescent prior (Fig. 2). The MCC is the tree with the maximum product of the posterior clade probabilities. To the extent that the posterior probabilities of different clades are additive, this definition is an estimate of the total probability of the given tree topology (http://beast.bio.ed.ac.uk/Summarizing_posterior_trees). To compare

Please cite this article in press as: Gray, R.R., et al. Evolutionary characterization of the West Nile Virus complete genome. Mol. Phylogenet. Evol. (2010), doi:10.1016/j.ympev.2010.01.019

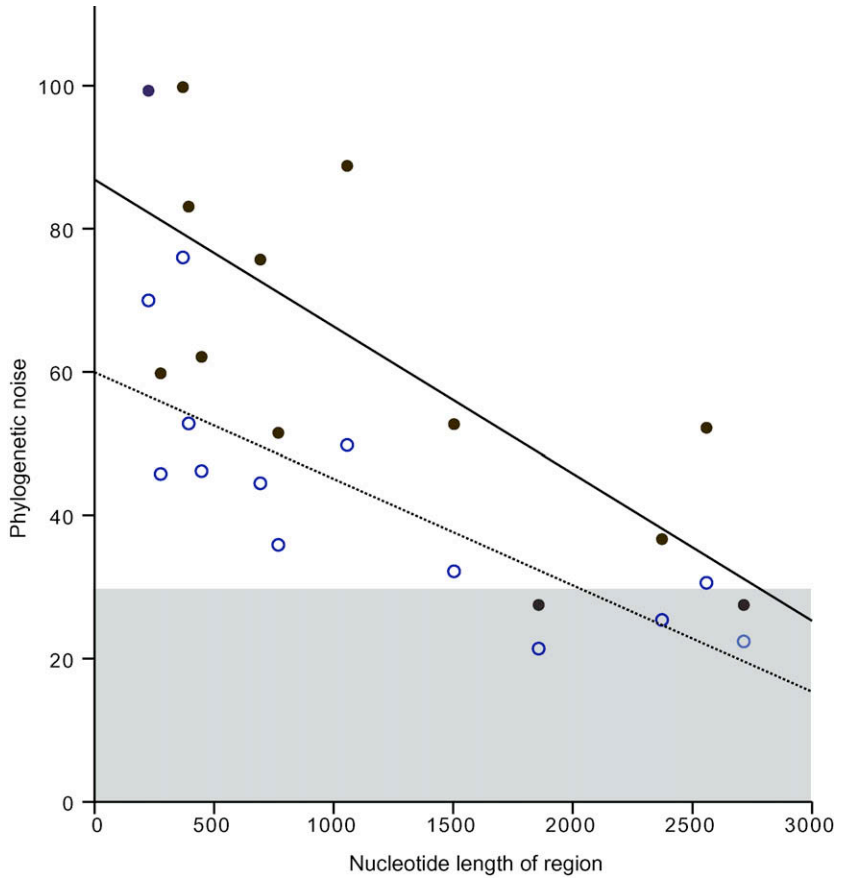


Fig. 1. Correlation between nucleotide length and phylogenetic noise. Length of each region tested for phylogenetic signal ($n = 13$) is plotted on the x axis, and phylogenetic noise as measured using Tree Quartet Puzzling is plotted on the y-axis. Open circles indicate measurements from the worldwide dataset, and filled circles indicate measurements from the North American dataset. A linear regression was performed on each dataset and the best-fit line is shown (dashed = worldwide, solid = North American). The grey filled region at the bottom of the graph denotes phylogenetic noise <30%.

Table 2
Bayesian estimation of the molecular clock, evolutionary rate, and height of the US clade.

Dataset	Gene	Clock Model	Bayes Factor	Evolutionary rate	Lower CI	Upper CI	Height of US clade	Lower CI	Upper CI
Worldwide	Env	Strict		2.11E-04	1.33E-04	2.90E-04	34.3	20.4	53.6
		Relaxed	77.6	2.01E-03	1.09E-03	3.02E-03	12.8	7.8	20.5
	NS3	Strict		2.78E-04	2.11E-04	3.50E-04	20.9	14.8	29.4
		Relaxed	81.6	1.50E-03	8.97E-04	2.20E-03	12.7	8.3	18.6
	NS5	Strict		2.37E-04	1.80E-04	3.00E-04	15.0	10.7	20.2
		Relaxed	74.4	1.18E-03	6.79E-04	1.77E-03	10.2	7.2	14.1
North American	Env	Strict		7.99E-04	5.16E-04	1.10E-03	7.8	3.9	12.2
		Relaxed	16.6	8.18E-04	5.22E-04	1.16E-03	8.7	4.5	13.5
	NS3	Strict		7.70E-04	5.43E-04	1.00E-03	11.5	6.7	17.0
		Relaxed	12.2	7.67E-04	5.31E-04	1.02E-03	12.1	7.2	17.2
	NS5	Strict		7.52E-04	5.60E-04	9.47E-04	11.4	6.9	16.34
		Relaxed	19.8	7.88E-04	5.75E-04	1.00E-03	13.0	8.1	18.3

the degree of posterior support for the MCC tree inferred for each gene, we calculated the sum and the product of the posterior support for all internal nodes of the MCC tree. Each dataset comprised the same number of taxa ($n = 76$), and each tree is a bifurcating rooted tree, so the number of internal nodes (including the root) is the same in each case ($n - 1$). NS5 had the highest sum and product of posterior clade probabilities, while E had the lowest (Table 3). This suggests that the support for the topology based on the NS5 is the highest.

In order to compare particular aspects of the topologies of each genealogy, clades with relatively high posterior support (>0.80) were assigned based on the NS5 tree, A–G (Fig. 2). "A" contains the sequences from the main 2002 "replacement" clade, while non-A sequences were sampled from the northeastern US (NY, MD, CT) from 1999 to 2001). "A" is comprised of bifurcating lineages "B" (including six sequences from NY, FL, LA, OH and MX) and "C" (including sequences sampled from 2002 to 2004 from NY, TX, CO, CA, MX, IN, GA, FL, ND, IL, WI, AZ, and NM). A sequence from NY sampled in 2000 is basal to "B" and "C". Within "C" are two well-supported clades "D" and "E". "D" contains sequences from the western US and Mexico (TX, CO, CA and MX). "E" is comprised of two clades "F" (all MX sequences sampled in 2003) and "G" (western US: AZ, CA, NM, and CO). Because the same strains were used in all analyses, the presence or absence of these clades could be evaluated on the NS3 and E trees. Clades A, B, C, D, and F were found with high posterior values on the NS3 tree

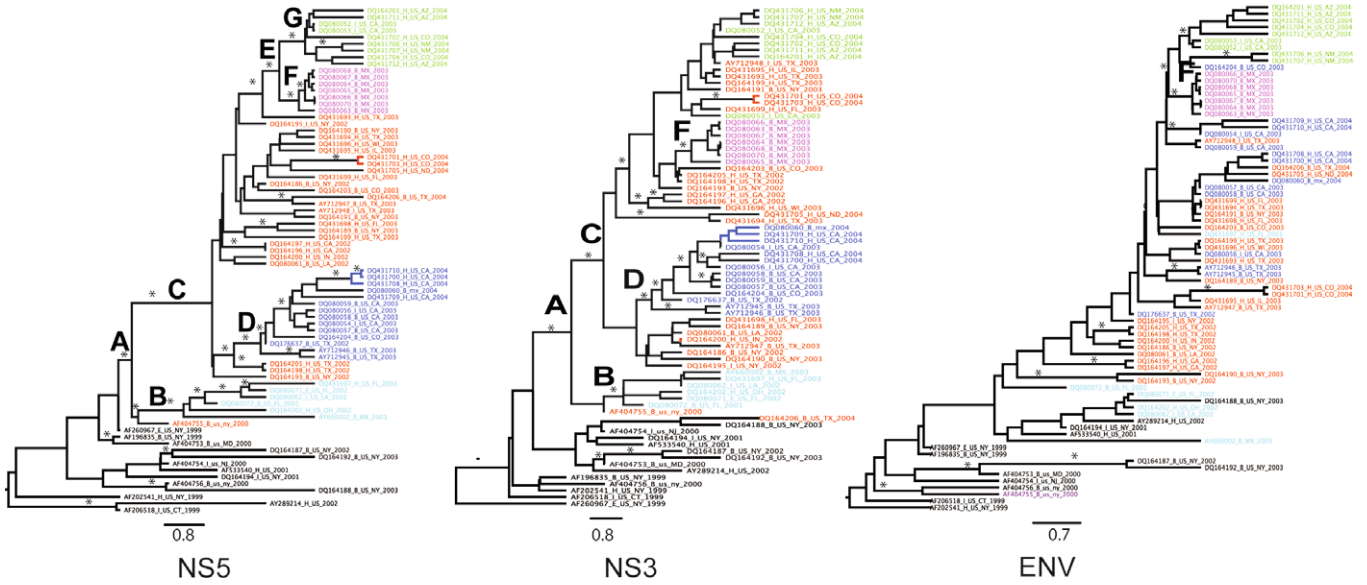


Fig. 2. Bayesian maximum clade credibility phylogenetic trees based on NS5 (left), NS3 (middle) and E (right) genes for the North American clade. The relaxed clock and constant population size model were assumed. Branch lengths are scaled in years. Branches with posterior support >0.80 are starred. Well-supported clades were assigned based on the NS5 tree and labeled A–G. Sequences within each clade are colored consistently across genes. Clades that are present in the E and NS3 trees are labeled consistently with the NS5 tree. Sequences are labeled with the accession number followed by a one-letter symbol indicating the host from which it was sampled (B, bird; H, human; I, insect), a two-letter symbol indicating the country (US or MX, Mexico), the standard two-letter code for the US state from which it was sampled, and the year. (For interpretation of color mentioned in this figure the reader is referred to the web version of the article.)

Table 3
Sum and product of the posterior probabilities for all internal nodes for major clades in the env, NS3, and NS5 Bayesian phylogenies of the North American datasets.

GENE	Sum of posterior probabilities	Product of posterior probabilities	A[a]	B	C	D	E	F	G
Env	20.52	4.85×10^{-112}	na[b]	na	na	na	na	1.0	na
NS3	25.91	1.21×10^{-78}	0.93	0.98	0.91	0.81	na	1.0	na
NS5	29.14	7.05×10^{-75}	0.81	1.0	0.99	0.94	1.0	0.99	0.98

[a] Clades A–G were assigned based on the NS5 tree and colored in Fig. 2.
[b] na, clade not present.

(Table 3). The exception was a 2004 sequence from TX, which was placed outside of the "A" clade in the NS3 tree. For the E tree, only clade "F" was present. The sequences in the other clades were either scattered throughout the tree (clades "B", "C", and "D") or contained additional sequences not present in the NS5 tree (clades "E", and "A"). Furthermore, the posterior support in general for the tree was very low, with most values for internal branches <0.1.

4. Discussion

Although many phylogenetic studies have been performed on WNV datasets, a comprehensive analysis has not been performed to determine which genes are the most appropriate for such studies. We evaluated all eleven genes separately for each of four datasets that included (1) the full dataset of all available complete genomes (worldwide); (2) only sequences in Lineage 1; (3) only Lineage 2; and (4) only North American sequences. We determined that in three of four datasets (worldwide, Lineage 1 and North America), the phylogenetic signal was >80% in the NS3 and NS5 genes. The diversity of the sequences in Lineage 2 provided sufficient signal for all genes. The rate matrix, proportion of invariable sites, and the alpha value for the gamma-distribution of rates was similar among genes for all datasets, in particular for E, NS3 and NS5, suggesting that differential evolutionary patterns among the genes was not responsible for the varying phylogenetic signal; rather the longer length of the NS3 and NS5 most likely contributes to the increased signal. However, length did not completely predict the phylogenetic signal, as two concatenated datasets contained >30% noise for the North American dataset despite the comparable length to NS5.

The lack of phylogenetic signal in E could potentially affect the estimation of the evolutionary rate and the topology of the genealogies. We found a slightly faster evolutionary rate in E as compared to NS3 and NS5 for both the worldwide and North American datasets, although the 95% HPDs overlapped. The estimate of the TMRCA for the North American clade was similar among all analyses, though slightly younger for E, and consistent with previously reported rates (Snapinn et al., 2007). We therefore conclude that the temporal signal in E is strong enough to infer the evolutionary rate with similar accuracy as the NS3 and NS5 genes. Interesting, analyses in which NS3 and NS5 were concatenated failed to converge even after 300,000,000 generations, highlighting the need to carefully select phylogenetically informative regions to represent the full genome.

To investigate how this difference affected the topology and the statistical support for the phylogeny, we compared the E, NS3, and NS5 MCC trees obtained using only the North American dataset. Similar clades with high posterior support were found for the NS3 and NS5 trees, indicating a consistent phylogenetic history and strong signal. However, the E tree was very poorly supported, and the clades present in the NS3 and NS5 trees were absent from the E tree. Furthermore, the clades present in the NS3 and NS5, but not E, trees contained sequences from specific geographic regions (i.e. west, southwest, and east), which corresponded to well-supported clades in a previous study that used full genome sequences

Please cite this article in press as: Gray, R.R., et al. Evolutionary characterization of the West Nile Virus complete genome. Mol. Phylogenet. Evol. (2010), doi:10.1016/j.ympev.2010.01.019

(Herring et al., 2007). Therefore, it is likely that the reports of no geographic structure in North America (Bertolotti et al., 2007, 2008; Tang et al., 2008) suffered from a lack of resolution in the dataset (which were based on E and prM).

Since 1999, WNV has spread from the initial point of entry in the northeastern United States to the rest of the US, Mexico, Canada and the Caribbean (Davis et al., 2005; Estrada-Franco et al., 2003; Komar et al., 2003; Lanciotti et al., 1999b). Phylogeographic inference of the routes and rates of the dissemination are important to understand in order to plan effective intervention, particularly for new and potentially much more deadly infectious diseases. Until now, only a few studies that used full genome sequences were able to define meaningful geographic structure in North America (Davis et al., 2005; Grinev et al., 2008; Herring et al., 2007). However, because of the lack of full genome sequences available, and the enormous computational time required for long sequences, comprehensive phylodynamic analyses have been unfeasible. The present study demonstrates that spatial information can be recovered from only the NS3 or NS5 genomic regions, which are shorter in length and more numerous in the database. This finding now allows the possibility for landscape phylodynamic analysis, which incorporates phylogenetics and geographic information system (GIS) frameworks (Gray et al., 2009), to provide a holistic interpretation of the causes and consequences of epidemics.

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at doi:10.1016/j.ympev.2010.01.019.

References

Anderson, J., Vossbrinck, C., Andreadis, T., Iton, A., Beckwith, W.r., Mayo, D., 2001. A phylogenetic approach to following West Nile virus in Connecticut. Proc. Natl. Acad. Sci. USA 98, 12885–12889.

Banet-Noach, C., Malkinson, M., Brill, A., Samina, I., Yadin, H., Weisman, Y., Pokamunski, S., King, R., Deubel, V., Stram, Y., 2003. Phylogenetic relationships of West Nile viruses isolated from birds and horses in Israel from 1997 to 2001. Virus Genes 26, 135–141.

Berthet, F.X., Zeller, H.G., Drouet, M.T., Rauzier, J., Digoutte, J.P., Deubel, V., 1997. Extensive nucleotide changes and deletions within the envelope glycoprotein gene of Euro-African West Nile viruses. J. Gen. Virol. 78 (Pt 9), 2293–2297.

Bertolotti, L., Kitron, U., Goldberg, T.L., 2007. Diversity and evolution of West Nile virus in Illinois and the United States, 2002–2005. Virology 360, 143–149.

Bertolotti, L., Kitron, U., Walker, E., Ruiz, M., Brawn, J., Loss, S., Hamer, G., Goldberg, T., 2008. Fine-scale genetic variation and evolution of West Nile virus in a transmission "hot spot" in suburban Chicago, USA. Virology 374, 381–389.

Blitvich, B., Fernández-Salas, I., Contreras-Cordero, J., Loroño-Pino, M., Marlenee, N., Díaz, F., González-Rojas, J., Obregón-Martínez, N., Chiu-García, J., Black, W.t., Beaty, B., 2004. Phylogenetic analysis of West Nile virus, Nuevo Leon State, Mexico. Emerg. Infect Dis. 10, 1314–1317.

Botha, E., Markotter, W., Wolfaardt, M., Paweska, J., Swanepoel, R., Palacios, G., Nel, L., Venter, M., 2008. Genetic determinants of virulence in pathogenic lineage 2 West Nile virus strains. Emerg. Infect Dis. 14, 222–230.

Briese, T., Rambaut, A., Pathmajeyan, M., Bishara, J., Weinberger, M., Pitlik, S., Lipkin, W., 2002. Phylogenetic analysis of a human isolate from the 2000 Israel West Nile virus epidemic. Emerg. Infect. Dis. 8, 528–531.

Davis, C., Beasley, D., Guzman, H., Raj, R., D'Anton, M., Novak, R., Unnasch, T., Tesh, R., Barrett, A., 2003. Genetic variation among temporally and geographically distinct West Nile virus isolates, United States, 2001, 2002. Emerg. Infect Dis. 9, 1423–1429.

Davis, C., Li, L., May, F., Bueno, R.J., Dennett, J., Bala, A., Guzman, H., Quiroga-Elizondo, D., Tesh, R., Barrett, A., 2007. Genetic stasis of dominant West Nile virus genotype, Houston, Texas. Emerg. Infect Dis. 13, 601–604.

Davis, C.T., Ebel, G.D., Lanciotti, R.S., Brault, A.C., Guzman, H., Siirin, M., Lambert, A., Parsons, R.E., Beasley, D.W., Novak, R.J., Elizondo-Quiroga, D., Green, E.N., Young, D.S., Stark, L.M., Drebot, M.A., Artsob, H., Tesh, R.B., Kramer, L.D., Barrett, A.D., 2005. Phylogenetic analysis of North American West Nile virus isolates, 2001–2004: evidence for the emergence of a dominant genotype. Virology 342, 252–265.

Drummond, A.J., Nicholls, G.K., Rodrigo, A.G., Solomon, W., 2002. Estimating mutation parameters, population history and genealogy simultaneously from temporally spaced sequence data. Genetics 161, 1307–1320.

Drummond, A.J., Rambaut, A., 2007. BEAST: Bayesian evolutionary analysis by sampling trees. BMC Evol. Biol. 7, 214.

Ebel, G., Carricaburu, J., Young, D., Bernard, K., Kramer, L., 2004. Genetic and phenotypic variation of West Nile virus in New York, 2000–2003. Am. J. Trop. Med. Hyg. 71, 493–500.

Estrada-Franco, J., Navarro-Lopez, R., Beasley, D., Coffey, L., Carrara, A., Travassos da Rosa, A., Clements, T., Wang, E., Ludwig, G., Cortes, A., Ramírez, P., Tesh, R., Barrett, A., Weaver, S., 2003. West Nile virus in Mexico: evidence of widespread circulation since July 2002. Emerg. Infect Dis. 9, 1604–1607.

Gray, R., Tatem, A., Lamers, S., Hou, W., Laeyendecker, O., Serwadda, D., Sewankambo, N., Gray, R., Wawer, M., Quinn, T., Goodenow, M., Salemi, M., 2009. Spatial phylodynamics of HIV-1 epidemic emergence in east Africa. AIDS.23, F9–F17.

Grinev, A., Daniel, S., Stramer, S., Rossmann, S., Caglioti, S., Rios, M., 2008. Genetic variability of West Nile virus in US blood donors, 2002–2005. Emerg. Infect. Dis. 14, 436–444.

Herring, B., Bernardin, F., Caglioti, S., Stramer, S., Tobler, L., Andrews, W., Cheng, L., Rampersad, S., Cameron, C., Saldanha, J., Busch, M., Delwart, E., 2007. Phylogenetic analysis of WNV in North American blood donors during the 2003–2004 epidemic seasons. Virology 363, 220–228.

Kass, R., Raftey, A., 1995. Bayes Factors. J. Am. Stat. Assoc. 90, 773–795.

Komar, O., Robbins, M., Klenk, K., Blitvich, B., Marlenee, N., Burkhalter, K., Gubler, D., Gonzálvez, G., Peña, C., Peterson, A., Komar, N., 2003. West Nile virus transmission in resident birds, Dominican Republic. Emerg. Infect. Dis. 9, 1299–1302.

Lanciotti, R., Roehrig, J., Deubel, V., Smith, J., Parker, M., Steele, K., Crise, B., Volpe, K., Crabtree, M., Scherret, J., Hall, R., MacKenzie, J., Cropp, C., Panigrahy, B., Ostlund, E., Schmitt, B., Malkinson, M., Banet, C., Weissman, J., Komar, N., Savage, H., Stone, W., McNamara, T., Gubler, D., 1999a. Origin of the West Nile virus responsible for an outbreak of encephalitis in the northeastern United States. Science 286, 2333–2337.

Lanciotti, R.S., Ebel, G.D., Deubel, V., Kerst, A.J., Murri, S., Meyer, R., Bowen, M., McKinney, N., Morrill, W.E., Crabtree, M.B., Kramer, L.D., Roehrig, J.T., 2002. Complete genome sequences and phylogenetic analysis of West Nile virus strains isolated from the United States, Europe, and the Middle East. Virology 298, 96–105.

Lanciotti, R.S., Roehrig, J.T., Deubel, V., Smith, J., Parker, M., Steele, K., Crise, B., Volpe, K.E., Crabtree, M.B., Scherret, J.H., Hall, R.A., MacKenzie, J.S., Cropp, C.B., Panigrahy, B., Ostlund, E., Schmitt, B., Malkinson, M., Banet, C., Weissman, J., Komar, N., Savage, H.M., Stone, W., McNamara, T., Gubler, D.J., 1999b. Origin of the West Nile virus responsible for an outbreak of encephalitis in the northeastern United States. Science 286, 2333–2337.

Murgue, B., Zeller, H., Deubel, V., 2002. The ecology and epidemiology of West Nile virus in Africa, Europe and Asia. Curr. Top Microbiol. Immunol. 267, 195–221.

Parreira, R., Severino, P., Freitas, F., Piedade, J., Almeida, A., Esteves, A., 2007. Two distinct introductions of the West Nile virus in Portugal disclosed by phylogenetic analysis of genomic sequences. Vector Borne Zoonotic Dis. 7, 344–352.

Savini, G., Monaco, F., Calistri, P., Lelli, R., 2008. Phylogenetic analysis of West Nile virus isolated in Italy in 2008. Eur. Surveill. 13.

Scherret, J.H., Poidinger, M., Mackenzie, J.S., Broom, A.K., Deubel, V., Lipkin, W.I., Briese, T., Gould, E.A., Hall, R.A., 2001. The relationships between West Nile and Kunjin viruses. Emerg. Infect. Dis. 7, 697–705.

Schmidt, H., Strimmer, K., Vingron, M., von Haeseler, A., 2002. TREE-PUZZLE: maximum likelihood phylogenetic analysis using quartets and parallel computing. Bioinformatics 18, 502–504.

Smithburn, K.C., Hughes, T.P., Burke, A.W., Paul, J.H., 1940. A neurotropic virus isolated from the blood of a Native Ugandan. Am. J. Trop. Med. Hyg. 20, 471–492.

Snapinn, K.W., Holmes, E.C., Young, D.S., Bernard, K.A., Kramer, L.D., Ebel, G.D., 2007. Declining growth rate of West Nile virus in North America. J. Virol. 81, 2531–2534.

Strimmer, K., von Haeseler, A., 1997. Likelihood-mapping: a simple method to visualize phylogenetic content of a sequence alignment. Proc. Natl. Acad. Sci. USA 94, 6815–6819.

Suchard, M., Weiss, R., Sincheimer, J., 2001. Bayesian selection of continuous-time Markov chain evolutionary models. Mol. Biol. Evol. 18, 1001–1013.

Swofford, D., 2002. PAUP*: Phylogenetic Analysis Using Parsimony (*and other methods).

Swofford, D., Sullivan, J., 2009. Phylogeny inference based on parsimony and other methods with PAUP*. In: Lemey, P., Salemi, M., Vandamme, A. (Eds.), The Phylogenetic Handbook: A Practical Approach to Phylogenetic Analysis and Hypothesis Testing. Cambridge University Press, New York.

Tang, Y., Liu, B., Hapip, C., Xu, D., Fang, C., 2008. Genetic analysis of West Nile virus isolates from US blood donors during 2002–2005. J. Clin. Virol. 43, 292–297.

Thompson, J.D., Gibson, T.J., Plewniak, F., Jeanmougin, F., Higgins, D.G., 1997. The CLUSTAL_X windows interface. Flexible strategies for multiple sequence alignment aided by quality analysis tools. Nucleic Acids Res. 25, 4876–4882.

Please cite this article in press as: Gray, R.R., et al. Evolutionary characterization of the West Nile Virus complete genome. Mol. Phylogenet. Evol. (2010), doi:10.1016/j.ympev.2010.01.019

6. DISCUSSÃO

A utilização das ferramentas de bioinformática tem crescido a cada dia, e com ela, o surgimento de novos e mais eficientes métodos e programas para responder diversas perguntas biológicas. Neste trabalho foi demonstrando o uso de algumas destas ferramentas para a melhor caracterização da diversidade, epidemiologia molecular, dinâmica populacional e determinação das relações temporal e geográfica dos vírus.

Antes de iniciar as análises filogenéticas é muito importante a escolha da região genômica que será utilizada. Esta região tem que fornecer informação suficiente para a que árvore filogenética seja o mais próximo possível da relação real das sequências. Para isto, dois fatores podem influenciar na escolha desta região: o tamanho da sequência e a sua diversidade genética. Uma sequência muito pequena irá fornecer pouca informação porém, o uso de sequências muito grandes pode aumentar muito o tempo da análise. Além disto, sequências muito conservadas também não fornecem informações suficientes, enquanto uma alta diversidade pode levar um chamado *ruído filogenético* muito alto, não sendo possível relacionar a história evolutiva dessas sequências.

Neste trabalho, o genoma completo do WNV foi utilizado para identificar a melhor região para ser utilizada, a que apresenta um melhor sinal filogenético. Foi encontrado que as regiões NS3 (1857nt) e NS5 (2715nt) apresentaram sinal filogenético superior a 70% em todos os grupos de sequências estudados (linhage 1, linhage 2, America do Norte e mundo). Estas regiões referem-se aos dois maiores genes virais sendo este um dos fatores que explicam este bom sinal. A região E (envelope) (1503nt) que é a região mais usada em diversos estudos (ANDERSON *et al*., 2001; BLITVICH *et al.,* 2004; DAVIS *et al*., 2003, 2007; LANCIOTTI *et al*., 1999; BERTOLOTTI *et al*., 2007, 2008; TANG *et al.*, 2008), é o terceiro maior gene, porém não foi considerada uma região adequada para ser usada apresentando um sinal filogenético inferior a 70%. Além disto, análises com genes concatenados demonstraram que não é apenas o tamanho da sequência que é importante, mas também a sua diversidade. Dos diversos grupos de sequências estudadas, as pertencentes a linhagem 2 do WNV apresentaram bom sinal filogenético para todas as regiões mostrando que a diversidade desta linhagem foi suficiente para obter o sinal.

Para determinar a qualidade das análises não é apenas a escolha da melhor região genômica, mas também o uso de ferramentas, métodos e modelos mais adequados para o conjunto de dados, que irão garantir a veracidade dos resultados encontrados. Neste trabalho, a hipótese do relógio molecular foi testada para identificar qual o modelo mais adequado, estrito ou relaxado, para cada conjunto de dados nas análises do WNV e da filodinâmica do HIV-1. A seleção do melhor modelo possibilita a inferência de como a história evolutiva de cada vírus ocorre ao longo do tempo. A hipótese do relógio molecular relaxado foi selecionado para todos os conjuntos de dados sugerindo uma heterogeneidade da taxa evolutiva no gene estudado ao longo dos ramos da árvore.

Árvores foram construídas utilizando as regiões do genoma do WNV com melhor sinal filogenético, NS3 e NS5, e com a região mais utilizada, E (envelope), utilizando o modelo do relógio molecular relaxado. A topologia dessas árvores mostrou que os clados similares, com um alto suporte da probabilidade posterior, foi encontrado nas árvores NS3 e NS5 demonstrando uma história filogenética consistente. Já a árvore construída com a região E apresentou um baixo suporte dos clados e não foi possível encontrar os clados encontrados nas árvores NS3 e NS5. Além disto, estes clados encontrados em NS3 e NS5 contém sequências de regiões geográficas especificas que corresponde a clados encontrados em um estudo anterior usando sequências do genoma completo (HERRING *et al*., 2007). Estes resultados mostram que provavelmente, a não estrutura geográfica nas sequências da América do Norte encontrada em outros estudos (BERTOLOTTI *et al*., 2007, 2008; TANG *et al.*, 2008) se deve a escolha da região genômica inadequada (baseados nas regiões E e prM).

Em análises com sequências de pares de mãe e filho infectados pelo HIV-1, nas fases aguda e crônica, de amostras de tempos diferentes foi possível observar a dinâmica da população viral intra-paciente. Na topologia das árvores Bayesianas, construídas utilizando o modelo do relógio molecular relaxado, não foi possível observar uma diferença entre a estrutura da mãe e a encontrada no filho. Porém, quando observamos o crescimento populacional e o padrão do tamanho da população efetiva, ao longo do tempo, foi possível concluir que as sequências provenientes de pares em fase crônica da infecção tem um crescimento mais constante. Em contrapartida, nas sequências dos pares na fase aguda da infecção se observa uma dinâmica das populações virais, com eventos de "gargalo de garrafa" (*bottleneck*), provavelmente devido à pressão do sistema imune e a não adaptação destes vírus.

Ferramentas de bioinformática também são de grande utilidade na detecção de sequências recombinantes, principalmente em retrovírus. O HIV-1 tem uma frequência de recombinação de $2{,}4x10^{-4}$ por nucleotídeo, por ciclo de replicação. Isto equivale a cerca de 2-3 eventos de recombinação, por ciclo de replicação, e pode ocorrer por todo o genoma (JETZ *et al*., 2000). Esta recombinação pode ocorrer entre diferentes genomas do mesmo subtipo ou de subtipos diferentes, recombinantes inter-subtipos, que infectam uma mesma célula, simultaneamente, ou até mesmo entre cepas diferentes do mesmo indivíduo gerando recombinantes intra-paciente (SUBBARAO; SCHOCHETMAN, 1996). A presença do genoma mosaico de recombinantes pode perturbar a construção da árvore filogenética por apresentar regiões com história evolutiva diferente (SCHIERUP & HEIN, 2000). Diante do exposto, é importante a detecção destes recombinantes antes de iniciar as análises. Nas análises, realizadas neste trabalho, avaliando a dinâmica do HIV-1, intra-paciente, com amostras de diferentes tempos foi utilizado o teste estatístico de PHI, que detecta recombinantes de sequências de relação próxima (SALEMI, 2008). Foi encontrada sequências recombinantes variando de 0% a 57,9% em cada tempo. A detecção e exclusão dessas sequências permitiram a melhor representação filogenética.

A identificação de recombinantes inter-subtipos é importante para traçar o perfil epidemiológico das cepas circulantes no mundo, além de fornecer informação quanto a diversidade viral por ser um evento responsável pelo aumento desta diversidade. Para identificar esses recombinantes são utilizados métodos de comparação com sequências referências dos diversos subtipos ou por inferência filogenética. Neste trabalho, foi realizada a caracterização molecular do gene *pol* do HIV na cidade de Salvador. Das 57 sequências estudadas, nove (15,8%) delas apresentaram recombinação dentro do gene *pol* e três (5,3%) apresentaram recombinação intergênica, com a subtipagem dos genes *gag* e *env* realizado em outro estudo (ARAUJO, 2010). Das três recombinantes intergênicas, uma foi caracterizada com subtipo B em *pol* e F1 em *env* ($B_{pol}/F1_{env}$) e os outros dois foram BF em *gag*, F1 em *pol* e B em *env* ($BF_{gag}/F1_{pol}/B_{env,}$). As três sequências foram caracterizadas com subtipo puro em *pol,* porém eram recombinantes, mostrando a importância da utilização de múltiplos genes na subtipagem de uma amostra e que o número de recombinantes circulantes pode está subestimado devido a maior parte dos estudos utilizarem apenas uma região genômica. Analisando o perfil dos recombinantes intragênicos encontramos seis (10,5%) sequências relacionadas com o CRF28/CRF29, duas (3,5%) com o CRF12 e uma (1,8%) com CRF39. O mosaico da região *pol* apresentou pontos de

recombinação semelhantes aos CRFs, porém quando observado subtipo da região gag e/ou *env*, o subtipo encontrado não era o mesmo da CRF com exceção da sequência BR42BA que apresentou o mesmo perfil de recombinação em *pol* e o mesmo subtipo de *gag* (F1), quando comparados à CRF39. Isto mostra que diferentes recombinantes BF estão co-circulando na cidade de Salvador, sob diferentes eventos de recombinação, aumentando a diversidade do HIV-1 na região.

A filogenia é importante também para detecção dos subtipos circulantes em uma região, além da identificação de recombinantes, traçando o perfil molecular do vírus em uma determinada região. Foi identificado que entre as 57 sequências analisadas neste estudo 44 (77,2%) pertencem ao subtipo B, 12 (21,0%) recombinantes BF e uma (1,8%) do subtipo F1 concordando com a epidemiologia da região nordeste. Conhecer o perfil molecular das sequências que circulam em cada região geográfica é importante para conhecer as características da epidemia e identificar a migração de diferente genótipos para outras regiões. Além disto, também é importante traçar as mutações que levam a resistência aos ARVs circulantes na região e se estas mutações estão sendo transmitidas, auxiliando estratégias de terapia e de saúde pública.

Neste trabalho, ao utilizarmos diversas ferramentas de bioinformática, em diferentes áreas de atuação, como o uso de ferramentas na construção filogenética e outras diferentes aplicações, foi possível construir um amplo espectro de alcance dessa ciência na obtenção de respostas para diversas perguntas biológicas.

7. CONCLUSÃO

As ferramentas de bioinformática são de grande utilidade para análises de sequências virais em estudos de epidemiologia molecular do HIV-1, filodinâmica das populações virais e das suas relações temporais e geográficas.

7.1 Caracterização Molecular do Gene *pol* do HIV-1 de Indivíduos Infectados de Salvador, Bahia, Brasil.

A epidemia do HIV-1 em Salvador apresenta uma alta diversidade genética apesar de ser sido encontrada predominância do subtipo B, seguidos de diferentes formas recombinantes BF, CRF_39 e subtipo F1.

Alta freqüência de eventos de recombinação entre os subtipos B e F, circulantes em cidade de Salvador, onde 5 padrões de recombinação, 2 intergênica e 3 intragênica, foram observados, mostrando uma alta variabilidade.

As sequências estudadas apresentaram uma alto número de mutações associadas à resistência aos antiretrovirais, mostrando a importância do monitoramento.

7.2 Avaliação filodinâmica de isolados do HIV-1 na transmissão materno-fetal.

O perfil da dinâmica do HIV-1 nos pares mãe-filho foi bem semelhante, não tendo sido encontradas diferenças na dinâmica evolutiva.

O par cronicamente infectado apresentou um perfil mais constante de crescimento

populacional, enquanto que os pares na fase aguda da infecção apresentam eventos de gargalo de garrafa, visualizados nas árvores e BSP, sugerindo uma dinâmica maior devido a processos de seleção.

7.3 Caracterização evolutiva do genoma total do Vírus do Oeste do Nilo.

As regiões NS3 e NS5 do genoma do WNV apresentaram melhor sinal filogenético sendo as regiões que podem fornecer mais informação para inferir relações geográficas e temporais entre as diferentes cepas que circulam no mundo.

REFERÊNCIAS BIBLIOGRÁFICAS

ANDERSON, J.; VOSSBRINCK, C.; ANDREADIS, T.; ITON, A.; BECKWITH, W.R.; MAYO, D. A. phylogenetic approach to following West Nile virus in Connecticut. **Proc. Natl. Acad. Sci.** USA 98, 12885–12889, 2001.

ARAUJO, A. F.; BRITES, C.; MONTEIRO, J.P.; SANTOS, L.A.; GALVAO–CASTRO, B., ALCANTARA, L.C.J. Lower Prevalence of Human Immunodeficiency Virus Type 1 Brazilian Subtype B Found in Northeastern Brazil with Slower Progression to AIDS. **AIDS RESEARCH AND HUMAN RETROVIRUSES.** *In press.* 2010.

AVISE, J. **Phylogeography: The History and Formation of Species**. USA: President and Fellows of Harvard College. 2000. ISBN 0-674-66638-0.

BARRE-SINOUSI, F; CHERMAN, JC; REY, F; NUGEYRE, MT; CHAMARET, S. *et al*.. Isolation of a T-lymphotropic retrovirus from a patient at risk for acquired immune deficiency syndrome (AIDS). **Science**, [s.l.], v. 200, p. 868-871, 1983.

BENSON, D.A.; KARSCH-MIZRACHI, I.; LIPMAN, D.J.; OSTEL, J.; WHEELER, D.L. GenBank. **Nucleic Acids Res**, 36:D25-30, 2008.

BERTHET, F.X.; ZELLER, H.G.; DROUET, M.T.; RAUZIER, J.; DIGOUTTE, J.P.; DEUBEL, V. Extensive nucleotide changes and deletions within the envelope glycoprotein gene of Euro-African West Nile viruses. **J. Gen. Virol**., 78 (Pt 9), 2293–2297, 1997.

BERTOLOTTI, L.; KITRON, U.; GOLDBERG, T.L. Diversity and evolution of West Nile virus in Illinois and the United States, 2002–2005**. Virology** 360, 143–149, 2007.

BERTOLOTTI, L., KITRON, U., WALKER, E., RUIZ, M., BRAWN, J., LOSS, S., HAMER, G., GOLDBERG, T. Fine-scale genetic variation and evolution of West Nile virus in a transmission ‘‘hot spot” in suburban Chicago, USA. **Virology,** 374, 381–389. 2008.

BETTE; KORBER, T.; BRANDER, M.; HAYNES, B.F.; KOUP, R.; MOORE, J.P.; WALKER, B.D.; WATKINS, D.I. **HIV Molecular Immunology 2006/2007**, Publisher: Los Alamos National Laboratory, Theoretical Biology and Biophysics, Los Alamos, New Mexico. LA-UR 07-4752.

BLITVICH, B.; FERNÁNDEZ-SALAS, I.; CONTRERAS-CORDERO, J.; LOROÑO-PINO, M.; MARLENEE, N.; DÍAZ, F. *et al.*. Phylogenetic analysis of West Nile virus, Nuevo Leon State, Mexico. **Emerg. Infect Dis**., 10, 1314–1317, 2004.

BRUEN, T.; PHILIPPE, H.; BRYANT, D. A simple and robust statistical test for detecting the presence of recombination. **Genetics**, 172:2665-2681, 2006.

CARIDE, E.; BRINDEIRO, R. ; HERTOGS, K. ; LARDER, B. ; DEHERTOGH, P. ; MACHADO, E. *et al*.. Drug resistance revers transcriptase genotyping and fenotyping B and non B subtypes (F and A) of human Imunodeficiency vírus type 1 found in brazilian patients falling HAART. **Virology,** [s.l.], v. 275, p. 107-115, 2001.

CDC. Vírus do Oeste do Nilo. Disponível em: <http://www.cdc.gov/ncidod/dvbid/westnile/portuguese/index.htm>

CHAN, D.C.; KIM, O. S. HIV entry and it's inhibition. **Cell**, 93:681-684, 1998.

COFFIN, J. M. HIV population dynamics *in vivo*: Implictions for genetics variation, pathogenesis, and therapy. **Science**, 267:483-489, 1995.

COHEN, O. J.; KINTER, A.; FAUCI, A. S. Host factors in the pathogenesis of HIV disease. **Immunol Rev**, 159:31-48, 1997.

DAVIS, C.; BEASLEY, D.; GUZMAN, H.; RAJ, R.; D'ANTON, M.; NOVAK, R.; UNNASCH, T.; TESH, R.; BARRETT, A. Genetic variation among temporally and geographically distinct West Nile virus isolates, United States, 2001, 2002. **Emerg. Infect Dis.**, 9, 1423–1429, 2003.

DAVIS, C.; LI, L.; MAY, F.; BUENO, R.J.; DENNETT, J.; BALA, A.; GUZMAN, H. *et al*.. Genetic stasis of dominant West Nile virus genotype, Houston, Texas. **Emerg. Infect Dis**., 13, 601–604, 2007.

DAVIS, C.T.; EBEL, G.D.; LANCIOTTI, R.S.; BRAULT, A.C.; GUZMAN, H.; SIIRIN, M. *et al.*. Phylogenetic analysis of North American West Nile virus isolates, 2001– 2004: evidence for the emergence of a dominant genotype. **Virology**, 342, 252– 265, 2005.

DE OLIVEIRA, T.; DEFORCHE, K.; CASSOL, S.; SALMINEN, M.; PARASKEVIS, D.; SEEBREGT, C. *et al*. An automated genotyping system for analysis of HIV-1 and other microbial sequences. **Bioinformatics**, 21(19):3797-800, 2005.

DOURADO, I.; MILROY, C.A.; MELLO, M.A.; FERRARO, G.A.; CASTRO-LIMA FILHO, H. *et al*.. HIV-1 seroprevalence in the general population of Salvador, Bahia State, Northeast Brazil. **Cadernos de Saúde Pública**, 23(1):25–32, 2007.

DRUMMOND, A. J.; RAMBAUT, A. BEAST: Bayesian evolutionary analysis by sampling trees. **BMC Evol Biol**, 7**:**214. 2007.

DRUMMOND, A.J.; RAMBAUT, A.; SHAPIRO, B. e PYBUS, O.G. Bayesian coalescent inference of past population dynamics from molecular sequences. **Molecular Biology Evolution**, 22**,** 1185-92, 2005.

FAUCI, A. S. Multifactorial nature of human immunodeficiency vírus disease: implications for therapy. **Science**, 262:1011-1018, 1993.

FELSENSTEIN, J. PHYLIP: phylogenetic inferencepackage (version 3.2). **Cladistics**, 5:164-166, 1989.

GREENE, W. C.; PETERLIN, B. M. Charting HIV's remarkable voyage through the cell: Basic science as a passport to future therapy. **Nat Med**, 8:673-680, 2002.

GRENFELL, B.; PYBUS, O.G.; GOG, J. WOOD, J.L.; DALY, J.M.; MUMFORD, J.A.; HOLMES, E.C. Unifying the epidemiological and evolutionary dynamics of pathogens. **Science**, 303:327-32, 2004.

GUIMARAES, M. L.; BASTOS, F. I.; TELLES, P. R.; GALVAO-CASTRO, B.; DIAZ, R. S. *et al*. Retrovirus infections in a sample of injecting drug users in Rio de Janeiro City, Brazil: prevalence of HIV-1 subtypes, and co-infection with HTLVI/II. **J Clin Virol**, 21:143-151, 2001.

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

HERRING, B.; BERNARDIN, F.; CAGLIOTI, S.; STRAMER, S.; TOBLER, L.; ANDREWS, W. *et al*.. Phylogenetic analysis of WNV in North American blood donors during the 2003–2004 epidemic seasons. **Virology**, 363, 220–228, 2007.

HOFFMANN, F.G.; HE, X.; WEST, J.T. *et al*.. Genetic variation in mother-child acute seroconverter pair from Zambia. **AIDS**, 22:817-824, 2008.

HOLMES, E. C. The phylogeography of human viruses. **Molecular Ecology**, 13 (4): 745–756, 2004.

HOPE, T. J. Viral RNA export. **Chem Biol**, 4:335-44, 1997.

HUDSON, R.R.; BOOS, D.D.; KAPLAN, N.L. A statistical test for detecting geographic subdivision. **Molecular Biology Evolution**, 9(1):138-51, 1992.

HUSON, D.; BRYANT, D. Application of phylogenetic network in evolutionary studies. **Molecular Biology Evolution,** 9:138-151, 2006.

JETZ, A.E.; YU, H.; KLARMANN, G.J.; RON, Y.; PRESTON, B.D.; DOUGHERTY, J.P. High rate recombination throughout the human immunodeficiency virus type 1 genome. **J. Virol**, v.

74, p. 1234-1240, 2000.

JOHNSON, V. A.; BRUN-VEZINET, F.; CLOTET, B.; KURITZKES, D. R.; PILLAY, D. *et al*.. Update of the drug resistance mutations in HIV-1: Fall 2006. **Top HIV Med,** 14:125-30, 2006.

KAHN, J. O.; WALKER, B.D. Acutehuman imunodeficiency virus type 1 infection. **N. Engl. J. Med**, 339:33-39, 1998.

KASS, R.; RAFTEY, A. Bayes Factors. **Journal of the American Statistical Association,** 90**,** 773-95, 1995.

KEDZIERSKA, K.; CROWE, S.M.; TURVILLE, S.; CUNNINGHAM, A.L. The influence of cytokines, chemokines and their receptors on HIV-1 replication in monocytes and macrophages. **Rev Med Virol**,13(1):39-56, 2003.

KIMURA, M. A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. **J Mol Evol,** 16:111-20, 1980.

LANCIOTTI, R.; ROEHRIG, J.; DEUBEL, V.; SMITH, J.; PARKER, M.; STEELE, K. *et al.*. Origin of the West Nile virus responsible for an outbreak of encephalitis in the northeastern United States. **Science**, 286, 2333–2337, 1999.

LANCIOTTI, R.S.; EBEL, G.D.; DEUBEL, V.; KERST, A.J.; MURRI, S.; MEYER, R. *et al.*. Complete genome sequences and phylogenetic analysis of West Nile virus strains isolated from the United States, Europe, and the Middle East. Virology. 298, 96–105. 2002.

LEHMAN, D.A.; FARGUHAR, C. Biological mechanisms of vertical human immunodeficiency virus (HIV-1) transmission. **Rev. Med. Virol**, 17:381-403, 2007.

LEMEY, P.; RAMBAUT, A.; PYBUS, O. HIV evolutionary dynamics within and among hosts. **AIDS Reviews**, 8:125-40, 2006.

LOS ALAMOS. **HIV Circulating Recombinant Forms (CRFs).** Disponível em: < http://www.hiv.lanl.gov/>

MCCUTCHAN, F. E. Understanding the genetic diversity of HIV-1. **AIDS 14**, [s.l.], p. S31-S44, 2000. Supplement 3.

MINISTÉRIO DA SAÚDE. **Epidemiologia HIV**. Disponível em: <http://www.aids.gov.br/data/Pages/LUMIS72418C70PTBRIE.htm.>

MORGADO, M. G.; GUIMARAES, M. L.; GRIPP, C. B. G.; NEVES, I.; COSTA, C. I. *et al*.. High prevalence of HIV-1 subtype B and identification of na HIV-1 Subtype D infection in the city of Rio de Janeiro, Brazil. **Journal of AIDS and Human Retroviruses**, [s.l.], v. 18, p. 488-494, 1998.

MORGADO, M. G.; GUIMARAES, M.L; GALVAO-CASTRO, B. HIV-1 Polymorphism: a challenge for Vaccine Develoment. **Mem Inst Oswaldo Cruz**, 97:143-150, 2002.

MORGADO, M. G.; SABINO, E.; SPHAER, E.; BONGERTZ, V.; BRIGIDO, L.; GUIMARÃES, M. D. C. *et al*. Polymorphism in the V3 region on the envelope protein of HIV-1 in Brazil: divergence from prevalent North American/European subtype B strains and identification of newly described F subtype. **AIDS Res Hum Retroviruses**, 10:569-576, 1994.

MURGUE, B.; ZELLER, H.; DEUBEL, V. The ecology and epidemiology of West Nile virus in Africa, Europe and Asia. **Curr. Top Microbiol. Immunol**. 267, 195–221. 2002.

NICHOLAS, K. B.; NICHOLAS, H. B. J.; DEERFIELD, D. W. GeneDoc: Analysis and visualization of genetic variation. **EMB News,** 14:30, 1997.

PALELLA, F.J.JR; DELANEY, K.M.; MOORMAN, A.C.; LOVELESS, M.O.; FUHRER, J.; SATTEN, G.A.; ASCHMAN, D.J.; HOLMBERG, S.D. Declining morbidity and mortality among patients with advanced human immunodeficiency virus infection. HIV Outpatient Study Investigators. **N Engl J Med**., 26;338(13):853-60, 1998.

PANTALEO, O.; FAUCI, S. Immunopathogenesis of HIV infection. **Annu Rev Microbiol**, 50:825-54, 1996.

PARREIRA, R.; SEVERINO, P.; FREITAS, F.; PIEDADE, J.; ALMEIDA, A.; ESTEVES, A. Two distinct introductions of the West Nile virus in Portugal disclosed by phylogenetic analysis of genomic sequences. **Vector Borne Zoonotic Dis**, 7, 344-352, 2007.

PERELSON, A.S.; NEUMANN, A.U.; MARKOWITZ, M.; LEONARD, J.M.; HO, D.D. HIV-1 dynamics in vivo: virion clearance rate, infected cell life-span, and viral generation time. **Science**, 15;271(5255):1582-6, 1996.

POND, S.L.K.; POSADA, D.; STAWISKI, E.; CHAPPEY, C.; POON, A.F.Y.; HUGHES, G. *et al*.. An Evolutionary Model-Based Algorithm for Accurate Phylogenetic Breakpoint Mapping and Subtype Prediction in HIV-1. **PLoS Computational Biology**, 5(11): e1000581, 2009.

PYBUS, O.G.; RAMBAUT, A. Evolutionary analysis of the dynamics of viral infectious disease. **Nat Rev Genet**, 10(8):540-50, 2009.

SABINO, E.; SPHAER, E.; MORGADO, M. G.; BORBER, B. T.; DIAS, R.S.; BONGERTZ, V. *et al*.. Identification of na HIV-1 proviral genome recombinant between subtype B and F in PBMCs obtained from an individual in Brazil. **J Virol**, 68:6340-6346, 1994.

SALEMI, M; GRAY, RR; GOODENOW, MM. An exploratory algoruthm to identify intra-host recombinat viral sequences. **Molecular Phylogenetics and Evolution**, 49(2):618-28, 2008.

SALMINEN, M. O.; CARR, J. K.; BURKE, D. S.; MCCUTCHAN, F. E. Identification of breakpoints in intergenotypic recombinants of HIV type 1 by bootscanning. **AIDS Res Hum Retroviruses,** 11:1423-5, 1995.

SCHERRET, J.H.; POIDINGER, M.; MACKENZIE, J.S.; BROOM, A.K.; DEUBEL, V.; LIPKIN, W.I. *et al*.. The relationships between West Nile and Kunjin viruses. **Emerg. Infect.**

Dis, 7, 697–705, 2001.

SCHIERUP, M.H.; HEIN, J. Consequences of recombination on traditional phylogenetic analysis. **Genetics**, 156(2):879-91, 2000.

SCHMIDT, H.; STRIMMER, K.; VINGRON, M.; VON HAESELER, A. TREE-PUZZLE: maximum likelihood phylogenetic analysis using quartets and parallel computing. **Bioinformatics**, 18, 502–504, 2002.

SEPKOWITZ, K.A. AIDS--the first 20 years. **N Engl J Med**, 7;344(23):1764-72, 2001.

SHAFER, R. W.; JUNG, D.R.; BETTS, B. J.; XI, Y.; GONZALES, M. J.. Human immunodeficiency virus reverse transcriptase and protease sequence database. **Nucleic Acids Res**, 28:346-8, 2000.

SILVA, A. C. M.; BARONE, A.A. Risk factors for HIV infection among patients infected with hepatitis C virus. **Revista de Saúde Pública**, 40(3):482–488, 2006.

SMITHBURN, K.C.; HUGHES, T.P.; BURKE, A.W.; PAUL, J.H. A neurotropic virus isolated from the blood of a Native Ugandan. **Am. J. Trop. Med. Hyg**, 20, 471–492, 1940.

SOARES, M.A.; DE OLIVEIRA, T.; BRINDEIRO, M.; DIAZ, R.S.; SABINO, E.C. *et al*.. A specific subtype C of human immunodeficiency virus type 1 circulates in Brazil. **AIDS**, [s.l.], v. 17, p. 11-21, 2003.

STRIMMER, K.; VON HAESELER, A. Likelihood-mapping: a simple method to visualize phylogenetic content of a sequence alignment. **Proc. Natl. Acad. Sci.,** USA 94, 6815–6819, 1997.

SUBBARAO, S.; SCHOCHETMAN, G. Genetic variability of HIV-1. **AIDS**. Suppl A, p. 13-23, 1996.

SUCHARD, MR; WEISS e SINCHEIMER J. Bayesian selection of continuous-time Markov chain evolutionary models. **Molecular Biology Evolution**, 18**,** 1001-13, 2001.

SWOFFORD, D. PAUP*: Phylogenetic analysis using parsimony. Version 4.0b10. **Smithsonian Institution**, Washington, D.C. 1997.

TANG, Y.; LIU, B.; HAPIP, C.; XU, D.; FANG, C. Genetic analysis of West Nile virus isolates from US blood donors during 2002–2005. **J. Clin. Virol**., 43, 292–297, 2008.

THOMPSON, J.D.; GIBSON, T. J.; PLEWNIAK, F.; JEANMOUGIN, F. e HIGGINS, D. G. The ClustalX windows interface: flexible strategies for multiple sequence alignment aided by quality analysis tools. **Nucleic Acids Research**, 25: 4876-4882, 1997.

UGOLINI, S.; MONDOR, I.; SATTENTAU, Q.J. HIV-1 attachment: another look. **Trends Microbiol**, v. 7, p. 144-149, 1999.

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

VANDAMME, A.M. Basic concepts of molecular evolution. **The Phylogenetic Handbook.** 2 ed. UK: Cambridge. 2009. p. 3-30.

WILLIAMSON, S. Adaptation in the env gene of HIV-1 and evolutionary theories of disease progresion. **Molecular Biology and Evolution**, 20(8):1318-1325. 2003.

WOLINSKY, S.M.; WIKE, C.M.; KORBER, B.T. *et al.*. Selective transmission of human immunodeficiency virus type-1 variants from mothers to infants. **Science**, 255(5048): 1134-7, 1992.

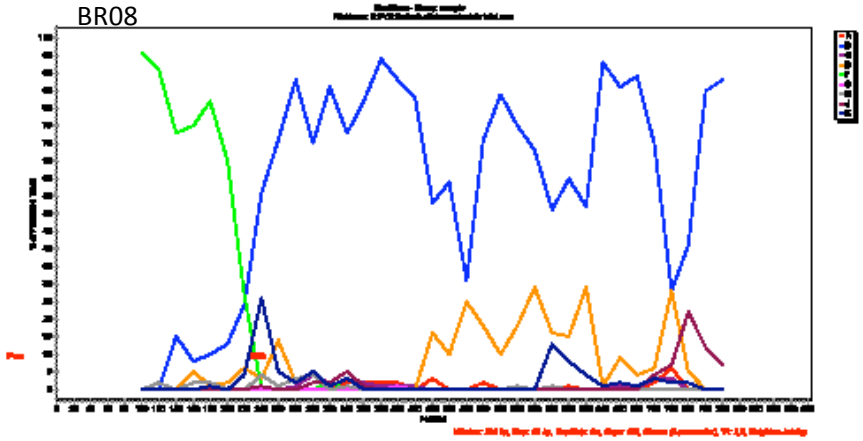
YARCHOAN, R.; KLECKER, R.W.; WEINHOLD, K.J.; MARKHAM, P.D.; LYERLY, H.K.;

DURACK, D.T. *et al*.. Administration of 3'-azido-3'-deoxythymidine, an inhibitor of HTLV-III/LAV replication, to patients with AIDS or AIDS-related complex. **Lancet**, 15;1(8481):575-80, 1986.

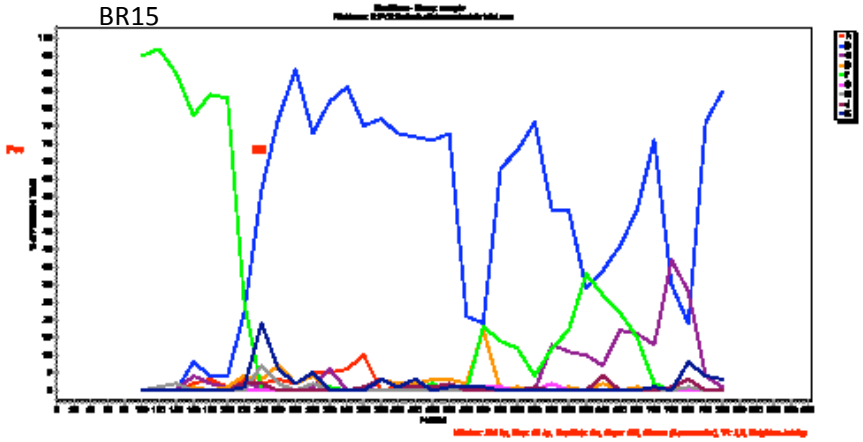
Apêndice A

Amostras que apresentaram o padrão CRF28/CRF29

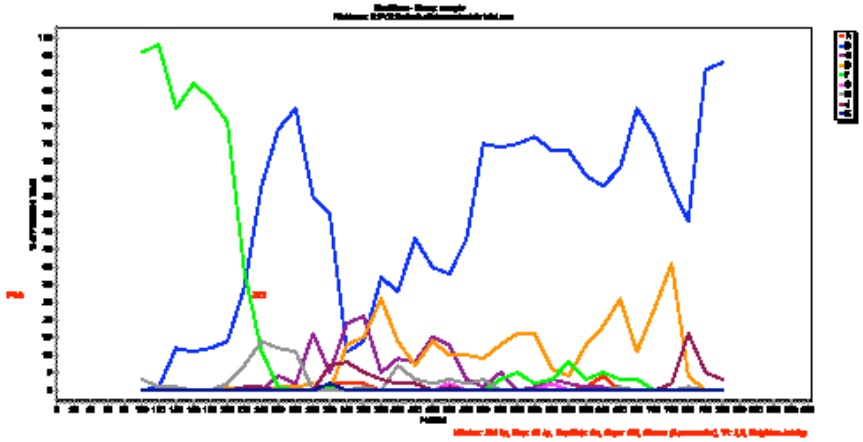
BR08



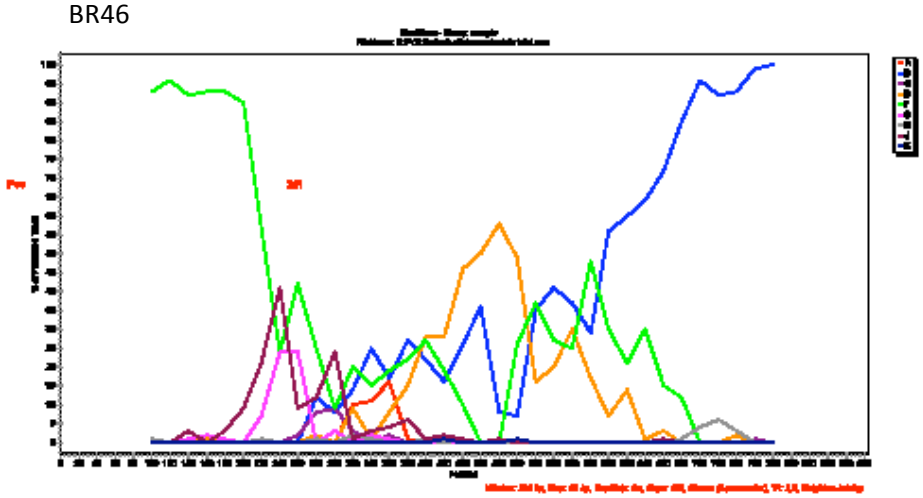
BR15



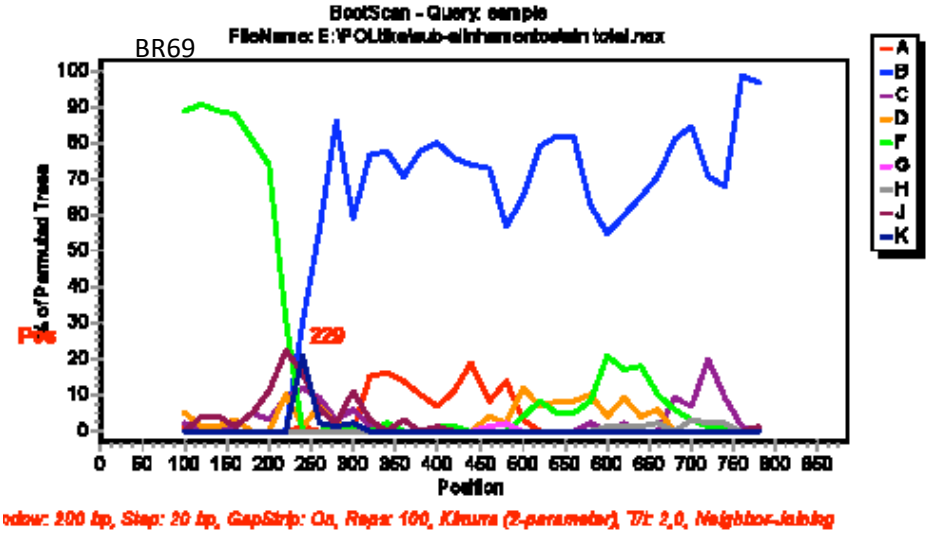
BR23



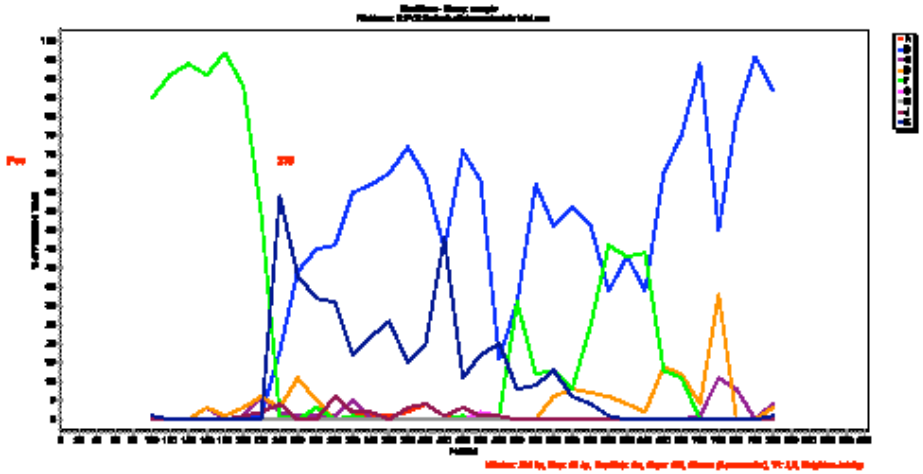
BR46



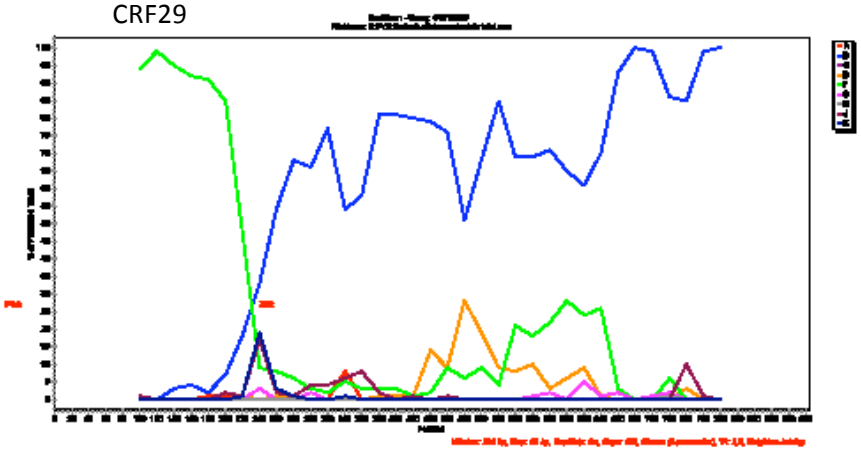
BR69



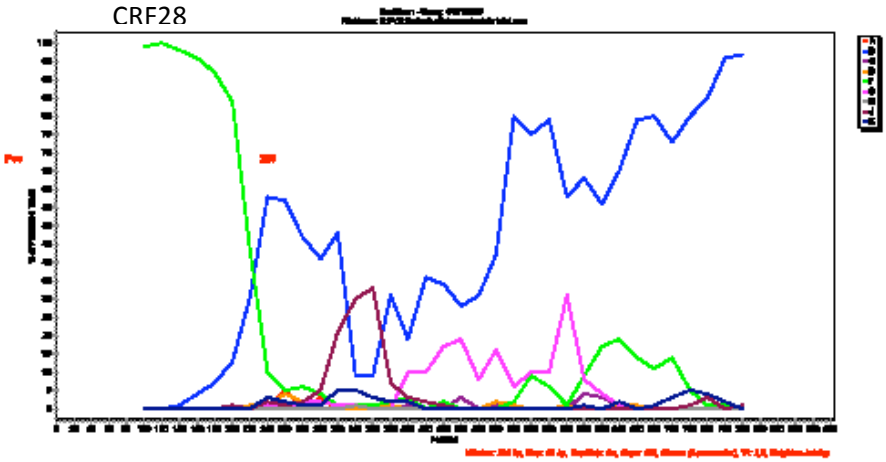
BR79



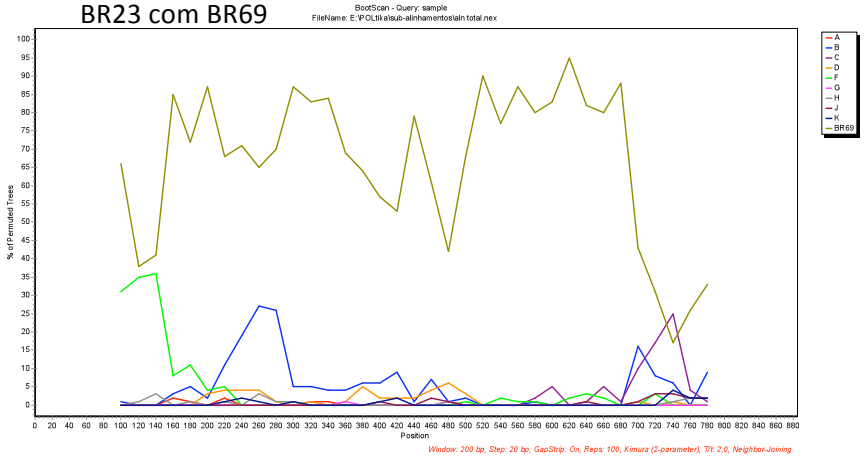
CRF29



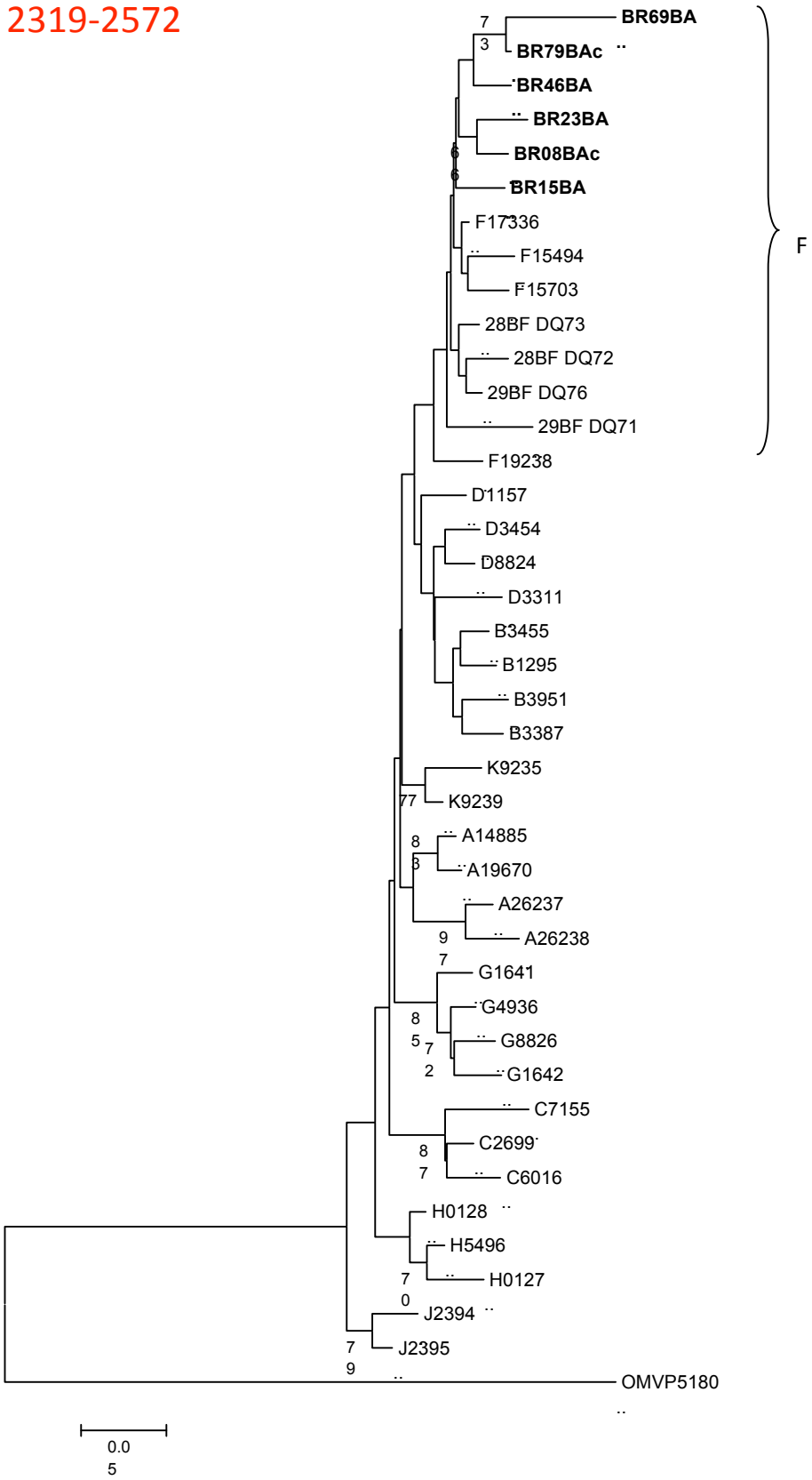
CRF28



BR23 com BR69

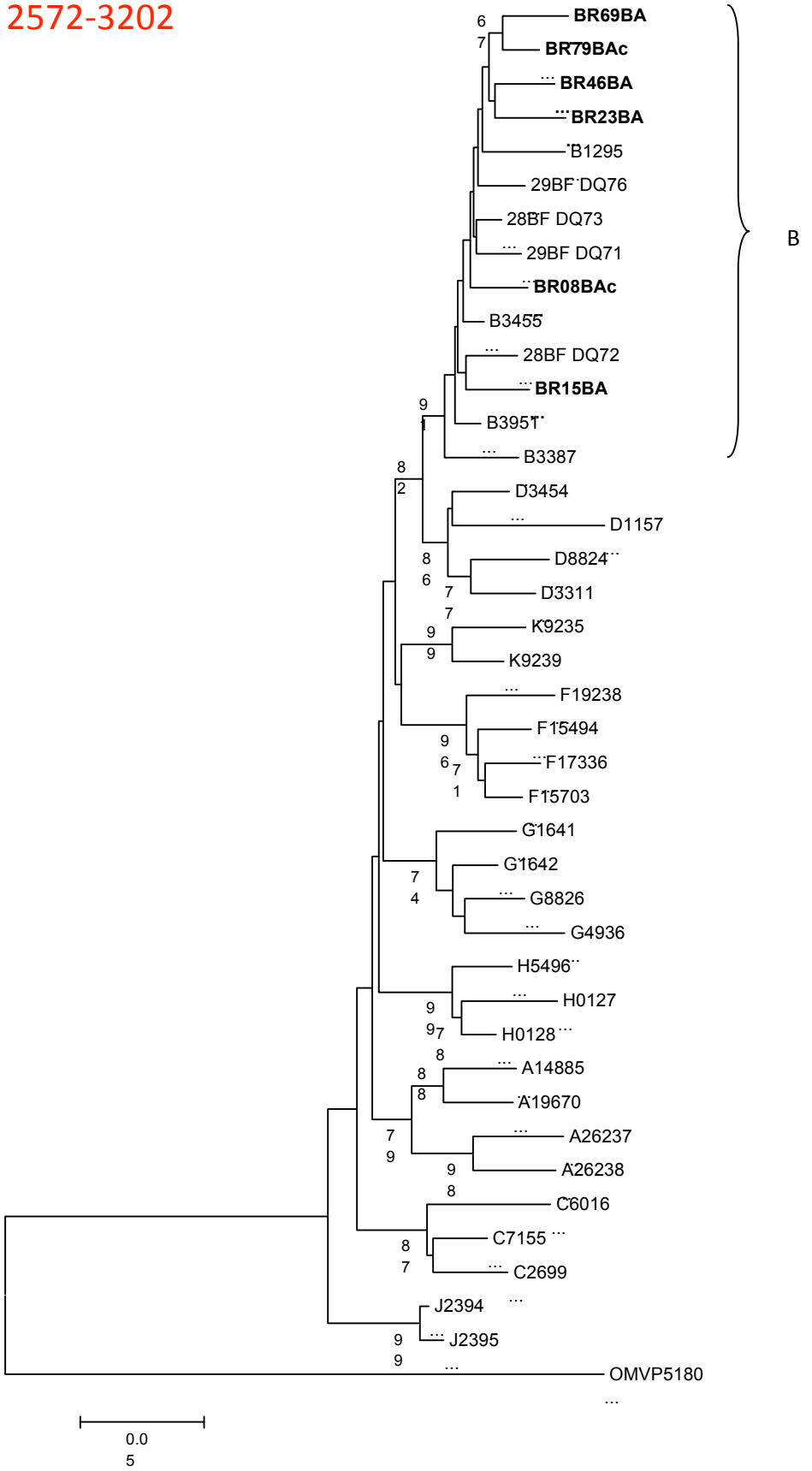


2319-2572



Árvore NJ com base no fragmento 2319-2572 dentro do gene *pol* do HIV-1. Seis amostras da Bahia agruparam dentro do grupo F, juntamente com seqüências das CRF28 e 29.

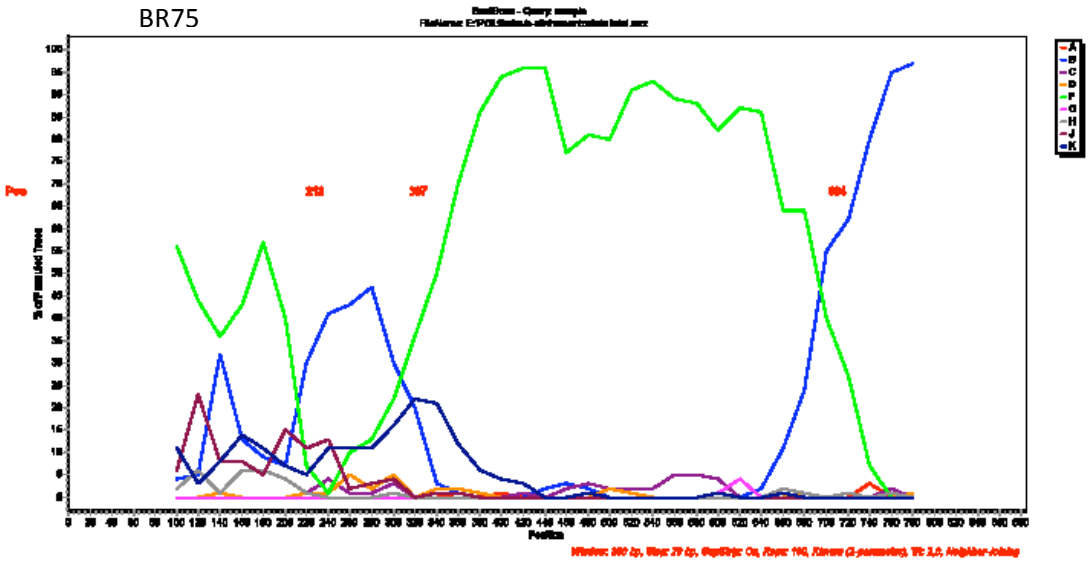
2572-3202



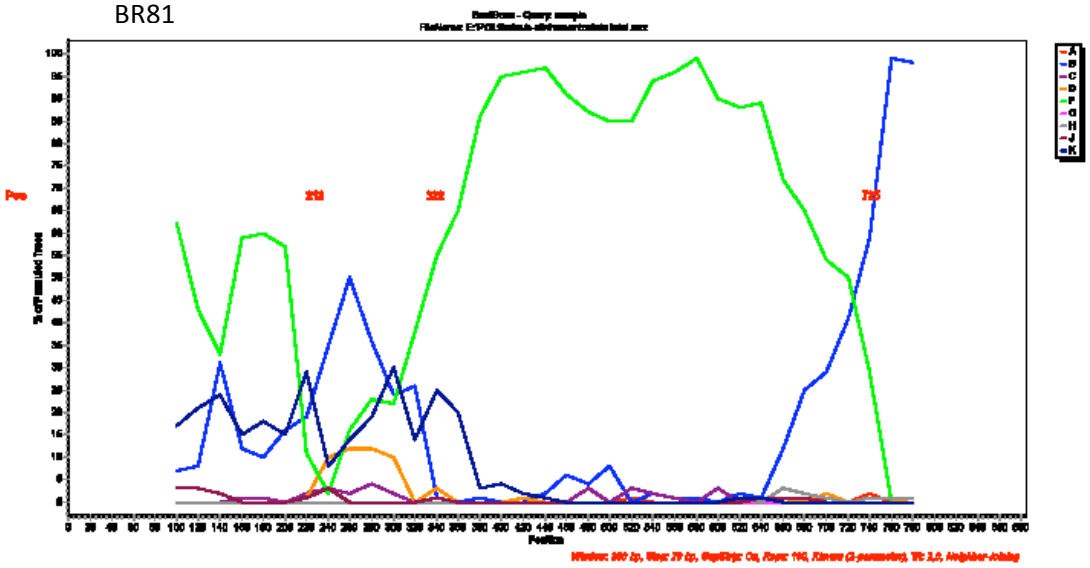
Árvore NJ com base no fragmento 2572-3202 dentro do gene *pol* do HIV-1. Seis amostras da Bahia agruparam dentro do grupo B, juntamente com seqüências das CRF28 e 29.

Amostras que apresentaram padrão CRF12

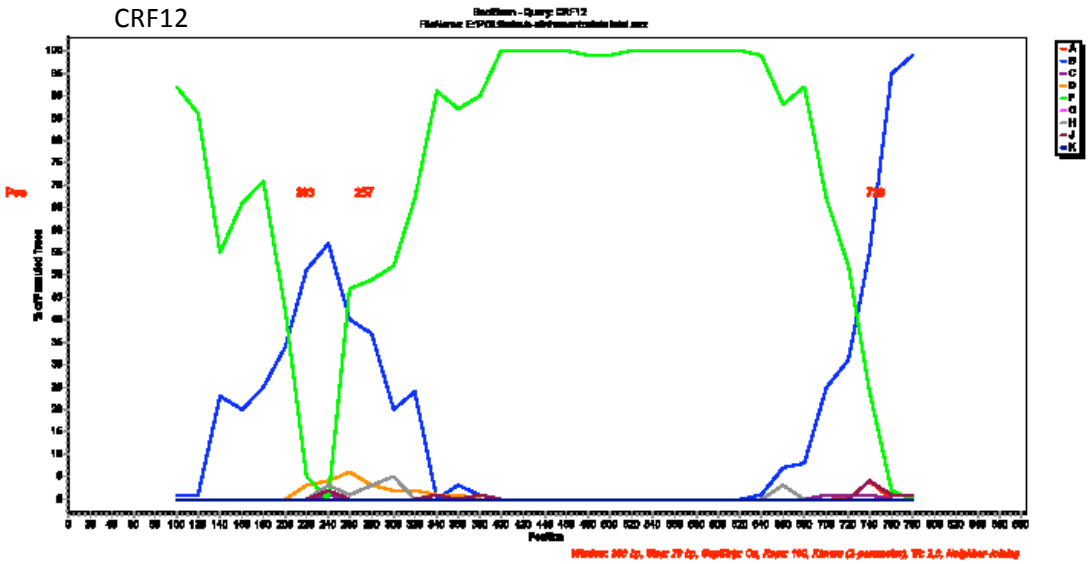
BR75



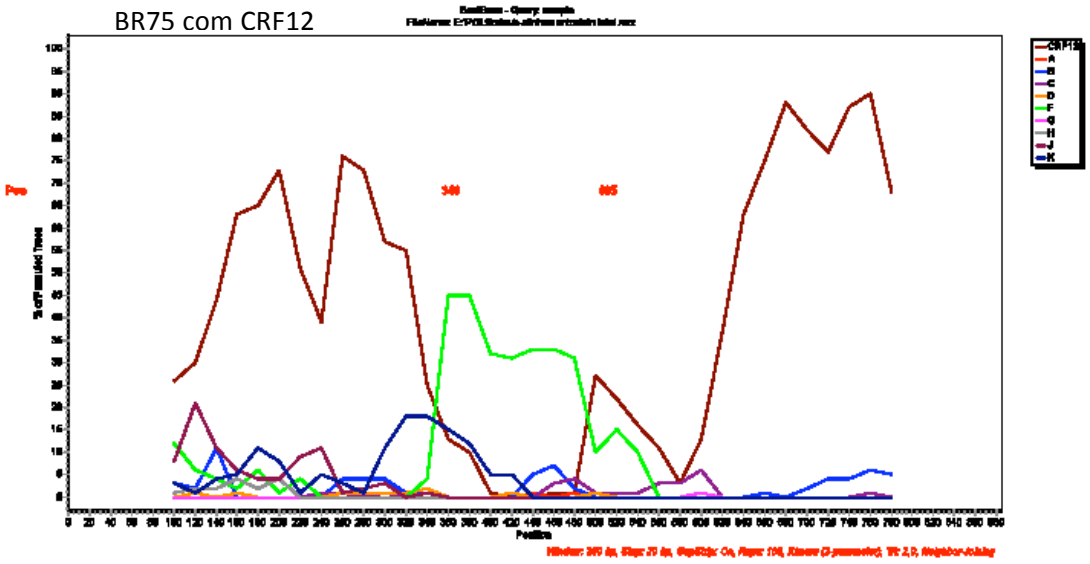
BR81



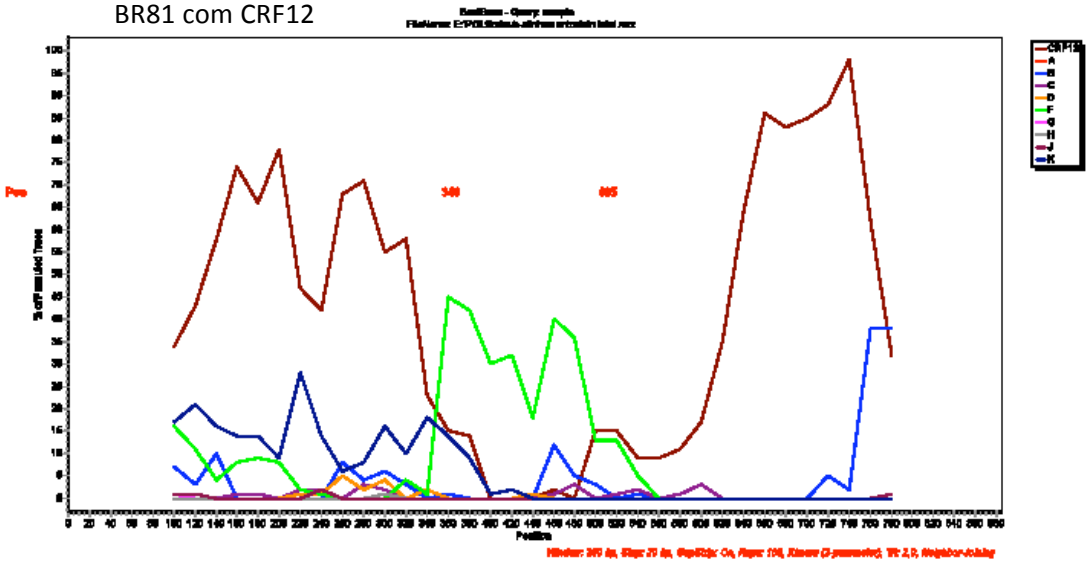
CRF12



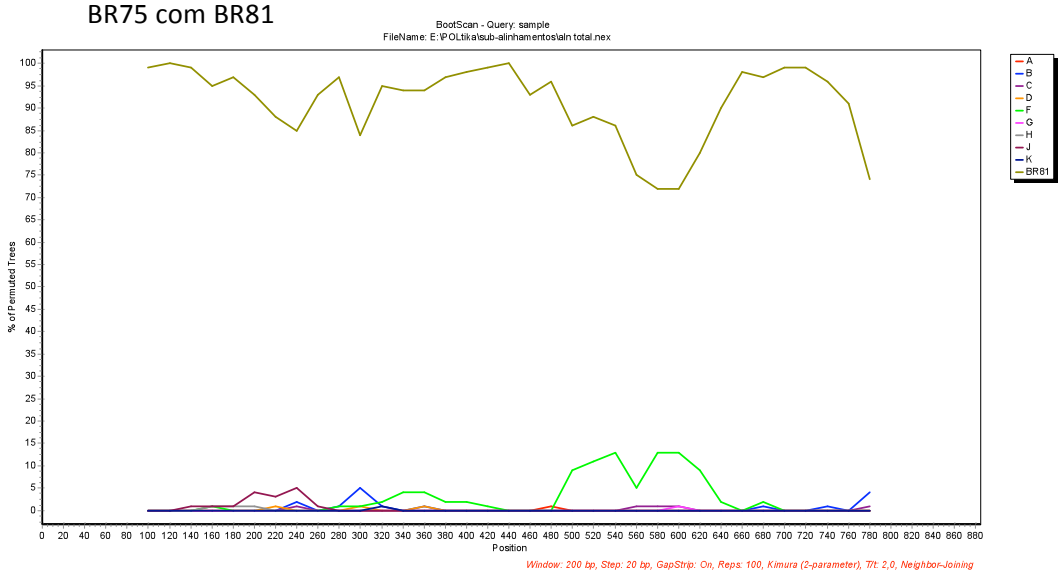
BR75 com CRF12

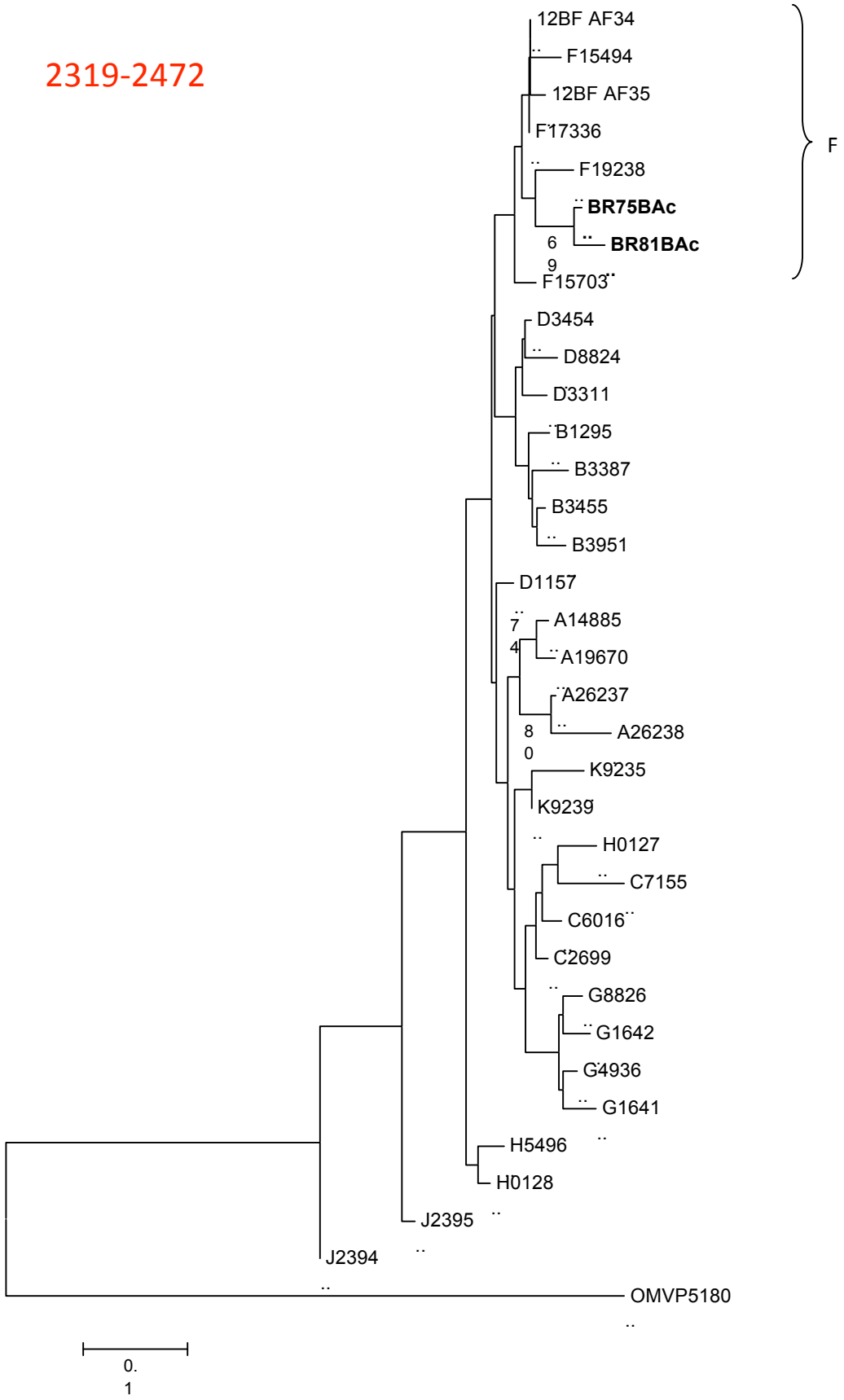


BR81 com CRF12

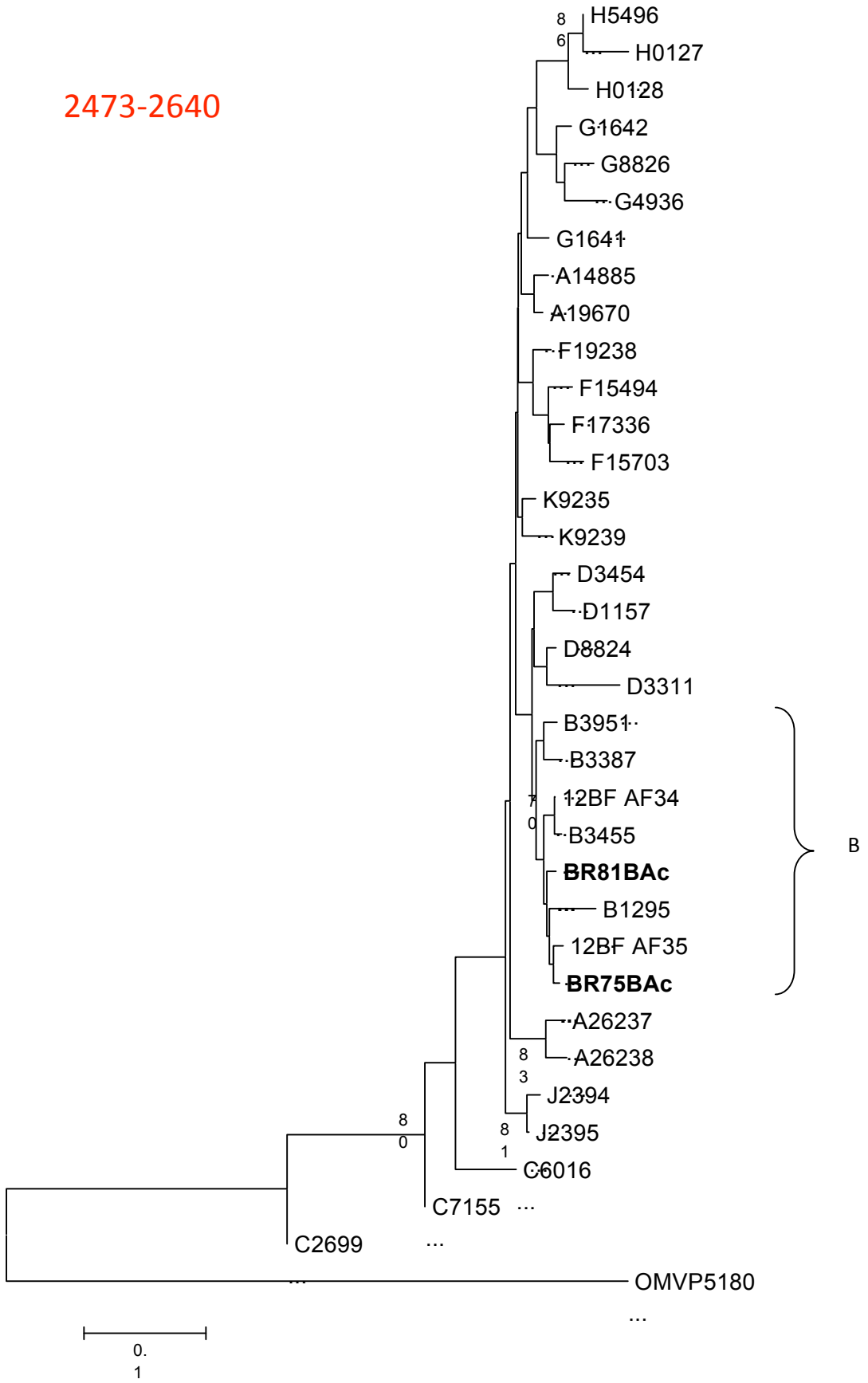


BR75 com BR81

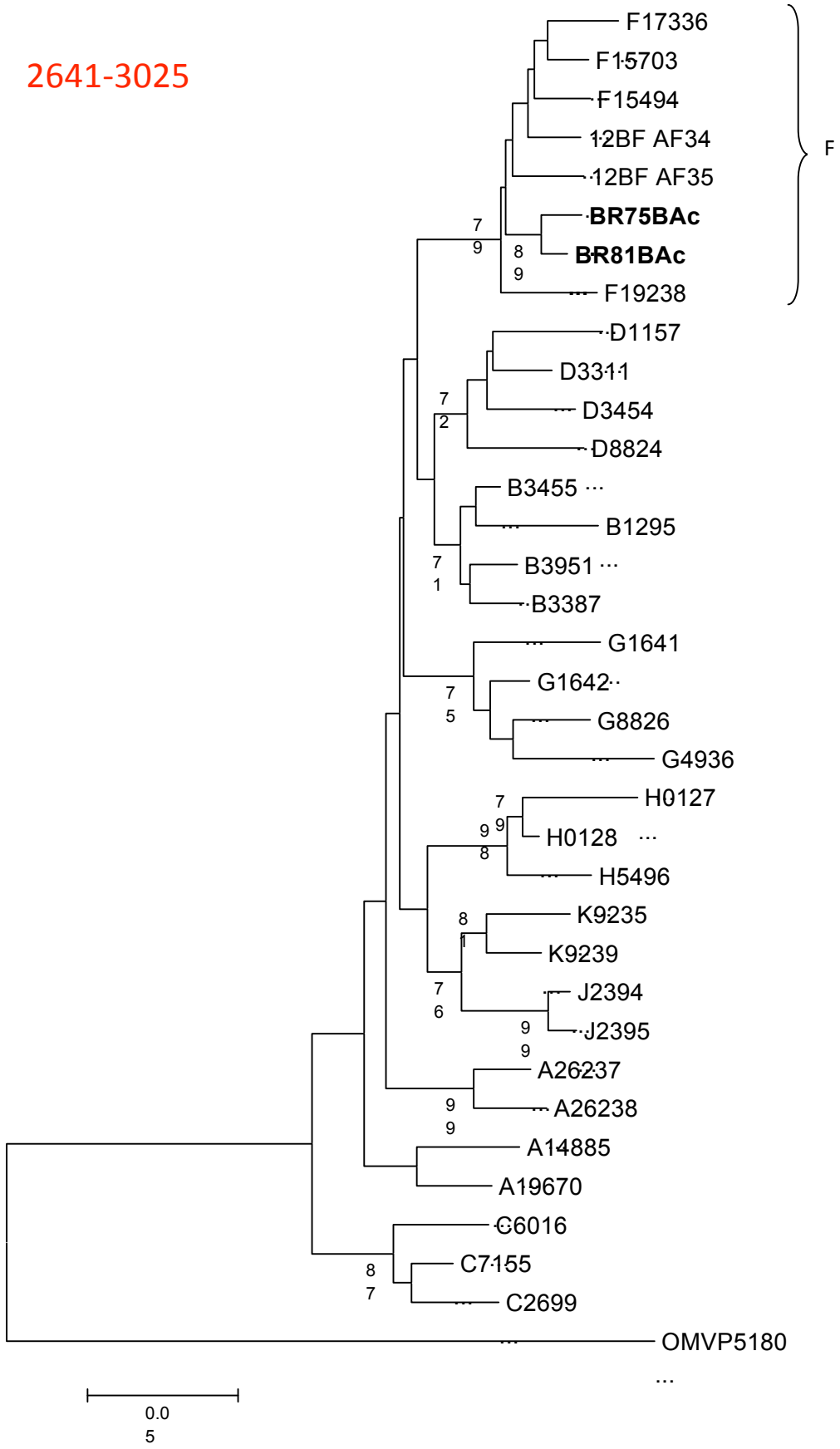




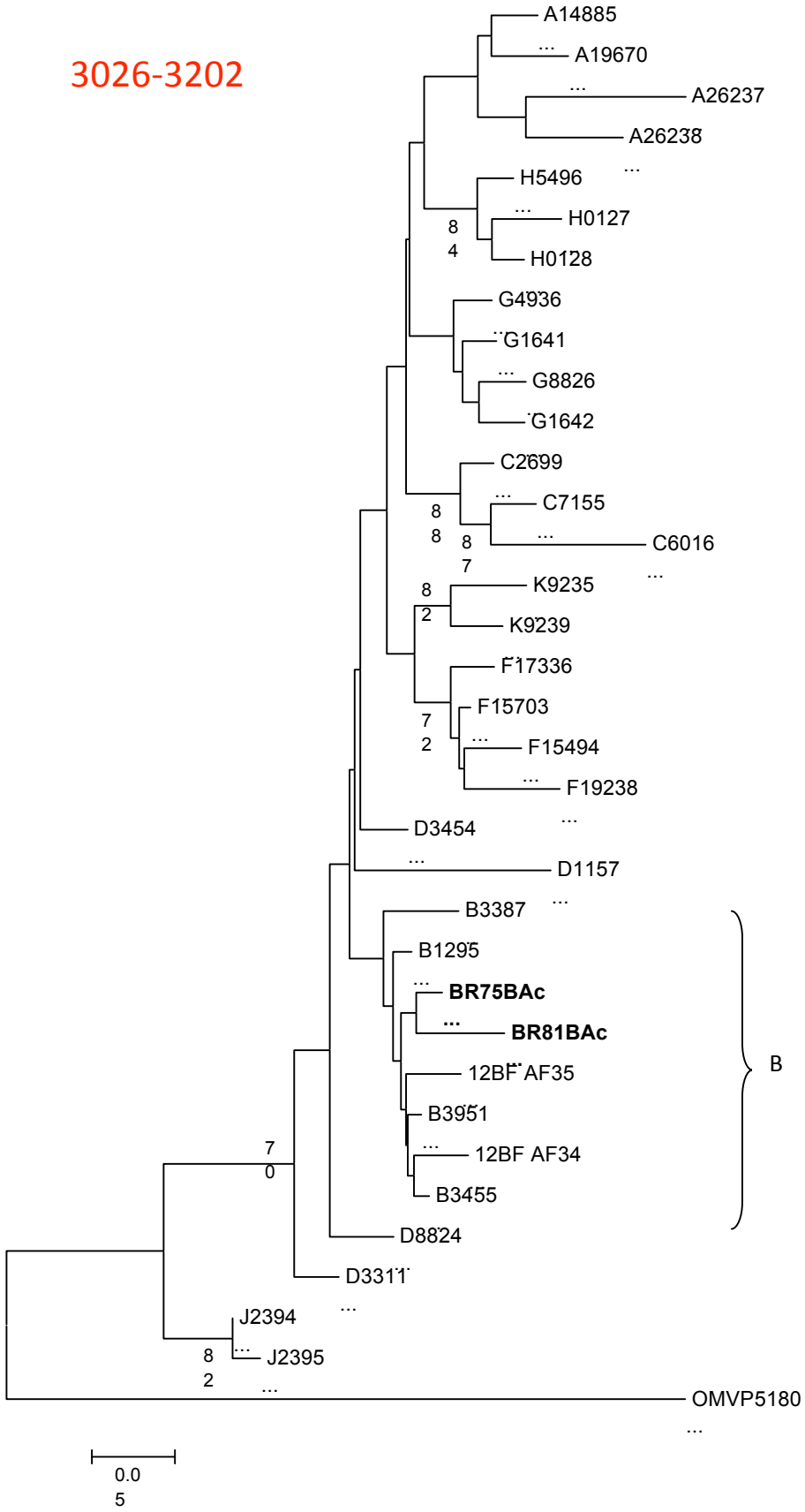
Árvore NJ com base no fragmento 2319-2472 dentro do gene *pol*. As amostras BR75 e BR81 agruparam com o grupo do subtipo F, no qual as seqüências da CRF12 estão incluídas.



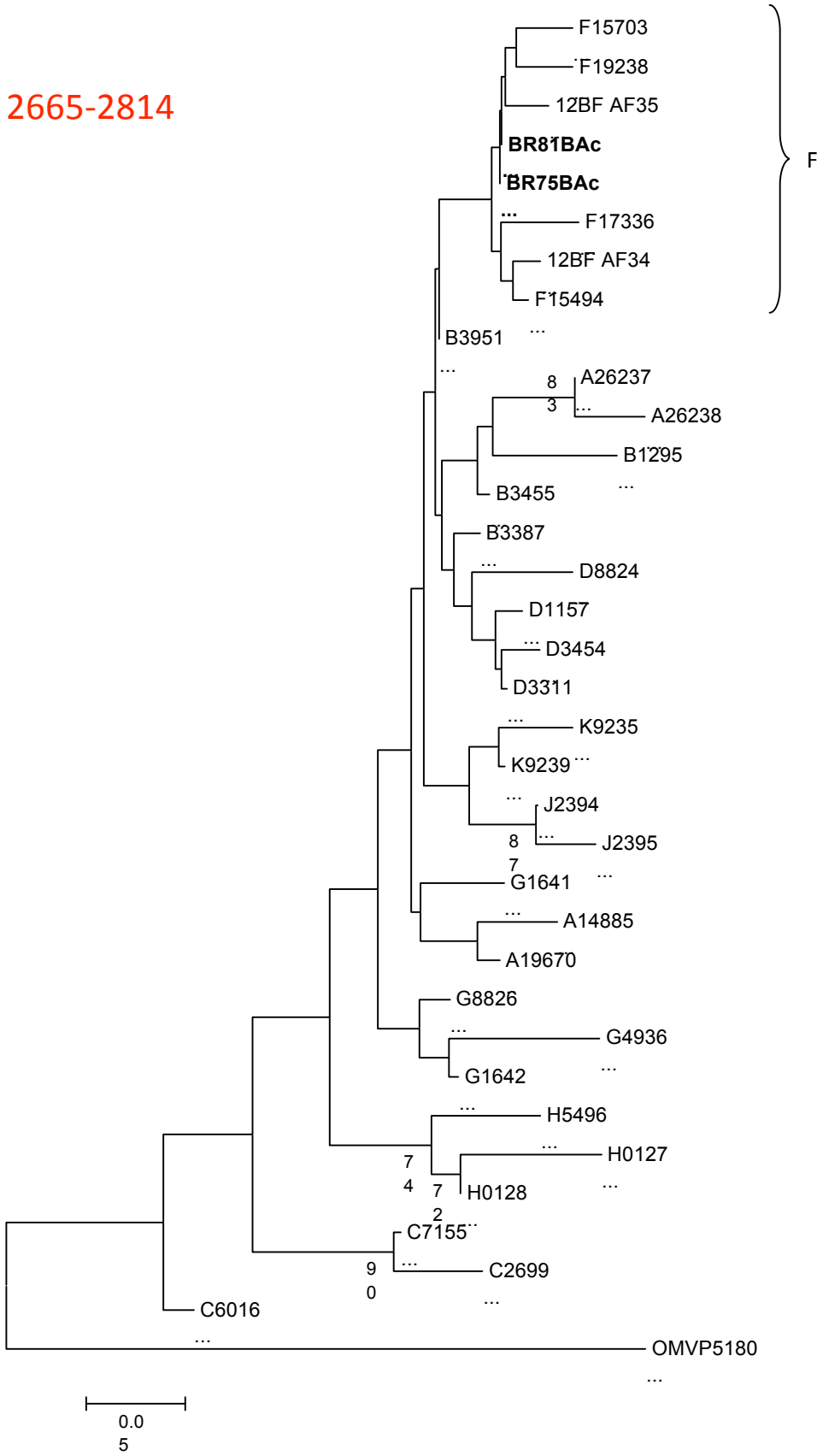
Árvore NJ com base no fragmento 2473-2640 dentro do gene *pol*. As amostras BR75 e BR81 agruparam com o grupo do subtipo B, no qual as seqüências da CRF12 estão incluídas.



Árvore NJ com base no fragmento 2641-3025 dentro do gene *pol*. As amostras BR75 e BR81 agruparam com o grupo do subtipo F, no qual as seqüências da CRF12 estão incluídas.



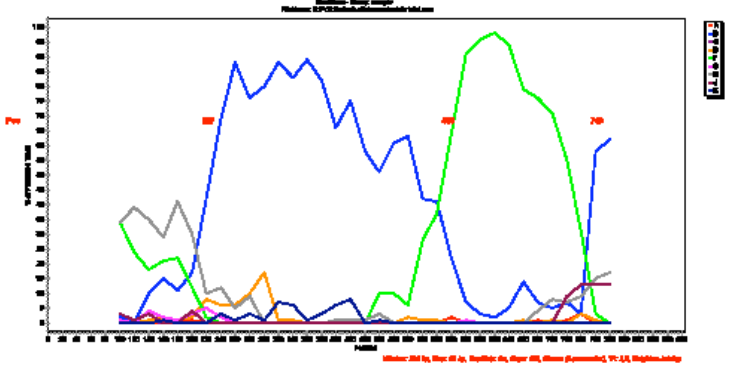
Árvore NJ com base no fragmento 3026-3202 dentro do gene *pol*. As amostras BR75 e BR81 agruparam com o grupo do subtipo B, no qual as seqüências da CRF12 estão incluídas.



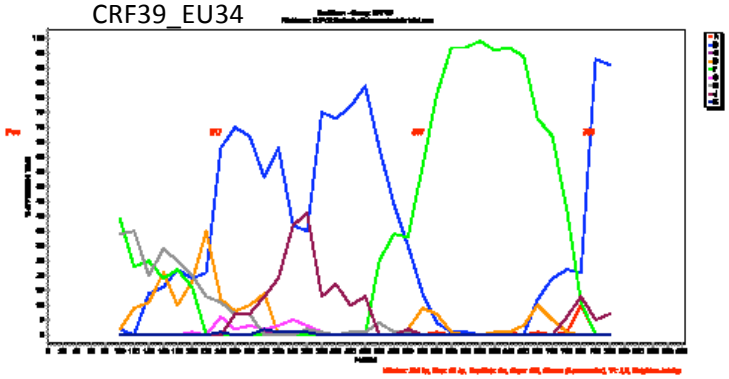
Árvore NJ com base no fragmento 2665-2814 dentro do gene *pol*. As amostras BR75 e BR81 agruparam com o grupo do subtipo F, no qual as sequências da CRF12 estão incluídas.

Comparação do perfil de recombinação entre amostra BR42BA e duas sequências da CRF39

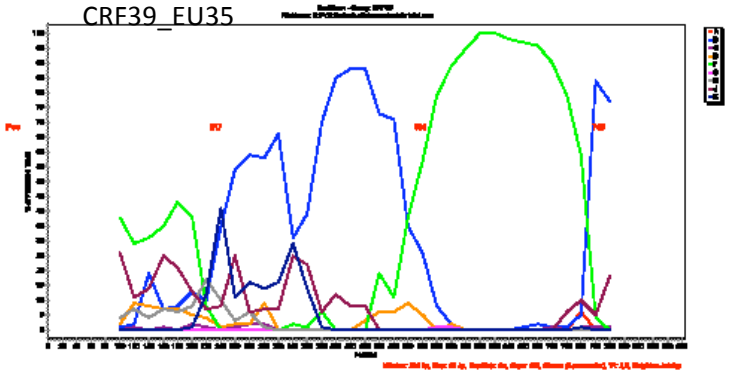
BR42



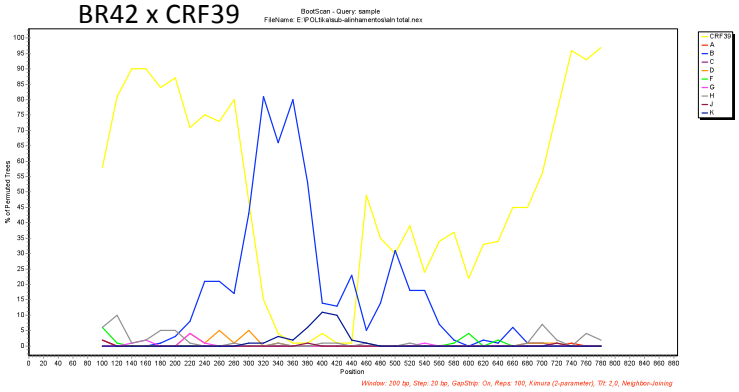
CRF39_EU34



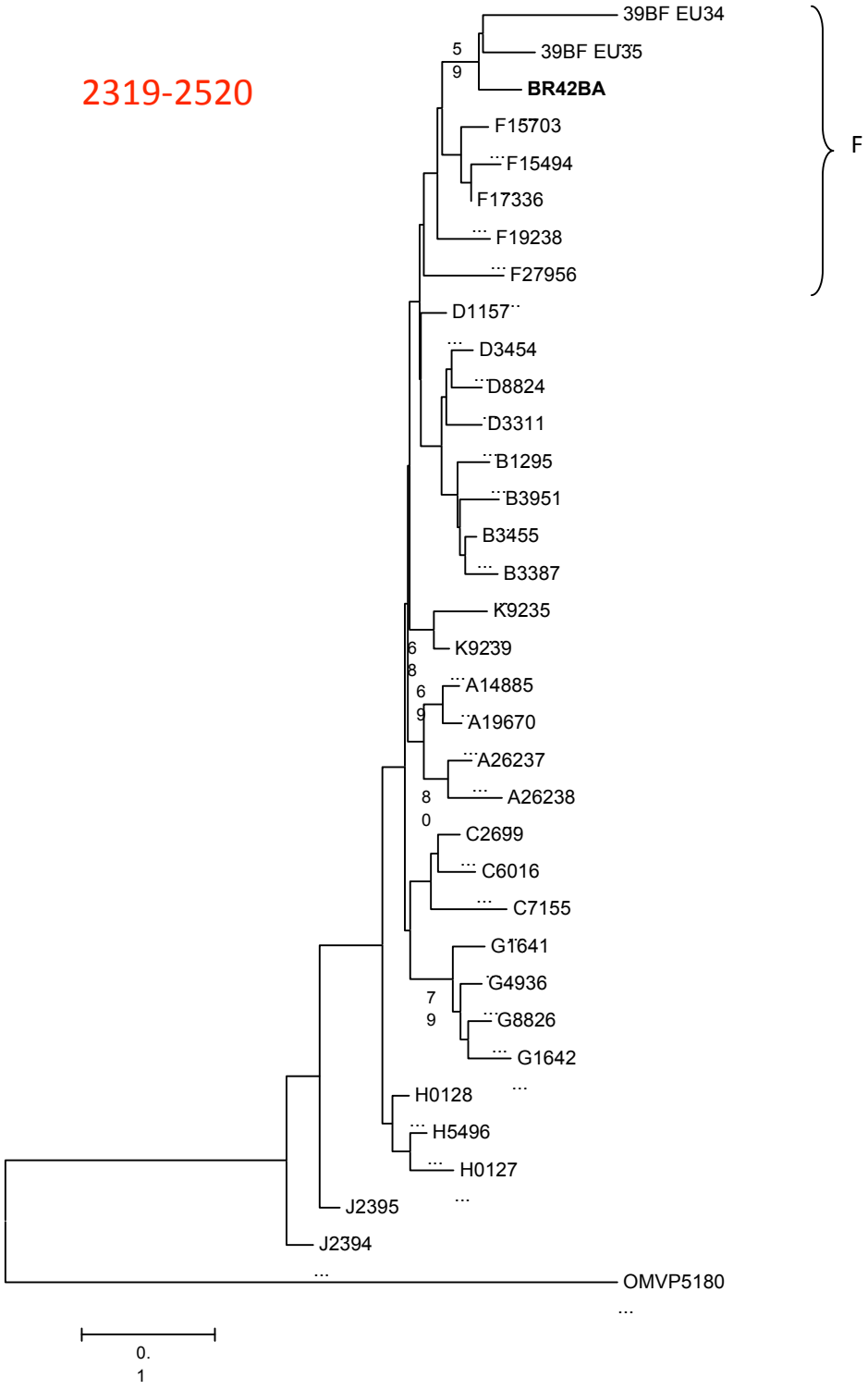
CRF39_EU35



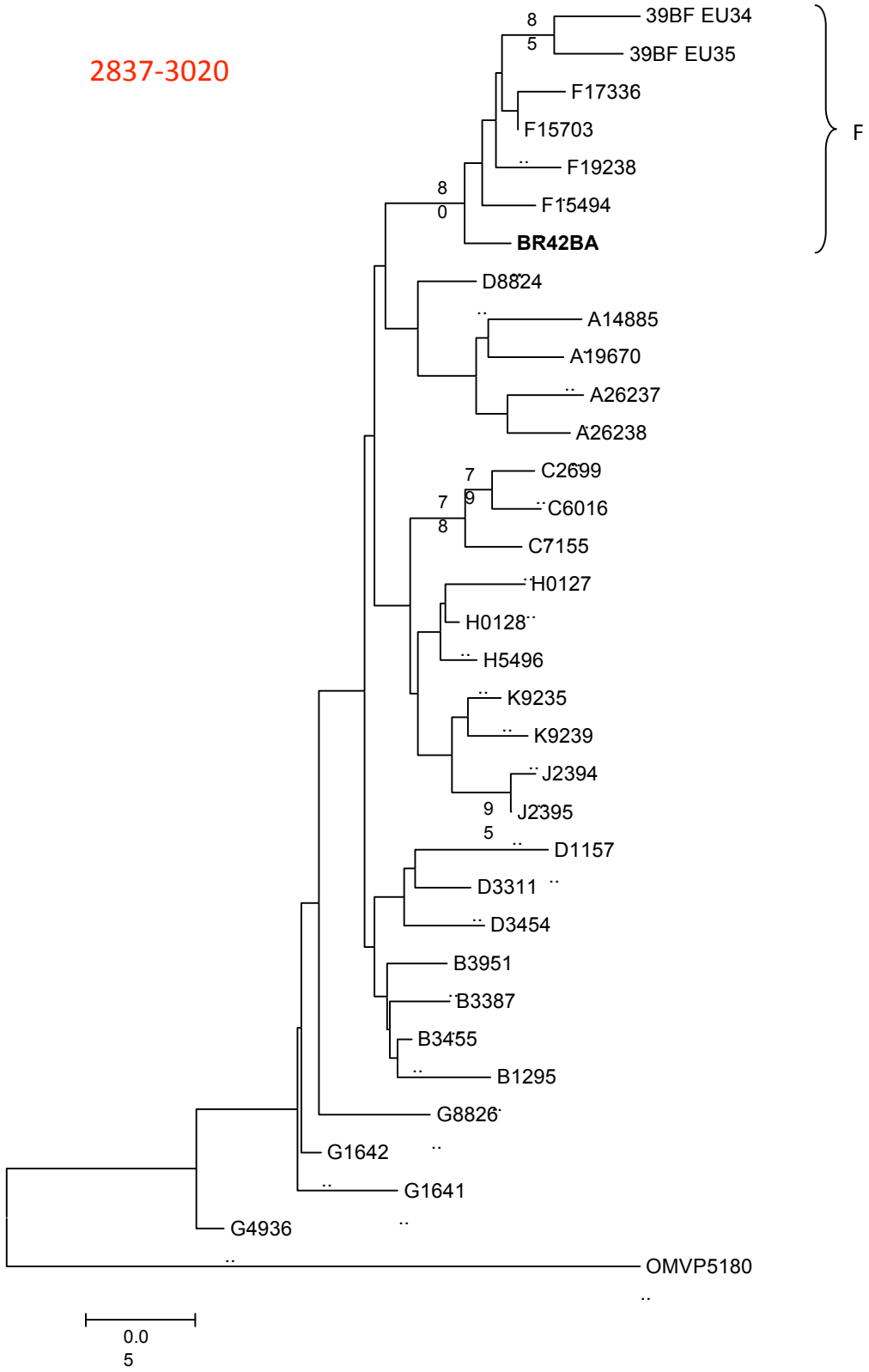
BR42 x CRF39



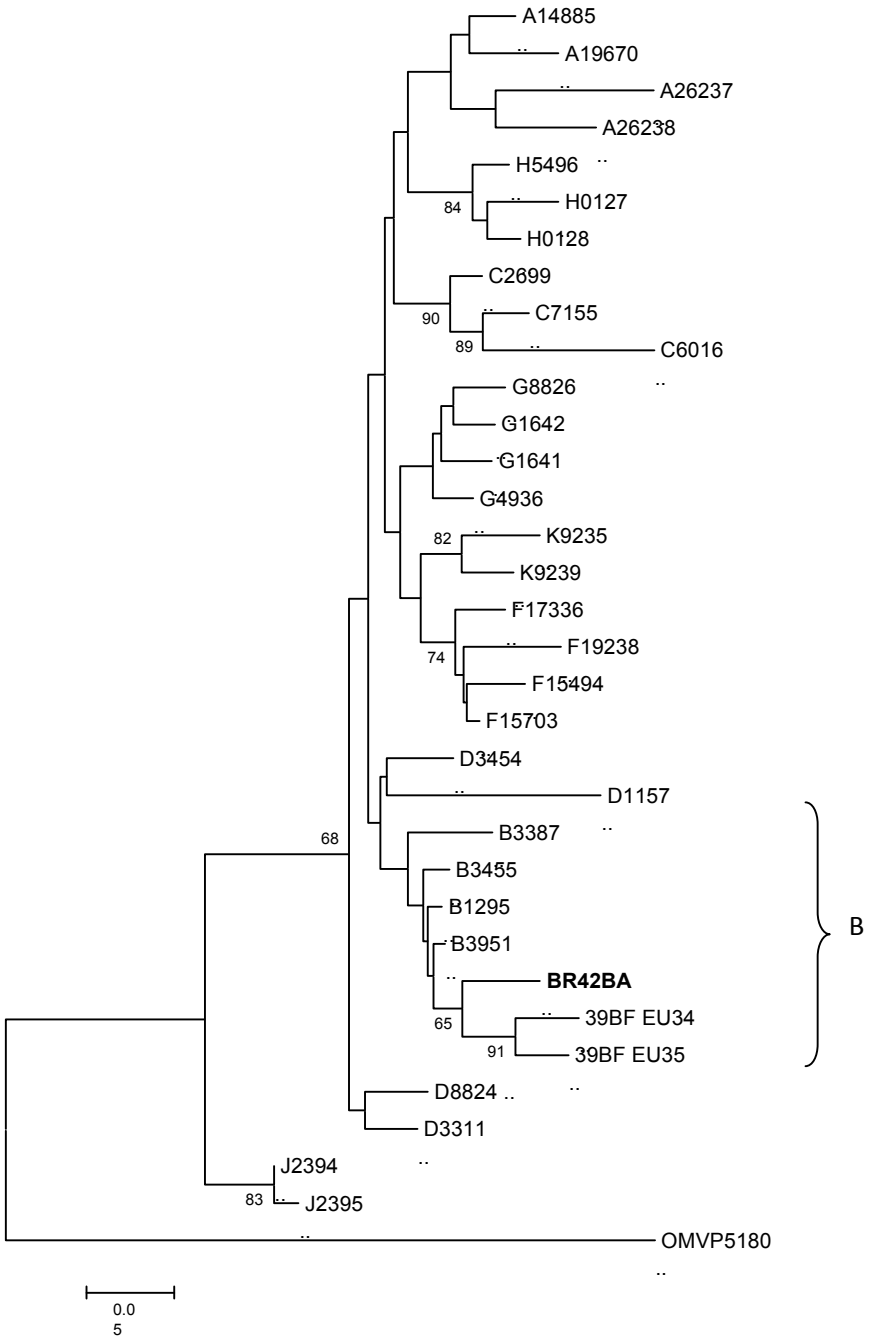
Árvore NJ com base no fragmento 2319-2836 dentro de *pol*. CRF39 é classificada como subtipo B ao longo de toda a extensão deste fragmento. BR42BA agrupou junto com CRF39 (Bootstrap = 95) dentro do cluster do subtipo B.



Árvore NJ com base no fragmento 2319-2520, dentro de *pol*. Nesta analise, tanto a amostra BR42BA como as sequências CRF29 agruparam junto com o subtipo F, confirmando o perfil encontrado no SIMPLOT. Entretanto, a CRF29 neste fragmento, esta classificada como B segundo a publicação original. Porem, o SIMPLOT desta amostra confirma que há recombinação neste ponto.



Árvore NJ com base no fragmento 2837-3020, dentro de *pol*. Nesta análise a amostra BR42BA agrupou com o subtipo F, mas não no mesmo subgrupo da CRF39.



Árvore NJ com base no fragmento 3020-3202 dentro do gene pol. Neste fragmento, a amostra BR42BA agrupou dentro do grupo B, assim como a CRF39.

Anexo A

Parecer do comiter de ética:

"Caracterização étnica/geográfica da população de Salvador e de portadores do HIV-1 e a correlação entre o índice de ancestralidade africana e vulnerabilidade a HIV/AIDS"

Ministério da Saúde

FIOCRUZ
Fundação Oswaldo Cruz
Centro de Pesquisas Gonçalo Moniz

PARECER Nº 84/2006

Protocolo: 180

Projeto de Pesquisa: Caracterização étnica/geográfica da população de Salvador e de portadores do HIV-1 e a correlação entre o índice de ancestralidade africana e vulnerabilidade a HIV/AIDS

Pesquisador Responsável: Dr. Bernardo Galvão Castro Filho

Instituição ou Departamento: LASP/FIOCRUZ

Considerações:

Após análise ética do projeto e realização dos esclarecimentos solicitados pelo responsável, o CEP considera que o projeto atende aos princípios éticos de autonomia, beneficiência, não maleficiência, equidade e justiça.

Diante do exposto, o Comitê de Ética em Pesquisas do Centro de Pesquisas Gonçalo Moniz da Fundação Oswaldo Cruz (CEP-CPqGM/FIOCRUZ), conforme atribuições conferidas pela CONEP/CNS/MS (Carta Doc.32/04/97), com base na Resolução 196/96, julga **aprovado** o projeto supracitado.

Salvador, 20 de março de 2006

Dra. Marilda de Souza Gonçalves
Coordenadora
CEP – CPqGM/FIOCRUZ

Comitê de Ética em Pesquisa - Rua Waldemar Falcão, nº 121, Candeal, Salvador, Bahia
CEP 40296-710, Brasil
Tel: (71) 3176-2239 Fax: (71) 3176-2326
e-mail: cep@cpqgm.fiocruz.br

Anexo B

Parecer do comiter de ética:

"Variabilidade genética dos isolados do HIV-1 em mulheres e crianças infectadas de Feira de Santana"

BAHIANA
ESCOLA BAHIANA DE MEDICINA E SAÚDE PÚBLICA

FBDC Fundação Bahiana para Desenvolvimento das Ciências

140

COMITÊ DE ÉTICA EM PESQUISA

OFÍCIO Nº 139/2008

Salvador, 20 de agosto de 2008.

Senhor Orientador,

Com referência ao seu prezado ofício, datado de 19 de agosto corrente, no qual solicita a este CEP a inclusão dos nomes das estudantes Luciana Amorim Santos e Giselle de Souza Costa no Protocolo de Pesquisa nº 86/2007 e, também, acrescentar ao citado protocolo de pesquisa os seguintes objetivos específicos:" *estudar a dinâmica evolutiva do gene env do HIV de mães e filhos infectados através da utilização de amostras de sangue destes indivíduos* e *analisar mutações no gene da glicoproteina Langerina e das regiões N-terminal, C-Terminal e no segundo loop extracelular (ECL2) do gene da proteína CCR5 nos pares de mães e filhos infectados pelo HIV*," vimos informar que esta Coordenadora, depois de revisar o Parecer dado no citado protocolo, concluiu que o mesmo é um Protocolo de Pesquisa Viral, podendo ser ACEITAS as modificações solicitadas.

Cordialmente,

Profa. Dra. Lucíola Maria Lopes Crisóstomo
Coordenadora do CEP/FBDC.

Ilmo. Sr.
PROF. DR. LUIZ CARLOS JUNIOR ALCÂNTARA
Rua Cícero Simões, 225 – Apart. 301 – Pituba
CEP.41.830-475 – Salvador-Bahia.