UNIVERSIDADE FEDERAL DO RIO GRANDE DO SUL

Instituto de Biociências

Programa de Pós-Graduação em Genética e Biologia Molecular

Dinâmica epidemiológica das transmissões do HIV-1 subtipo B

Dennis Maletich Junqueira

Tese submetida ao Programa de Pós-Graduação em Genética e Biologia Molecular da UFRGS como requisito parcial para a obtenção do grau de Doutor em Ciências (Genética e Biologia Molecular).

Orientadora: Dr.ª Sabrina Esteves de Matos Almeida

Porto Alegre, maio de 2015

Este trabalho foi realizado nas instalações do Centro de Desenvolvimento Científico e Tecnológico (CDCT) da Fundação Estadual de Produção e Pesquisa em Saúde (FEPPS) com financiamento do Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq), da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES) e da Fundação de Amparo à Pesquisa do Estado do Rio Grande do Sul (FAPERGS).

Querem que vos ensine o modo de chegar à ciência verdadeira? Aquilo que se sabe, saber que se sabe; aquilo que não se sabe, saber que não se sabe; na verdade é este o saber.

Confúcio

AGRADECIMENTOS:

- À Dra. Sabrina Almeida pela orientação, confiança e, especialmente, pela liberdade de trabalho em todo este tempo. Foram anos muito importantes para minha formação como pessoa e como profissional;
- À Rubiazinha, acima de tudo, por toda a amizade. As conversas filosóficas, o incentivo, a cobrança e a vontade de ajudar constantemente foram detalhes fundamentais para construção do meu trabalho. Nestes oito anos, em um misto de amizade e profissional, a presença dela foi fundamental para que muito acontecesse. A vontade de aprender e de fazer a diferença é contagiante. Além do exemplo de sensibilidade e sensatez, foi um exemplo de capacidade, competência e profissionalismo;
- Ao Tiago Gräf por toda a ajuda durante o doutorado, pelas discussões de técnicas e artigos, pelos comentários e pelos ensinamentos;
- À Dra. Maria Lucia Rosa Rossetti por permitir a acolhida e utilização de toda a estrutura dos laboratórios do CDCT desde a iniciação científica;
- Aos colegas do CDCT, especialmente à Regina, por toda a colaboração;
- Ao Núcleo de Bioinformática do Laboratório de Imunogenética pela intensa colaboração no trabalho de mutações;
- Ao Dr. José Artur Bogo Chies pela disponibilidade a atenção e, especialmente, pela orientação durante os anos iniciais do doutorado;
- A este Programa de Pós-Graduação pela permissão do vínculo empregatício na Uniritter em concomitância com a bolsa de estudos;
- Ao Dr. Hugo Verli pelo convite para fazer parte da equipe de autores do livro de Bioinformática e ao Dr. Rodrigo Braun pela parceria nos capítulos;
- Ao Dr. Gonzalo Bello pelo convite de colaboração;
- À Dra. Vanessa Rodrigues Paixão-Côrtes que esteve ali desde muito antes da graduação, me acompanhou, influenciou, aconselhou e ajudou. Espero que façamos ainda muitas parcerias de trabalho no futuro aí em Salvador!
- À Pinguinzada, estas seis bravas mulheres que vieram, por sorte do destino, se juntar a mim na Uniritter e se tornaram pessoas muito especiais. Agradeço pelo encorajamento, pelas trocas e pela parceria;

- À Sheila e a Teresinha, por estarem presentes, pela amizade, pelas risadas e pelos vários cafés na Escola Técnica. Vocês foram grandes incentivadoras disto tudo!
- Aos amigos de graduação, Juliana, Gabriela, Mauro e Jorge por estes bons anos de amizade, conversas e risadas;
- Aos meus fieis escudeiros, Rodrigo e Bruno, por estarem sempre presentes dispostos a ouvir e a ajudar;
- Aos amigos de infância, Dedé e Tiago, pela paciência nestes anos;
- Ao Marcelo, pelo exemplo, ajuda e incentivo durante mestrado e boa parte do tempo de doutorado;
- Ao meu sobrinho, Névil, meu irmãozão, sempre disposto, sempre ouvidos, sempre atento, sempre presente. Inspiração de educação, humanidade e respeito!
- Às minhas irmãs, Dania, Deise e Denise, e também à agregada da família, Daiane (risos!), pelo apoio, pelo carinho e pelo incentivo. A participação de vocês nisto tudo foi muito importante!
- Ao meu pai, que sempre apoiou e torceu. Apesar de não entender o que exatamente são estes anos de dedicação para mestrado e doutorado, sempre esteve incentivando e vibrando orgulhoso com as minhas conquistas;
- Em especial à minha mãe, por ser um exemplo de pessoa. Todo o amor, a doçura e a dedicação para a família foram o suporte maior para a construção do que eu e minhas irmãs somos hoje. Desde muito cedo, foi a principal incentivadora dos estudos. Lembro de nós, sentados juntos, estudando história, geografia, ciências e português ainda lá no ensino fundamental. A paciência e a facilidade com que aqueles momentos ocorriam entre nós ficarão pra sempre marcados! Obrigado, mãe, pelo amor, pela sensibilidade, pela dedicação, pelo carinho, pelos estudos, pelo incentivo e por estar sempre comigo! Te amo;

SUMÁRIO

LISTA DE APÊNDICES

ABREVIATURAS

AIDS	Acquired Immunodeficiency Syndrome (Síndrome da Imunodeficiência Adquirida)
ART	(Antiretroviral Therapy) Terapia Antirretroviral
CRF	Circulating Recombinant Form (Forma Recombinante Circulante)
DST	Doenças Sexualmente Transmissíveis
HIV	Human Immunodeficiency Virus (Vírus da Imunodeficiência Humana)
HIV-1B	Vírus da Imunodeficiência Humana Tipo 1 grupo M subtipo B
HSH	Homens que fazem Sexo com Homens
INTR	Inibidores Nucleosídeos da Transcriptase Reversa
PS	Profissionais do Sexo
SIV	Simian Immunodeficiency Virus (Vírus da Imunodeficiência Símia)
SIVcpzPtt	SIV infectante de chimpanzés da espécie *Pan troglodytes troglodytes*
SIVgor	SIV infectante de gorilas
UNAIDS	Joint United Nations Program on HIV/AIDS (Programa de HIV/Aids da Organização das Nações Unidas
URF	Unique Recombinant Form (Forma Recombinante Única)

RESUMO:

O deslocamento humano e o comportamento sexual são os principais fatores que impulsionaram a pandemia do Vírus da Imunodeficiência Humana Tipo 1 (HIV-1) para o perfil atual. Dentro da extensa variabilidade genética do HIV-1, o subtipo B (HIV-1 B) é a variante mais disseminada geograficamente, relacionando-se a aproximadamente 11% de todas as infecções no mundo. A estrutura intrínseca da transmissão do HIV-1B entre diferentes indivíduos tem valiosa importância para a compreensão da epidemia e para as intervenções em saúde pública. Assim, o presente estudo tem como objetivo caracterizar epidemiologicamente a dinâmica de transmissão e disseminação do HIV-1 subtipo B. Duas abordagens metodológicas foram empregadas: (1) Caracterização genética e temporal da diversidade molecular do HIV-1B em diferentes pontos no Brasil, através de coleta, amplificação e sequenciamento do material genético viral e, (2) Identificação das cadeias de transmissão existentes na epidemia do HIV-1B no Brasil através da seleção de sequências disponíveis em bancos de dados e posterior reconstrução filogenética. Nossos resultados mostram que a epidemia de HIV-1B é largamente influenciada por tendências regionais e sugerem uma maior probabilidade de transmissão do HIV entre indivíduos de mesma origem geográfica. A aparente diferença na prevalência da variante B”-GWGR em distintas regiões do Brasil e a relação de assinaturas genéticas virais com grupos específicos de exposição em determinados pontos do país corrobora a dinâmica epidemiológica localizada. Este isolamento na epidemia, aparentemente particular para o Brasil e demais países da América do Sul, pode ser um reflexo da discreta conexão nodal entre as diferentes cidades no continente, garantindo importante limitação local da epidemia de HIV-1B. Além disso, identificamos um curto prazo na dinâmica de transmissão do vírus entre indivíduos, envolvendo em média 29 meses. No grupo de indivíduos homossexuais/bissexuais (HSH), especificamente, o intervalo foi estimado em um ano. Estes resultados revelam uma dinâmica particular para o grupo HSH na epidemia do HIV-1B, em que o intervalo de tempo entre as novas infecções é muito estreito e possui ampla participação de indivíduos recém-infectados. A ampla coleta de dados e a utilização de métodos estatísticos robustos permitirão melhor entendimento da dinâmica epidemiológica do HIV e, através de programas ativos de saúde pública, podem influenciar no direcionamento de campanhas de intervenção para populações específicas.

ABSTRACT:

The human displacement and sexual behavior are the main factors driving the human immunodeficiency virus type 1 (HIV-1) pandemic to the current profile. Within the extensive genetic variability of HIV-1, subtype B (HIV-1 B) is the most widespread variant being related to approximately 11% of all infections worldwide. The intrinsic structure of the HIV transmission among different individuals has valuable importance for the understanding of the epidemic and for the public health response. Thus, this study aims to characterize the epidemiological dynamics of HIV-1B transmission and dissemination. Two methodological approaches are required: (1) genetic and temporal characterization of molecular diversity of HIV-1B at different points in Brazil, through collection, amplification, and sequencing of viral genetic material, and (2) identification of transmission clusters within the HIV-1B epidemic in Brazil by reconstruction of phylogenetic trees using sequences selected from public databases. Our results show that the HIV-1B epidemic is largely influenced by regional trends and suggest a higher probability of HIV transmission between individuals from the same geographic origin. The apparent difference in the prevalence of the B"-GWGR variant in different regions of Brazil and the relationship of viral genetic signatures with specific exposure groups in certain parts of the country also supports the main influence of local factors in the HIV-1B epidemic. This epidemiological isolation apparently particular for Brazil and other South American countries may be a reflection of the discrete nodal connection between different cities within the continent. In addition, we identified a short-term dynamic spread within the transmission clusters in which the mean transmission time is approximately two years. In the group of homosexual/bisexual (MSM) specifically the intertransmissions interval has been estimated at about 1 year. These results show a specific dynamic in the HIV epidemic for the MSM group, and show a narrow interval between new infections with extensive involvement of newly infected individuals. These data highlight the need for better characterization of the HIV epidemic in Brazil and, in addition, show the need for specific prevention measures for concentrated epidemics. Public health services can be broadly benefited from this kind of information in order to guide intervention programs in public health.

Capítulo 1: Introdução

1.1 Considerações Iniciais:

Após mais de três décadas da descoberta do vírus da imunodeficiência humana (HIV) e do seu desenfreado alastramento pelos diferentes grupos populacionais, refletindo em mudanças nos contextos sociais, econômicos e culturais, o mundo se deparou recentemente com uma reversão da pandemia (Hemelaar 2012; UNAIDS 2014a). Apesar de aproximadamente 35 milhões de pessoas conviverem com o vírus, a epidemia, desde 2001, apresenta uma queda de 38% no número de novos casos. O número de mortes relacionadas à consequente Síndrome da Imunodeficiência Adquirida (do inglês *Acquired Immunodeficiency Syndrome*, Aids), causada pelo vírus, não decresceu, mas se encontra em um patamar estável. Por outro lado, a qualidade de vida dos pacientes tem sido amplamente beneficiada pela crescente disponibilidade e adesão à terapia antirretroviral (UNAIDS 2013).

A ação global conjunta em busca da retração na expansão da epidemia demonstra a abrangência do impacto de ações da iniciativa pública (UNAIDS 2013). Em poucas décadas, a associação de políticas engajadas a mudanças sociais, inovação tecnológica e injeção de recursos financeiros transformou a resposta à epidemia em um modelo de vanguarda em saúde (Bello et al. 2011). Ações políticas de informação, conscientização e prevenção contribuíram sobremaneira para as estimativas

promissoras. Além disso, as políticas de combate à epidemia estimularam alterações sociais e comportamentais na população em geral, em especial, entre jovens adultos, profissionais do sexo (PS) e seus clientes, indivíduos que utilizam drogas injetáveis (UDI), homens que fazem sexo com homens (HSH) e transgêneros (UNAIDS 2011).

Em países onde há uma epidemia generalizada, a combinação de mudanças sociais e comportamentais (incluindo a redução no número de parceiros sexuais, aumento no número de indivíduos que utilizam camisinha e atraso para iniciação da vida sexual) foi fundamental para a redução na incidência de HIV (UNAIDS 2013). À exemplo da circuncisão, diversos países tem iniciado programas de intervenção cirúrgica voluntária para jovens meninos ainda não em idade reprodutiva devido às estimativas de redução em até 60% das taxas de transmissão do HIV durante o contato sexual (Auvert et al. 2009). Modelos matemáticos sugerem que sem as mudanças comportamentais e sociais ocorridas ao longo dos trinta anos da epidemia, a incidência mundial de HIV poderia atingir o dobro dos valores atuais, adicionando 35 mil novas infecções por ano (UNAIDS 2012).

Acima de tudo, a introdução da terapia antirretroviral (ART) como método de tratamento propiciou uma dramática redução no número de mortes associadas à AIDS e garantiu uma melhoria na qualidade de vida dos indivíduos infectados (Thompson et al. 2012). Ainda, recentemente, diversos estudos têm mostrado que a supressão viral ocasionada pelo uso efetivo da terapia é capaz de diminuir as chances de transmissão do vírus entre indivíduos e mesmo, quando usada regularmente por indivíduos não infectados pelo HIV, pode fornecer proteção contra a infecção (Donnell et al. 2010; Grant et al. 2010). Como resultado global, a terapia antirretroviral beneficia atualmente aproximadamente 13 milhões de indivíduos, abrangendo mais de 37% daqueles que deveriam estar em tratamento (UNAIDS 2013; UNAIDS 2014a). Estima-se que, nos

próximos anos, todos os pacientes elegíveis terão acesso universal aos medicamentos, representando um investimento de mais de 24 bilhões de dólares por parte dos governos (UNAIDS 2012).

Apesar dos recentes avanços em relação às novas tecnologias, ao tratamento e às mudanças culturais e sociais nos mais diversos contextos, a epidemia de HIV ainda é desigual ao redor do mundo e atinge aproximadamente dois milhões de novos indivíduos todos os anos (UNAIDS 2014a). As promissoras estimativas internacionais, nacionais, ou mesmo regionais devem ser interpretadas com cautela, já que refletem números médios de uma população e acabam por mascarar pontos negativos de microepidemias em determinadas regiões. Assim, a ineficácia das diversas medidas em bloquear completamente a ainda corrente disseminação do vírus cria um desafio para as políticas de manejo de saúde pública e põe em questionamento a eficiência de tais métodos para toda a heterogeneidade e atual extensão da epidemia de HIV. Esse cenário ainda enfatiza a necessidade de abordagens inovadoras para melhor entender a transmissão do HIV em um nível populacional (Dennis et al. 2012).

Segundo a UNAIDS, a rápida resposta contra as altas taxas de mortalidade, morbidade e transmissão relacionadas ao HIV, em certas localidades ao redor do mundo, exige um amplo e intenso esforço para coleta de dados a respeito do vírus e dos próprios pacientes (UNAIDS 2014b). Complementarmente, a aplicação de métodos de análise, adequados para a interpretação destas informações clínico-demográficas, podem guiar esforços intensos e focados em saúde pública, proporcionando o máximo impacto sobre as epidemias locais (UNAIDS 2014b). Estas informações estratégicas, além disso, contribuem para a manutenção de uma vigilância epidemiológica eficiente e podem garantir formas de beneficiar um maior número de pacientes infectados (Bello et al. 2011; Thompson et al. 2012).

No contexto metodológico, análises filogenéticas vêm sendo amplamente utilizadas para a investigação sobre origem, disseminação e transmissão de HIV entre diferentes indivíduos (Brenner et al. 2007; Lewis et al. 2008; Junqueira et al. 2011; Faria et al. 2014). Abordagens específicas, especialmente sobre dispersão e transmissão do vírus, se mostraram poderosos métodos para a compreensão de questões sociais, demográficas e geográficas pertinentes à epidemia (Rothenberg et al. 1998; Lewis et al. 2008; Almeida et al. 2012; Junqueira et al. 2013). Ainda, estas análises se propõem a correlacionar contextos locais com vias de transmissão, resistência às drogas, comportamentos de risco e cadeias de transmissão (Lama et al. 2006; Hué et al. 2009; Chalmet et al. 2010; Jia et al. 2014). A ampla coleta de dados e a utilização de métodos estatísticos robustos permitem um melhor entendimento da dinâmica epidemiológica do HIV e, através de programas ativos de saúde pública, podem influenciar no direcionamento das campanhas para populações específicas com o objetivo de reduzir as taxas de transmissão e, consequentemente, retardar o aumento do número de novos casos (Brenner and Wainberg 2013).

1.2 Surgimento e diversificação do HIV:

O início da história da infecção pelo HIV em humanos é um exemplo catastrófico da relação entre diferentes espécies (Yamaguchi et al. 2000; Santiago et al. 2005; Nau et al. 2009). O impacto da colonização de potências europeias e a consequente imposição de um modelo econômico nas comunidades africanas colaboraram sobremaneira com a origem, expansão e difusão do vírus pela população humana (Perrin et al. 2003). Aproximadamente um século após o aparecimento do HIV, oito mil pessoas morrem diariamente vítimas da infecção e, apesar da redução nas

estimativas, milhares de novos casos são detectados anualmente (Worobey et al. 2008; UNAIDS 2013; UNAIDS 2014a).

O surgimento de duas linhagens de HIV, HIV-1 e HIV-2, é o resultado de, no mínimo, doze transmissões zoonóticas independentes de vírus símios para humanos (Wertheim and Worobey 2009; Nau et al. 2009). Atualmente, mais de 30 espécies de primatas não-humanos abrigam infecções pelo vírus da imunodeficiência símia (SIV) no continente africano e, embora espécie-específicos, a transmissão destes vírus entre as diferentes espécies já foi observada (Hahn et al. 2000; Sharp and Hahn 2010). Apesar dos relatos, a frequência destes eventos e o impacto para a evolução do vírus e para o desenvolvimento da patogenia ainda estão sendo estudados (Trivedi 2010).

O HIV-1, especificamente, se originou através do contato interespécies entre humanos-chimpanzés e humanos-gorilas na região centro-ocidental da África (Gao et al. 1999; D'arc et al. 2015). Através de eventos de transmissão, o vírus parece ter cruzado a barreira entre espécies algumas vezes e em quatro ocasiões foi capaz de se estabelecer no organismo humano em níveis suficientes para se tornar detectável e transmissível (Gao et al. 1999; Sharp et al. 2001; Kalish et al. 2005; Wertheim and Worobey 2009; Sharp et al. 2010; Trivedi 2010). Essas efetivas passagens geraram quatro grupos filogeneticamente relacionados de HIV-1, denominados M (major), N (new), O (outlier) e P (Charneau et al. 1994; Simon et al. 1998; Gao et al. 1999; Nau et al. 2009; Hemelaar et al. 2011; Faria et al. 2014).

Os grupos M e N do HIV-1 são filogeneticamente relacionados e muito provavelmente originaram-se de eventos de transmissão independentes do SIV infectante de chimpanzés da espécie *Pan troglodytes troglodytes* (SIVcpzPtt) para humanos na África Central (Gao et al. 1999; Bailes et al. 2003). Os grupos O e P estão intimamente relacionados ao SIV circulante na população de gorilas (SIVgor) e muito

provavelmente surgiram a partir da transmissão gorila-humano (Nau et al. 2009; D'arc et al. 2015). O fato de os vírus de símios terem sido transmitidos à outra espécie em inúmeras ocasiões não é surpreendente devido ao contato próximo entre macacos e humanos em algumas regiões na África, seja como animais de estimação ou mesmo como caça (Bibollet-Ruche et al. 2004; Kalish et al. 2005; Keele et al. 2006).

Apesar de um local de origem relacionado, os grupos filogenéticos do HIV-1 se disseminaram de forma diferente ao redor do mundo (Vidal et al. 2000; Berry et al. 2001; Rambaut et al. 2004; Kalish et al. 2004; Vallari et al. 2010). Enquanto os vírus dos grupos N, O e P são restritos principalmente a Camarões e seus países vizinhos, o grupo M foi capaz de estabelecer uma pandemia e hoje é responsável pela grande maioria das infecções por HIV, afetando globalmente aproximadamente 35 milhões de pessoas (Gürtler et al. 1996; Nau et al. 2009; Vallari et al. 2010; Delaugerre et al. 2011; GHO 2013). A diferença na propagação dos vírus de diferentes grupos do HIV-1 pode estar relacionada a variações na proteína viral Vpu. Enquanto nos vírus do grupo M esta proteína, ao longo do processo evolutivo, se tornou um antagonista totalmente funcional dos fatores de restrição intracelulares humanos, especialmente de teterinas, nos demais grupos, esta proteína seria incapaz de inibir eficazmente estes fatores e criaria uma barreira para a disseminação efetiva dos vírus na população humana (Sauter et al. 2009). Adicionalmente, a disseminação diferencial dos grupos de HIV-1 pode ter sido uma consequência de eventos aleatórios que impediram a entrada do vírus em uma população suficientemente propícia para a expansão da população viral.

Após a transmissão do SIVcpzPtt de chimpanzés para humanos no centro-sul de Camarões (Keele et al. 2006), o grupo M provavelmente se expandiu em Kinshasa, capital da República Democrática do Congo, por volta de 1920 (1909-1930) (Faria et al. 2014). A mudança comportamental das profissionais do sexo na região, o crescimento

urbano e a mobilidade humana através das ligações ferroviárias tiveram um papel fundamental na propagação inicial do vírus. O crescimento exponencial da população viral em conjunto com as altas taxas de mutação e as elevadas taxas de replicação do HIV no hospedeiro humano permitiram a geração de extrema variabilidade genética no vírus (Vidal et al. 2000; Worobey et al. 2008). Em conjunto, estes fatores explicam a diversificação do grupo M em variantes virais geneticamente distintas (denominadas subtipos) e formas geneticamente recombinantes.

Atualmente, a epidemia do grupo M é composta por nove subtipos (A, B, C, D, F, G, H, J, e K) e mais de 65 formas circulantes recombinantes (http: //www.hiv.lanl .gov / content / seqüência / HIV / CRFs / CRFs.html). A alta variação genética intrasubtipo, em alguns casos, levou à identificação de sub-subtipos. O subtipo A é filogeneticamente dividido em A1, A2, A3, A4, A5 e A6 enquanto que o subtipo F é classificado em F1 e F2 (Lihana et al. 2012). A África, por tratar-se do local de origem, é o continente com a maior diversidade de grupos e subtipos do HIV-1 (Figura 1.1). O subtipo C, responsável por metade das infecções por HIV no mundo, é predominante no sul e no leste do continente, além de ter grande prevalência na Índia (Hemelaar et al. 2011). Infecções pelos subtipos A, D, F, G, H, J, K e ainda pelo CRF01_AE e CRF02_AG são também descritas na região (Taylor et al. 2008).

Nas Américas, o subtipo B circula predominantemente na região e, especificamente na América do Sul, o subtipo F e recombinantes BF possuem grande importância epidemiológica (Figura 1.2) (Russell et al. 2000; Nadai et al. 2009; Mehta et al. 2010). Outros subtipos e formas circulantes recombinantes como o CRF12_BF, URFs BF, subtipo A e o CRF02_AG já foram também notificados na região (Thomson et al. 2002; Hierholzer et al. 2002; Carrion et al. 2003; Aulicino et al. 2005; Leal et al. 2007; Bello et al. 2010). Recentemente, o subtipo C tem se disseminado pela região sul

da América do Sul e propiciado o surgimento de recombinantes BC (Brígido et al. 2007; Bello et al. 2008; Fontella et al. 2008). América Central e Caribe possuem também uma epidemia baseada na circulação do subtipo B, além de abrigar menor quantidade dos subtipos C, D, F, G, H e J (Osmanov et al. 2002; Cuevas et al. 2002; Hemelaar et al. 2011).

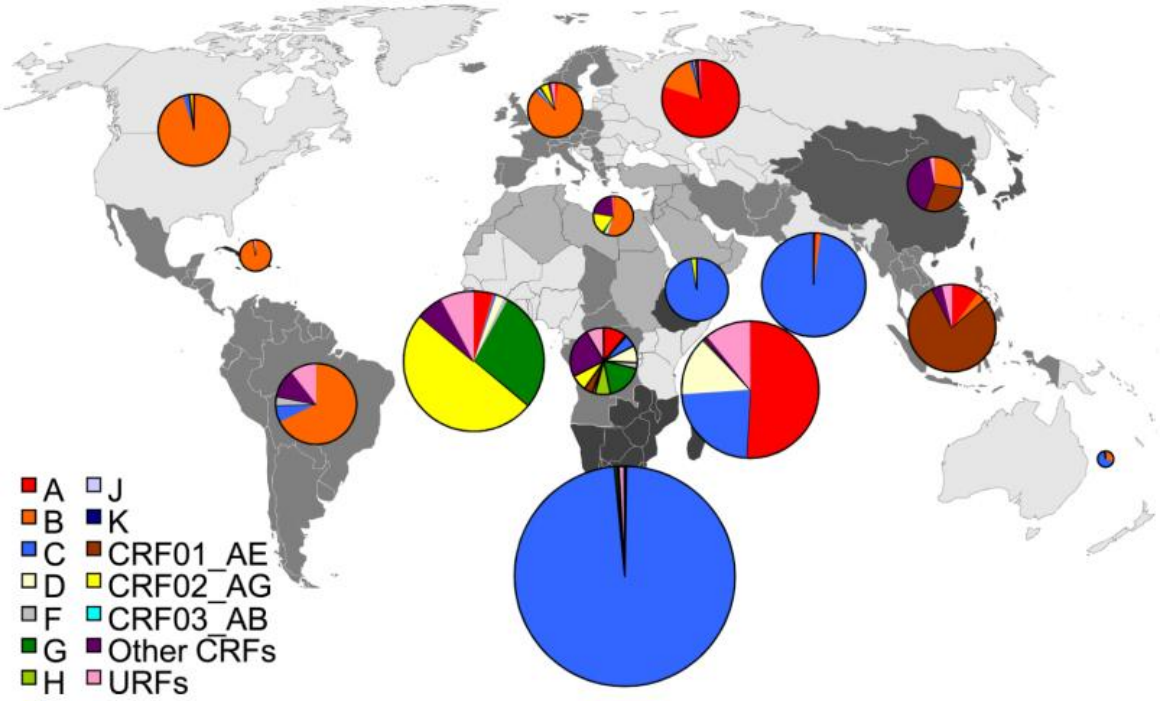


Figura 1.1. Distribuição regional dos subtipos do HIV-1 e suas formas recombinantes. Fonte: Hemelaar, 2011 (40).

As epidemias da Europa Oriental e da Europa Ocidental parecem diferir e abrigar diferentes subtipos de HIV-1. A epidemia da Europa Ocidental, assim como as Américas e a Oceania, é composta principalmente pelo subtipo B (88% das infecções). O restante das infecções é dividido pelos subtipos C, A, G e pela forma recombinante CRF02_AG. Na Europa Oriental e na Ásia Central, porém, 79% das infecções são causadas pelo subtipo A e apenas 15% pelo subtipo B. Além disso, esta região é a única no mundo onde o CRF03_AB parece circular significativamente. Finalmente, a Ásia

parece ser epidemiologicamente dominada pelo CRF01_AE e em menores quantidades circulam o subtipo B e diversos recombinantes, em especial as formas CRF07_BC e CRF08_BC (Hemelaar et al. 2011; Abecasis et al. 2013; Ye et al. 2014).

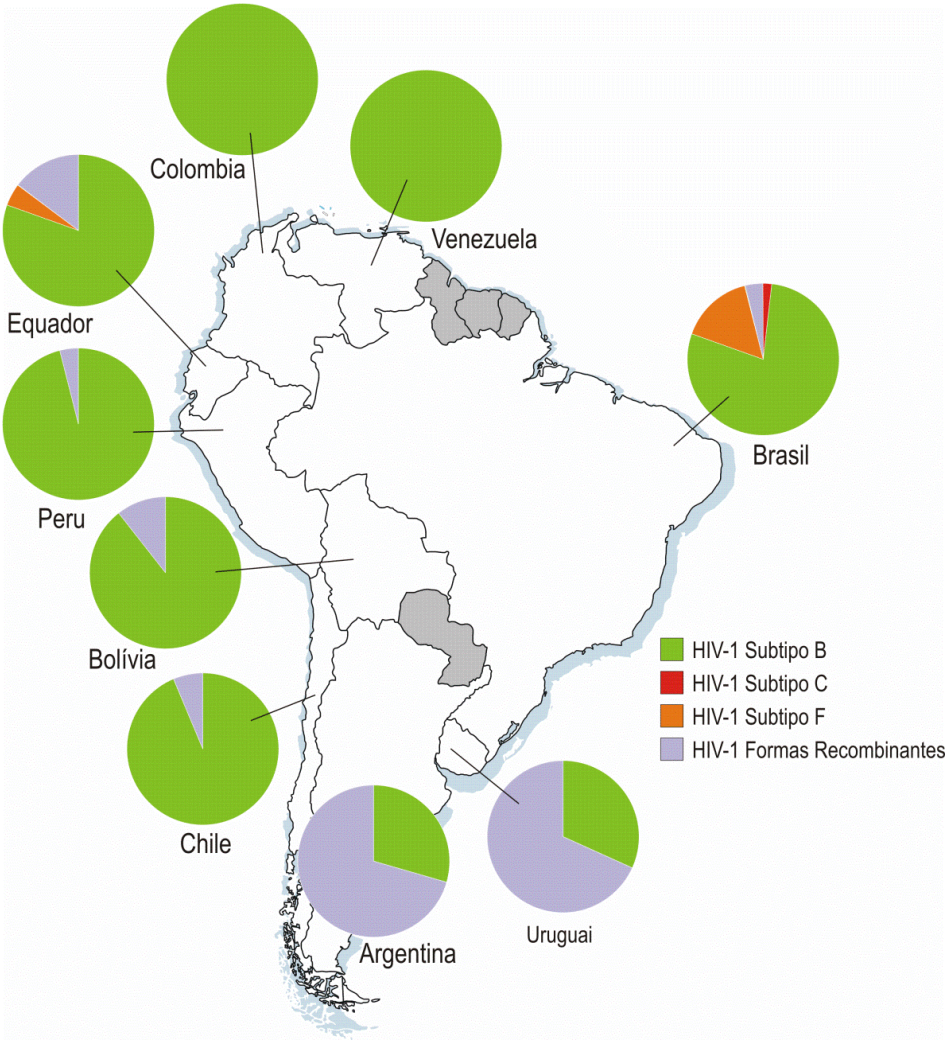


Figura 1.2. Epidemiologia Molecular do HIV-1 na América do Sul. Os diagramas mostram as porcentagens dos subtipos circulantes em cada país. As porcentagens referentes ao Equador, Peru, Bolívia, Uruguai e Argentina são baseadas na subtipagem das regiões *Gag*, *Pol*, *Int e Env* (Hierholzer et al. 2002). Para Brasil (Bongertz et al. 2000) e Chile (Rios et al. 2005) as porcentagens são referentes à análise da região *Env*. Para a Venezuela os dados de subtipagem foram obtidos de *Pol*, *Env*, *Vif ou Nef* (Rangel et al. 2009). Para a Colômbia os dados são resultados de análise da região *Pol* (Sanchez et al. 2006). Para as áreas destacadas em cinza não foram apresentados dados de epidemiologia molecular do HIV-1.

Durante 60 anos, o HIV-1 se espalhou silenciosamente entre a população humana. Os primeiros casos de infecção por HIV só foram notificados em 1981, nos Estados Unidos, em um grupo de jovens gays que apresentavam grave deficiência imunológica (Gottlieb et al. 1981). A notificação foi seguida por um número crescente de casos que, mais tarde, foram relacionados à Síndrome da Imunodeficiência Adquirida (Prusiner 2002). A relação entre aids e seu agente etiológico, o HIV, só foi resolvida em 1983, por um grupo de pesquisadores coordenador pelo Dr. Luc Montagnier, no Instituto Pasteur (França), e pela equipe do Dr. Robert Gallo, no Instituto Nacional de Saúde (Estados Unidos) (Barré-Sinoussi et al. 1983; Gallo et al. 1983).

1.3 Subtipo B:

O HIV-1 grupo M subtipo B (HIV-1B) teve um papel importante para a história da epidemia de HIV/aids. Além de ser a forma genética viral mais disseminada no mundo, as amostras isoladas por Gallo, Montagnier e colaboradores para identificar o agente causador da aids eram *quasispecies* do vírus relacionadas ao subtipo B. Como primeiro clado a ser identificado e caracterizado, tornou-se um modelo para estudos e hoje parece ser o subtipo mais estudado no mundo inteiro (Leitner 1996). A maioria dos medicamentos antirretrovirais, utilizados atualmente, foi concebida através de experimentos cujo modelo viral é baseado no HIV-1B. De forma semelhante, diversos experimentos *in vitro,* incluindo modelos de infecção, são também primariamente derivados desta variante (Parkin and Schapiro 2004; Santos et al. 2008).

Evidências filogenéticas e históricas sugerem que o subtipo B surgiu como uma linhagem viral após a disseminação do HIV-1 a partir da África (Gilbert et al. 2007). O ancestral comum do subtipo B foi originado provavelmente em Kinshasa e já circulava

nesta região na metade da década de 1940 (Faria et al. 2014). A partir da África, este subtipo emergiu nas ilhas do Caribe, por volta de 1966, introduzido por emigrantes Haitianos que retornavam da região do Congo (Gilbert et al. 2007; Junqueira et al. 2011). A expansão da epidemia local envolveu a disseminação do vírus para diferentes ilhas na região a partir da epidemia instaurada no Haiti (Junqueira et al. 2011; Pagán and Holguín 2013; Cabello et al. 2014; Cabello et al. 2015). Estudos recentes sugerem que várias linhagens emergiram deste país e foram disseminadas para outros países do Caribe e da América Latina, onde, puderam estabelecer epidemias locais efetivas (Cabello et al. 2014; Cabello et al. 2015). Ainda hoje, vários países apresentam uma epidemia composta primariamente por estas linhagens (Cabello et al. 2015).

Historicamente, do Caribe, o vírus foi introduzido diretamente na população dos Estados Unidos por volta de 1969 (Gilbert et al. 2007). A conexão direta Haiti-EUA mudou a epidemia e aumentou exponencialmente o número de novos casos de infecção pelo vírus. A circulação do HIV-1B em um grupo com comportamento de alto risco com altas taxas de trocas de parceiros, como usuários de drogas injetáveis ou grupos homossexuais, pode explicar o êxito da epidemia instaurada neste país (Pape et al. 1983; Kuiken et al. 2000). As extensas relações internacionais entre países (especialmente através do tráfego aéreo) foram fundamentais para a disseminação do HIV-1B para além dos limites da América e, potencialmente refletem o elevado grau de conexão de uma sociedade globalizada (Perrin et al. 2003). A partir dos Estados Unidos, o vírus foi disseminado para diferentes partes do mundo, incluindo Oceania, Ásia, Europa, América Latina e Caribe (Gilbert et al. 2007). Atualmente, o HIV-1B é responsável por aproximadamente 11% de todas as infecções por HIV ao redor do mundo.

1.4 Subtipo B no Brasil:

Dados filogenéticos sugerem que o HIV-1B foi introduzido na América do Sul, a partir de vírus circulantes na América do Norte e no Caribe, em meados da década de 1960 (Gilbert et al. 2007; Junqueira et al. 2011). No Brasil, o vírus emergiu através de diferentes entradas, provavelmente entre os anos de 1965 e 1970, de forma semelhante à origem das epidemias de regiões desenvolvidas, como Estados Unidos e Europa (Bello et al. 2007). A uniformidade nas datas de origem de diferentes epidemias do HIV-1B ao redor do mundo sugere a disseminação concomitante do vírus para diferentes países a partir de um mesmo foco, provavelmente nos Estados Unidos (Gilbert et al. 2007).

O epicentro da epidemia no Brasil se deu provavelmente na região sudeste, já que as primeiras notificações de aids dentro do país ocorreram nestes municípios, especialmente São Paulo e Rio de Janeiro (Ministério da Saúde et al. 2005). Dados epidemiológicos, revelam que 89% dos casos de aids notificados no Brasil até 1985, haviam sido detectados em municípios da região sudeste e análises retrospectivas demonstram que os primeiros casos da infecção ocorreram no início da década de 1980 na mesma região. Em conjunto, estes elementos sugerem que a introdução inicial do HIV-1B na população brasileira ocorreu através da região sudeste, em algum momento antes de 1980 (Ministério da Saúde et al. 2005; Bello et al. 2006).

Após a introdução, provavelmente em grupos de transmissão altamente interligados, a epidemia do HIV-1B cresceu exponencialmente nos primeiros anos. Na metade da década de 1980, no entanto, verifica-se uma queda no número de novas infecções e a epidemia desacelera seu crescimento (Bello et al. 2006; Bello et al. 2007). Este padrão pode ser explicado pela mudança no comportamento sexual da população ou pela saturação das principais cadeias de transmissão difusoras do vírus no país (Blower 1991; Levi and Vitória 2002). As altas taxas de crescimento descritas para o

subtipo B podem ser atribuídas à introdução do vírus em indivíduos HSH e redes de usuários de drogas injetáveis, em especial durante os primeiros anos da epidemia (Bello et al. 2007).

Estudos moleculares têm mostrado ao menos duas linhagens geneticamente distintas de subtipo B circulando no Brasil, de acordo com a região V3 do genoma viral (Morgado et al. 1994). Nos países da Europa e da América do Norte, é notável a prevalência de linhagens do HIV-1B que apresentam o motivo GPGR no topo da alça V3 (região responsável pela interação com as proteínas humanas no momento da infecção), enquanto em outros países como Brasil, China, França, República Tcheca, Filipinas e Cuba estudos epidemiológicos têm detectado a variante GWGR (Potts et al. 1993; Louwagie et al. 1994; Morgado et al. 1994; Carr et al. 1998; Morgado et al. 1998; Casseb et al. 1998; Junqueira et al. 2013). Essa variante é comumente identificada como B", B'-GWGR, BBr ou B brasileiro (Casseb et al. 2002; Leal et al. 2008; Arruda et al. 2011; Junqueira et al. 2013). Apesar de a variante ter sido isolada pela primeira vez no Japão e ser detectada em 23 países diferentes ao redor do mundo, o Brasil possui uma epidemia de HIV-1B com prevalência de B" que varia entre 17% e 50%, dependendo do estado analisado (Casseb et al. 1998; Leal and Villanova 2010; Araujo et al. 2010; Franca et al. 2011; Junqueira et al. 2013). Estes dados de prevalência sugerem que o Brasil atuou como local de origem da variante B", mas sua origem monofilética no país ainda é alvo de contínua discussão na literatura (Covas et al. 1998; Bello et al. 2007; Diaz et al. 2008; Leal and Villanova 2010). Diversos estudos têm tentado desvendar as condições iniciais da epidemia e consideraram diferentes cenários para explicar a origem e disseminação da variante na população humana (Bello et al. 2007; Pinto et al. 2008; Diaz et al. 2008; Leal and Villanova 2010).

A presença do triptofano (W) na segunda posição do tetrapeptídeo parece afetar a eficácia na neutralização do vírus pelo sistema imune (Casseb et al. 2004). O motivo GWGR, desta forma, aumentaria a avidez dos anticorpos pela região V3 nas proteínas virais, inibindo a infecção de novas células e contribuindo para a progressão mais lenta da doença em comparação com a infecção pela variante B-GPGR (Casseb et al. 2004; Brito et al. 2006). Os pacientes infectados com esta variante tendem a ter uma progressão mais lenta para a aids, contagem de células T-CD4+ superior e menor carga viral (Santoro-lopes et al. 2000; Brito et al. 2006; Araujo et al. 2010). Ainda, estes pacientes são mais propensos a sobreviver à infecção do que indivíduos portadores da variante GPGR do subtipo B (Brito et al. 2006). Estes resultados sugerem que a diferença clínica na progressão para a aids pode ser associada à linhagem do HIV-1 infectante e demonstra a importância dos estudos de epidemiologia molecular para a compreensão da doença.

Apesar da relevância epidemiológica do subtipo B no Brasil, os subtipos F e C são também detectados na epidemia instaurada no país (Bongertz et al. 2000; Brindeiro et al. 2003; de Medeiros et al. 2011; Almeida et al. 2012). Diversas formas recombinantes, incluindo o CRF28_BF, CRF29_BF, CRF39_BF, CRF40_BF, CRF31_BC, além de diversas formas recombinantes entre os subtipos B/F e B/C são também descritos na epidemia brasileira (Guimarães et al. 2002; Soares et al. 2003; Sa Filho et al. 2006; Brígido et al. 2007; Guimarães et al. 2008).

1.5 Cadeias de Transmissão:

A diversidade genética do HIV dentro de um indivíduo infectado é determinada pelas altas taxas de mutação e recombinação, pela alta taxa de replicação, pelo subtipo do vírus, pela pressão do sistema imune e potencialmente pela pressão da terapia

antirretroviral (Wilbe 2004). Como consequência destes fatores, a composição genética do HIV é relativamente única para cada indivíduo infectado, mas potencialmente similar para indivíduos epidemiologicamente ligados, especialmente na relação transmissor-receptor. Assim, as variações genéticas entre as populações virais de cada indivíduo são marcadores da evolução intrapaciente e refletem, em última análise, os processos evolutivos resultantes da disseminação do vírus (Lemey et al. 2006). Diversos estudos tem demonstrado a utilidade de dados genéticos do HIV para a inferência a respeito dos contatos entre indivíduos e na reconstrução das principais teias de transmissão envolvidas na disseminação do patógeno em determinado local. Essas teias, atualmente, têm sido referidas como cadeias de transmissão (Hué et al. 2005; Lewis et al. 2008; Yerly et al. 2009; Zehender et al. 2010; Callegaro et al. 2011; Ng et al. 2013; Ragonnet-Cronin et al. 2013; Grabowski et al. 2014).

A impossibilidade de uma amostragem completa de indivíduos infectados na população e as respectivas informações de contato entre estes inviabiliza a aplicação de métodos clássicos de epidemiologia e abre espaço para a utilização de modernas técnicas moleculares. As análises filogenéticas, neste sentido, fornecem um método único para capturar os processos genéticos intrinsecamente envolvidos na disseminação do HIV entre diferentes indivíduos (Lemey et al. 2003; Dean et al. 2005; Brenner et al. 2007; Lewis et al. 2008; Salemi et al. 2008; Guimarães et al. 2009; Gray et al. 2009; de Oliveira et al. 2010; Leigh Brown et al. 2011; Junqueira et al. 2011; Brenner and Wainberg 2013; Faria et al. 2014; Cabello et al. 2015). Aliadas a informações clínicas e sociodemográficas, as análises filogenéticas, recentemente, tornaram-se importantes ferramentas na identificação das cadeias de transmissão existentes na epidemia de HIV (Brenner et al. 2007; Lewis et al. 2008; Yerly et al. 2009; Chalmet et al. 2010; Rieder et al. 2010; Callegaro et al. 2011). O método permite a identificação de pacientes

epidemiologicamente relacionados através da análise das relações genéticas entre as sequências de nucleotídeos do vírus (Prosperi et al. 2011; Ragonnet-Cronin et al. 2013). Ao mesmo tempo em que as altas taxas de mutação e a extrema diversidade genética são um obstáculo para o desenvolvimento de vacinas e drogas antivirais, a diversidade acumulada ao longo dos anos nos genomas virais possibilita a reconstrução histórica da expansão da epidemia e o entendimento da dinâmica de transmissão entre hospedeiros (Rambaut et al. 2004; Grabowski and Redd 2014).

Umas das principais dificuldades relacionadas à análise de sequências para identificação das relações epidemiológicas é a falta de uma metodologia consenso, particularmente quando consideradas filogenias que incluem um grande número de amostras (Chalmet et al. 2010; Prosperi et al. 2011). Cadeias de transmissão têm sido associadas à subclados filogenéticos, delimitados por um determinado número de indivíduos (de 2 à mais de 10 pacientes) e pelo grau de confiabilidade na identificação dos clados (comumente *bootstrap* ou probabilidade posterior) (Brenner et al. 2007; Lewis et al. 2008; Hué et al. 2009; Yerly et al. 2009; Hughes et al. 2009; Kouyos et al. 2010; Ragonnet-cronin et al. 2010; Audelin et al. 2013; Bezemer et al. 2014). Alguns autores, além da observação do valor de confiança dos clados, levam em consideração a diversidade genética dentro do grupo de transmissão (Prosperi et al. 2011; Ragonnet-Cronin et al. 2013). Restrições geográficas e distâncias patrísticas foram também adicionadas como fatores capazes de definir estas relações (Prosperi et al. 2011). Uma provável explicação para a incerteza na definição metodológica e formal a respeito do agrupamento de amostras em cadeias de transmissão é provavelmente relacionada à extensão e heterogeneidade de amostragem dos diferentes estudos realizados até o momento (Lewis et al. 2008; Yerly et al. 2009; Zehender et al. 2010). Recentemente, em um extenso estudo sobre a disseminação de HIV pelo mundo, os autores inovaram e

propuseram a utilização de *networks* para avaliar as transmissões do vírus entre os indivíduos (Wertheim et al. 2014).

As características das cadeias de transmissão afetam diretamente a dinâmica da infecção e prevalência da doença em uma determinada população (Eames and Keeling 2004). Fatores como geografia, modo de transmissão fatores de risco podem influenciar nos parâmetros das filogenias (Pybus and Rambaut 2009). Além disso, recentemente, foi demonstrado que os diferentes padrões de transmissão podem também afetar a estrutura (topologia) de filogenias (Colijn and Gardy 2014). Embora as transmissões de HIV ocorram preferencialmente entre indivíduos pertencentes a uma mesma categoria de exposição (heterossexual, homossexual e usuário de droga injetável), interações entre grupos pode desempenhar um importante papel na disseminação da epidemia (Lewis et al. 2008; Kouyos et al. 2010; Pennings et al. 2014). Adicionalmente, a complexidade no estudo da dinâmica das doenças infecciosas emerge da complexa rede de migrações estabelecidas pela população humana (Perrin et al. 2003; Kouyos et al. 2010). Assim, a identificação eficiente de grupos epidemiologicamente relacionados (relação geográfica ou de exposição) pode ser fundamental para desvendar as principais teias propulsoras da epidemia e permite revelar diferentes informações de interesse para saúde pública, incluindo a transmissão de mutações de resistência, principais grupos expostos e, até mesmo, pontos geográficos importantes na dispersão (Yerly et al. 2009; Chalmet et al. 2010; Rieder et al. 2010; Mehta et al. 2010).

Capítulo 2: Objetivos

2.1 Objetivo Geral:

Caracterizar epidemiologicamente a dinâmica de transmissão e disseminação do HIV-1 subtipo B.

2.2 Objetivos Específicos:

2.2.1. Caracterizar historicamente a origem e disseminação do HIV-1 subtipo B no mundo;

2.2.2. Caracterizar epidemiologicamente a dinâmica de transmissão da linhagem B"-GWGR do HIV-1 subtipo B circulante no Brasil;

2.2.3. Identificar e descrever os pares e cadeias de transmissão do HIV-1 subtipo B associados ao Brasil, de maneira a atribuir a relação epidemiológica entre os diferentes estados e mesmo entre os diferentes países da América do Sul.

Capítulo 3: Artigo 01

"HIV-1 Subtype B: Traces of a pandemic"

Dennis Maletich Junqueira, Sabrina Esteves de Matos Almeida

Manuscrito a ser submetido

Virology

HIV-1 Subtype B: traces of a pandemic

Dennis Maletich Junqueira[1, 2, 3], Sabrina Esteves de Matos Almeida[1, 2, 4]

[1] Centro de Desenvolvimento Científico e Tecnológico (CDCT), Fundação Estadual de Produção e Pesquisa em Saúde (FEPPS), Porto Alegre, RS, Brazil.

[2] Programa de Pós-Graduação em Genética e Biologia Molecular, Universidade Federal do Rio Grande do Sul (UFRGS), Porto Alegre, RS, Brazil.

[3] Uniritter Laureate International Universities, Departamento de Ciências da Saúde, Porto Alegre, RS, Brazil.

[4] Instituto de Ciências da Saúde, Universidade FEEVALE, Novo Hamburgo, RS, Brazil

Corresponding Author:

Dennis Maletich Junqueira

Key words:

HIV-1, Subtype B, dissemination, epidemiology

Abstract:

Human migration is one of the major processes shaping the origin and dissemination of HIV viruses. Within HIV-1, Subtype B (HIV-1B) is the most disseminated variant and is assumed to be the causative agent of approximately 11% of all infections around the world. Data from phylogenetic studies revealed that HIV-1B emerged in Kinshasa, Africa, and due to human displacement have been introduced into the Caribbean region through Haiti in 1966. After a local dispersion, the virus was brought to the United States via homosexual/bisexual contact around 1969. Inside the country the HIV-1B epidemic experienced an exponential growth and became generalized in the population, affecting not only homosexual but also heterosexual individuals and injected drug users. Soon after, the virus was disseminated into several regions of the world establishing successful epidemics in countries of Europe, Asia, Latin America, and Australia. Recent studies are suggesting that in addition to this pandemic clade several lineages emerged from Haiti and reached other Caribbean and Latin American countries as the result of short-distance disseminations. Concomitantly, the presence of genetic variants related to subtype B has been demonstrated in the epidemic. Until now four genetic variants are described: subtype B' which mainly circulates in Thailand and other Asian countries, the Trinidadian and Tobagian variant, the GPGS variant which mainly circulates in Korea, and finally, the GWGR variant mainly detected in Brazil. This paper reviews the history of the HIV-1B covering its evolutionary history since the emergence and the impact caused by this strain in the human population.

Introduction:

Over the past 30 years, HIV/AIDS has evolved into a highly heterogeneous epidemic structured in multiple sub epidemics each influenced by several variables such as biological, behavioral, and cultural factors [1–4]. The impact of HIV-1 in the human population is catastrophic and the numbers related to the infection reveal the need for multisectorial efforts on different fronts to combat and reduce the number of new infections, expand the access to health services and guarantee the access to the antiretroviral therapy for the general population [5,6]. Today, HIV-1 affects more than 35 million individuals around the world and is actually the most important virus related to AIDS [5].

Genetically, HIV is characterized by enormous diversification and rapidly evolution [7,8]. The current HIV-1 pandemic is phylogenetically divided in four distinct groups, termed M (for Main), N (for non-M/non-O), O (for outlier), and P [9–12]. Despite the innumerous theories for the emergence of HIV, each of these groups is the result of an inadvertent cross-species transmission event of a Simian Immunodeficiency Virus (SIV) from African primates to humans even though contundent evidences of a direct transmission of a simian-to-human is still missing [13–17]. Species-specific strains of SIV are divided in six different lineages and were identified in more than 30 species of African primates [18–20]. Recent studies revealed that SIV has been present in monkeys and apes for at least 32,000 years and as in HIV, genetic recombination seems to be an important mechanism for the generation of variability [21,22].

HIV-1 groups M and N are closely related and most likely resulted from independent transmission events of SIV from infected chimpanzees of the subspecies *Pan troglodytes troglodytes* (SIVcpz) to humans in western-central Africa [15]. It is now confirmed that these two groups arose from geographically distinct chimpanzee

populations in Cameroon [23]. Group P is very closely related to SIV infecting gorillas (SIVgor), and is most likely resulted from gorilla-to-human transmission [11]. Finally, group O is also related to SIVgor, but it is unclear whether gorillas were also the immediate source of this variant and its origin remains to be identified [24–26]. The fact that primate viruses were transmitted to humans in innumerous occasions is not surprising due to the close contact between monkeys and humans in some regions on Africa, whether as a pet or as a hunting [18,23,27]. Recent studies investigating current zoonotic SIV infections in humans revealed that persons who hunt and butcher wild non-human primates are subject to ongoing exposure and potential infection with SIV or other primate retroviruses [27,28]. These results are alarming and of substantial importance for global public health since simians may represent potential reservoirs for the emergence of new lentiviral diseases in the human population.

Despite the common viral source, HIV-1 groups spread differentially throughout the world [29–33]. Groups N, O, and P viruses are mainly restricted to Cameroon and its neighboring countries despite the detection of some of these variants outside of this region [11,29,34,35]. In opposition, group M established a pandemic spread and is responsible for the great majority of all HIV infections worldwide. Currently, group M is detected in five continents and almost 35 million people are infected worldwide [36]. Several hypotheses were formulated to explain the difference in the spread of HIV group viruses. The most accepted theory argues that the HIV group M Vpu protein evolved to become a fully functional antagonist of human intracellular restriction factors. In constrast, Vpu proteins of groups O and N are unable to efficiently antagonize these factors creating a barrier to the effective spread of these viruses into the human population [37].

After the transmission from chimpanzees to humans in southeastern and south central Cameroon, group M viruses most likely spawned in Kinshasa, capital of the Democratic Republic of Congo, around 1920 (1909-1930) [23,38] (Figure 1). Subsequent to localized transmission in that city, the virus probably reached neighboring cities, such as Brazzaville, Lubumbashi, and Mbuji-Mayi in less than 20 years later [38]. The changing behavior of sex workers, urban growth and human mobility by railway connections had a fundamental role in the early spread of the virus into its host population and may explain the linkage among these cities. During this early phase of dissemination, group M viral population had an exponential growth comparable to the population growth of Kinshasa. After 1960, however, the growth rate of the viral population almost triplicates. This change in rates is most likely explained by the introduction of the virus in a high risk group and by the exposure of individuals to contaminated injections [38]. Such situation allowed the virus to disseminate in a highly interconnected chain of transmission and explains the ignition of the epidemic in Africa.

The localized dissemination of HIV-1 group M in Kinshasa and around cities allowed the generation of striking genetical diversity [30,39]. In addition to the number of hosts, such extreme variability is the result of the high mutation and recombination rates of the HIV reverse transcriptase enzyme in combination with the high rates of viral replication in the human hosts [33]. These factors together supported the diversification of HIV-1 group M into genetically distinct viral subtypes and recombinant forms. Currently, the HIV-1 Group M epidemic is composed of nine subtypes (A, B, C, D, F, G, H, J, and K) and more than 65 recombinant forms (http://www.hiv.lanl.gov/content/sequence/HIV/CRFs/CRFs.html). Within some subtypes, the high genetic variation leaded to the classification into sub-subtypes.

Subtype A has been subdivided into A1, A2, A3, A4, A5, and A6 and subtype F has been subdivided into F1 and F2 [40,41].

Subtypes within the HIV-1 Group M have been powerful epidemiological markers to track the course of the pandemic [30]. Despite the potential use in the epidemiological investigation of origin and dissemination, it is not clear the role of these genetic variants in the clinical outcome of HIV infection [42–44]. Several studies have identified punctual differences between infections by distinct subtypes; however, due to the great genetic diversity of both the host and the virus these questions are yet opened to debate. There are currently suggestions in the literature that the viral subtype can affect the transmission, the disease progression, the evolution rate, and may potentially affect antiviral drug resistance development [45–51].

During sixty years HIV-1 has spread silently within the human population. The first cases of HIV infection were only notified in 1981 in a group of young gay men presenting a severe immune deficiency in the United States [52]. That report was followed by a growing number of cases of what were later called Acquired Immune Deficiency Syndrome (AIDS) [53]. The relationship between AIDS and its cause were only solved in 1983 by a group of researchers coordinated by Dr. Luc Montagnier at the Pasteur Institute (France), and the team of Dr. Robert Gallo at the National Institutes of Health (United States) [54,55]. In 1986 at the conference of the International Committee on the Taxonomy of Viruses the virus was officially named Human Immunodeficiency Virus [56].

The HIV-1 Group M subtype B played an important role to the history of HIV/AIDS epidemic. The samples isolated by Gallo, Montagnier and colleagues to identify the causative agent of AIDS were *quasispecies* related to subtype B. As the first clade to be identified and characterized, HIV-1B has become a model for studies on the

virus and today seems to be the most studied subtype worldwide [57]. Most current HIV antiretroviral drugs were designed using HIV-1B and similarly, i*n-vitro* experiments are mainly based in such variant [58,59]. Due to the epidemiological and clinical importance of the subtype B around the world, several studies aimed to understand the origin and dissemination routes of this clade into the human population adding a new piece into this epidemiological puzzle [38,60–62].

Early HIV-1 Group M Subtype B epidemic:

Phylogenetic and historical evidences suggest that subtype B arose as a viral lineage after the spread of the HIV-1 out of Africa [61]. The common ancestor of subtype B was originated in Kinshasa and was already circulating in the region by 1944 [38]. The initial spread of the epidemic outside Africa began around 1966 [60,61] ignited by the returning to the home country of Haitian professionals who went to Africa in the beginning of 1960s and worked in the Congo mainly as bureaucrats or teachers [63,64] (Figure 1). About 4,500 Haitians were sent to Africa at that time [64]. This massive immigration back to Haiti is related to the independence of Congo, colony of Belgium since 1908, and the subsequent political crises in that country [65,66]. This major process of human immigration into Haiti may have introduced the virus in a new population of a different country and effectively started the international dissemination of HIV-1B. It is believed that the current worldwide HIV-1B pandemic resulted from this single introduction of a few samples from Africa into the Haitian population [61] (Figure 1).

The first case of AIDS in Haiti was only notified in 1983 [67], but retrospective analyzes identified a hospitalized patient with AIDS symptoms in 1979 [68]. Besides the inevitable incapacity of ascertain the sexual practices of the index cases, the initial

dissemination of the virus into Haiti was drove with most likely predominance by men who have had sex with other men (MSM) individuals [61,69–73]. The nature of the epidemic has changed over time and reached heterosexual individuals and recipients of blood transfusions inside Haiti [6]. In 1988, more than 90% of HIV infected patients living in Haiti acquired the virus through a heterosexual route and today it is the predominantly mode of HIV transmission [6,74]. Currently, Haiti has one of the higher prevalence rate of HIV/AIDS in the world reaching almost 2%, but incidences have declined around 49% recently [75].

Several outbreaks were seeded by the epidemic instaured at Haiti [60,61,76–78] (Figure 1). Due to the geographical proximity it is assumed that the virus spread through the Hispaniola Island (shared by the Dominican Republic and Haiti) and reached the Dominican Republic. Previous phylogenetic studies found sequences of this two nations intermixed with each other showing a uniform epidemic inside the island [60,76]. After its introduction, the virus may have been spreading slowly in the Hispaniola Island, most likely circulating in the heterosexual population [61,62,70]. Following this local epidemic, phylogenetic and historical evidences suggest that the virus moved out of Haiti on several independent occasions [60,62,76,78] (Figure 1).

Historically, from Haiti the virus was directly introduced into the United States [61] (Figure 1). Reaching the North America the virus was amplified in this population and subsequently has also established successful epidemics in Europe, Asia, Latin America, and Australia [12,57,61]. In addition, recent studies suggest that concomitantly to the pandemic clade, several lineages emerged from Haiti and reached other Caribbean and Latin American countries as the result of a short-distance spread. Due to the non epidemiological explosion caused by these strains, such lineages have been jointly identified as a non-pandemic clade [62,76] (Figure 1).

Several studies have identified genetic variants related to subtype B also circulating in the HIV-1B epidemic (Figure 2). Today four genetic variants are recognized and currently characterized in the epidemic: the B-Thai, the Trinidadian and Tobagian B, the Korean B, and finally the B"-GWGR. These variants represent well-established subclades of the HIV-1 subtype B that are circulating in specific regions around the world [46,79–86].

HIV-1 Subtype B pandemic Clade:

From the Hispaniola Island, a single or a few lineages of HIV-1 subtype B were carried out to the United States. The time to the most recent common ancestor of this U.S. HIV-1 epidemic is estimated to be 1969 (1966–1972) [60,61]. However, the first cases of AIDS inside the country were only reported in 1981 approximately twelve years after the introduction of the virus. This silently spread may be explained by the median time interval between infection with HIV and progression to AIDS of ten years in this region [87]. Retrospective studies identified AIDS cases among American and Haitian individuals living in the United States before 1981 [88–92], and the early evidence of HIV inside the country is dated to 1977 [93].

The direct connection Haiti-USA changed the HIV-1B epidemic and exponentially increased the number of new cases of HIV infection. In the United States, the virus was introduced early into the MSM population. Due to the higher prevalence of the infection in Haitian male patients (~85%) in the beginning of the epidemic, it is commonly assumed that homosexual contacts formed the "HIV bridge" between the two countries [68,94]. Moreover, Port-au-Prince, capital of Haiti, in the mid-1970s was a prime destination for American gay sex tourism, especially for male homosexuals from the New York City metropolitan area [64]. This link between Haiti-USA may have

facilitated the direct introduction of the virus into the MSM population inside the United States. The number of infected individuals early in the epidemic suggest no lag time in the spread of HIV upon introduction into the country, but rather, an exponentially growing proliferation into the susceptible population [92]. The use of injectable drugs may also have facilitated the proliferation of the virus inside the country [14].

In the 1980s, Haitian ancestry was identified as a risk factor for AIDS along with heroin addiction, hemophilia, and homosexuality [95]. The high number of AIDS cases among Haitians individuals in the United States in the beginning of the epidemic resulted in a generalized national stigmatization [96–98]. Similarly, many MSM individuals have suffered prejudice due to the association between homosexuality and HIV [99]. The prevalence of the infection became higher among these individuals than among heterosexuals most likely because of the number of different sexual partners that allows the virus to spread in higher rates [2,100,101].

There is some evidence for the importance of MSM individuals to the epidemic in the form of the notorious "patient 0". In March 1984, a CDC study tracking the sexual networks of MSM individuals in several states of the United States found one man to be the center of these transmission chains [102]. This individual was popularly pointed as the "patient 0" of the epidemic in the United States. However, later this theory were disproven [61].

Air travel and international relations between countries were fundamental to the spread of HIV-1 beyond the limits of America and potentially reflects the social connection of a globalized society [103]. This spread have been molded by a conjunction of factors that in addition to the viral characteristics reflects the social behavior of the human host [94]. The intense connection with other countries and the introduction of the virus in a high-risk behavior group may explain the successfulness of

the HIV-1B epidemic instaured at the United States. From that country the virus were disseminated to different parts of the world including Oceania, Asia, Europe, Latin America and the Caribbean effectively generating regional epidemics inside these places [61] (Figure 1).

HIV-1 Subtype B in Latin America and the Caribbean:

Subtype B remains the most prevalent HIV-1 variant throughout the Latin America and the Caribbean accounting for about 70% of infections in the region [104,105]. Cuba seems to be an exception due to the circulation of several HIV-1 subtypes and recombinant forms [106]. The HIV epidemic in Latin America is concentrated in and around networks of MSM individuals, although heterosexual contact is increasing and in Caribbean/Central America is the main mode of HIV transmission [107–110]. The Caribbean and Central American HIV-1 epidemics are unique since subtype B is the most prevalent form in the heterosexual individuals being fueled by commercial sex work in societies with widespread poverty [73,111]. A study conducted in Port of Spain (capital of Trinidad and Tobago) developed an ethnographic model that links sex between older males, younger females, crack, cocaine and sex for money as risk factors in the heterosexual spread of HIV-1 in the region [85,112].

Between the late 1970s and early 1980s HIV-1B became widely disseminated in the Caribbean Islands most likely via MSM contact with US foreigners [113–115]. The HIV-1B epidemic in the Caribbean initially started in the homosexual/bisexual community, but since then, it has moved to the heterosexual population most likely by a "bisexual bridge". Today the heterosexual contact is the main mode of HIV transmission in the region [73,85]. For the rest of Latin America the largest prevalence of HIV remains in the MSM group directly reflecting the beginning of the epidemic.

Exceptions to this rule are Bermuda and Puerto Rico were IDUs are mainly contributing to the spread of the virus [70,116–118].

The HIV-1B epidemic in this region seems to be the result of multiple introductions from virus circulating in the United States [60,78,119]. Back introductions in the Caribbean region are also suggested by previous phylogenetic analysis [60,78]. Pagán et al. (2013) indicates that the pandemic B clade was reintroduced in the Caribbean from the US through Puerto Rico in the beginning of the 1980's (1960-1986) and gave rise to the HIV-1 epidemic in the region. The authors state that the current epidemic of subtype B in the Caribbean is primarily composed of genetic variants entirely derived from this reintroduction in the 1980s whereas the strains introduced directly from Africa had not succeeded in establishing an epidemic in the region [78]. However, different studies have shown the presence of the non-pandemic clade (strains circulating in the Caribbean since the introduction of the virus from Africa) in the region and in some cases in higher prevalences than the pandemic clade (strains from the United States) [60,62,76]. Furthermore, it seems to be highly improbable that a simple monophyletic origin would explain the circulation of the B pandemic throughout the Caribbean and Central America region. The real scenario most likely reflects the existence of multiple introductions from virus circulating in different regions of the world that maintain intensive migration/travel with the countries in the Caribbean [76]. Today, the epidemic of HIV-1B in Latin America and the Caribbean is composed of genetic variants derived both from the pandemic clade as from the non-pandemic clade [76]. The Cuban epidemic is an exception in the region since it is the result of multiple introductions in the early 1990s from virus circulating in North America and Europe reflecting an increase in the number of tourists in the region [120].

The HIV-1B epidemic in Central America seems to be the result of a polyphyletic origin. However, Murillo et al. (2013) found one single introduction estimated to be at 1966 (1955-1977) accounting for most current cases. Despite the early origin, predating the introduction of the virus in the United States these strains seems to be related to the B pandemic clade. The narrow time interval of the sequences used in the study to date the tMRCA of the Central American epidemic may have underestimate the year of introduction. Four additional monophyletic clades within different countries were detected and suggest a highly compartmentalized epidemic in Central America [77].

South American countries and Mexico also have epidemics derived from the United States. With the exception of the Brazilian epidemic little data was found to the other countries of Latin America. Multiple introductions starting from the late 1960s characterizes the current HIV-1B epidemic in Brazil [119]. The demographic pattern where HIV was initially transmitted through homosexual intercourse and injecting drug use is very similar to that reported for the subtype B epidemics in high-income countries [61,119].

HIV-1 Subtype B in Europe:

Subtype B accounts for approximately 66% of the HIV-1 infections in Europe [121]. The HIV-1B epidemic, with few exceptions, seems to be the result of multiple introductions caused mainly by homosexual contact or needle sharing in or from virus circulating in the United States [100,122]. Some of these introductions resulted in local dispersion of the virus, while others lead to dead end infections [100]. Data from the early epidemic identified that the majority of the infected male European homosexuals had travelled to the USA where they had had contact with US homosexuals [122]. In

addition, several infections were directly imported from the Haiti by European MSM individuals [122]. As an exception, the epidemic in Poland seems to be the result of a monophyletic introduction most likely trough injection drug users (IDU) networks associated with extensive local epidemics [100].

Epidemiologically, the European continent exhibited an intriguing difference in the distribution of HIV/AIDS cases between distinct risk groups [123]. In north–central Europe, MSM individuals constitute the main group of infected patients (65% of aids cases), while in southern Europe most of the cases involves IDUS [124]. It is important to note that the epidemic among IDUS for several European countries presents a monophyletic origin [125–129]. These data point to a non-uniform pattern of introduction and dissemination of the virus inside Europe and with the exception of Poland reflects a complex migratory events of the human population inside the continent [100]. Today, the HIV-1 molecular epidemiology in Europe is still highly stratified according to gender and risk group. Subtype B was significantly more diagnosed in men than in women and is still proportionally higher in MSM patients than in IDUs and in heterosexual individuals [121]. This stratification by risk group is most likely a result from the beginning of the epidemic and may reflect the epidemiological persistence of the founder effect.

HIV-1 Subtype B in Asia:

Despite the rapidly expanding of the epidemic among MSM individuals, Asia has a different pattern in the HIV-1 epidemic [130,131]. While in several other parts of the world the HIV-1B is mainly associated to MSM infection, Asia, with some exceptions, recently witnessed a dramatic shift in genotype distribution from subtype B to CRF01_AE among MSM [132–135]. A recent study suggest that a major HIV-1B

lineage (JP.MSM.B-1) circulating among Japanese MSM individuals has been disseminated globally causing infections in China, Canada, United States, Deutschland, and the United Kingdom [133,136].

The HIV-1B circulation in Asia seems to be mainly imported from the epidemic in the United States [61]. Until now no studies have yet tried to understand the origin of the epidemic and the initial year of the HIV dissemination in the region remains unknown. As opposed, a few studies aimed to understand the origin and dissemination of the HIV-1B viruses specifically among MSM individuals [133,136–138]. These studies found different dates for the viral introduction in the Asian MSM population but all these primary links are dated to the beginning of the 1980s [136]. Ye et al. (2014) studying samples from the general population found four different clades of the HIV-1B circulating in China: B' (Thai-B), BJ-B (Beijing-B), Pan-B (Pandemic-B), and TW-B (Taiwan-B), according to the origin of the sequences. These results suggest different introduction events at different time points in the region [139].

HIV-1 Subtype B in Oceania:

The lack of molecular data for the HIV-1 epidemic in Oceania undermines the understanding on origin and dissemination of the virus inside the continent [105]. Despite the small amount of information, there is a clear association between subtype B and exposure through sex between men in the region [140]. Indeed, the majority of infections are associated with homosexual-bisexual contact (78%) and as a consequence, more men (94%) than women (6%) are infected with HIV-1 in Australia [141]. Subtype B is the causative agent of approximately 68% of the HIV infections in the region [141,142].

HIV-1 Subtype B Non-pandemic Clade:

Aside from the lineage that gave rise to HIV-1B pandemic in the USA, other secondary outbreaks emerged from the Hispaniola Island (shared by the Dominican Republic and Haiti) and allowed the circulation of the HIV-1 subtype B in several countries. The dispersion of the virus mainly occurred through short-distance movements and reached countries from South America and Central America. While the HIV-1 subtype B epidemic that started at the United States fueled the epidemic of several other countries and became a generalized pandemic throughout the world this local non-pandemic epidemic most likely reflects the circulation of the virus in concentrated transmission networks in culturally related countries that in most cases resulted in dead-end infections due to a combination of chance effects and complex socio-ecological factors [62]. Other lineages of the non-pandemic B clade from the Caribbean fed the epidemic of the US; however, these failed to ignite significant outbreaks [76].

Non-pandemic strains started to spread from the Hispaniola Island in the beginning of the 1970 and until the 1980s would have reached Trinidad and Tobago, Jamaica, Mexico, Venezuela, Panama, Colombia, Ecuador, El Salvador, Honduras, Suriname and Brazil [62]. The Trinidad and Tobago epidemic is derived from a monophyletic introduction and can be viewed as a secondary hub in this epidemic seeding tertiary outbreaks in short-distanced countries such as Jamaica, Venezuela, Guyana, Brazil, and also a back dissemination to the Hispaniola Island. In the following decades after the initial dissemination, Hispaniola Island HIV-1B epidemic also seeded Suriname, Colombia, El Salvador, and Ecuador epidemics [60,62,76]. Still today several countries present an epidemic mainly composed of strains belonging to the non-pandemic B clade including Haiti and Dominican Republic (~75%), Jamaica (~50%),

Trinidad and Tobago (~95%), other Lesser Antilles (~40-75%), French Guyana (~40-50%), and Suriname (~40-50%) [62,76]. The clade is also detected at lower prevalences (<1%-10%) in the following countries: Brazil, Colombia, Ecuador, Mexico, Panama, Venezuela, Argentina, El Salvador, Honduras, and Peru [62]. Although the pandemic B clade and the non-pandemic B clade would have reached the Latin America at the same time, the pandemic clade was most likely introduced in a high-risk group increasing the rate of transmission between different hosts, specially MSM individuals [61,62].

Subtype B Variants:

The high evolutionary rates described for HIV [33] are mainly a consequence of the reverse transcription process [33]. Infidelity in the addition of complementary nucleotides in reverse transcription generates insertions, deletions, duplications and nucleotide pairing errors in the new proviral genomes [143]. The frequency of these evolutionary processes ensures that each new virus is unique and different from their parental genome and associated to the emergence of about 10^{10}-10^{12} different virions per day within the same host, the large number of infected individuals and the persistence of infection allowed the arising and expansion of the current HIV diversity [144–148].

Due to the globally extension of the HIV-1B epidemic and specially the strong potential of the founder effect to evidence the genetic variation in different parts of the world several variants of the subtype B are being noticed [80,85,149,150]. Most of them were clearly identified by phylogenetic analysis and not necessarily by some clinical variation on the pathogenesis of AIDS. It is accepted that the great majority of these genetic variations arose from an ancestral pandemic subtype B disseminated out of the United States. Of note, these pandemic strains characteristically have the GPGR motif at

the central portion of the V3 loop in the gp120 protein [151]. Some lineages, however, have been found circulating in specific regions worldwide with alternate signature patterns or distinguished mutations. Four lineages of subtype B aside from the pandemic form could efficiently establish regional epidemics and are currently identified circulating in the HIV-1B pandemic (Figure 2).

Brazilian Subtype B: Some strains of the HIV-1 subtype B in Brazil have found to harbour an alternate signature GWGR at the tip of the V3 loop on gp120 [149,152–156]. These strains are routinely called B", B'-GWGR, Bbr or Brazilian B [46,156–158]. Despite B" have been first isolated in Japan and is currently notified in 23 different countries around the world, Brazil have found this variant at frequencies ranging from 17% to 50% [79,152,155,156,159–162]. This country is accepted to be the epicentre of the B" variant; however, the monophyly of the epidemic is a target of continuing debate [119,162–164]. Several studies have attempted to unravel the initial conditions of the epidemic and considered different scenarios to explain the dissemination through the Brazilian population [79,119,162,163]

It has been previously suggested that the GWGR variant of subtype B is less pathogenic than other HIV-1 isolates [163]. The GWGR motif seems to increase the avidity of V3 antibodies for the virus and contributes to slower disease progression in comparison to B-GPGR infection [165,166]. Patients infected with this variant tend to have a slower progression to AIDS, higher CD4+ T cell counts and lower viral load compared with patients infected with viruses harboring the GPGR motif [160,166,167]. Patients infected with B" are more likely to survive than individuals infected with the pandemic B variant [166]. These results suggests that the clinical difference in the

progression to AIDS can be associated to the HIV infecting strain and shows the importance of molecular epidemiological studies to the understanding of the disease.

Korean Subtype B: In South Korea, 80% of the HIV-1 epidemic is attributed to subtype B [168]. Based on phylogenetic analysis of *nef* and *env* genes several studies reported that South Korean HIV epidemic is composed (~88%) of a distinct monophyletic clade than the pandemic form of subtype B [150,168,169]. These isolates are usually called "Korean B" (B^K). In addition to the higher proportion of the amino acid signature GPGS at the tip of the V3 loop on gp120, the clearly distinction of B^K from the pandemic B is attributed to a 32 amino acid signature pattern dispersed throughout gp160 [150,168]. It is accepted that the higher prevalence of this variant in the South Korean population is the result of a founder effect since B^K is consistently transmitted since it was first introduced in the country most likely from a single source [150,170]. The introduction of HIV in the Korean population occurred a few years before 1985 when the first case of AIDS was reported. Early in the epidemic seafarers were assumed as the higher-risk group and with the change in the epidemic pattern today MSM is being identified as the major high-risk group [171]. None study until now tried to correlate the infection with B^K with differences in the disease progression or pathogenesis.

Thailand Subtype B: The B' (B' Thailand, Thai-B or Bb) variant was first detected in 1988 in injecting drug users (IDUs) around the golden triangle, the main illicit opium-producing areas covering the boundaries of Myanmar, Laos, Thailand and China [80,83,172–174]. The variant is characterized to exhibit a distinctive GPGQ motif on the crown region of the V3 loop in addition to characteristically mutations in the V3 and

on the p17. In Asia, B' variant emerged monophyletically in Thailand most likely from the B pandemic form around 1983 (1975-1990) especially in IDU individuals [139]. The epidemic seems to have been established very fast taking about 3 years to be transmitted all over South-East Asia [80,81,83,175]. This is consistent with findings from earlier studies suggesting that the B' variant is closely associated with the initial phase of the blood-born HIV-1transmission in Asia infecting mainly IDUs and paid blood donors [82,174]. Years later (~1991), B' emerged in a series of HIV-1 outbreaks among former plasma donors in Central China [174,176]. In addition to the clearly epidemiologically importance of the B' in the HIV-1 epidemic, this variant is also a constituent of several different CRFs circulating in Asia: five CRFs encompassing portions of the B' and portions of the subtype C are circulating in China (CRF07_BC, CRF08_BC, CRF61_BC, CRF62_BC, and CRF57_BC), two CRFs encompassing portions of the B' and portions of the CRF01_AE are circulating in Thailand (CRF15_01B and CRF34_01B), five CRFs encompassing portions of the B' and portions of the CRF01_AE are circulating in Malaysia (CRF33_01B, CRF48_01B, CRF53_01B, CRF54_01B, and CRF58_01B), one recombinant form also including portions of the B' and the CRF01_AE is dispersed throughout the Southeast Asia (CRF52_01B) and one CRF encompassing portions of the B', CRF01_A and C (CRF65_cpx) is circulating in China [177–190]. Several URFs (unique recombinant forms) involving the B' have been described [191–194].

Trinidadian and Tobagonian Subtype B: HIV-1 Subtype B is mainly transmitted by heterosexual contact in the Trinidad and Tobago epidemic despite the original introduction in the MSM group in the late 1970s to early 1980s [85,111,113]. Most of the current epidemic is explained by a single or few sources in the past [85,195].

Examination of the V3 loop sequences of the Trinidadadian subtype B isolates suggests the presence of a slightly different form from the pandemic B. These strains present a signature threonine deletion just C terminal to the crown of the loop and are currently jointly called Trinidad and Tobago Clade (B^{TT}). It seems to be a consistent feature of the virus circulating in the region and is highly conserved among and between infected individuals [85]. In addition, Collins-Fairclough *et al.* found these isolates presenting longer and more glycosylated V2 loops, shorter V3 loops (due to T319- deletion) and amino acid substitution R315K within V3 [195]. The fact that Trinidad and Tobago is a twin island country and approximately 80% of the strains in the current epidemic harbor the threonine deletion provides strong support for a founder effect in the beginning of the epidemic. An isolate from a MSM individual infected before 1983 in Trinidad and Tobago already presents the signature pattern. No specific demographic or epidemiologic factor was statistically associated with the variant in this population [85]. None study until now tried to correlate the infection with B^{TT} with differences in the disease progression or pathogenesis.

Conclusion:

Subtype B is the most widespread HIV-1 variant and accounts for approximately 11% of all infections around the world. Thirty four years since the first cases were detected there is still a clear association between homosexual contact and the infection by HIV-1 subtype B in most subepidemics around the world. This segregation of subtype by risk factor is most likely a reflection of the founder effect occurred in the beginning of the epidemic than any reflection on viral advantage in this type of transmission. Despite the narrow time interval since its emergence, HIV-1B established successful epidemics in many countries from five continents and continues to expand.

Several studies have addressed the origins and spread of the HIV-1B and tried to understand the evolution of this pandemic through phylogenetic means. Today it seems clear the importance of the Caribbean and the United States to the onset of the epidemic out of Africa and how the human movements between different countries and the change in the sexual behavior influenced in this process. The epidemic became more complex due to the intricate migration pattern in the current world. Despite this intense movement of the human population some genetic structure in the HIV-1B epidemic can still be observed in different regions around the globe. In recent years the multisectorial efforts on different fronts of the epidemic had some effect in the numbers related to HIV/AIDS but more commitment from the global community will be necessary to achieve the UNAIDS goal of ending the epidemic by 2030.

Figures:

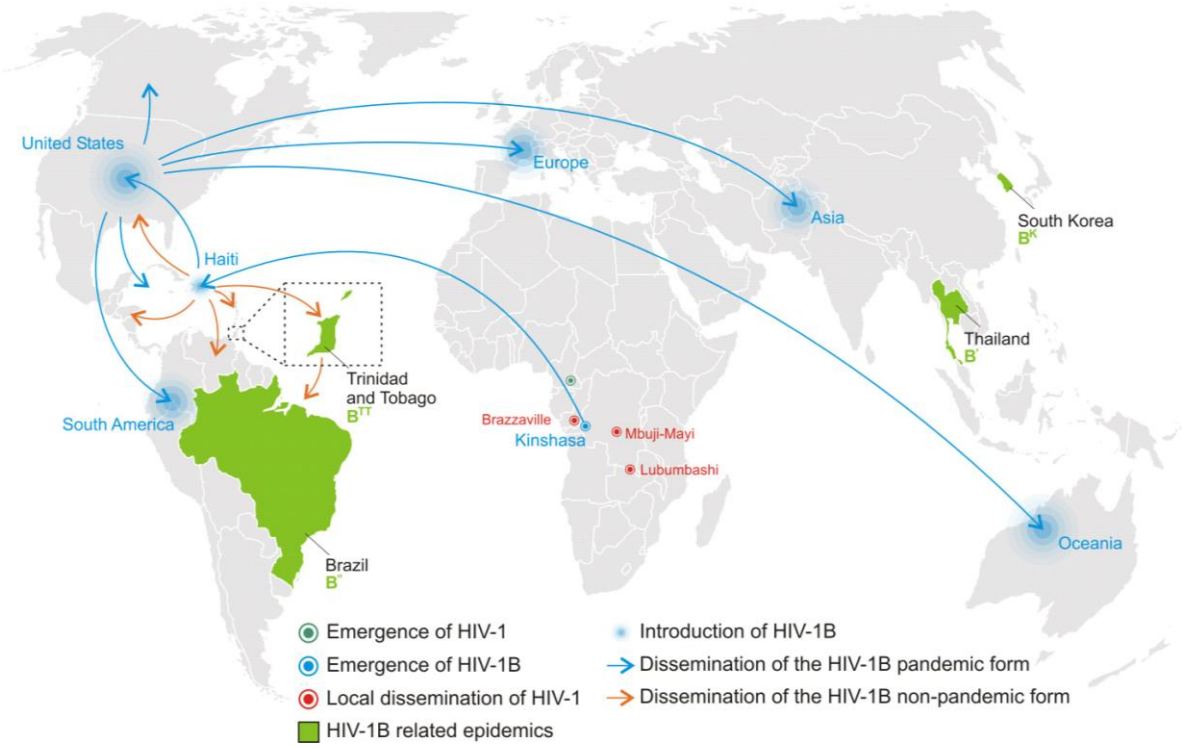


Figure 1. Estimated spatial dynamics of HIV-1 Subtype B around the world. Blue lines represent the main dissemination routes of HIV-1 subtype B pandemic form and the main disease epicenters. The epidemic starts in the city of Kinshasa, Democratic Republic of the Congo (Africa). Orange lines represent the direction of the B non-pandemic lineage dissemination. Green regions demarcate the countries where a specific lineage of the HIV-1 subtype B is circulating.

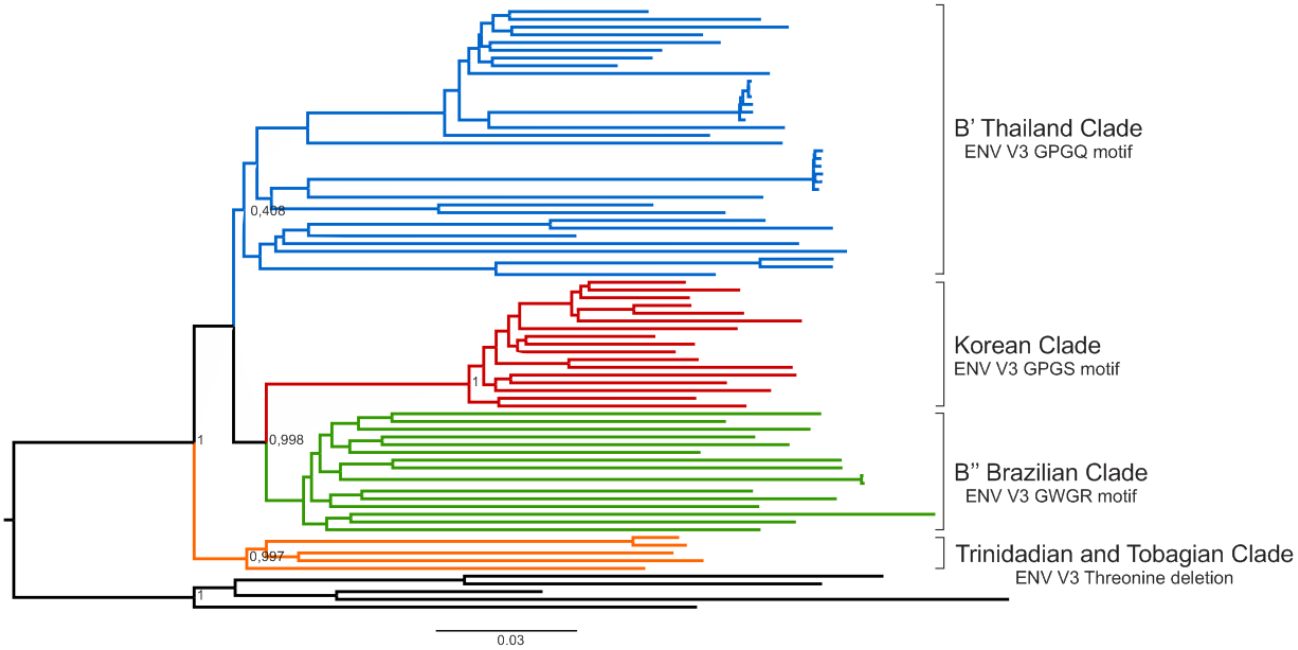


Figure 2. Maximum Likelihood phylogenetic tree for HIV-1 subtype B complete sequences. The color of the branches represent the geographic region from where the subtype B strains were sampled. Brackets indicate the subtype B different clades in relation to the subtype D outgroup. The aLRT support values are indicated only at key nodes. Horizontal branch lengths are drawn to scale with the bar at the bottom indicating nucleotide substitutions per site.

References:

1. Tebit DM, Arts EJ. Tracking a century of global expansion and evolution of HIV to drive understanding and to combat disease. Lancet Infect Dis. 2011;11: 45–56.

2. Beyrer C, Baral SD, van Griensven F, Goodreau SM, Chariyalertsak S, Wirtz AL, et al. Global epidemiology of HIV infection in men who have sex with men. Lancet. 2012;380: 367–77.

3. Brenner BG, Wainberg M a. Future of phylogeny in HIV prevention. J Acquir Immune Defic Syndr. 2013;63 Suppl 2: S248–54.

4. Oguntibeju OO, Van Schalkwyk FE, Van Den Heever WMJ. The HIV Epidemic : factors responsible for the epidemic and the impact of HIV / AIDS. RMJ. 2003; 1–13.

5. UNAIDS. Global AIDS Response Progress Reporting. 2015; Available: http://www.unaids.org/sites/default/files/media_asset/JC2702_GARPR2015guidelines_en.pdf

6. UNAIDS. Report on the global AIDS epidemic [Internet]. 2013. Available: http://www.unaids.org/en/media/unaids/contentassets/documents/epidemiology/2013/gr2013/UNAIDS_Global_Report_2013_en.pdf

7. Lemey P, Rambaut A, Pybus OG. HIV evolutionary dynamics within and among hosts. AIDS Rev. 2006;8: 125–40.

8. Hemelaar J. The origin and diversity of the HIV-1 pandemic. Trends Mol Med. 2012;18: 182–92.

9. Charneau P, Borman AM, Quillent C, Guétard D, Chamaret S, Cohen J, et al. Isolation and envelope sequence of a highly divergent HIV-1 isolate: definition of a new HIV-1 group. Virology. 1994;205: 247–53.

10. Simon F, Mauclère P, Roques P, Loussert-Ajaka I, Müller-Trutwin MC, Saragosti S, et al. Identification of a new human immunodeficiency virus type 1 distinct from group M and group O. Nat Med. 1998;4: 1032–7.

11. Nau J-Y, Plantier J-C, Leoz M, Dickerson JE, De Oliveira F, Cordonnier F, et al. A new human immunodeficiency virus derived from gorillas. Rev Med Suisse. 2009;5: 1741.

12. Aldrich C, Hemelaar J. Global HIV-1 diversity surveillance. Trends Mol Med. 2012;18: 691–4.

13. Worobey M, Santiago ML. Contaminated polio vaccine theory refuted. Nature. 2004;428: 820.

14. Holmes EC. When HIV spread afar. Proc Natl Acad Sci USA. 2007;104: 18351–18352.

15. Gao F, Bailes E, Robertson DL, Chen Y, Rodenburg CM, Michael SF, et al. Origin of HIV-1 in the chimpanzee Pan troglodytes troglodytes. Nature. 1999;397: 436–41.

16. Hahn BH, Shaw GM, De Cock KM, Sharp PM. AIDS as a zoonosis: scientific and public health implications. Science. 2000;287: 607–14.

17. Heeney JL, Dalgleish AG, Weiss R a. Origins of HIV and the Evolution of Resistance to AIDS. Science. 2006;313: 462–466.

18. Bibollet-Ruche F, Bailes E, Gao F, Pourrut X, Barlow KL, Clewley JP, et al. New simian immunodeficiency virus infecting De Brazza's monkeys (Cercopithecus neglectus): evidence for a cercopithecus monkey virus clade. J Virol. 2004;78: 7748–62.

19. Yamaguchi J, Devare SG, Brennan CA. Identification of a new HIV-2 subtype based on phylogenetic analysis of full-length genomic sequence. AIDS Res Hum Retroviruses. 2000;16: 925–930.

20. Georges-Courbot MC, Lu CY, Makuwa M, Telfer P, Onanga R, Dubreuil G, et al. Natural infection of a household pet red-capped mangabey (Cercocebus torquatus torquatus) with a new simian immunodeficiency virus. J Virol. 1998;72: 600–8.

21. Worobey M, Telfer P, Souquière S, Hunter M, Coleman C a, Metzger MJ, et al. Island biogeography reveals the deep history of SIV. Science. 2010;329: 1487.

22. Bailes E, Gao F, Bibollet-Ruche F, Courgnaud V, Peeters M, Marx P a, et al. Hybrid origin of SIV in chimpanzees. Science. 2003;300: 1713.

23. Keele BF, Van Heuverswyn F, Li Y, Bailes E, Takehisa J, Santiago ML, et al. Chimpanzee reservoirs of pandemic and nonpandemic HIV-1. Science. 2006;313: 523–526.

24. Sharp PM, Hahn BH. The evolution of HIV-1 and the origin of AIDS. Philos Trans R Soc B Biol Sci. 2010;365: 2487–2494.

25. Van Heuverswyn F, Li Y, Bailes E, Neel C, Lafay B, Keele BF, et al. Genetic diversity and phylogeographic clustering of SIVcpzPtt in wild chimpanzees in Cameroon. Virology. 2007;368: 155–71.

26. Van Heuverswyn F, Li Y, Neel C, Bailes E, Keele BF, Liu W, et al. Human immunodeficiency viruses: SIV infection in wild gorillas. Nature. 2006;444: 164.

27. Kalish ML, Wolfe ND, Ndongmo CB, McNicholl J, Robbins KE, Aidoo M, et al. Central African hunters exposed to simian immunodeficiency virus. Emerg Infect Dis. 2005;11: 1928–30.

28. Wolfe ND, Switzer WM, Carr JK, Bhullar VB, Shanmugam V, Tamoufe U, et al. Naturally acquired simian retrovirus infections in central African hunters. Lancet. 2004;363: 932–937.

29. Vallari A, Holzmayer V, Harris B, Yamaguchi J, Ngansop C, Makamche F, et al. Confirmation of Putative HIV-1 Group P in Cameroon. J Virol. 2010;85: 1403–7.

30. Vidal N, Peeters M, Mulanga-Kabeya C, Nzilambi N, Robertson D, Ilunga W, et al. Unprecedented degree of human immunodeficiency virus type 1 (HIV-1) group M genetic diversity in the Democratic Republic of Congo suggests that the HIV-1 pandemic originated in Central Africa. J Virol. 2000;74: 10498–507.

31. Kalish ML, Robbins KE, Pieniazek D, Schaefer A, Nzilambi N, Quinn TC, et al. Recombinant viruses and early global HIV-1 epidemic. Emerg Infect Dis. 2004;10: 1227–34.

32. Berry N, Davis C, Jenkins A, Wood D, Minor P, Schild G, et al. Phylogeny and the origin of HIV-1. Nature. 2001;410: 1047–1048.

33. Rambaut A, Posada D, Crandall KA, Holmes EC. The causes and consequences of HIV evolution. Nat Rev Genet. 2004;5: 52–61.

34. Gürtler LG, Zekeng L, Tsague JM, van Brunn A, Afane Ze E, Eberle J, et al. HIV-1 subtype O: epidemiology, pathogenesis, diagnosis, and perspectives of the evolution of HIV. Arch Virol Suppl. 1996;11: 195–202.

35. Delaugerre C, De Oliveira F, Lascoux-Combe C, Plantier J-C, Simon F. HIV-1 group N: travelling beyond Cameroon. Lancet. 2011;378: 1894.

36. GHO. Global Health Observatory data. 2013.

37. Sauter D, Schindler M, Specht A, Landford WN, Münch J, Kim K, et al. Tetherin-driven adaptation of Vpu and Nef function and the evolution of pandemic and nonpandemic HIV-1 strains. Cell Host Microbe. 2009;6: 409–21.

38. Faria NR, Rambaut A, Suchard MA, Baele G, Bedford T, Ward MJ, et al. The early spread and epidemic ignition of HIV-1 in human populations. Science. 2014;346: 56–61.

39. Worobey M, Gemmel M, Teuwen DE, Haselkorn T, Kunstman K, Bunce M, et al. Direct evidence of extensive diversity of HIV-1 in Kinshasa by 1960. Nature. 2008;455: 661–664.

40. Lihana RW, Ssemwanga D, Abimiku A, Ndembi N. Update on HIV-1 diversity in Africa: a decade in review. AIDS Rev. 2012;14: 83–100.

41. Sharp P, Bailes E, Robertson D, Gao F, Hahn B. Origins and evolution of AIDS viruses.pdf. Biol Bull. 1999; 338–342.

42. Taylor BS, Hammer SM, Mccutchan FE, D P. The challenge of HIV-1 subtype diversity. N Engl J Med. 2008;359: 1965–6.

43. Hu DJ, Buvé a, Baggs J, van der Groen G, Dondero TJ. What role does HIV-1 subtype play in transmission and pathogenesis? An epidemiological perspective. AIDS. 1999;13: 873–81.

44. Respess R, Parekh B, Phillips S, Granade TC, Baggs J, Hu DJ. Impact of HIV type 1 subtype variation on viral RNA quantitation. AIDS Res Hum Retroviruses. 1999;15: 133–42.

45. Renjifo B, Gilbert P, Chaplin B, Msamanga G, Mwakagile D, Fawzi W, et al. Preferential in-utero transmission of HIV-1 subtype C as compared to HIV-1 subtype A or D. AIDS. 2004;18: 1629–36.

46. Casseb J, Komninakis S, Abdalla L, Brigido L, Rodrigues R, Araujo F, et al. HIV disease progression: is the Brazilian variant subtype B' (GWGR motif) less pathogenic than US/European subtype B (GPGR)?1, 2. Int J Infect Dis. 2002;6: 164–169.

47. Kanki PJ, Hamel DJ, Sankalé JL, Hsieh C c, Thior I, Barin F, et al. Human immunodeficiency virus type 1 subtypes differ in disease progression. J Infect Dis. 1999;179: 68–73.

48. Kaleebu P, French N, Mahe C, Yirrell D, Watera C, Lyagoba F, et al. Effect of human immunodeficiency virus (HIV) type 1 envelope subtypes A and D on disease progression in a large cohort of HIV-1-positive persons in Uganda. J Infect Dis. 2002;185: 1244–50.

49. Baeten JM, Chohan B, Lavreys L, Chohan V, McClelland RS, Certain L, et al. HIV-1 subtype D infection is associated with faster disease progression than subtype A in spite of similar plasma HIV-1 loads. J Infect Dis. 2007;195: 1177–80.

50. Abecasis AB, Vandamme A-M, Lemey P. Quantifying differences in the tempo of human immunodeficiency virus type 1 subtype evolution. J Virol. 2009;83: 12917–24.

51. Camacho RJ, Vandamme A-M. Antiretroviral resistance in different HIV-1 subtypes: impact on therapy outcomes and resistance testing interpretation. Curr Opin HIV AIDS. 2007;2: 123–9.

52. Gottlieb MS, Schroff R, Schanker HM, Weisman JD, Fan PT, Wolf RA, et al. Pneumocystis carinii pneumonia and mucosal candidiasis in previously healthy homosexual men: evidence of a new acquired cellular immunodeficiency. N Engl J Med. 1981;305: 1425–31.

53. Prusiner SB. Historical essay. Discovering the cause of AIDS. Science. 2002;298: 1726.

691 54. Barré-Sinoussi F, Chermann JC, Rey F, Nugeyre MC, Chamaret S, Gruest J, et al. Isolation of a
692 T-lymphotropic retrovirus from patient at risk for AIDS. Science. 1983;220: 868–870.

693 55. Gallo RC, Sarin PS, Gelmann EP, Robert-Guroff M, Richardson E, Kalyanaraman VS, et al.
694 Isolation of human T-cell leukemia virus in acquired immune deficiency syndrome (AIDS).
695 Science. 1983;220: 865–867.

696 56. Coffin J, Haase A, Levy JA, Montagnier L, Oroszlan S, Teich N, et al. What to call the AIDS
697 virus? Nature. 1986;321: 10.

698 57. Leitner T. Genetic subtypes of HIV-1. In: Myers G, Foley B, Mellors JW, Korber B, Jeang KT,
699 Wain-Hobson S, editors. Human Retroviruses and AIDS. Theor Biol Biophys Los Alamos Natl
700 Lab Los Alamos,. 1996; III28–40.

701 58. Santos AF a, Lengruber RB, Soares E a, Jere A, Sprinz E, Martinez AMB, et al. Conservation
702 patterns of HIV-1 RT connection and RNase H domains: identification of new mutations in
703 NRTI-treated patients. PLoS One. 2008;3: e1781.

704 59. Parkin NT, Schapiro JM. Antiretroviral drug resistance in non-subtype B HIV-1, HIV-2 and SIV.
705 Antivir Ther. 2004;9: 3–12.

706 60. Junqueira, de Medeiros RM, Matte MCC, Araújo LAL, Chies JAB, Ashton-Prolla P, et al.
707 Reviewing the History of HIV-1: Spread of Subtype B in the Americas. Martin DP, editor. PLoS
708 One. 2011;6: e27489.

709 61. Gilbert MTP, Rambaut A, Wlasiuk G, Spira TJ, Pitchenik AE, Worobey M. The emergence of
710 HIV/AIDS in the Americas and beyond. Proc Natl Acad Sci U S A. 2007;104: 18566–70.

711 62. Cabello M, Junqueira DM, Bello G. Dissemination of nonpandemic Caribbean HIV-1 subtype B
712 clades in Latin America. AIDS. 2015;29: 483–92.

713 63. Piot P, Taelman H, Bila Minlangu K, Mbendi N, Ndangi K, Kalambayi K, et al. Acquired
714 Immunodeficiency Syndrome in a Heterosexual Population in Zaire. Lancet. 1984;324: 65–69.

715 64. Pepin J. The Origins of AIDS [Internet]. Cambridge: Cambridge University Press; 2011.

716 65. Stengers J. Congo: Mythes et réalités. Editions R. Lannoo R, editor. Brussels; 2005.

717 66. Crowder M, editor. The Cambridge History of Africa [Internet]. Cambridge: Cambridge
718 University Press; 1984.

719 67. Malebranche R, Guérin J., Laroche A., Elie R, Spira T, Drotman P, et al. Acquired
720 Immunodeficiency Syndrome with severe gastrointestinal manifestaions in Haiti. Lancet.
721 1983;322: 873–878.

722 68. Pape JW, Liautaud B, Thomas F, Mathurin JR, St Amand MM, Boncy M, et al. Characteristics of
723 the acquired immunodeficiency syndrome (AIDS) in Haiti. N Engl J Med. 1983;309: 945–50.

724 69. Worobey M, Pitchenik AE, Gilbert MTP, Wlasiuk G, Rambaut A. Reply to Pape et al.: the
725 phylogeography of HIV-1 group M subtype B. Proc Natl Acad Sci U S A. 2008;105: E16.

726 70. Figueroa JP. The HIV epidemic in the Caribbean: meeting the challenges of achieving universal
727 access to prevention, treatment and care. West Indian Med J. 2008;57: 195–203.

728 71. Adrien A, Cayemittes M, Bergevin Y. AIDS-related knowledge, attitudes, beliefs, and practices
729 in Haiti. Bull Pan Am Health Organ. 1993;27: 234–43.

72. Deschamps M-M. Heterosexual Transmission of HIV in Haiti. Ann Intern Med. 1996;125: 324.

73. Wheeler VW, Radcliffe KW. HIV infection in the Caribbean. Int J STD AIDS. 1994;5: 79–89.

74. Pape J, Johnson WD. AIDS in Haiti: 1982-1992. Clin Infect Dis. 1993;17 Suppl 2: S341–5.

75. De Boni R, Veloso VG, Grinsztejn B. Epidemiology of HIV in Latin America and the Caribbean. Curr Opin HIV AIDS. 2014;9: 192–8.

76. Cabello M, Mendoza Y, Bello G. Spatiotemporal Dynamics of Dissemination of Non-Pandemic HIV-1 Subtype B Clades in the Caribbean Region. PLoS One. 2014;9: e106045.

77. Murillo W, Veras N, Prosperi M, de Rivera IL, Paz-Bailey G, Morales-Miranda S, et al. A single early introduction of HIV-1 subtype B into Central America accounts for most current cases. J Virol. 2013; 7463–7470.

78. Pagán I, Holguín Á. Reconstructing the Timing and Dispersion Routes of HIV-1 Subtype B Epidemics in The Caribbean and Central America: A Phylogenetic Story. PLoS One. 2013;8: e69218.

79. Pinto ME, Schrago CG, Miranda a B, Russo C a M. A molecular study on the evolution of a subtype B variant frequently found in Brazil. Genet Mol Res. 2008;7: 1031–44.

80. Ou CY, Takebe Y, Weniger BG, Luo CC, Kalish ML, Auwanit W, et al. Independent introduction of two major HIV-1 genotypes into distinct high-risk populations in Thailand. Lancet. 1993;341: 1171–4.

81. Brown TM, Robbins KE, Sinniah M, Saraswathy TS, Lee V, Hooi LS, et al. HIV type 1 subtypes in Malaysia include B, C, and E. AIDS Res Hum Retroviruses. 1996;12: 1655–7.

82. Cassol S, Weniger BG, Babu PG, Salminen MO, Zheng X, Htoon MT, et al. Detection of HIV type 1 env subtypes A, B, C, and E in Asia using dried blood spots: a new surveillance tool for molecular epidemiology. AIDS Res Hum Retroviruses. 1996;12: 1435–41.

83. Kusagawa S, Sato H, Watanabe S, Nohtomi K, Kato K, Shino T, et al. Genetic and serologic characterization of HIV type 1 prevailing in Myanmar (Burma). AIDS Res Hum Retroviruses. 1998;14: 1379–85.

84. Chen J, Young NL, Subbarao S, Warachit P, Saguanwongse S, Wongsheree S, et al. HIV type 1 subtypes in Guangxi Province, China, 1996. AIDS Res Hum Retroviruses. 1999;15: 81–4.

85. Cleghorn FR, Jack N, Carr JK, Edwards J, Mahabir B, Sill A, et al. A distinctive clade B HIV type 1 is heterosexually transmitted in Trinidad and Tobago. Proc Natl Acad Sci U S A. 2000;97: 10532–7.

86. Kim EY, Cho YS, Maeng SH, Kang C, Nam JG, Lee JS. Characterization of V3 loop sequences from HIV type 1 subtype B in South Korea: predominance of the GPGS motif. AIDS Res Hum Retrovir. 1999;15: 681–6.

87. Bacchetti P, Moss AR. Incubation period of AIDS in San Francisco. Nature. 1989;338: 251–3.

88. Laverdiere M, Tremblay J. AIDS in Haitian immigrants and in a Caucasian woman closely associated with Haitians. Can Med Assoc J. 1983;129: 1209–1212.

89. Masur H, Michelis MA, Greene JB, Onorato I, Stouwe RA, Holzman RS, et al. An outbreak of community-acquired Pneumocystis carinii pneumonia: initial manifestation of cellular immune dysfunction. N Engl J Med. 1981;305: 1431–8.

90. Noel GE. Another case of AIDS in the pre-AIDS era. Rev Infect Dis. 1988;10: 668–9.

91. Selik RM, Haverkos HW, Curran JW. Acquired immune deficiency syndrome (AIDS) trends in the United States, 1978-1982. Am J Med. 1984;76: 493–500.

92. Robbins KE, Lemey P, Pybus OG, Jaffe HW, Youngpairoj AS, Brown TM, et al. U . S . Human Immunodeficiency Virus Type 1 Epidemic : Date of Origin , Population History , and Characterization of Early Strains. Society. 2003;77: 6359–6366.

93. Thomas P, O'Donnell R, Williams R, Chiasson MA. HIV infection in heterosexual female intravenous drug users in New York City, 1977-1980. N Engl J Med. 1988;319: 374.

94. Kuiken C, Thakallapalli R, Esklid a, de Ronde a, Eskild A, Ronde A De. Genetic analysis reveals epidemiologic patterns in the spread of human immunodeficiency virus. Am J Epidemiol. 2000;152: 814–22.

95. Cohen J. HIV/AIDS: Latin America & Caribbean. HAITI: making headway under hellacious circumstances. Science (80-). 2006;313: 470–3.

96. Pape JW, Farmer P, Koenig S, Fitzgerald D, Wright P, Johnson W. The epidemiology of AIDS in Haiti refutes the claims of Gilbert et al. Proc Natl Acad Sci U S A. 2008;105: E13.

97. Centers for Disease Control. Opportunistic Infections and Kaposi's Sarcoma among Haitians in the United States. MMWR Morb Mortal Wkly Rep. 1982: 353– 354, 360–361.

98. Pitchenik AE. Opportunistic Infections and Kaposi's Sarcoma Among Haitians: Evidence of a New Acquired Immunodeficiency State. Ann Intern Med. 1983;98: 277.

99. Herek GM, Capitanio JP, Widaman KF. HIV-Related Stigma and Knowledge in the United States: Prevalence and Trends, 1991–1999. Am J Public Health. 2002;92: 371–377.

100. Paraskevis D, Pybus O, Magiorkinis G, Hatzakis A, Wensing AMJ, Vijver DA Van De, et al. Tracing the HIV-1 subtype B mobility in Europe: a phylogeographic approach. Retrovirology. 2009;6: 49.

101. Beyrer C, Sullivan P, Sanchez J, Baral SD, Collins C, Wirtz AL, et al. The increase in global HIV epidemics in MSM. AIDS. 2013;27: 2665–78.

102. Auerbach DM, Darrow WW, Jaffe HW, Curran JW. Cluster of cases of the acquired immune deficiency syndrome. Patients linked by sexual contact. Am J Med. 1984;76: 487–92.

103. Perrin L, Kaiser L, Yerly S, Ag CRF. Travel and the spread of HIV-1 genetic variants. Lancet Infect Dis. 2003;3: 22–7.

104. Osmanov S, Pattou C, Walker N, Schwardländer B, Esparza J. Estimated global distribution and regional spread of HIV-1 genetic subtypes in the year 2000. J Acquir Immune Defic Syndr. 2002;29: 184–90.

105. Hemelaar J, Gouws E, Ghys PD, Osmanov S. Global trends in molecular epidemiology of HIV-1 during 2000-2007. AIDS. 2011;25: 679–89.

106. Cuevas MT, Ruibal I, Villahermosa ML, Díaz H, Delgado E, Parga EV, et al. High HIV-1 genetic diversity in Cuba. AIDS. 2002;16: 1643–53.

107. UNAIDS. World AIDS Day Report [Internet]. 2011. Available: http://www.unaids.org/sites/default/files/media_asset/JC2216_WorldAIDSday_report_2011_en_1.pdf

108. Sanchez J, Lama JR, Peinado J, Paredes A, Lucchetti A, Russell K, et al. High HIV and ulcerative sexually transmitted infection incidence estimates among men who have sex with men in Peru: awaiting for an effective preventive intervention. J Acquir Immune Defic Syndr. 2009;51 Suppl 1: S47–51.

109. Bastos FI, Cáceres C, Galvão J, Veras MA, Castilho EA. AIDS in Latin America: assessing the current status of the epidemic and the ongoing response. Int J Epidemiol. 2008;37: 729–37.

110. Arán-Matero D, Amico P, Arán-Fernandez C, Gobet B, Izazola-Licea JA, Avila-Figueroa C. Levels of spending and resource allocation to HIV programs and services in Latin America and the Caribbean. PLoS One. 2011;6: e22373.

111. Nadai Y, Eyzaguirre LM, Sill A, Cleghorn F, Nolte C, Charurat M, et al. HIV-1 epidemic in the Caribbean is dominated by subtype B. PLoS One. 2009;4: e4814.

112. Cleghorn FR, Jack N, Murphy JR, Edwards J, Mahabir B, Paul R, et al. HIV-1 prevalence and risk factors among sexually transmitted disease clinic attenders in Trinidad. AIDS. 1995;9: 389–94.

113. Bartholomew C. Transmission of HTLV-I and HIV Among Homosexual Men in Trinidad. JAMA J Am Med Assoc. 1987;257: 2604.

114. Murphy EL, Gibbs WN, Figueroa JP, Bain B, LaGrenade L, Cranston B, et al. Human immunodeficiency virus and human T-lymphotropic virus type I infection among homosexual men in Kingston, Jamaica. J Acquir Immune Defic Syndr. 1988;1: 143–9.

115. Noel RJ, Chaudhary S, Rodriguez N, Kumar A, Yamamura Y. Phylogenetic relationships between Puerto Rico and continental USA HIV-1 pol sequences: a shared HIV-1 infection. Cell Mol Biol (Noisy-le-grand). 2003;49: 1193–8.

116. UNAIDS. Report on the global AIDS epidemic [Internet]. 2012. Available: http://www.unaids.org/en/resources/documents/2012/20121120_UNAIDS_Global_Report_2012

117. Pérez L, Thomson MM, Bleda MJ, Aragonés C, González Z, Pérez J, et al. HIV Type 1 molecular epidemiology in cuba: high genetic diversity, frequent mosaicism, and recent expansion of BG intersubtype recombinant forms. AIDS Res Hum Retroviruses. 2006;22: 724–33.

118. De Arazoza H, Joanes J, Lounes R, Legeai C, Clémençon S, Pérez J, et al. The HIV/AIDS epidemic in Cuba: description and tentative explanation of its low HIV prevalence. BMC Infect Dis. 2007;7: 130.

119. Bello G, Eyer-silva W a, Couto-Fernandez JC, Guimarães ML, Chequer-Fernandez SL, Teixeira SLM, et al. Demographic history of HIV-1 subtypes B and F in Brazil. Infect Genet Evol. 2007;7: 263–70.

120. Delatorre E, Bello G. Phylodynamics of the HIV-1 Epidemic in Cuba. Zhang C, editor. PLoS One. 2013;8: e72448.

121. Abecasis AB, Wensing AMJ, Paraskevis D, Vercauteren J, Theys K, Van de Vijver DAMC, et al. HIV-1 subtype distribution and its demographic determinants in newly diagnosed patients in Europe suggest highly compartmentalized epidemics. Retrovirology. 2013;10: 7.

122. Glauser MP, Infeetieuses M, Hospitatier C. Clinical and Epidemiological Survey of Acquired Immune Deficiency Syndrome in Europe. Eur J Clin Microbiol. 1984;3: 55–58.

123. Liebert MA, Casado C, Urtasun I, Saragosti S, Chaix M, Rossi ADE, et al. Different Distribution of HIV Type 1 Genetic Variants in European Patients with Distinct Risk Practices. AIDS Res Hum Retroviruses. 2000;16: 299–304.

124. HIV/AIDS Surveillance in Europe. European Center for the Epidemiological Monitoring of AIDS. Q Rep. 1998; 59.

125. Casado C, Urtasun I, Saragosti S, Chaix ML, De Rossi A, Cattelan AM, et al. Different distribution of HIV type 1 genetic variants in European patients with distinct risk practices. AIDS Res Hum Retroviruses. 2000;16: 299–304.

126. Lukashov V V., Kuiken CL, Vlahov D, Coutinho R a., Goudsmit J. Evidence for HIV Type 1 Strains of U.S. Intravenous Drug Users as Founders of AIDS Epidemic among Intravenous Drug Users in Northern Europe. AIDS Res Hum Retroviruses. 1996;12: 1179–1183.

127. Bobkov A, Kazennova E, Selimova L, Bobkova M, Khanina T, Ladnaya N, et al. A sudden epidemic of HIV type 1 among injecting drug users in the former Soviet Union: identification of subtype A, subtype B, and novel gagA/envB recombinants. AIDS Res Hum Retroviruses. 1998;14: 669–76.

128. Holmes EC, Zhang LQ, Robertson P, Cleland A, Harvey E, Simmonds P, et al. The Molecular Epidemiology Of Human Immunodeficiency Virus Type 1 In Edinburgh. J Infect Dis. 1995;171: 45–53.

129. Nabatov AA, Kravchenko ON, Lyulchuk MG, Shcherbinskaya AM, Lukashov V V. Simultaneous introduction of HIV type 1 subtype A and B viruses into injecting drug users in southern Ukraine at the beginning of the epidemic in the former Soviet Union. AIDS Res Hum Retroviruses. 2002;18: 891–5.

130. Van Griensven F, de Lind van Wijngaarden JW, Baral S, Grulich A. The global epidemic of HIV infection among men who have sex with men. Curr Opin HIV AIDS. 2009;4: 300–7.

131. Baral S, Sifakis F, Cleghorn F, Beyrer C. Elevated risk for HIV infection among men who have sex with men in low- and middle-income countries 2000-2006: a systematic review. PLoS Med. 2007;4: e339.

132. Kato S, Saito R, Hiraishi Y, Kitamura N, Matsumoto T, Hanabusa H, et al. Differential prevalence of HIV type 1 subtype B and CRF01_AE among different sexual transmission groups in Tokyo, Japan, as revealed by subtype-specific PCR. AIDS Res Hum Retroviruses. 2003;19: 1057–63.

133. Kondo M, Lemey P, Sano T, Itoda I, Yoshimura Y, Sagara H, et al. Emergence in Japan of an HIV-1 variant associated with transmission among men who have sex with men (MSM) in China: first indication of the International Dissemination of the Chinese MSM lineage. J Virol. 2013;87: 5351–61.

134. Wang W, Xu J, Jiang S, Yang K, Meng Z, Ma Y, et al. The dynamic face of HIV-1 subtypes among men who have sex with men in Beijing, China. Curr HIV Res. 2011;9: 136–9.

135. Tee KK, Saw TL, Pon CK, Kamarulzaman A, Ng KP. The evolving molecular epidemiology of HIV type 1 among injecting drug users (IDUs) in Malaysia. AIDS Res Hum Retroviruses. 2005;21: 1046–50.

136. Takebe Y, Naito Y, Raghwani J, Fearnhill E, Sano T, Kusagawa S, et al. Intercontinental dispersal of HIV-1 subtype B associated with transmission among men who have sex with men in Japan. J Virol. 2014;88: 9864–76.

137. Chen JH-K, Wong K-H, Chan KC-W, To SW-C, Chen Z, Yam W-C. Phylodynamics of HIV-1 subtype B among the men-having-sex-with-men (MSM) population in Hong Kong. PLoS One. 2011;6: e25286.

138. Ng KT, Ong LY, Lim SH, Takebe Y, Kamarulzaman A, Tee KK. Evolutionary history of HIV-1 subtype B and CRF01_AE transmission clusters among men who have sex with men (MSM) in Kuala Lumpur, Malaysia. PLoS One. 2013;8: e67286.

139. Ye J, Lu H, Su X, Xin R, Bai L, Xu K, et al. Phylogenetic and temporal dynamics of human immunodeficiency virus type 1B in China: four types of B strains circulate in China. AIDS Res Hum Retroviruses. 2014;30: 920–6.

140. Chibo D, Birch C. Increasing Diversity of Human Immunodeficiency Virus Type 1 Subtypes Circulating in Australia. AIDS Res Hum Retroviruses. 2012;28: 578–583.

141. National Centre in HIV Epidemiology and Clinical Research. The National HIV Database. Aust HIV Surv Rep. 2002; 1–23.

142. Hawke KG, Waddell RG, Gordon DL, Ratcliff RM, Ward PR, Kaldor JM. HIV non-B subtype distribution: emerging trends and risk factors for imported and local infections newly diagnosed in South Australia. AIDS Res Hum Retroviruses. 2013;29: 311–7.

143. Spira S, Wainberg MA, Loemba H, Turner D, Brenner BG. Impact of clade diversity on HIV-1 virulence, antiretroviral drug sensitivity and drug resistance. J Antimicrob Chemother. 2003;51: 229–240.

144. Ho DDD, Huang Y. The HIV-1 vaccine race. Cell. Elsevier; 2002;110: 135–138.

145. Nelson M, Portsmouth S, Stebbing J, Atkins M, Barr A, Matthews G, et al. An open-label study of tenofovir in HIV-1 and Hepatitis B virus co-infected individuals. AIDS. 2003;17: F7–10.

146. Perelson a S, Neumann a U, Markowitz M, Leonard JM, Ho DD. HIV-1 dynamics in vivo: virion clearance rate, infected cell life-span, and viral generation time. Science. 1996;271: 1582–6.

147. Preston BD, Poiesz BJ, Loeb LA. Fidelity of HIV-1 reverse transcriptase. Science. 1988;242: 1168–1171.

148. Walker BD, Korber BT. Immune control of HIV: the obstacles of HLA and viral diversity. Nat Immunol. 2001;2: 473–475.

149. Potts KE, Kalish ML, Lott T, Orloff G, Luo CC, Bernard MA, et al. Genetic heterogeneity of the V3 region of the HIV-1 envelope glycoprotein in Brazil. Brazilian Collaborative AIDS Research Group. AIDS. 1993;7: 1191–1197.

150. Daniels RS, Kang C, Patel D, Xiang Z, Douglas NW, Zheng NN, et al. An HIV type 1 subtype B founder effect in Korea: gp160 signature patterns infer circulation of CTL-escape strains at the population level. AIDS Res Hum Retroviruses. 2003;19: 631–41.

151. Carr JK, Foley BT, Leitner T, Salminen M, Korber B, McCutchan F. Reference sequences representing the principal genetic diversity of HIV-1 in the pandemic. Hum retroviruses AIDS. Los Alamos, NM: Theoretical Biology and Biophysics Group; 1998; 10–19.

152. Morgado MG, Sabino EC, Shpaer EG, Bongertz V, Brigido L, Guimaraes MD, et al. V3 region polymorphisms in HIV-1 from Brazil: prevalence of subtype B strains divergent from North American/European prototype and detection of subtype F. AIDS Res Hum Retroviruses. 1994;10: 569–76.

153. Morgado MG, Guimarães ML, Neves Júnior I, dos Santos VG, Linhares-de-Carvalho MI, Castello-Branco LR, et al. Molecular epidemiology of HIV in Brazil: polymorphism of the antigenically distinct HIV-1 B subtype strains. The Hospital Evandro Chagas AIDS Clinical Research Group. Mem Inst Oswaldo Cruz. 1998;93: 383–6.

154. Louwagie J, Delwart EL, Mullins JI, McCutchan FE, Eddy G, Burke DS. Genetic analysis of HIV-1 isolates from Brazil reveals presence of two distinct genetic subtypes. AIDS Res Hum Retroviruses. 1994;10: 561–7.

155. Casseb J, Hong MA, Gonsalez C, Brígido LF, Duarte AJ, Michael-Hendry R. Two variants of HIV-1 B serotype are transmitted heterosexually in São Paulo, Brazil. Braz J Med Biol Res. 1998;31: 1243–6.

156. Junqueira DM, Medeiros RM De, Leite TCNF, Guimarães ML, Gräf T, Pinto AR, et al. Detection of the B"-GWGR variant in the southernmost region of Brazil: unveiling the complexity of the human immunodeficiency virus-1 subtype B epidemic. Mem Inst Oswaldo Cruz. 2013;108: 735–40.

157. Leal E, Silva WP, Sucupira MC, Janini LM, Diaz RS. Molecular and structural characterization of HIV-1 subtype B Brazilian isolates with GWGR tetramer at the tip of the V3-loop. Virology. 2008;381: 222–9.

158. Arruda L, Romano C, Martinez M, Araújo M, Costa F, Oliveira K, et al. The HIV-1 Subtype B variant (B'-GWGR motif) was introduced by founder effect among the HIV-1-infected subjects in São Paulo city, Brazil. 6 IAS Conference on HIV Pathogenesis, Treatment and Prevention. 2011.

159. Shimizu N, Takeuchi Y, Naruse T, Inagaki M, Moriyama E, Gojobori T, et al. Six strains of human immunodeficiency virus type 1 isolated in Japan and their molecular phylogeny. J Mol Evol. 1992;35: 329–36.

160. Araujo AF, Brites C, Monteiro-Cunha J, Santos LA, Galvao-Castro B, Alcantara LCJ. Lower prevalence of human immunodeficiency virus type 1 Brazilian subtype B found in northeastern Brazil with slower progression to AIDS. AIDS Res Hum Retroviruses. 2010;26: 1249–54.

161. Franca RFO, Castro-Jorge LA, Neto RJP, Jorge DMM, Lima DM, Colares JKB, et al. Genotypic characteristics of HIV type 1 based on gp120 hypervariable region 3 of isolates from Southern Brazil. AIDS Res Hum Retroviruses. 2011;27: 903–9.

162. Leal É, Villanova FE. Diversity of HIV-1 subtype B: implications to the origin of BF recombinants. PLoS One. 2010;5: e11833.

163. Diaz RS, Leal E, Sanabani S, Sucupira MC a, Tanuri AAA, Sabino EC, et al. Selective regimes and evolutionary rates of HIV-1 subtype B V3 variants in the Brazilian epidemic. Virology. Elsevier Inc.; 2008;381: 184–193.

164. Covas DT, Bíscaro TA, Kashima S, Duarte G, Machado AA. High frequency of the GWG (Pro Trp) envelope variant of HIV-1 in Southeast Brazil. J Acquir immune Defic Syndr Hum retrovirology. 1998;19: 74–9.

165. Casseb J, Montanheiro P, Komninakis S, Brito A, Duarte AJS. Human immunodeficiency virus type 1 Brazilian subtype B variant showed an increasing avidity of the anti-V3 antibodies over time compared to the subtype B US/European strain in São Paulo, Brazil. Mem Inst Oswaldo Cruz. 2004;99: 69–71.

166. Brito A De, Komninakis SC V, Oliveira RM De, Fonseca LAM, Duarte AJS, Casseb J. Women Infected with HIV Type 1 Brazilian Variant , Subtype B (B -GWGR Motif) Have Slower

977 Progression to AIDS , Compared with Patients Infected with Subtype B (B-GPGR Motif). Clin
978 Infect Dis. 2006;43: 0–5.

979 167. Santoro-lopes G, Harrison LEEH, Tavares MD, Xexéo A, Santos ANACEDOS, Schechter M, et
980 al. HIV disease progression and V3 serotypes in Brazil: is B different from B-Br? AIDS Res Hum
981 Retroviruses. 2000;16: 953–958.

982 168. Kim GJ, Nam J-G, Shin BG, Kee MK, Kim E-J, Lee J-S, et al. National survey of prevalent HIV
983 strains: limited genetic variation of Korean HIV-1 clade B within the population of Korean men
984 who have sex with men. J Acquir Immune Defic Syndr. 2008;48: 127–32.

985 169. Kim YB, Cho YK, Lee HJ, Kim CK, Kim YK, Yang JM. Molecular phylogenetic analysis of
986 human immunodeficiency virus type 1 strains obtained from Korean patients: env gene
987 sequences. AIDS Res Hum Retroviruses. 1999;15: 303–7.

988 170. Kim YB, Cho YK. Monophyletic clade of HIV-1 subtype B in Korea: evolutionary pressure or
989 single introduction? AIDS Res Hum Retroviruses. 2003;19: 619–23.

990 171. KCDC. Korea Center for Disease Control and Prevention - Periodical Report. Seoul; 2007.

991 172. Deng X, Liu H, Shao Y, Rayner S, Yang R. The epidemic origin and molecular properties of B':
992 a founder strain of the HIV-1 transmission in Asia. AIDS. 2008;22: 1851–8.

993 173. Kalish ML, Baldwin A, Raktham S, Wasi C, Luo CC, Schochetman G, et al. The evolving
994 molecular epidemiology of HIV-1 envelope subtypes in injecting drug users in Bangkok,
995 Thailand: implications for HIV vaccine trials. AIDS. 1995;9: 851–7.

996 174. Zhang L, Chen Z, Cao Y, Yu J, Li G, Yu W, et al. Molecular characterization of human
997 immunodeficiency virus type 1 and hepatitis C virus in paid blood donors and injection drug users
998 in china. J Virol. 2004;78: 13591–9.

999 175. Li Z, He X, Wang Z, Xing H, Li F, Yang Y, et al. Tracing the origin and history of HIV-1
1000 subtype B' epidemic by near full-length genome analyses. AIDS. 2012;26: 877–84.

1001 176. Li Y, Uenishi R, Hase S, Liao H, Li X, Tsuchiura T, et al. Explosive HIV-1 subtype B '
1002 epidemics in Asia driven by geographic and risk group founder events. Virology. Elsevier Inc.;
1003 2010;402: 223–227.

1004 177. Su L, Graf M, Zhang Y, von Briesen H, Xing H, Köstler J, et al. Characterization of a virtually
1005 full-length human immunodeficiency virus type 1 genome of a prevalent intersubtype (C/B')
1006 recombinant strain in China. J Virol. 2000;74: 11367–76.

1007 178. Piyasirisilp S, McCutchan FE, Carr JK, Sanders-Buell E, Liu W, Chen J, et al. A recent outbreak
1008 of human immunodeficiency virus type 1 infection in southern China was initiated by two highly
1009 homogeneous, geographically separated strains, circulating recombinant form AE and a novel BC
1010 recombinant. J Virol. 2000;74: 11286–95.

1011 179. Tovanabutra S, Watanaveeradej V, Viputtikul K, De Souza M, Razak MH, Suriyanon V, et al. A
1012 new circulating recombinant form, CRF15_01B, reinforces the linkage between IDU and
1013 heterosexual epidemics in Thailand. AIDS Res Hum Retroviruses. 2003;19: 561–7.

1014 180. Tovanabutra S, Kijak GH, Beyrer C, Gammon-Richardson C, Sakkhachornphop S, Vongchak T,
1015 et al. Identification of CRF34_01B, a second circulating recombinant form unrelated to and more
1016 complex than CRF15_01B, among injecting drug users in northern Thailand. AIDS Res Hum
1017 Retroviruses. 2007;23: 829–33.

181. Tee KK, Li X-J, Nohtomi K, Ng KP, Kamarulzaman A, Takebe Y. Identification of a novel circulating recombinant form (CRF33_01B) disseminating widely among various risk populations in Kuala Lumpur, Malaysia. J Acquir Immune Defic Syndr. 2006;43: 523–9.

182. Li Y, Tee KK, Liao H, Hase S, Uenishi R, Li X-J, et al. Identification of a novel second-generation circulating recombinant form (CRF48_01B) in Malaysia: a descendant of the previously identified CRF33_01B. J Acquir Immune Defic Syndr. 2010;54: 129–36.

183. Liu Y, Li L, Bao Z, Li H, Zhuang D, Liu S, et al. Identification of a novel HIV type 1 circulating recombinant form (CRF52_01B) in Southeast Asia. AIDS Res Hum Retroviruses. 2012;28: 1357–61.

184. Chow WZ, Al-Darraji H, Lee YM, Takebe Y, Kamarulzaman A, Tee KK. Genome sequences of a novel HIV-1 CRF53_01B identified in Malaysia. J Virol. 2012;86: 11398–9.

185. Li X, Ning C, He X, Yang Y, Li F, Xing H, et al. Genome Sequences of a Novel HIV-1 Circulating Recombinant Form (CRF61_BC) Identified among Heterosexuals in China. Genome Announc. 2013;1.

186. Wei H, His J, Feng Y, Xing H, He X, Liao L, et al. Identification of a novel HIV-1 circulating recombinant form (CRF62_BC) in western Yunnan of China. AIDS Res Hum Retroviruses. 2014;30: 380–3.

187. Han X, An M, Zhao B, Duan S, Yang S, Xu J, et al. High prevalence of HIV-1 intersubtype B'/C recombinants among injecting drug users in Dehong, China. PLoS One. 2013;8: e65337.

188. Ng KT, Ong LY, Takebe Y, Kamarulzaman A, Tee KK. Genome sequence of a novel HIV-1 circulating recombinant form 54_01B from Malaysia. J Virol. 2012;86: 11405–6.

189. Chow WZ, Takebe Y, Syafina NE, Prakasa MS, Chan KG, Al-Darraji HAA, et al. A newly emerging HIV-1 recombinant lineage (CRF58_01B) disseminating among people who inject drugs in Malaysia. PLoS One. 2014;9: e85250.

190. Feng Y, Wei H, Hsi J, Xing H, He X, Liao L, et al. Identification of a novel HIV Type 1 circulating recombinant form (CRF65_cpx) composed of CRF01_AE and subtypes B and C in Western Yunnan, China. AIDS Res Hum Retroviruses. 2014;30: 598–602.

191. Kijak GH, Tovanabutra S, Sanders-Buell E, Watanaveeradej V, de Souza MS, Nelson KE, et al. Distinguishing molecular forms of HIV-1 in Asia with a high-throughput, fluorescent genotyping assay, MHAbce v.2. Virology. 2007;358: 178–91.

192. Takebe Y, Motomura K, Tatsumi M, Lwin HH, Zaw M, Kusagawa S. High prevalence of diverse forms of HIV-1 intersubtype recombinants in Central Myanmar: geographical hot spot of extensive recombination. AIDS. 2003;17: 2077–87.

193. Wang B, Lau KA, Ong L-Y, Shah M, Steain MC, Foley B, et al. Complex patterns of the HIV-1 epidemic in Kuala Lumpur, Malaysia: evidence for expansion of circulating recombinant form CRF33_01B and detection of multiple other recombinants. Virology. 2007;367: 288–97.

194. Yang R, Xia X, Kusagawa S, Zhang C, Ben K, Takebe Y. On-going generation of multiple forms of HIV-1 intersubtype recombinants in the Yunnan Province of China. AIDS. 2002;16: 1401–7.

195. Collins-Fairclough AM, Charurat M, Nadai Y, Pando M, Avila MM, Blattner WA, et al. Significantly longer envelope V2 loops are characteristic of heterosexually transmitted subtype B HIV-1 in Trinidad. PLoS One. 2011;6: e19995.

Capítulo 4: Artigo 02

"HIV- Detection of the B"-GWGR variant in the southernmost region of Brazil: unveiling the complexity of the human immunodeficiency virus-1 subtype B epidemic"

Dennis Maletich Junqueira, Rúbia Marília de Medeiros, Thaysse Cristina Neiva Ferreira Leite, Monick Lindenmeyer Guimarães, Tiago Gräf, Aguinaldo Roberto Pinto, Sabrina Esteves de Matos Almeida

Memórias do Instituto Oswaldo Cruz, 2013

Detection of the B”-GWGR variant in the southernmost region of Brazil: unveiling the complexity of the human immunodeficiency virus-1 subtype B epidemic

Dennis Maletich Junqueira[1,2]/+, Rúbia Marília de Medeiros[1,2], Thaysse Cristina Neiva Ferreira Leite[3], Monick Lindenmeyer Guimarães[3], Tiago Gräf[1,4], Aguinaldo Roberto Pinto[4], Sabrina Esteves de Matos Almeida[1]

[1]Centro de Desenvolvimento Científico e Tecnológico, Fundação Estadual de Produção e Pesquisa em Saúde, Porto Alegre, RS, Brasil
[2]Programa de Pós-Graduação em Genética e Biologia Molecular, Universidade Federal do Rio Grande do Sul, Porto Alegre, RS, Brasil
[3]Laboratório de AIDS e Imunologia Molecular, Instituto Oswaldo Cruz-Fiocruz, Rio de Janeiro, RJ, Brasil
[4]Laboratório de Imunologia Aplicada, Departamento de Microbiologia, Imunologia e Parasitologia, Centro de Ciências Biológicas, Universidade Federal de Santa Catarina, Florianópolis, SC, Brasil

Typical human immunodeficiency virus-1 subtype B (HIV-1B) sequences present a GPGR signature at the tip of the variable region 3 (V3) loop; however, unusual motifs harbouring a GWGR signature have also been isolated. Although epidemiological studies have detected this variant in approximately 17-50% of the total infections in Brazil, the prevalence of B”-GWGR in the southernmost region of Brazil is not yet clear. This study aimed to investigate the C2-V3 molecular diversity of the HIV-1B epidemic in southernmost Brazil. HIV-1 seropositive patients were analysed at two distinct time points in the state of Rio Grande do Sul (RS98 and RS08) and at one time point in the state of Santa Catarina (SC08). Phylogenetic analysis classified 46 individuals in the RS98 group as HIV-1B and their molecular signatures were as follows: 26% B”-GWGR, 54% B-GPGR and 20% other motifs. In the RS08 group, HIV-1B was present in 32 samples: 22% B”-GWGR, 59% B-GPGR and 19% other motifs. In the SC08 group, 32 HIV-1B samples were found: 28% B”-GWGR, 59% B-GPGR and 13% other motifs. No association could be established between the HIV-1B V3 signatures and exposure categories in the HIV-1B epidemic in RS. However, B-GPGR seemed to be related to heterosexual individuals in the SC08 group. Our results suggest that the established B”-GWGR epidemics in both cities have similar patterns, which is likely due to their geographical proximity and cultural relationship.

Key words: HIV - subtype B - B”-GWGR - southernmost Brazil - molecular epidemiology

The third hypervariable region 3 (V3) of the human immunodeficiency virus-1 (HIV-1) gp120 protein consists of 35 amino acids and plays an important role in viral infection by promoting the interaction between the virus and its co-receptor in the host membrane (Hwang et al. 1991). Despite the recognised potential of HIV-1 to escape from neutralising antibodies through the extensive variability of its viral envelope glycoproteins, especially gp120, it is well known that the four amino acids at the tip of the V3 loop are subjected to strong purifying selection pressure due to their functional importance (Kwong et al. 2002, Liu et al. 2008). Although the vast majority of HIV-1 sequences present a GPGX signature at this position regardless of subtype, unusual patterns have been reported around the world (Shimizu et al. 1992, Brown et al. 1996, Kim et al. 1999, Leal & Villanova 2010).

Subtype B is the most geographically widespread variant of HIV-1 (HIV-1B) (Hemelaar 2012). The pandemic form of subtype B, which is prevalent in European, American and Asian countries, is typically characterised as having a GPGR motif (B-GPGR) at the tip of the V3 loop. However, several molecular studies have shown that various genetically and antigenically distinct V3 motifs, which are diversified particularly at the second position of the tetramer, co-circulate in the HIV-1B epidemic (Shimizu et al. 1992, Morgado et al. 1994, Candotti et al. 1999, Kim et al. 1999, Leal et al. 2008, Franca et al. 2011). In particular, some strains have been found to harbour an alternative signature in which the second residue of the tetrapeptide, proline, is substituted with tryptophan (B”-GWGR) (Potts et al. 1993, Casseb et al. 1998, 2002, Morgado et al. 1998, Santoro-Lopes et al. 2000, Brito et al. 2006, Araujo et al. 2010, Franca et al. 2011). Clinical studies support the hypothesis that the B”-GWGR motif is correlated with slower disease progression in infected patients when compared with those infected with the B-GPGR variant (Santoro-Lopes et al. 2000, Brito et al. 2006, Araujo et al. 2010).

Brazil is accepted as the epicentre of the B”-GWGR epidemic (Diaz et al. 2008, Pinto et al. 2008). While viruses presenting this motif are sporadically observed in other countries, several studies in Brazil have found this variant at frequencies ranging from 17-50% (Morgado et

doi: 10.1590/0074-0276108062013010
Financial support: FAPESC, CNPq, FIOCRUZ
+ Corresponding author: dennismaletich@hotmail.com
Received 1 April 2013
Accepted 26 June 2013

al. 1994, Casseb et al. 1998, Araujo et al. 2010, Franca et al. 2011). Despite these results, recent findings about the temporal trends of the B"-GWGR epidemic in this country suggest a decline in the prevalence of this variant (Araujo et al. 2010, Franca et al. 2011). Several studies have highlighted the complexity of the HIV-1 epidemic in the southernmost region of Brazil, where subtype C, subtype B and several recombinant forms have been detected (Soares et al. 2003, Santos et al. 2006, de Medeiros et al. 2011, Gräf et al. 2011, Gräf & Pinto 2013), but no studies have attempted to detect B"-GWGR in this region. The current extent of the B"-GWGR epidemic in Brazil remains an unresolved question. Thus, the present study aims to investigate the molecular diversity of the HIV-1 subtype B epidemic in distinct exposure categories in the southernmost region of Brazil.

SUBJECTS, MATERIALS AND METHODS

The present study investigated 278 samples from three distinct groups of HIV-1-positive individuals from various outpatient clinics in the southernmost region of Brazil. The first group (RS98) contained 83 blood samples that were obtained from HIV-1-positive individuals recruited from a health reference centre in Porto Alegre, the capital city of the state of Rio Grande do Sul, in 1998. The second group (RS08) was comprised of 97 samples from HIV-infected individuals that were collected in Porto Alegre between 2005-2008. The third group (SC08) was composed of blood samples that were collected from 98 HIV-positive patients at follow-up appointments at a reference centre in the city of Florianópolis, the capital city of the state of Santa Catarina, between 2004-2008. The demographic data of the patients included in this study were extracted from clinical records or were obtained through direct interview. All of the individuals from Porto Alegre were antiretroviral treatment-naïve, while the individuals from Florianópolis were either naïve or under antiretroviral therapy at the time of blood collection. This study was approved by the ethical committees of the institutions involved and all patients provided written informed consent.

DNA was extracted from 200 µL of each whole blood sample using a QIAamp DNA kit (Qiagen, CA, USA) according to the manufacturer's protocol. The partial C2-V3 region of the *env* gene (nucleotides 6921-7283, relative to strain HXB2) was amplified by polymerase chain reaction (PCR) using nested primers as previously described (Delwart et al. 1993). The products were purified using a PureLink PCR Purification kit (Invitrogen, CA, USA) according to the manufacturer's directions. The purified DNA was sequenced using the ABI BigDye Terminator v.3.1 Cycle Sequencing Ready Reaction kit (Applied Biosystems, CA, USA) and processed with an automated ABI 3130xl Genetic Analyzer (Applied Biosystems). The sequences were edited and then aligned with reference sequences retrieved from the Los Alamos Sequence Database using CLUSTALX (Larkin et al. 2007).

Subtypes were assigned based on maximum likelihood (ML) phylogenetic reconstruction conducted on MEGA 5 software under the GTR+G+I model of nucleotide substitution (Tamura et al. 2011). A bootstrap test with 1,000 replicates was used to estimate the confidence of the branching patterns of the phylogenetic tree. The sequences were also submitted to the REGA Subtyping Tool of the BioAfrica Database to corroborate the subtypes assigned by the ML analysis.

The molecular signatures of the HIV-1 subtype B V3 loop were identified through visual inspection from amino acid positions 15-18 (nucleotides 7158-7169, relative to strain HXB2). The typical subtype B signature was identified as B-GPGR, whereas viruses harbouring the alternative W (tryptophan) signature were assigned as B"-GWGR. Assuming that the epidemiological, serological and clinical differences observed between the B and B" signatures in previous studies are accurate (Hendry et al. 1996, Santoro-Lopes et al. 2000, Casseb et al. 2002, Brito et al. 2006, Leal et al. 2008, Pinto et al. 2008), sequences presenting related motifs that retained the P or W at second position of the tetrapeptide (XPXX and XWXX) were also considered as B-GPGR or B"-GWGR, respectively, because these most likely evolved from an ancestral sequence containing one of these motifs (Diaz et al. 2008). Sequences depicting an amino acid other than W or P at position 16 of the V3 loop were evaluated separately.

Statistical comparisons between and within groups were made using Pearson's χ^2-test and Fisher's exact test when appropriate. The statistical analyses were performed using WinPepi v.11.22 (Abramson 2004) and the significance level was set at $p < 0.05$. Due to their low frequency, the sequences harbouring amino acids other than P or W in the second position of the V3 loop were excluded from the statistical analyses.

RESULTS

Phylogenetic analysis of group RS98 revealed the co-circulation of four HIV-1 subtypes: B (55%), C (39%), F1 (5%) and A (1%). Of the 46 HIV-1 subtype B samples in RS98, 54% were typed as B-GPGR, 26% as B"-GWGR and 20% presented an amino acid other than W or P at position 16 of the V3 loop (Tables I, II). In this study, B"-GWGR viruses accounted for 15% of all of the HIV infections evaluated in Porto Alegre in 1998. Based on their medical records, the RS98 HIV-1B-infected patients were categorised according to the route of probable infection: heterosexual (HET) (50%), men who have sex with men (MSM) (44%), people who inject drugs (4%) and blood transfusion (2%). Their classification according to gender was 70% male and 30% female (Table I). The HIV-1B signatures according to the patients' gender and exposure category are shown in Table I. No association could be established between the subtype B molecular signatures and the exposure category or gender.

Of the 97 HIV-1 individuals in group RS08, 33% were subtyped as B and 67% were subtyped as C. The subtype B V3 loop signatures were as follows: 59% B-GPGR, 22% B"-GWGR and 19% other motifs (Tables I, II). In total, B"-GWGR was observed in 8% of all of the HIV-1 infections evaluated in Porto Alegre between 2005-2008. Regarding the exposure categories of the subtype B samples, group RS08 showed (53%) HET (6 males and 11 females) and (31%) MSM individuals. The exposure

category was not determined for five individuals. No statistically significant associations between the subtype B signatures and the exposure category or gender were observed for RS08. After verifying the homogeneity of the gender and number of the individuals in each exposure category, statistical comparison based on Pearson's χ^2-test revealed no temporal differences between the RS98 and RS08 groups regarding the subtype B epidemic.

In group SC08, subtype B was detected in 32 samples with a differential distribution of the molecular signatures: 59% of the samples presented the B-GPGR motif, 28% presented the B"-GWGR motif and 13% presented other motifs (Tables I, II). The remaining 68% samples were assigned as subtype C. Epidemiologically, the B"-GWGR variant was found to be responsible for approximately 9% of the total infections in Florianópolis. Regarding the risk factors for HIV-1B infection, 72% of the patients were HET (15 females and 8 males) and 28% were MSM. Similar distributions of the HIV-1B molecular signatures were observed in Porto Alegre (RS08) and Florianópolis (SC08) during the same time period. A statistically significant association ($p = 0.035$) was observed between the HET exposure category and the B-GPGR variant in the SC08 group.

Nineteen samples presented 13 different motifs that were unrelated to the B-GPGR or B"-GWGR motifs in this analysis (9 samples from RS98, 6 samples from RS08 and 4 from SC08) (Table II). In addition, six motifs that were related to the B-GPGR motif and two motifs that were related to the B"-GWGR motif were detected. Three samples from group RS98 exhibited unique motifs: GRGA, GRGR and RRGG.

DISCUSSION

Now present in at least 23 countries around the world (Pinto et al. 2008, Leal & Villanova 2010), the B"-GWGR variant of HIV-1 seems to have originated in Brazil (Diaz et al. 2008, Pinto et al. 2008, Leal & Villanova 2010) and since its isolation in the early 1990s, it has been widely studied in this country (Potts et al. 1993, Louwagie et al. 1994, Casseb et al. 1998, 2002, Morgado et al. 1998, Santoro-Lopes et al. 2000, Brito et al. 2006, Pinto et al. 2008, Araujo et al. 2010, Franca et al. 2011). However, the vast majority of studies concerning the B"-GWGR motif have focused on the study of the HIV-1 epidemic in southeastern Brazil, where the prevalence of subtype B is extremely high. In addition, information about the B"-GWGR epidemic in other regions of Brazil is scarce. This is the first report of the circulation of B"-GWGR viruses in the southernmost region of Brazil.

The current study shows that B"-GWGR motif-containing viruses play an important role in the HIV-1 subtype B epidemic in southernmost Brazil (Table I). Our molecular analysis of HIV-1 blood samples from Porto Alegre showed that the B"-GWGR motif was present in approximately 24% of the subtype B samples. These results revealed no significant difference in the distribution of molecular signatures within the subtype B

TABLE I

Human immunodeficiency virus-1 subtype B variable region 3 loop motifs frequencies according to patient's gender and exposure category

	Group RS98 (n = 46)			Group RS08 (n = 32)			Group SC08 (n = 32)		
	B"-GWGR and related (n = 12)	B-GPGR and related (n = 25)	Other (n = 9)	B"-GWGR and related (n = 7)	B-GPGR and related (n = 19)	Other (n = 6)	B"-GWGR and related (n = 9)	B-GPGR and related (n = 19)	Other (n = 4)
Gender [n (%)]									
Male	8 (67)	17 (68)	7 (78)	4 (57)	11 (58)	3 (5)	7 (78)	7 (37)	3 (75)
Female	4 (33)	8 (32)	2 (22)	2 (29)	7 (37)	2 (33)	2 (22)	12 (63)	1 (25)
ND	-	-	-	1 (14)	1 (5)	1 (17)	-	-	-
Exposure category [n (%)]									
HET	6 (50)	12 (48)	5 (56)	5 (71)	8 (42)	4 (66)	4 (44)	17 (89)[a]	2 (50)
MSM	5 (42)	13 (52)	2 (22)	1 (14)	8 (42)	1 (17)	5 (56)	2 (11)	2 (50)
PWID	1 (8)	-	1 (11)	-	-	-	-	-	-
BT	-	-	1 (11)	-	-	-	-	-	-
ND	-	-	-	1 (14)	3 (16)	1 (17)	-	-	-
Prevalence (within group) (%)									
	26	54	20	22	59	19	28	59	13

a: significant association ($p = 0.035$) between heterosexual (HET) individuals and B-GPGR motif infection; BT: blood transfusion; MSM: men who have sex with men; ND: not declared; PWID: people who inject drugs. Values in the brackets are the relative percentages according to gender or exposure category.

epidemic between 1998-2008 and suggest that the epidemic has been stable in this city since the mid-1990s. In contrast, epidemiological studies have shown that the prevalence of B"-GWGR has been decreasing over time in other regions of Brazil (Hendry et al. 1996, Casseb et al. 1998, Brito et al. 2006, Araujo et al. 2010). B"-GWGR viruses were previously estimated (by anti-V3 serologic assay) to account for approximately 27%, 48% and 64% of all HIV infections from samples isolated in the 1990s in the cities of Salvador, Rio de Janeiro and São Paulo, respectively (Hendry et al. 1996). However, more recently, molecular investigations detected this motif in 18%, 23% and 34% of the investigated HIV-infected individuals in the same cities in the years of 2006-2010 (Araujo et al. 2010, Arruda et al. 2011, Pimentel et al. 2011). Nevertheless, the differences observed in the other Brazilian regions could be related to the varied methods used to detect the B"-GWGR variant or could be due to the sampling of individuals from different exposure categories. In addition, the observed temporal difference in the B"-GWGR epidemic could be the result of the influence of a random effect on the local transmission networks. However, further studies that use molecular analysis to examine representative sample sizes that include individuals from the same gender, exposure category and acquired immune deficiency syndrome (AIDS) progression stage are needed to assess the temporal trends of the B"-GWGR epidemic in Brazil.

TABLE II

Human immunodeficiency virus-1 subtype B variable region 3 (V3) motifs diversity according to the local of sample collection and year

V3 motif				Studied groups		
				RS98 (n = 46)	RS08 (n = 32)	SC08 (n = 32)
B-(GPGR) and related motifs						
G	P	G	R	19	14	11
G	P	G	K	2	2	1
A	P	G	R	-	1	1
G	P	G	S	2	-	2
G	P	G	Q	1	-	2
A	P	G	S	-	-	1
G	P	G	G	1	2	1
B"-(GWGR) and related motifs						
G	W	G	R	8	6	7
G	W	R	R	-	1	1
A	W	G	R	4	-	1
Other V3 motifs						
A	F	G	R	-	1	-
G	F	G	R	3	-	1
G	L	G	R	2	1	-
A	M	G	R	-	1	-
G	M	G	R	-	-	1
G	Q	G	R	-	-	1
G	S	G	R	-	1	-
G	T	G	R	1	-	-
G	G	G	R	-	1	1
G	V	G	R	-	1	-
G	R	G	A	1	-	-
G	R	G	R	1	-	-
R	R	G	G	1	-	-

The V3 loop motifs of the viruses in the subtype B epidemic in the city of Florianópolis (SC08) were not significantly different from those of the epidemic in Porto Alegre (RS08) in the same time period. Moreover, due to the circulation of non-B subtypes in southernmost Brazil, especially subtype C, the B"-GWGR variant may be evaluated as part of the total epidemic. In this case, B"-GWGR was responsible for 15% and 8% of the HIV-1 infections in the years 1998 and 2005-2008 in Porto Alegre and for 9% of the HIV-1 infections in 2004-2008 in Florianópolis. Nevertheless, these results are not comparable with the results of other related studies because most of the other studies were performed in states where non-B subtypes are more infrequent and subtype B predominates (Santos et al. 2011, Alcalde et al. 2012, Pilotto et al. 2012). Taken together, our results suggest that the HIV-1 subtype B epidemics in Florianópolis and Porto Alegre have a similar pattern (Table I). The cultural relationship and the geographical proximity of Porto Alegre and Florianópolis may have influenced the dynamic of the transmission chains and could explain the similar results within these cities (Bello et al. 2012). However, comparing the entire HIV-1 epidemic in Southern (which consists of subtype C, subtype B and recombinant forms) with the HIV-1 epidemic in other states of Brazil, it seems that B"-GWGR is not as prevalent in the more southern states of Brazil.

Although many studies have attempted to unravel the origin and prevalence of the B"-GWGR variant, only one study has sought to understand the spreading pattern of this variant and its relationship with the patient's exposure category (Pimentel et al. 2011). A recent study carried out in Rio de Janeiro found that the B"-GWGR variant was most likely introduced into the local epidemic by bisexual individuals (Pimentel et al. 2011). Analysis of the results for the RS98 and RS08 groups revealed no significant association between the exposure category and the V3 loop motif of HIV-1 subtype B. The lack of an association found here may be explained by the complete intermixing of local transmission chains (Almeida et al. 2012). Alternatively, this result may suggest the inexistence of V3 loop motif stratification by exposure category in the initial spread of HIV-1 subtype B in Porto Alegre. Although previous results have found a significant association ($p < 0.05$) between the MSM exposure category and subtype B in the 1990s (Almeida et al. 2012), our results suggest that there is no difference in regard to the V3 motifs in the subtype B epidemic in MSM and HET individuals between 1998-2008. Analysis of group SC08 suggests that the B"-GWGR motif is also not associated with any exposure category. In

contrast, it seems that the dissemination of B-GPGR in Florianópolis is associated with the HET population. In a previous study based on the *pol* gene, a significant difference in the subtype distribution among distinct exposure categories in the HIV epidemic in Florianópolis was observed (Gräf et al. 2011). Together, these results suggest the existence of limited events of transmission between MSM and HET individuals due to a reduced overlap of the transmission chains in Florianópolis.

A remarkable feature of the HIV-1 epidemic in the southernmost region of Brazil is the co-circulation of subtypes B and C in high proportions (de Medeiros et al. 2011, Gräf et al. 2011, Almeida et al. 2012, Araújo et al. 2012). In addition, this region encompasses the 10 cities with the highest AIDS incidence rates in Brazil (MS 2012). This scenario becomes even more complex with the identification of a co-circulating molecular variant of subtype B that seems to have clinical particularities (Santoro-Lopes et al. 2000, Casseb et al. 2002, Brito et al. 2006, Leal et al. 2008). The GWGR motif seems to increase the avidity of V3 antibodies for the virus and contributes to slower disease progression in comparison to B-GPGR infection (Casseb et al. 2004, Brito et al. 2006). As longer periods of HIV-1 infection are expected for B"-GWGR-containing viruses, the chance of HIV-1 transmission to other individuals should be greater. Consequently, an increased number of infections caused by B"-GWGR viruses is anticipated over time in Brazil. In contrast, the increase in the avidity of V3 antibodies could contribute to a decrease in the viral load in B"-GWGR infected patients, thereby reducing the chances of HIV-1 transmission to new hosts and consequently decreasing the number of infections caused by this variant. The results presented here for the HIV-1 epidemic in the southernmost region of Brazil demonstrate that the number of infections caused by B"-GWGR within the subtype B epidemic is not increasing, but is being maintained. Several hypotheses can explain the stabilisation of this epidemic. However, other studies assessing behavioural as well as biological and clinical data will be needed to answer these questions and predict the future of the HIV-1 epidemic in Brazil. Despite the complex HIV-1 epidemic in the southernmost region of Brazil, the frequency of B"-GWGR within subtype B viruses is comparable to that found in other Brazilian states. These results add another layer to the already complex HIV epidemic of southernmost Brazil and highlight the importance of surveillance studies in monitoring the dissemination of HIV-1 variants, specifically B"-GWGR, which seems to confer a differential clinical prognosis.

REFERENCES

Abramson JH 2004. WINPEPI (PEPI-for-Windows): computer programs for epidemiologists. *Epidemiol Perspect Innov 1*: 6.

Alcalde R, Guimarães ML, Duarte AJ, Casseb J 2012. Clinical, epidemiological and molecular features of the HIV-1 subtype C and recombinant forms that are circulating in the city of São Paulo, Brazil. *Virol J 9*: 156.

Almeida SEM, de Medeiros RM, Junqueira DM, Gräf T, Passaes CP, Bello G, Morgado MG, Guimarães ML 2012. Temporal dynamics of HIV-1 circulating subtypes in distinct exposure categories in southern Brazil. *Virol J 9*: 306.

Araujo AF, Brites C, Monteiro-Cunha J, Santos LA, Galvao-Castro B, Alcantara LCJ 2010. Lower prevalence of human immunodeficiency virus type 1 Brazilian subtype B found in northeastern Brazil with slower progression to AIDS. *AIDS Res Hum Retroviruses 26*: 1249-1254.

Araújo LA, Junqueira DM, de Medeiros RM, Matte MC, Almeida SE 2012. Naturally occurring resistance mutations to HIV-1 entry inhibitors in subtypes B, C and CRF31_BC. *J Clin Virol 54*: 6-10.

Arruda LB, Romano C, Martinez M, Araújo M, Costa F, Oliveira K, Gonsalez C, Duarte A, Casseb J 2011. The HIV-1 subtype B variant (B'-GWGR motif) was introduced by founder effect among the HIV-1-infected subjects in São Paulo city, Brazil. Proceedings of the 6th IAS Conference on HIV Pathogenesis, Treatment and Prevention, 17-20 July 2011, International AIDS Society, Rome. Available from: pag.ias2011.org/abstracts.aspx?aid=844.

Bello G, Zanotto PMA, Iamarino A, Gräf T, Pinto AR, Couto-Fernandez JC, Morgado MG 2012. Phylogeographic analysis of HIV-1 subtype C dissemination in southern Brazil. *PLoS ONE 7*: e35649.

Brito A, Komninakis SCV, Oliveira RM, Fonseca LAM, Duarte AJS, Casseb J 2006. Women infected with HIV type 1 Brazilian variant, subtype B (B -GWGR motif) have slower progression to AIDS, compared with patients infected with subtype B (B-GPGR Motif). *Clin Infect Dis 43*: 1-5.

Brown TM, Robbins KE, Sinniah M, Saraswathy TS, Lee V, Hooi LS, Vijayamalar B, Luo CC, Ou CY, Rapier J, Schochetman G, Kalish ML 1996. HIV type 1 subtypes in Malaysia include B, C and E. *AIDS Res Hum Retroviruses 12*: 1655-1657.

Candotti D, Tareau C, Barin F, Joberty C, Rosenheim M, M'Pele P, Huraux JM, Agut H 1999. Genetic subtyping and V3 serotyping of HIV type 1 isolates in Congo. *AIDS Res Hum Retroviruses 15*: 309-314.

Casseb J, Hong MA, Gonsalez C, Brígido LF, Duarte AJ, Michael-Hendry R 1998. Two variants of HIV-1 B serotype are transmitted heterosexually in São Paulo, Brazil. *Braz J Med Biol Res 31*: 1243-1246.

Casseb J, Komninakis S, Abdalla L, Brigido L, Rodrigues R, Araujo F, Veiga AP, de Almeida A, Flannery B, Hendry RM, Duarte AJ 2002. HIV disease progression: is the Brazilian variant subtype B' (GWGR motif) less pathogenic than US/European subtype B (GPGR). *Int J Infect Dis 6*: 164-169.

Casseb J, Montanheiro P, Komninakis S, Brito A, Duarte AJS 2004. Human immunodeficiency virus type 1 Brazilian subtype B variant showed an increasing avidity of the anti-V3 antibodies over time compared to the subtype B US/European strain in São Paulo, Brazil. *Mem Inst Oswaldo Cruz 99*: 69-71.

de Medeiros RM, Junqueira DM, Matte MCC, Barcellos NT, Chies JAB, Almeida SEM 2011. Co-circulation HIV-1 subtypes B, C and CRF31-BC in a drug-naïve population from southernmost Brazil: analysis of primary resistance mutations. *J Med Virol 83*: 1682-1688.

Delwart EL, Shpaer EG, Louwagie J, McCutchan FE, Grez M, Rübsamen-Waigmann H, Mullins JI 1993. Genetic relationships determined by a DNA heteroduplex mobility assay: analysis of HIV-1 env genes. *Science 262*: 1257-1261.

Diaz RS, Leal E, Sanabani S, Sucupira MC, Tanuri A, Sabino EC, Janini LM 2008. Selective regimes and evolutionary rates of HIV-1 subtype B V3 variants in the Brazilian epidemic. *Virology 381*: 184-193.

Franca RF, Castro-Jorge LA, Neto RJ, Jorge DM, Lima DM, Colares JK, Paula SO, da Fonseca BA 2011. Genotypic characteristics of HIV type 1 based on gp120 hypervariable region 3 of isolates from southern Brazil. *AIDS Res Hum Retroviruses 27*: 903-909.

Gräf T, Passaes CP, Ferreira LG, Grisard EC, Morgado MG, Bello G, Pinto AR 2011. HIV-1 genetic diversity and drug resistance among treatment naïve patients from Southern Brazil: an association of HIV-1 subtypes with exposure categories. *J Clin Virol 51*: 186-191.

Gräf T, Pinto AR 2013. The increasing prevalence of HIV-1 subtype C in southern Brazil and its dispersion through the continent. *Virology 435*: 170-178.

Hemelaar J 2012. The origin and diversity of the HIV-1 pandemic. *Trends Mol Med 18*: 182-192.

Hendry RM, Hanson CV, Bongertz V, Morgado M, Duarte A, Casseb J, Brigido L, Sabino E, Diaz R, Galvão-Castro B 1996. Immunoreactivity of Brazilian HIV isolates with different V3 motifs. *Mem Inst Oswaldo Cruz 91*: 347-348.

Hwang SS, Boyle TJ, Lyerly HK, Cullen BR 1991. Identification of the envelope V3 loop as the primary determinant of cell tropism in HIV-1. *Science 253*: 71-74.

Kim EY, Cho YS, Maeng SH, Kang C, Nam JG, Lee JS 1999. Characterization of V3 loop sequences from HIV type 1 subtype B in South Korea: predominance of the GPGS motif. *AIDS Res Hum Retroviruses 15*: 681-686.

Kwong PD, Doyle ML, Casper DJ, Cicala C, Leavitt SA, Majeed S, Steenbeke TD, Venturi M, Chaiken I, Fung M, Katinger H, Parren PWIH, Robinson J, Van Ryk D, Wang L, Burton DR, Freire E, Wyatt R, Sodroski J, Hendrickson WA, Arthos J 2002. HIV-1 evades antibody-mediated neutralization through conformational masking of receptor-binding sites. *Nature 420*: 678-682.

Larkin MA, Blackshields G, Brown NP, Chenna R, McGettigan PA, McWilliam H, Valentin F, Wallace IM, Wilm A, Lopez R, Thompson JD, Gibson TJ, Higgins DG 2007. CLUSTALW and CLUSTALX version 2.0. *Bioinformatics 23*: 2947-2948.

Leal E, Silva WP, Sucupira MC, Janini LM, Diaz RS 2008. Molecular and structural characterization of HIV-1 subtype B Brazilian isolates with GWGR tetramer at the tip of the V3-loop. *Virology 381*: 222-229.

Leal E, Villanova FE 2010. Diversity of HIV-1 subtype B: implications to the origin of BF recombinants. *PLoS ONE 5*: e11833.

Liu Y, Curlin ME, Diem K, Zhao H, Ghosh AK, Zhu H, Woodward AS, Maenza J, Stevens CE, Stekler J, Collier AC, Genowati I, Deng W, Zioni R, Corey L, Zhu T, Mullins JI 2008. Env length and N-linked glycosylation following transmission of human immunodeficiency virus type 1 subtype B viruses. *Virology 374*: 229-233.

Louwagie J, Delwart EL, Mullins JI, McCutchan FE, Eddy G, Burke DS 1994. Genetic analysis of HIV-1 isolates from Brazil reveals presence of two distinct genetic subtypes. *AIDS Res Hum Retroviruses 10*: 561-567.

Morgado MG, Guimarães ML, Neves Jr I, dos Santos VGV, Linhares-de-Carvalho MI, Castello-Branco LR, Bastos FI, Castilho EA, Galvão-Castro B, Bongertz V, The Hospital Evandro Chagas AIDS Clinical Research Group 1998. Molecular epidemiology of HIV in Brazil: polymorphism of the antigenically distinct HIV-1 B subtype strains. *Mem Inst Oswaldo Cruz 93*: 383-386.

Morgado MG, Sabino EC, Shpaer EG, Bongertz V, Brigido L, Guimaraes MD, Castilho EA, Galvão-Castro B, Mullins JI, Hendry RM 1994. V3 region polymorphisms in HIV-1 from Brazil: prevalence of subtype B strains divergent from North American/European prototype and detection of subtype F. *AIDS Res Hum Retroviruses 10*: 569-576.

MS - Ministério da Saúde 2012. Boletim Epidemiológico AIDS-DST. Available from: aids.gov.br/sites/default/files/anexos/publicacao/2012/52654/boletim_jornalistas_pdf_22172.pdf.

Pilotto JHS, Grinztejn B, Veloso VG, Velasque L, Friedman RK, Moreira RI, Rodrigues-Pedro A, Oliveira SM, Currier J, Morgado MG 2012. Moderate prevalence of transmitted drug-resistance mutations among antiretroviral-naïve HIV-infected pregnant women in Rio de Janeiro, Brazil. *AIDS Res Hum Retroviruses 29*: 1-6.

Pimentel VF, Morgado MG, Guimarães MDC, Castilho E, Veloso VG, Gimarães ML 2011. Temporal trends of the HIV-1 subtype B among heterosexual and bisexual men in Brazil. Proceedings of the 6th IAS Conference on HIV Pathogenesis, Treatment and Prevention, 17-20 July 2011, International AIDS Society, Rome. Available from: iasociety.org/Abstracts/A200743204.aspx.

Pinto ME, Schrago CG, Miranda AB, Russo CA 2008. A molecular study on the evolution of a subtype B variant frequently found in Brazil. *Genet Mol Res 7*: 1031-1044.

Potts KE, Kalish ML, Lott T, Orloff G, Luo CC, Bernard MA, Alves CB, Badaro R, Suleiman J, Ferreira O 1993. Genetic heterogeneity of the V3 region of the HIV-1 envelope glycoprotein in Brazil. *AIDS 7*: 1191-1197.

Santoro-Lopes G, Harrison LEEH, Tavares MD, Xexéo A, Santos CEA, Schechter M, dos Santos AC 2000. HIV disease progression and V3 serotypes in Brazil: is B different from B-Br? *AIDS Res Hum Retroviruses 16*: 953-958.

Santos AF, Sousa TM, Soares EA, Sanabani S, Martinez AM, Sprinz E, Silveira J, Sabino EC, Tanuri A, Soares MA 2006. Characterization of a new circulating recombinant form comprising HIV-1 subtypes C and B in southern Brazil. *AIDS 20*: 2011-2019.

Santos LA, Monteiro-Cunha JP, Araujo AF, Brites C, Galvao-Castro B, Alcantara LCJ 2011. Detection of distinct human immunodeficiency virus type 1 circulating recombinant forms in Northeast Brazil. *J Med Virol 83*: 2066-2072.

Shimizu N, Takeuchi Y, Naruse T, Inagaki M, Moriyama E, Gojobori T, Hoshino H 1992. Six strains of human immunodeficiency virus type 1 isolated in Japan and their molecular phylogeny. *J Mol Evol 35*: 329-336.

Soares EA, Santos RP, Pellegrini JA, Sprinz E, Tanuri A, Soares MA 2003. Epidemiologic and molecular characterization of human immunodeficiency virus type 1 in southern Brazil. *J Acquir Immune Defic Syndr 34*: 520-526.

Tamura K, Peterson D, Peterson N, Stecher G, Nei M, Kumar S 2011. MEGA5: molecular evolutionary genetics analysis using maximum likelihood, evolutionary distance, and maximum parsimony methods. *Mol Biol Evol 28*: 2731-2739.

Capítulo 5: Artigo 03

"Short-term dynamics and local epidemiological trends in the HIV-1B epidemic"

Dennis Maletich Junqueira, Rúbia Marília de Medeiros, Tiago Gräf,

Sabrina Esteves de Matos Almeida

Manuscrito a ser submetido

PLOS One

Short-term dynamics and local epidemiological trends in the HIV-1B epidemic

Dennis Maletich Junqueira[1, 2, 3], Rubia Marília de Medeiros[1, 2], Tiago Gräf[1, 4], Sabrina Esteves de Matos Almeida[1, 2, 5]

[1] Centro de Desenvolvimento Científico e Tecnológico (CDCT), Fundação Estadual de Produção e Pesquisa em Saúde (FEPPS), Porto Alegre, RS, Brazil.

[2] Programa de Pós-Graduação em Genética e Biologia Molecular, Universidade Federal do Rio Grande do Sul (UFRGS), Porto Alegre, RS, Brazil.

[3] Departamento de Ciências da Saúde, Uniritter Laureate International Universities, Porto Alegre, RS, Brazil.

[4] Programa de Pós-graduação em Biotecnologia e Biociências, Universidade Federal de Santa Catarina, Florianópolis, SC, Brazil.

[5] Instituto de Ciências da Saúde, Universidade FEEVALE, Novo Hamburgo, RS, Brazil

Corresponding Author:

Dennis Maletich Junqueira

Key words:

HIV-1B, phylogenetics, transmission clusters, transmission interval, drug resistance.

Abstract:

The human displacement and sexual behavior are the main factors driving the HIV-1 pandemic to the current profile. The intrinsic structure of the HIV transmission among different individuals has valuable importance for the understanding of the epidemic and for the public health response. The aim of this study was to characterize the HIV-1 subtype B (HIV-1B) epidemic in South America through the identification of transmission links and infer trends about geographical patterns and median time of transmission between individuals. Sequences of the protease and reverse transcriptase coding regions from 4,810 individuals were selected from GenBank. Maximum likelihood phylogenies were inferred and submitted to ClusterPicker to identify transmission links. Bayesian analyses were applied only for clusters including ≥5 dated samples in order to estimate the median maximum inter-transmission interval. This study analyzed sequences sampled from 12 South American countries, from individuals of different exposure categories, under different antiretroviral profiles, and from a wide period of time (1989-2013). Continentally, Brazil, Argentina and Venezuela were revealed important sites for the spread of HIV-1B among countries inside South America. Of note, about 70% of the HIV-1B infections are primarily occurring among individuals living in the same geographic region. In addition, these transmissions seem to occur early after the infection of an individual, taking in average 2.39 years (95% CI 1.48 - 3.30) to succeed. Homosexual/Bisexual individuals transmit the virus as quickly as almost half time of that estimated for the general population sampled here. Public health services can be broadly benefitted from this kind of information whether to focus on specific programs of response to the epidemic whether as guiding of prevention campaigns to specific risk groups.

Introduction:

The spread of HIV-1 have been molded by a conjunction of factors that reflects the clinic specificities of the viral infection and the social behavior of the human host [1–3]. In addition to the long duration of the asymptomatic stage and the high replication rate, the human sexual behavior and displacement drove the HIV-1 pandemic to the current profile [3,4]. Despite the progressive effort to control the epidemic and the recent global developments regarding the drop in the number of new infections [5], the statistics related to the mortality associated with AIDS are still worrying.

Today, nearly one million individuals are infected by HIV-1 in South America and approximately 35,000 individuals dye per year victims of the symptoms of aids [6]. A quick answer to the high mortality rate, morbidity and transmission related to HIV in certain specific regions of the South America may require a broad and efficient collection of data on the virus and the host [5]. Complementarily, application of robust analysis using this demographic and clinic information can guide intensive and focused efforts providing the maximum impact on local epidemics [7–14]. This strategic information contributes to generate an efficient epidemic surveillance and in addition to the prevention campaigns and the use of highly active antiretroviral therapy (HAART) can ensure ways to benefit public health [15,16]. In this context, phylogenetic analyses have been used to investigate the origin, spread and transmission of HIV between individuals and can be a powerful method in the understanding of the social, demographic and geographical issues in the epidemic [2,7,13,17–22].

Transmission clusters among HIV-infected individuals are identified here as phylogenetic inferences of the viral transmission between different patients. Through the genetic and evolutionary relationship these epidemiological chains allow the

inference of social contacts between individuals and provide a way to reconstruct the main transmission links involved in the spread of the virus in a certain place [13,23,24]. These findings improves the correlation of local HIV epidemic with transmission pathway, drug resistance, risk behavior and cluster size and may influence the direction focus of public campaigns to specific populations aiming to reduce the rate of transmissions and consequently delaying the increase of new infected cases [25]. While several countries in Europe, Asia, Africa, Australia and North America already have impactful information about the transmission between individuals, South American countries are clearly deficient for this kind of information [9,10,12–14,26–31]. Addressing the specific issues within local epidemics is crucial to a greatly improved in the HIV-1 epidemic [5].

Several studies have exploited the HIV-1 high rates of mutation and its intrinsic rapid evolution to understand the spread of the virus in a certain population of a specific place [17,32–34]. Here we used phylogenetic analysis to identify and characterize the transmission pairs and transmission clusters among all HIV-1 South American sequences available in Genbank. As South America has an epidemic primarily based on HIV-1 subtype B, all analysis were performed using *pol* sequences of this subtype.

Material and Methods:

Dataset Compilation:

All available HIV-1 sequences from *pol* gene including the protease and partial segment of the reverse transcriptase (nucleotides 2253–3252 relative to strain HXB2) were selected from the Los Alamos HIV Sequence Database (http://www.hiv. lanl.gov/). Only sequences from South American countries were downloaded. Additionally,

PubMed (http://www.ncbi.nlm.nih.gov/pubmed) was used to consult all studies of HIV-1 conducted in South American countries through the use of the search key "HIV-1 AND country" (where "country" was substituted for the name of each country in the South America) to select sequences deposited in GenBank (http://www.ncbi.nlm.nih.gov/nucleotide/) but not included in the Los Alamos Database. This approach guarantees that all published sequences were selected to compose our dataset. When available, geographical, demographic and clinical information were also retrieved from Los Alamos or directly from the study describing the sequence. All sequences were individually renamed and identified by a code including year and region of sampling and, in addition, the codification included patient's gender and exposure category when available. Around 200 articles were revised for patient's information.

All sequences selected were submitted to REGA Subtyping Tool v2.0 [35], and RIP 3.0 [36] to confirm the subtype. Sequences presenting discordances in the subtype assigned by these online tools were submitted to bootscanning analysis using Simplot 3.5.1 software [37]. For the final alignment we excluded sequences presenting premature stop codons, replicates from the same patient, and intersubtype recombinants.

A total of 4,810 HIV-1 Subtype B *pol* sequences from 12 South American countries (all South American countries except for French Guyana) were included in the analysis. Additionally, four sequences of subtype D were used as outgroup. Sequence alignments were generated using MUSCLE [38].

Alignments:

In order to compare the influence of convergent evolution of antiretroviral drug resistance mutations on the transmission cluster detection, two versions of the HIV-1 subtype B dataset were built: (i) a set including complete sequences (complete set,

1000bp), and (ii) a codon-stripped dataset (codon-stripped set) where the main sites associated with major antiretroviral drug resistance in protease (30, 32, 46, 47, 48, 50, 54, 76, 82, 84, 88, and 90) and reverse transcriptase (41, 65, 67, 69,70, 74, 100, 101, 103, 106, 115, 138, 151, 181, 184, 188, 190, 210, 215, 219, and 230) were excluded, resulting in a 901bp alignment [39]. All alignments are available from the authors upon request.

Identification of Transmission Clusters:

Maximum Likelihood (ML) phylogeny was inferred in RAxML [40] on the CIPRES Science Gateway [41] incorporating the best-fitted nucleotide substitution model (GTR+I+Γ) as determined in MEGA6 [42]. The approximate likelihood-ratio test (aLRT) based on Shimodaira-Hasegawa-like procedure were used to assess confidence in topology [43]. The maximum likelihood phylogenetic tree was draw using FigTree v1.4.2 [44].

Transmission pairs (including two individuals) and larger clusters (≥3 individuals) were identified by using Cluster Picker [45] with a SH-aLRT support threshold of ≥90. Different maximum pairwise genetic distances (1.0%, 1.5%, 2.0%, 2.5%, 3.0%, 3.5%, 4.0%, 4.5%, 5.0%, 5.5%, 6.0%, 6.5%, 7.0%, and 7.5%) within the clusters were evaluated to identify a cut-off value that identifies a transmission cluster inside the South America. Transmission links previously known from the literature (09 transmission pairs and 01 transmission triads) were used as controls for the genetic distance evaluation [46–48]. The transmission clusters identified by Cluster Picker were then classified in local (clusters involving sequences sampled in the same state for Brazilian sequences or in the same country for non-Brazilian sequences), interstate (clusters involving sequences sampled in different states within Brazil), or international

type (clusters involving sequences sampled in different countries) according to the sampling region of the sequences involved in the transmission cluster.

Time-scaled phylogenies:

In order to estimate the median time of HIV-1 transmission between individuals, dated phylogenies were reconstructed using a Bayesian MCMC method implemented in BEAST v1.8 [49]. This approach allows the inference of maximum estimates of inter-transmission intervals in years [13,14]. Only clusters including ≥5 dated samples were selected for further analysis. All clusters presenting these characteristics were assembled to form a new dataset (one for complete sequences and one for codon-stripped sequences).

The Bayesian analyses assumed an uncorrelated lognormal relaxed molecular clock under the GTR+I+Γ nucleotide substitution model and a Bayesian skyline coalescent tree prior. Previous estimates of the evolutionary rate for subtype B were used as normal mean prior [50]. The MCMC chain were run for 5.0×10^8 chain steps and the convergence was evaluated in TRACER v1.5 excluding an initial 10% for burn-in [51]. Maximum clade credibility trees (MCC) were summarized using TreeAnnotator v1.8.0 in BEAST package after 50% of the burn-in was discarded and the resulting tree was visualized with FigTree v.1.4.2.

Resistance Mutation Analysis:

Alignments including the complete sequences were submitted to CPR tool to detect the presence of Surveillance Drug Resistance Mutations (SDRM) [52]. Clusters that included sequences presenting the same antiretroviral drug resistance mutation were submitted to an ancestral reconstruction analysis in order to understand the transmission

of drug resistance among different individuals within a transmission group. Sequences eligible for this analysis were compiled in a new dataset which was posteriorly submitted to maximum likelihood phylogeny inference in PhyML program using an online web server and incorporating the best-fitted nucleotide substitution model (GTR+I+Γ). The resulting maximum likelihood tree were used to reconstruct the ancestral nucleotide sequence of each cluster using the FastML web server software [53].

Statistical Analysis:

Statistical comparisons between datasets were made using Pearson's χ^2-test and Fisher's exact test when appropriate. Statistical analyses were performed using WinPepi v.11.22 and the significance level was set at $P < 0.05$.

Results:

Dataset:

Among 7,600 *pol* sequences available in public databases, 4,810 HIV-1 subtype B isolates from South American countries were selected to set up a dataset for the identification of transmission pairs and transmission clusters. Twelve of the 13 South American countries were represented in the 26 different regions included in this study. Due to the large area occupied by Brazil in South America, samples isolated in this country were identified by the specific state of sampling (Figure 1).

Brazil is by far the most represented country in our dataset, exhibiting 2,826 sequences (58.7%) representing 14 geographical states (Figure 1). Due to the unavailability of the specific sampling region information in the online databases, 780

sequences from Brazil could not have an attribution of the exact state of sampling. Within Brazil, São Paulo state encompasses almost 43% of the Brazilian sequences included here. Argentina is the second best represented country with 1,490 sequences. Bolivia, Ecuador, Guyana, Paraguay, Suriname, and Uruguay are misrepresented regions with less than 10 sequences per country.

Alignments:

Traditionally, transmission clusters studies have used datasets including codon-stripped sequences from which codons associated with major resistance in protease and reverse transcriptase are removed [9,12,13,27,45]. As one of the goals of this study was to evaluate the influence and transmission of drug resistance mutations in transmission clusters, we constructed two different datasets (complete and codon-stripped datasets) to test whether the analysis of the complete set could distort the results. Despite the differences in the composition of sequences inside the transmission clusters identified, all the results obtained in this study are relatively similar for both datasets, including geographic distribution, resistance mutation and time between infections (Supplementary Material).

Within-cluster maximum genetic distance thresholds:

We first evaluated the effect of different maximum genetic distances on cluster identification among HIV-1 sequences from South America. The branch support threshold was fixed at 90 (SH-aLRT) with a within-cluster maximum genetic distance varying between 1.5% and 7.5% (using intervals of 0.5%). The number of clusters detected increased with the genetic distance threshold, reaching a maximum at 6.5% (Figure 2A and Figure S1). At genetic distances of 7.0% and 7.5% the number of

226 detected clusters was smaller than for 6.5%. Beyond the within-cluster maximum
227 genetic distance of 4.5%, the proportion of new transmission clusters detected gradually
228 decreases.

229 In the other way, the proportion of sequences included in the clusters detected
230 here enlarged with the increase in the genetic distance threshold (Figure 2B and Figure
231 S1). At the genetic distance of 7.5%, more than 80% of the sequences analyzed here
232 were included in transmission clusters. These results indicate that as the maximum
233 genetic distance cut-off is relaxed more sequences are being added to the clusters.

234

235 Identification of transmission clusters:

236 Independently, each dataset was submitted to Cluster Picker to infer the main
237 linkages between HIV-1 infected individuals in South America. In order to understand
238 the HIV-1 epidemic patterns and identify the main transmission chains, transmission
239 pairs and transmission clusters were identified using a within-cluster maximum genetic
240 distance of 4.5% for all analysis. This cut-off effectively detected all transmission links
241 included as controls in this study and was previously used in other studies [13,14,45].

242

243 HIV-1 Subtype B transmission pairs:

244 A total of 506 transmission pairs identified by a SH-aLRT support threshold of
245 ≥90 and a maximum genetic distance of 4.5% were found. Around 70% of the
246 transmissions between two individuals inside the South America occur within the same
247 geographical region (Table 1 and Table S1). As expected, transmissions including
248 individuals from different states (within Brazil) or countries were detected to a lesser
249 extent (8% and 7% respectively).

In Brazil, our results showed that the states of Goiás, São Paulo, and Rio de Janeiro seem to be important sites of interstate transmissions in absolute number of links (Table S2). These states presented the highest number of interstate links in our analysis. In addition, São Paulo, and Rio de Janeiro are the main sites within Brazil to link the epidemic with other South American countries. Together with Brazil, Argentina also plays an important role in the HIV-1 subtype B epidemic among countries inside the continent (Figure 1 and Table S2).

HIV-1 Subtype B transmission clusters (≥3 individuals):

We found 163 transmission clusters including more than 2 individuals (Table 1 and Table S1). On average, 66% of the groups including more than three individuals are reflecting local transmissions, corroborating our findings for local transmission pairs. In addition, around 15% of the clusters are linking individuals from different states, and 7% represent infections linked by international contacts.

Analysis of the interstate links within Brazil pointed São Paulo, Goiás, Rio de Janeiro, Mato Grosso, and Paraná as important sites of transmission inside the country (Figure 1 and Table S2). Within Brazil, São Paulo is an important site of HIV dissemination to other countries inside South America and in opposition to our results of transmission pairs, Rio de Janeiro had none links to countries in South America. We also found Venezuela and Argentina as important sites of transmission of HIV-1B among countries.

Only four transmission clusters linked more than seven individuals (Table 1). It is important to highlight that two of these larger clusters were associated exclusively to MSM transmission, and one included both MSM and heterosexual individuals. We

verified that approximately 41% of all clusters including more than 4 individuals were made up entirely of MSM individuals.

Dating Transmission events:

As each tip of the tree represents a single patient, the internal nodes represent the most recent common ancestors of these infections and include at least one transmission event. Therefore, the internode intervals are used to estimate the maximum inter-transmission times between patients scaled by calendar years. It is important to highlight that these time intervals are maximum estimates and the inclusion of more individuals in the cluster would lead to shorter average transmission intervals.

We determined the average period in which these clustered transmissions occurred for all groups including ≥5 individuals (95 sequences, 16 clusters) (Table 2 and Table S3). Analysis of the overall distribution of the internode intervals in South America revealed a median of 2.39 years (95% confidence interval [CI] 1.48 - 3.30). The great majority of the transmission clusters analyzed with the Bayesian approach were from Brazil allowing the determination of a country-specific average transmission interval of 3.31 years (95% CI 2.63 – 3.99).

Clusters primarily composed of MSM individuals were separately analyzed in order to calculate the average time of transmission between MSM individuals in South America. For MSM clusters, we found a median time of transmission between individuals of 0.97 years (6 clusters, 31 sequences, 95% CI 0 – 2.16).

Resistance Mutation Analysis:

The overall drug resistance mutation (DRM) prevalence of the samples included in clusters was 44.1% (720/1633; 95% CI 41.7 – 46.5; Table 3). The prevalence of

sequences presenting only drug resistance mutations against protease inhibitors (PI) was 1.53% (25/1633; 95% CI 0.94 – 2.13), against nucleoside RT inhibitors (NRTI) was 6.43% (105/1633; 95% CI 5.24 – 7.62), and against non-NRTI (NNRTI) was 3.74% (61/1633; 95% CI 2.82 – 4.66). In addition, 170 sequences (10.4%; 95% CI 8.93 – 11.89) presented mutations that conferred resistance to the three classes of antiretroviral drugs (PI, NRTI, and NNRTI), 171 sequences (10.5%; 95% CI 8.99 – 11.96) presented resistance mutations against PI and NRTI, 10 sequences (0.6%; 95% CI 0.23 – 0.99) presented PI and NNRTI-resistance mutations, and 178 sequences (10.9%; 95% CI 9.39 – 12.41) were identified as carrying resistance mutations against NRTI and NNRTI (Table 3). Within protease, L90M (17.5%) was the most frequent mutation, followed by M46I (13%), I54V (9.1%), V82A (9%), and I84V (9%) (Table S4). The revertants at RT position 184 (M184V) were the most prevalent (18%), followed by M41L (15%), T215Y (13%), L210W (9.7%), and D67N (9.6%) (Table S5 and Table S6). We found 400 clusters presenting samples with DRM.

We also analyzed the prevalence of DRM in the set of sequences not included in any cluster in order to understand the influence of the clustering in the transmission of drug resistance mutations. From 3,177 sequences not included in transmission clusters, 2,287 samples (72%; 95% CI 70.42 – 73.55) were identified as carrying drug resistance mutations (Table 3). Due to the high prevalence of resistance mutation in this set, we analyzed all sequences by searching for the patient antiretroviral therapy status directly from the study describing each sequence. The great majority of the sequences not included in any cluster (1,537 samples, 48.4%) could not have its antiretroviral status attributed (Table S7). The remaining 1,640 sequences were categorized in three groups: (i) sequences sampled from patients failing antiretroviral therapy (1271 samples, 40%), (ii) sequences sampled from treatment-naive patients (236 samples, 7.4%), and (iii)

sequences sampled from antiretroviral treated individuals (133 samples, 4.2%; Table S7). Statistical comparison revealed that clustered sequences included more samples from treatment-naive patients and fewer samples from patients failing antiretroviral therapy (P<0,001) than no clustered sequences. Analysis of the most frequent DRM in protease and reverse transcriptase genes revealed a similar pattern between the sequences not included in any cluster and those included in clusters (Tables S4, S5, and S6).

In addition, by maximum likelihood reconstruction we examined the ancestral states at all sites associated with drug resistance for 28 clusters including more than 3 individuals (129 total individuals) (Table S8). In all cases the reconstructed ancestral sequence of the transmission cluster harbored the same drug resistance mutation as the sequences within the cluster. However, the selection of the clusters to evaluate the ancestral amino acids may create a bias in this analysis since we only chose clusters that included more than three sequences presenting the same resistance mutation.

Discussion:

The dataset used here is the best possible approximation of the current HIV-1 epidemic in South America and reflects the great effort to select all available sequences from public databases. Here we analyzed sequences sampled from 12 South American countries, from individuals of different exposure categories, under different antiretroviral profiles, and from a wide period of time (1989-2013) seeming to be one of the first studies to mix so many samples from different patients and countries [54]. The heterogeneity of our sequences brings a new challenge to the identification of transmission clusters due to the difficulty in the definition and interpretation of each cluster, but more consistently explains the general epidemic of HIV-1 subtype B (HIV-

1B). In this sense, this study brings an innovative objective on the transmission dynamic of HIV-1B and adds another piece in the understanding of the epidemic in South America.

Traditionally, transmission clusters studies have analyzed *pol* gene to infer the dissemination of the virus in a certain population by stripping the codons associated to resistance mutation in order to remove the influence of convergent evolution [13,14,45,55]. Here, we tested the influence of the dataset on the identification of transmission clusters including or excluding the codons associated to resistance mutation and found no statistically significant difference between the number of transmission pairs and transmission clusters identified for both datasets. In adittion, we found similar results for both datasets in all analyses conducted here, including geographic distribution, resistance mutations and time interval between infections supporting the use of the complete dataset (Tables 1, 2, 3, S1, S2, and S3).

Due to the inexistence of an established consensus methodology to identify transmission clusters, we also tested the effect of different maximum genetic distances on cluster identification among HIV-1 samples from South America fixing the branch support threshold at 90 (SH-aLRT). We observed that beyond 6.5% the number of detected clusters decreased when compared to more restrictive genetic distances (Figure 2A). In the same way, from the within-cluster maximum genetic distance of 4.5% the proportion of new transmission clusters identified by Cluster Picker gradually decreases. This result is expected since the inclusion of more sequences in each cluster and/or the combination of smaller ones will allow the detection of increased size clusters. Based on this result we judged the within-cluster maximum genetic distance of 4.5% to be appropriate in detecting transmission clusters. This is not a permissive cut-off that will identify every sequences of a clade with a branch support threshold >90 as a

transmission cluster, nor a very rigorous trait that will only group very closely related sequences. In addition, previous studies have used this cut-off for the identification of transmission links among patients living in the same country [56].

At a genetic distance threshold of 4.5%, more than 500 individuals had a link to one other subject (Table 1). It is important to note that the vast majority of these infections (~70%) included individuals living in the same geographic region. Transmission clusters involving 3 or more sequences corroborated the great importance occupied by these local links within the South America epidemic and suggests that HIV-1B infections primarily occur among individuals living in the same geographic region despite the current mobility of the human populations throughout the world. This estimative represents the minimum proportion since several sequences in our dataset presented missing data for geographic origin. If extrapolated, this result may indicate that in average 70% of the epidemic from a particular location in South America can be explained only by local epidemiological trends. The remaining portion represents interstate or international transmissions (Table 1 and Table S1). We also observed a tendency relating geographical distance to epidemiological connection and it seems to be more likely the linking of individuals from closest states or countries. These estimates might be a reflection of the poor infrastructure integration relative to accessibility to and among urban centers in the South America. The isolation of countries and cities in South America, therefore might be viewed as potential factors driving the strength of the local trends in the HIV-1B epidemic.

In relation to the Brazilian interstate links, the geographic distribution revealed that Goiás, São Paulo, Rio de Janeiro, Paraná, and Mato Grosso are important sites of transmission within this country (Figure 1 and Table S3). Demographic data from the Official Brazilian Demographic Data Center (IBGE) shows that these states receive the

399 highest number of internal migrants in the country supporting the idea that migratory
400 relationship between different states creates potential networks to the diffusion of HIV
401 in the human population (IBGE, 2010). Moreover, São Paulo and Rio de Janeiro are
402 important sites for an international connection of the HIV-1B epidemic within Brazil
403 (Figure 1). Demographic data can support our results since São Paulo and Rio de
404 Janeiro are among the states in Brazil to receive the highest number of immigrants from
405 other countries (IBGE, 2010).

406 Our results revealed that Brazil, Argentina and Venezuela are important sites for
407 the spread of HIV-1B within South America. Migratory patterns in the continent
408 revealed that these three countries are important points of attraction of significant
409 numbers of regional migration (IOM, 2014). Supposedly the spread of HIV-1 in South
410 America may be associated to the migratory pathways established by the human
411 population in search of economic or social change (IBGE, 2010). These results highlight
412 the importance of association between demographic data and epidemiological
413 information to construct a full scenario of the current pandemic in a certain region [57].

414 We also analyzed the dynamics of dissemination within clusters to infer the time
415 between transmissions from an individual to another. Bayesian analysis revealed an
416 average time period of HIV-1B transmissions within South America of 2.39 years (95%
417 confidence interval [CI] 1.48 - 3.30) (Table 2). Our results propose a short-term
418 dynamic of the epidemic among HIV infected individuals in which after approximately
419 2 years being infected patients will transmit the virus to other individuals. A previous
420 study analyzing heterosexual transmission in a group of 11,071 individuals from the
421 United Kingdom epidemic of HIV-1 non-B subtypes found similar results [14]. The
422 agreement with our findings may indicate a similar pattern in the transmission of HIV-1
423 irrespective of the population analyzed. This information is of utmost importance to

understand the dynamics of infection and suggest that the early stage of infection potentially influence the onward spread of HIV [58,59].

The identification of clusters including only MSM individuals in our results allowed the estimation of an average time of transmission in this group of 0.97 years. We observed that MSM individuals transmit the virus to another subject in almost half time of that for the general population sampled here. The short time interval between transmission can be explained by the stage of infection since during the first year of HIV MSM individuals infection are eight times as infectious as during chronic phase [58]. Additionally, we encountered a larger number of transmission clusters (41%) including more than four individuals comprising only MSM individuals. It suggests a difference in the dynamics of the epidemics in different risk groups most likely reflecting social and sexual behavior of the human population [14]. These results should be taken as a warning data to the public health services since the transmission of the virus within the group is faster than for others groups and that in most countries the population of MSM individuals is still the most affected by HIV infection [60,61].

The elevated proportion (62%) of sequences presenting drug resistance mutation (DRM) in our dataset led us to investigate the antiretroviral therapy status of each patient included here (Table 3 and Table S7). The vast majority of sequences assigned for a therapy status were sampled from patients failing antiretroviral therapy (37%) which might explain the increased prevalence of drug resistance mutations in our sample [62–64]. When stratifying the data, we found lower prevalence of DRM in the set of sequences included in transmission clusters (Table 3). This result is consistent with the decreased replication fitness of viruses harboring DRM [22,54,65]. In addition, statistical comparisons based on Pearson's $\chi 2$-test revealed an increased number of sequences from treatment-naive patients included in the linked transmissions suggesting

449 that naive individuals might be transmitting the virus in higher rates than treated or
450 failing antiretroviral individuals. The reasons for this pattern may be related to
451 ignorance about HIV status and the high viral load characteristic of these untreated
452 patients.

453 The ancestral reconstructions within transmission clusters revealed a tendency to
454 the circulation of transmitted drug resistance mutations in South America since we
455 found all inferred ancestral sequences harboring the most prevalent mutation within
456 each cluster (Table S8). In addition, among the resistance mutations found in this study,
457 M184V (found in 38% of the sequences) and M41L (found in 28.5% of the sequences)
458 were the most frequent mutations found in our samples (Tables S1, S2, and S3). Both
459 mutations reflect resistance to NNRTI drugs e are described as the most common in
460 patients failing antiretroviral therapy [66].

461 Our results show that the HIV-1B epidemic is largely influenced by regional
462 trends and suggest a higher probability of HIV transmission between individuals from
463 the same geographic origin. This compartmentalized analysis of the epidemic (by states
464 or countries) and the consequent influence of local, interstate or international
465 transmissions had never been defined for the HIV-1 epidemic in South America. In
466 addition, we propose a short-term dynamics within clusters in which the mean time of
467 transmissions is two years. Information regarding the regional dynamic of HIV
468 infections allied to statistical methods may contribute to the understanding on
469 dissemination and expansion of new variants in certain places, including the entry of
470 new subtypes and the circulation of drug resistant strains. Despite the great sampling
471 effort our results are a direct reflection of the sequence dataset and to some extent are
472 limiting due to lack of patient information and even the lack of genetic information
473 about the HIV-1 epidemic in some regions. However, the perfect sampling is not

required to understand the general processes responsible for the spread of the virus within a certain region [22,67]. Therefore, investments in data collection on the virus and the host are important factors that can achieve better inferences about the spread and the patterns driving HIV epidemics. Public health services can be largely benefited using this strategic information to focus on more precisely and effective HIV programmes that can reach a great number of individuals and can help to improve the understanding of the epidemic and the response.

Figures:

Code	State	N°
AM	Amazonas	23
BA	Bahia	13
ES	Espírito Santo	56
GO	Goiás	176
MT	Mato Grosso	59
MS	Mato Grosso do Sul	43
MG	Minas Gerais	27
PR	Paraná	30
PE	Pernambuco	10
RJ	Rio de Janeiro	220
RS	Rio Grande do Sul	102
SC	Santa Catarina	36
SP	São Paulo	1208
TO	Tocantins	43
-	Unidentified	780

Figure 1. Geographic distribution and proportion of HIV-1 subtype B *pol* sequences from South American countries. Map shows locations from where HIV-1 subtype B sequences were included in the current dataset. The proportion (green bars) and the total number of sequences analyzed from each location is indicated. A compilation of all sequences included in the Brazilian set and its respective state of sampling is indicated at the table included in the figure. Black dots indicate the cities where sequences representing the genetic variability of each country could have its sampling region identified. Sequences from Venezuela were sampled at Caracas

(n=213), from Colombia were sampled at Medellín (n=32) and Bogotá (n=7), the unique sequence from Bolivia was sampled at La Paz, sequences from Argentina were sampled at Mendoza (n=2) and Buenos Aires (n=1238), and sequences from Uruguay were sampled at Montevideo. For the rest of the countries the sequences had no identification of sampling region. Gray-shaded areas indicate regions not included in this study.

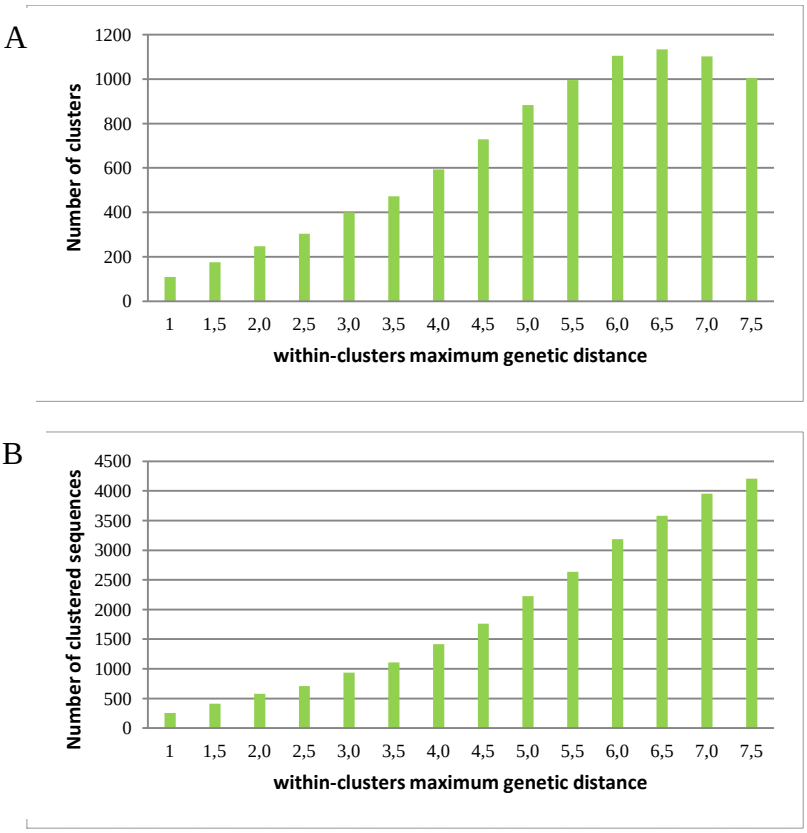


Figure 2. Number of transmission clusters and clustered sequences among 4,810 HIV-1 Subtype B codon-stripped *pol* sequences from South America. (A) Absolute number of transmission clusters identified using Cluster Picker with a SH-aLRT support threshold of ≥90 and under different within-maximum genetic distances. (B) Absolute number of clustered sequences under different within-cluster genetic distances.

Tables:

Table 1. Geographical type of HIV-1 Subtype B transmissions among clusters identified within South America for the codon-stripped dataset (901bp).

Clustered Individuals	Geographical Type of Transmission	Number of Clusters Identified	%
2	Local Transmission*	392	70.6
	Interstate Transmission (Brazil)	44	7.93
	International Transmission	31	5.59
	Unidentified	88	15.9
3	Local Transmission*	81	68.6
	Interstate Transmission (Brazil)	13	11
	International Transmission	6	5.08
	Unidentified	18	15.3
4	Local Transmission*	28	73.7
	Interstate Transmission (Brazil)	4	10.5
	International Transmission	2	5.26
	Unidentified	4	10.5
5	Local Transmission*	9	81.8
	Interstate Transmission (Brazil)	1	9.09
	International Transmission	1	9.09
6	Local Transmission*	1	100
7	Local Transmission*	1	50
	Interstate Transmission (Brazil)	1	50
≥8	Local Transmission*	4	100
	Total	729	-

* Transmission Clusters involving sequences sampled in the same state for Brazilian sequences or in the same country for non-Brazilian sequences

559 **Table 2.** Average time of HIV-1 subtype B transmission among South American
560 individuals for the codon-stripped dataset (901bp).

Cluster	Codon-Stripped Set			
	Number of Taxons	**Geographical Type**	**Exposure Category**	**Median Internal Branchs (years)**
1	4	Local	-	6,410
2	7	Interstate	HET/MSM	4.468
3	5	Local	MSM	0.411
4	5	Local	MSM	0.302
5	4	Local	HET	2,247
6	6	Local	MSM	0.394
7	5	Local	MSM	1,623
8	5	International	-	1,117
9	5	Local	-	1,540
10	5	Local	MSM	0.757
11	13	Local	MSM	1,596
12	6	Local	-	3,361
13	5	Local	MSM	0.517
14	7	Local	HET/MSM	2,970
				Median 1.98 (95% CI 1.04 - 2.92)

561 Abbreviations: HET: Heterosexual individual, MSM: men who have sex with men individual

562

563

564

565

566

567

568

569

570

571

572

Table 3. Drug resistance mutations identified among 4,810 sequences clustered or not clustered in transmission clusters within South America.

Drug Resistance Mutation (DRM)	Number of Sequences presenting DRM (n= 4810)	Clustered Sequences (n= 1633)		Not Clustered Sequences (n= 3177)	
		N	%	N	%
PI	64	25	1.53	39	1.23
NRTI	346	105	6.43	241	7.59
NNRTI	180	61	3.74	119	3.75
PI + NRTI	751	171	10.5	580	18.3
PI + NNRTI	31	10	0.61	21	0.66
NRTI + NNRTI	788	178	10.9	610	19.2
PI + NRTI + NNRTI	847	170	10.4	677	21.3
Total	**3007**	**720**	**44.1**	**2287**	**72.0**

Abbreviations: NRTI: nucleoside reverse transcriptase inhibitor, NNRTI: non- nucleoside reverse transcriptase inhibitor, PI: protease inhibitor

Supplementary Material:

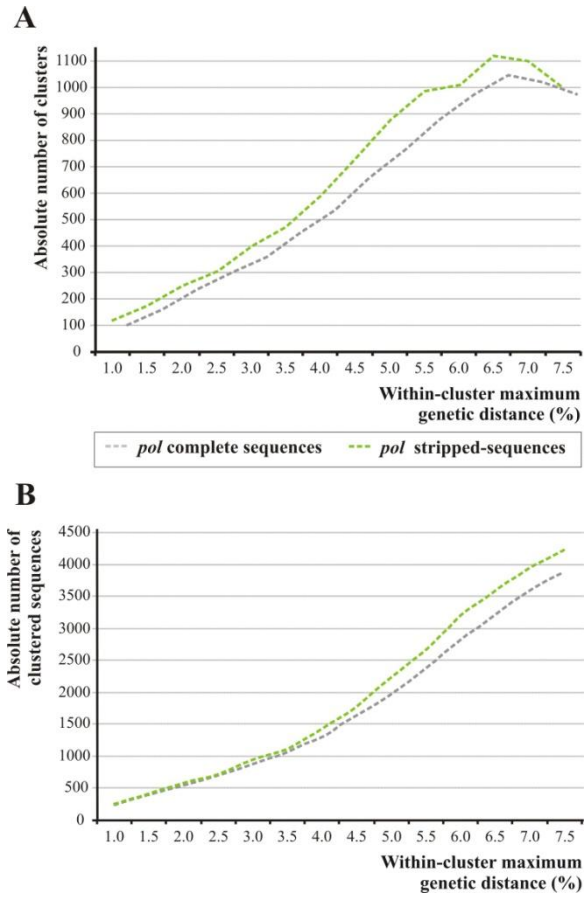


Figure S1. Number of transmission clusters and clustered sequences for the complete and codon-stripped datasets among 4,810 HIV-1 Subtype B *pol* sequences from South America. (A) Absolute number of transmission clusters identified with a SH-aLRT support threshold of ≥90. (B) Absolute number of clustered sequences under different within-cluster genetic distances.

Table S1. Geographical type of HIV-1 Subtype B transmissions among clusters identified within South America for the complete dataset (1000bp).

Clustered Individuals	Geographical Type of Transmission	Number of Clusters Identified	%
2	Local Transmission**	353	69.8
	Interstate Transmission (Brazil)	40*	7.91
	International Transmission	34	6.72
	Unidentified	79	15.6
3	Local Transmission**	69	66.3
	Interstate Transmission (Brazil)	13	12.5
	International Transmission	7	6.73
	Unidentified	15	14.4
4	Local Transmission**	23	60.5
	Interstate Transmission (Brazil)	9	23.7
	International Transmission	3	7.89
	Unidentified	3	7.89
5	Local Transmission**	11	78.6
	Interstate Transmission (Brazil)	1	7.14
	International Transmission	2	14.3
6	Local Transmission**	1	100
7	Local Transmission**	-	-
	Interstate Transmission (Brazil)	2	100
≥8	Local Transmission**	4	100
	Total	669	-

* Higher number of insterstate links within transmission pairs than within transmission clusters (including more than two individuals) (P=0.005).

** Transmission Clusters involving sequences sampled in the same state for Brazilian sequences or in the same country for non-Brazilian sequences

Table S2. Number of links identified for the states within Brazil involved in interstate transmissions and for the countries involved in international transmissions at South America (including results from transmission pairs and transmission clusters).

Geographical Type of Transmission	Region	Number of Links with other Sequences	
		Complete *pol* Sequences	Codon-stripped *pol* Sequences
Interstate	Amazonas (AM)	-	1
	Espírito Santo (ES)	4	2
	Goiás (GO)	8	5
	Minas Gerais (MG)	1	-
	Mato Grosso do Sul (MS)	2	2
	Mato Grosso (MT)	10	3
	Paraná (PR)	6	5
	Rio de Janeiro (RJ)	8	3
	Rio Grande do Sul (RS)	2	1
	Santa Catarina (SC)	2	-
	São Paulo (SP)	12	8
	Tocantins (TO)	3	2
	Unidentified	2	2
International	Argentina	5	6
	Brazil	7	4
	Chile	2	2
	Guiana	1	1
	Peru	2	2
	Paraguai	1	-
	Uruguai	2	1
	Venezuela	2	3
	Unidentified	-	-

Table S3. Average time of HIV-1 subtype B transmission among South American individuals for the complete dataset (1000bp).

Cluster	Complete Set			
	Number of Taxons	Geographical Type	Exposure Category	Median Internal Branchs (years)
1	5	Local	MSM	0.355
2	5	Local	MSM	0.468
3	4	Local	-	3,615
4	5	Local	-	2,350
5	5	International	-	6,613
6	5	Local	MSM	3,987
7	5	Local	MSM	0.273
8	4	International	-	0.802
9	5	Local	HET	2,355
10	5	Local	-	2,617
11	5	Local	MSM	0.239
12	6	Local	MSM	0.494
13	7	Interstate	-	4,035
14	7	Interstate	HET/MSM	4,161
15	9	Local	HET/MSM	2,973
16	13	Local	-	2,891
				Median 2.39 (95% CI 1.48 - 3.30)

Abbreviations: HET: Heterosexual individual, MSM: men who have sex with men individual

Table S4. Amino acid substitutions in HIV-1 Protease gene related to drug resistance to protease inhibitors (PI) identified among 4,810 sequences clustered or not clustered in transmission clusters within South America.

PI Major Mutation	Full Dataset (n= 4,810)		Clustered Sequences (n= 1,633)		Not Clustered Sequences (n= 3,177)	
	N	%	N	%	N	%
L90M	926	17.1	214	17.5	712	17.0
M46I	647	11.9	159	13.0	488	11.6
I54V	592	10.9	112	9.14	480	11.4
V82A	556	10.3	109	8.89	447	10.7
I84V	335	6.18	111	9.05	224	5.34
D30N	250	4.61	36	2.94	214	5.10
N88D	233	4.30	36	2.94	197	4.69
G73S	226	4.17	60	4.89	166	3.96
M46L	208	3.84	50	4.08	158	3.77
I85V	160	2.95	42	3.43	118	2.81
F53L	145	2.67	29	2.37	116	2.76
L24I	139	2.56	28	2.28	111	2.65
V32I	138	2.55	39	3.18	99	2.36
I47V	107	1.97	32	2.61	75	1.79
L76V	69	1.27	17	1.39	52	1.24
I54L	62	1.14	12	0.98	50	1.19
G73T	62	1.14	13	1.06	49	1.17
V82T	50	0.92	11	0.90	39	0.93
V82F	50	0.92	8	0.65	42	1.00
I50L	50	0.92	3	0.24	47	1.12
G48V	50	0.92	13	1.06	37	0.88
I54M	47	0.87	22	1.79	25	0.60
N88S	40	0.74	11	0.90	29	0.69
N83D	34	0.63	20	1.63	14	0.33
I50V	28	0.52	3	0.24	25	0.60
I54A	26	0.48	3	0.24	23	0.55
G73C	25	0.46	2	0.16	23	0.55
L23I	24	0.44	1	0.08	23	0.55
V82S	19	0.35	6	0.49	13	0.31
V82C	15	0.28	4	0.33	11	0.26
G73A	11	0.20	3	0.24	8	0.19
V82M	9	0.17	-	-	9	0.21
M46IL	9	0.17	1	0.08	8	0.19
I47A	9	0.17	-	-	9	0.21

F53Y	9	0.17	2	0.16	7	0.17
V82L	8	0.15	2	0.16	6	0.14
V82AT	8	0.15	3	0.24	5	0.12
I54T	8	0.15	3	0.24	5	0.12
I54S	7	0.13	-	-	7	0.17
G48M	7	0.13	2	0.16	5	0.12
G73AST	5	0.09	-	-	5	0.12
G73ST	4	0.07	1	0.08	3	0.07
I54AV	3	0.06	1	0.08	2	0.05
G73AT	3	0.06	-	-	3	0.07
I54MV	2	0.04	1	0.08	1	0.02
I54LV	2	0.04	-	-	2	0.05
V82AS	1	0.02	-	-	1	0.02
I84A	1	0.02	-	-	1	0.02
I54ST	1	0.02	-	-	1	0.02
I54LM	1	0.02	1	0.08	-	-
I54AS	1	0.02	-	-	1	0.02
Total	**5422**	**-**	**1226**	**-**	**4196**	**-**

Table S5. Amino acid substitutions in HIV-1 Reverse transcriptase gene related to drug resistance to nucleoside reverse transcriptase inhibitors (NRTI) identified among 4,810 sequences clustered or not clustered in transmission clusters within South America.

NRTI Major Mutation	Full Dataset (n=4,810)		Clustered Sequences (n= 1,633)		Not Clustered Sequences (n= 3,177)	
	N	%	N	%	N	%
D67G	79	0.79	18	0.87	61	0.76
M184V	1831	18.2	376	18.2	1455	18.2
L74V	293	2.92	67	3.25	226	2.83
M41L	1370	13.7	306	14.8	1064	13.3
D67N	1071	10.7	197	9.56	874	11.0
T215S	28	0.28	9	0.44	19	0.24
F116Y	98	0.98	23	1.12	75	0.94
K65R	60	0.60	11	0.53	49	0.61
V75M	140	1.40	21	1.02	119	1.49
K70R	698	6.96	118	5.73	580	7.27
T215Y	1229	12.2	268	13.0	961	12.1
T215F	342	3.41	61	2.96	281	3.52
L74I	142	1.42	20	0.97	122	1.53
T69D	209	2.08	53	2.57	156	1.96
T215D	23	0.23	5	0.24	18	0.23
F77L	101	1.01	23	1.12	78	0.98
K219Q	354	3.53	57	2.77	297	3.72
K219N	134	1.34	16	0.78	118	1.48
L210W	899	8.96	200	9.70	699	8.77
Y115F	73	0.73	28	1.36	45	0.56
D67EG	1	0.01	1	0.05	-	-
V75T	43	0.43	8	0.39	35	0.44
V75S	8	0.08	1	0.05	7	0.09
T215SY	35	0.35	5	0.24	30	0.38
K70E	11	0.11	2	0.10	9	0.11
K219E	298	2.97	53	2.57	245	3.07
K219R	87	0.87	31	1.50	56	0.70
D67E	13	0.13	4	0.19	9	0.11
T215E	2	0.02	-	-	2	0.03
V75A	27	0.27	5	0.24	22	0.28
M184IV	8	0.08	1	0.05	7	0.09
T215I	41	0.41	9	0.44	32	0.40
L74IV	14	0.14	6	0.29	8	0.10
T215C	22	0.22	2	0.10	20	0.25

M184I	25	0.25	6	0.29	19	0.24
Q151M	124	1.24	29	1.41	95	1.19
T215DY	2	0.02	1	0.05	1	0.01
T215FY	26	0.26	5	0.24	21	0.26
T215V	17	0.17	4	0.19	13	0.16
T215CF	2	0.02	1	0.05	1	0.01
T215FV	6	0.06	2	0.10	4	0.05
T215FS	4	0.04	-	-	4	0.05
K219QR	2	0.02	1	0.05	1	0.01
K219EQ	10	0.10	-	-	10	0.13
V75AT	7	0.07	1	0.05	6	0.08
T215CY	8	0.08	4	0.19	4	0.05
T215FIS	11	0.11	2	0.10	9	0.11
T215FI	6	0.06	-	-	6	0.08
T215IV	1	0.01	-	-	1	0.01
Total	**10035**	**-**	**2061**	**-**	**7974**	**-**

673 **Table S6.** Amino acid substitutions in HIV-1 reverse transcriptase gene related to drug
674 resistance to non-nucleoside reverse transcriptase inhibitors (NNRTI) identified among
675 4,810 sequences clustered or not clustered in transmission clusters within South
676 America.

NNRTI Major Mutation	Full Dataset (n= 4,810)		Clustered Sequences (n= 1,633)		Not Clustered Sequences (n= 3,177)	
	N	%	N	%	N	%
190A	13	0.42	-	-	13	0.54
190S	2	0.07	-	-	2	0.08
225H	3	0.10	-	-	3	0.12
230L	1	0.03	-	-	1	0.04
G190A	450	14.7	105	16.0	345	14.3
G190AS	2	0.07	-	-	2	0.08
G190E	13	0.42	5	0.76	8	0.33
G190S	75	2.44	16	2.44	59	2.45
K101E	205	6.68	37	5.63	168	6.96
K101P	57	1.86	11	1.67	46	1.91
K103N	1095	35.7	244	37.1	851	35.3
K103NS	18	0.59	3	0.46	15	0.62
K103S	58	1.89	8	1.22	50	2.07
L100I	165	5.37	30	4.57	135	5.59
M230L	32	1.04	3	0.46	29	1.20
P225H	146	4.76	25	3.81	121	5.01
V106A	47	1.53	7	1.07	40	1.66
V106M	31	1.01	6	0.91	25	1.04
V179F	4	0.13	-	-	4	0.17
Y181C	482	15.7	115	17.5	367	15.2
Y181I	21	0.68	7	1.07	14	0.58
Y181V	14	0.46	3	0.46	11	0.46
Y188C	5	0.16	-	-	5	0.21
Y188H	7	0.23	2	0.30	5	0.21
Y188HL	7	0.23	2	0.30	5	0.21
Y188L	117	3.81	28	4.26	89	3.69
Total	**3070**	**-**	**657**	**-**	**2413**	**-**

677

678

Table S7. Antiretroviral therapy (ART) status of the patients and clustering behavior of the respective sequences.

Patients' Antiretroviral Therapy Status	Full Dataset (n=4,810)		Clustered Sequences (n=1,633)		Not Clustered Sequences (n=3,177)	
	N	%	N	%	N	%
Naive	511	10.6	275*	16.8	236	7.4
Treated	221	4.6	88	5.4	133	4.2
Failing ART	1777	36.9	506*	31	1271	40
Unidentified	2301	47.8	764	46.8	1537	48.4
Total	**4810**	**-**	**1633**	**-**	**3177**	**-**

*P<0.001

695 **Table S8.** Ancestral reconstruction analysis of the sequences included in clusters and presenting resistance drug mutations.

Cluster	Number of individuals	Region	Type	Resistance Mutations Analyzed			Ancestral Reconstruction*	
				PI	NRTI	NNRTI	PT	RT
33	3	Chile	Local	-	-	K103N	-	K103N
267	4	Venezuela	Local	M46I, I84V, L90M	D67N, T215F, K219Q	K101P, K103N	M46I, I84V, L90M	D67N, T215F, K219Q
305	5	Argentina	Local	-	M184V	K103NS	-	K103N
380	3	Argentina	Local	-	D67N, L74IV, M184V, L210W, T215Y	Y181C	-	D67N, Y181C, M184V, L210W, T215Y
445	4	Argentina	Local	-	M41L, T215F	-	-	M41L, T215F
469	24	Argentina	Local	M46I, I84V, L90M	M41L, D67N, M184V, L210W, T215Y	Y188L	M46I, I84V, L90M	M41L, D67N
476	4	São Paulo	Local	-	-	K103N	-	K103N
507	3	Argentina	Local	V32I, M46I, I47V, I54M, V82A, L90M	K65R, K70R, Q151M, K219E	Y181C	V32I, M46I, I47V, I54M, V82A, L90M	K65R, K70R, Q151M, Y181C, K219E
508	3	Venezuela	Local	I84V, L90M	Y115F, Q151M, M184V	-	I84V, L90M	Y115F, Q151M, M184V
510	3	Argentina	Local	L90M	M41L, L210W, T215Y	G190A	L90M	M41L, L90M, L210W, T215Y
511	3	Argentina	Local	M46L, V82A, L90M	M41L, L210W, T215Y	-	M46L, V82A, L90M	M41L, L210W, T215Y
526	4	Argentina	Local	L90M	-	K103N	-	K103N
552	3	Argentina	Local	-	M41L, D67N, T69D, K70R, L210W, T215Y, K219Q	-	-	M41L, D67N, T69D, K70R, L210W, T215Y, K219Q
560	3	Argentina	Local	M46I, L90M	T215Y	-	M46I, L90M	T215Y
575	3	Argentina	Local	-	T215SY	-	-	T215Y

578	3	Argentina	Local	M46L, I54V, V82A	M41L, D67N, T215Y	K103N, G190A	M46L, I54V, V82A	M41L, D67N, K103S, G190A, T215Y
632	3	Argentina	Local	G48V, I54V, V82A, L90M	M41L, T215FY	-	G48V, I54V, V82A, L90M	M41L, T215FY
642	3	Argentina	Local	M46L, L90M	M41L, L74V, L210W, T215Y	-	M46L, L90M	M41L, L74V, L210W, T215Y
643	20	Argentina	Local	M46I, I84V	M41L, T69D, M184V, T215Y, K219R	K103N	-	M41L, T69D, T215Y
648	3	Argentina	Local	I84V, L90M	M41L, D67G, T69D, L210W, T215Y, K219R	Y181C	I84V, L90M	M41L, D67G, T69D, Y181C, L210W, T215Y, K219R
653	3	Argentina	Local	M46I, I54V, V82A, L90M	M41L, L210W, T215Y	K103N	M46I, I54V, L90M	M41L, K103N, L210W, T215Y
658	3	Argentina	Local	M46L, G48V, I50V, I54V, V82A	M41L, K70R, T215FY, K219E	Y181C, G190A	M46L, G48V, I50V, I54V, V82A	M41L, K70R, T215F, K219E
665	3	Argentina	Local	M46I, I47V, I54V, I84V, L90M	L74I, T215Y	K103N	M46I, I47V, I54V, I84V, L90M	L74I, K103N, T215Y
666	3	Argentina	Local	M46L, I47V, I54M, L76V, I84V, I85V, L90M	M41L, D67N, T69D, L210W, T215Y, K219N	L100I, K103N	M46L, I47V, I54M, L76V, I84V, I85V, L90M	M41L, D67N, T69D, L100I, K103N, L210W, T215Y, K219N
668	3	Argentina	Local	M46L, I54V, V82A, L90M	M41L, D67N, L74V, L210W, T215Y, K219R	101E, Y181C, G190A	M46L, I54V, V82A, L90M	M41L, D67N, L74V, 101E, Y181C, G190A, L210W, T215Y, K219R
156	3	Argentina	Local	-	M41L, D67N, K70R, T215F, K219E	-	-	M41L, D67N, K70R, T215F, K219E
184	3	Brasil	Interestate	D30N, N88D	-	-	D30N, N88D	-

206	4	Argentina	Local	V32I, M46I, V82A	D67G, T69D, K70R, M184V, T215F, K219Q	-	V32I, M46L, V82A	D67G, T69D, K70R, M184V, T215F, K219Q

* Drug resistance mutations harbored by the reconstructed ancestral sequence which were detected in the sequences included in the cluster.

References:

1. Kuiken C, Thakallapalli R, Esklid a, de Ronde a, Eskild A, Ronde A De. Genetic analysis reveals epidemiologic patterns in the spread of human immunodeficiency virus. Am J Epidemiol. 2000;152: 814–22.

2. Faria NR, Rambaut A, Suchard MA, Baele G, Bedford T, Ward MJ, et al. The early spread and epidemic ignition of HIV-1 in human populations. Science. 2014;346: 56–61.

3. Rambaut A, Posada D, Crandall KA, Holmes EC. The causes and consequences of HIV evolution. Nat Rev Genet. 2004;5: 52–61.

4. Duffy S, Shackelton L a, Holmes EC. Rates of evolutionary change in viruses: patterns and determinants. Nat Rev Genet. 2008;9: 267–76.

5. UNAIDS. Local Epidemics Issues Brief [Internet]. 2014. Available: http://www.unaids.org/en/resources/documents/2014/20140707_JC2559_local-epidemics

6. UNAIDS. Report on the global AIDS epidemic [Internet]. 2013. Available: http://www.unaids.org/en/media/unaids/contentassets/documents/epidemiology/2013/gr2013/UNAIDS_Global_Report_2013_en.pdf

7. Rothenberg RB, Potterat JJ, Woodhouse DE, Muth SQ, Darrow WW, Klovdahl a S. Social network dynamics and HIV transmission. AIDS. 1998;12: 1529–36.

8. Salazar-Gonzalez JF, Bailes E, Pham KT, Salazar MG, Guffey MB, Keele BF, et al. Deciphering human immunodeficiency virus type 1 transmission and early envelope diversification by single-genome amplification and sequencing. J Virol. 2008;82: 3952–70.

9. Bezemer D, Faria NR, Hassan A, Hamers RL, Mutua G, Anzala O, et al. HIV Type 1 transmission networks among men having sex with men and heterosexuals in Kenya. AIDS Res Hum Retroviruses. 2014;30: 118–26.

10. Yerly S, Junier T, Gayet-Ageron A, Amari EB El, von Wyl V, Günthard HF, et al. The impact of transmission clusters on primary drug resistance in newly diagnosed HIV-1 infection. AIDS. 2009;23: 1415–23.

11. Rieder P, Joos B, von Wyl V, Kuster H, Grube C, Leemann C, et al. HIV-1 transmission after cessation of early antiretroviral therapy among men having sex with men. AIDS. 2010;24: 1177–83.

12. Chalmet K, Staelens D, Blot S, Dinakis S, Pelgrom J, Plum J, et al. Epidemiological study of phylogenetic transmission clusters in a local HIV-1 epidemic reveals distinct differences between subtype B and non-B infections. BMC Infect Dis. 2010;10: 262.

13. Lewis F, Hughes GJ, Rambaut A, Pozniak A, Leigh Brown AJ, Brown AJL. Episodic sexual transmission of HIV revealed by molecular phylodynamics. PLoS Med. 2008;5: e50.

14. Hughes GJ, Fearnhill E, Dunn D, Lycett SJ, Rambaut A, Leigh Brown AJ. Molecular phylodynamics of the heterosexual HIV epidemic in the United Kingdom. PLoS Pathog. 2009;5: e1000590.

15. Bello G, Simwaka B, Ndhlovu T, Salaniponi F, Hallett TB. Evidence for changes in behaviour leading to reductions in HIV prevalence in urban Malawi. Sex Transm Infect. 2011;87: 296–300.

16. Thompson MA, Aberg JA, Hoy JF, Telenti A, Benson C, Cahn P, et al. Antiretroviral treatment of adult HIV infection: 2012 recommendations of the International Antiviral Society-USA panel. JAMA. 2012;308: 387–402.

17. Junqueira, de Medeiros RM, Matte MCC, Araújo LAL, Chies JAB, Ashton-Prolla P, et al. Reviewing the History of HIV-1: Spread of Subtype B in the Americas. Martin DP, editor. PLoS One. 2011;6: e27489.

18. Bello G, Guimarães ML, Morgado MG, Guimara ML. Evolutionary history of HIV-1 subtype B and F infections in Brazil. AIDS. 2006;20: 763–8.

19. Almeida SE, de Medeiros RM, Junqueira DM, Gräf T, Passaes CP, Bello G, et al. Temporal dynamics of HIV-1 circulating subtypes in distinct exposure categories in southern brazil. Virol J. 2012;9: 306.

20. Frost SDW, Pillay D. Understanding Drivers of Phylogenetic Clustering in Molecular Epidemiological Studies of HIV. J Infect Dis. 2015;211: 856–8.

21. Poon AFY, Joy JB, Woods CK, Shurgold S, Colley G, Brumme CJ, et al. The impact of clinical, demographic and risk factors on rates of HIV transmission: a population-based phylogenetic analysis in British Columbia, Canada. J Infect Dis. 2015;211: 926–35.

22. Pennings PS, Holmes SP, Shafer RW. HIV-1 transmission networks in a small world. J Infect Dis. 2014;209: 180–2.

23. Hué S, Pillay D, Clewley JP, Pybus OG. Genetic analysis reveals the complex structure of HIV-1 transmission within defined risk groups. Proc Natl Acad Sci U S A. 2005;102: 4425–9.

24. Zehender G, Ebranati E, Lai A, Santoro MM, Alteri C, Giuliani M, et al. Population dynamics of HIV-1 subtype B in a cohort of men-having-sex-with-men in Rome, Italy. J Acquir Immune Defic Syndr. 2010;55: 156–60.

25. Brenner BG, Wainberg M a. Future of phylogeny in HIV prevention. J Acquir Immune Defic Syndr. 2013;63 Suppl 2: S248–54.

26. Ragonnet-cronin M, Ofner-agostini M, Merks H, Pilon R, Rekart M, Archibald CP, et al. Longitudinal Phylogenetic Surveillance Identifies Distinct Patterns of Cluster Dynamics. J Acquir Immune Defic Syndr. 2010;55: 102–108.

27. Kaye M, Chibo D, Birch C. Phylogenetic investigation of transmission pathways of drug-resistant HIV-1 utilizing pol sequences derived from resistance genotyping. J Acquir Immune Defic Syndr. 2008;49: 9–16.

28. Brenner BG, Roger M, Routy J, Moisi D, Ntemgwa M, Matte C, et al. High rates of forward transmission events after acute/early HIV-1 infection. J Infect Dis. 2007;195: 951–9.

29. Brenner BG, Roger M, Moisi DD, Oliveira M, Hardy I, Turgel R, et al. Transmission networks of drug resistance acquired in primary/early stage HIV infection. AIDS. 2008;22: 2509–15.

30. Audelin AM, Cowan S a, Obel N, Nielsen C, Jørgensen LB, Gerstoft J. Phylogenetics of the Danish HIV epidemic: the role of very late presenters in sustaining the epidemic. J Acquir Immune Defic Syndr. 2013;62: 102–8.

31. Mehta SR, Delport W, Brouwer KC, Espitia S, Patterson T, Pond SK, et al. The relatedness of HIV epidemics in the United States-Mexico border region. AIDS Res Hum Retroviruses. 2010;26: 1273–7.

32. Lemey P, Rambaut A, Bedford T, Faria N, Bielejec F, Baele G, et al. Unifying viral genetics and human transportation data to predict the global transmission dynamics of human influenza H3N2. PLoS Pathog. 2014;10: e1003932.

33. Paraskevis D, Pybus O, Magiorkinis G, Hatzakis A, Wensing AMJ, Vijver DA Van De, et al. Tracing the HIV-1 subtype B mobility in Europe: a phylogeographic approach. Retrovirology. 2009;6: 49.

34. Gilbert MTP, Rambaut A, Wlasiuk G, Spira TJ, Pitchenik AE, Worobey M. The emergence of HIV/AIDS in the Americas and beyond. Proc Natl Acad Sci U S A. 2007;104: 18566–70.

35. Alcantara LCJ, Cassol S, Libin P, Deforche K, Pybus OG, Van Ranst M, et al. A standardized framework for accurate, high-throughput genotyping of recombinant and non-recombinant viral sequences. Nucleic Acids Res. 2009;37: 1–9.

36. Kosakovsky Pond SL, Posada D, Stawiski E, Chappey C, Poon AFY, Hughes G, et al. An evolutionary model-based algorithm for accurate phylogenetic breakpoint mapping and subtype prediction in HIV-1. PLoS Comput Biol. 2009;5: e1000581.

37. Lole KS, Bollinger RC, Paranjape RS, Gadkari D, Kulkarni SS, Novak NG, et al. Full-length human immunodeficiency virus type 1 genomes from subtype C-infected seroconverters in India, with evidence of intersubtype recombination. J Virol. 1999;73: 152–60.

38. McWilliam H, Li W, Uludag M, Squizzato S, Park YM, Buso N, et al. Analysis Tool Web Services from the EMBL-EBI. Nucleic Acids Res. 2013;41: W597–600.

39. Johnson V a, Calvez V, Gunthard HF, Paredes R, Pillay D, Shafer RW, et al. Update of the drug resistance mutations in HIV-1: March 2013. Top Antivir Med. 2013;21: 6–14.

40. Stamatakis A. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics. 2014;30: 1312–3.

41. Miller, M.A., Pfeiffer, W., and Schwartz T. Creating the CIPRES Science Gateway for inference of large phylogenetic trees. Proceedings of the Gateway Computing Environments Workshop (GCE). New Orleans; 2010. pp. 1 – 8.

42. Tamura K, Stecher G, Peterson D, Filipski A, Kumar S. MEGA6: Molecular Evolutionary Genetics Analysis version 6.0. Mol Biol Evol. 2013;30: 2725–9.

43. Anisimova M, Gascuel O. Approximate likelihood-ratio test for branches: A fast, accurate, and powerful alternative. Syst Biol. 2006;55: 539–52.

44. Rambaut A. FigTree v1.4: Tree Figure Drawing Tool [Internet]. 2009. Available: http://tree.bio.ed.ac.uk/software/figtree/

45. Ragonnet-Cronin M, Hodcroft E, Hué S, Fearnhill E, Delpech V, Brown AJL, et al. Automated analysis of phylogenetic clusters. BMC Bioinformatics. 2013;14: 317.

46. Eyer-Silva WA, Morgado MG. Autochthonous horizontal transmission of a CRF02_AG strain revealed by a human immunodeficiency virus type 1 diversity survey in a small city in inner state of Rio de Janeiro, Southeast Brazil. Mem Inst Oswaldo Cruz. 2007;102: 809–815.

47. Eyer-Silva WA, Couto-Fernandez JC, Morgado MG. Molecular epidemiology of HIV type 1 in inner Rio De Janeiro State, Brazil. AIDS Res Hum Retroviruses. 2007;23: 303–8.

48. Eyer-Silva WA, Morgado MG. Molecular epidemiology of HIV-1 infection in a small Brazilian county: usefulness of envelope and polymerase sequences to epidemiologic studies. J Acquir Immune Defic Syndr. 2006;41: 664–70.

49. Drummond AJ, Suchard MA, Xie D, Rambaut A. Bayesian phylogenetics with BEAUti and the BEAST 1.7. Mol Biol Evol. 2012;29: 1969–73.

50. Bello G, Eyer-silva W a, Couto-Fernandez JC, Guimarães ML, Chequer-Fernandez SL, Teixeira SLM, et al. Demographic history of HIV-1 subtypes B and F in Brazil. Infect Genet Evol. 2007;7: 263–70.

51. Rambaut A, Drummond A. Tracer v1.6 [Internet]. 2007. Available: http://tree.bio.ed.ac.uk/software/tracer/

52. Gifford RJ, Liu TF, Rhee S-Y, Kiuchi M, Hue S, Pillay D, et al. The calibrated population resistance tool: standardized genotypic estimation of transmitted HIV-1 drug resistance. Bioinformatics. 2009;25: 1197–8.

53. Ashkenazy H, Penn O, Doron-Faigenboim A, Cohen O, Cannarozzi G, Zomer O, et al. FastML: a web server for probabilistic reconstruction of ancestral sequences. Nucleic Acids Res. 2012;40: W580–4.

54. Wertheim JO, Leigh Brown AJ, Hepler NL, Mehta SR, Richman DD, Smith DM, et al. The global transmission network of HIV-1. J Infect Dis. 2014;209: 304–13.

55. Prosperi MCF, Ciccozzi M, Fanti I, Saladini F, Pecorari M, Borghi V, et al. A novel methodology for large-scale phylogeny partition. Nat Commun. 2011;2: 321.

56. Leigh Brown AJ, Lycett SJ, Weinert L, Hughes GJ, Fearnhill E, Dunn DT. Transmission network parameters estimated from HIV sequences for a nationwide epidemic. J Infect Dis. 2011;204: 1463–9.

57. Beyrer C, Baral SD, Weir BW, Curran JW, Chaisson RE, Sullivan PS. A call to action for concentrated HIV epidemics. Curr Opin HIV AIDS. 2014;9: 95–100.

58. Volz EM, Ionides E, Romero-Severson EO, Brandt M-G, Mokotoff E, Koopman JS. HIV-1 Transmission during Early Infection in Men Who Have Sex with Men: A Phylodynamic Analysis. Hallett TB, editor. PLoS Med. 2013;10: e1001568.

59. Hallett TB. Early HIV infection in the United States: a virus's eye view. PLoS Med. 2013;10: e1001569.

60. Beyrer C, Sullivan P, Sanchez J, Baral SD, Collins C, Wirtz AL, et al. The increase in global HIV epidemics in MSM. AIDS. 2013;27: 2665–78.

61. Jia Y, Aliyu MH, Jennifer Huang Z. Dynamics of the HIV epidemic in MSM. Biomed Res Int. 2014;2014: 497543.

62. Sinha S, Shekhar RC, Ahmad H, Kumar N, Samantaray JC, Sreenivas V, et al. Prevalence of HIV drug resistance mutation in the northern Indian population after failure of the first line antiretroviral therapy. Curr HIV Res. 2012;10: 532–8.

63. Varella RB, Ferreira SB, Castro MB de, Tavares MD, Zalis MG. Prevalence of resistance-associated mutations in Human Immunodeficiency Virus type 1-positive individuals failing HAART in Rio de Janeiro, Brazil. Brazilian J Infect Dis. 2008;12: 380–384.

64. WHO. The HIV drug resistance report. 2012; Available: http://apps.who.int/iris/bitstream/10665/75183/1/9789241503938_eng.pdf

65. Biesinger T, Kimata JT. HIV-1 Transmission, Replication Fitness and Disease Progression. Virology (Auckl). 2008;2008: 49–63.

66. Van de Vijver DAMC, Wensing AMJ, Åsjö B, Bruckova M, Jorgensen LB, Camacho R, et al. HIV-1 drug-resistance patterns among patients on failing treatment in a large number of European countries. Acta dermatovenerologica Alpina, Pannonica, Adriat. 2010;19: 3–9.

67. Goodreau SM, Cassels S, Kasprzyk D, Montaño DE, Greek A, Morris M. Concurrent partnerships, acute infection and HIV epidemic dynamics among young adults in Zimbabwe. AIDS Behav. 2012;16: 312–22.

Capítulo 6: Discussão Final

Após pouco mais de trinta anos do descobrimento do agente etiológico da aids, em meio à complexa e avassaladora pandemia, o Brasil estabelece um cenário voltado para ações de saúde pública que se tornaram modelo em todo o mundo (Okie 2006). Desde 1996, o Brasil é um pioneiro na oferta gratuita de medicamentos antirretrovirais e no acesso à assistência médica e laboratorial relacionada à infecção (Brindeiro et al. 2003; Ministério da Saúde et al. 2014). O programa beneficiou milhões de brasileiros e aumentou a expectativa de vida dos pacientes, além de evitar uma perda de aproximadamente 2,2 bilhões de dólares com custos de hospitalização dos infectados (Okie 2006). No entanto, apesar de todo o esforço para conter a epidemia e dos recentes avanços mundiais em relação à queda nas taxas de incidência (UNAIDS 2014a), as estatísticas brasileiras quanto ao número de novas infecções e à mortalidade associada à aids ainda são alarmantes (Ministério da Saúde et al. 2014).

Recentemente, a epidemia de HIV/aids, apesar de ainda afetar a população em geral, tornou-se majoritariamente concentrada em grupos específicos. Nestes indivíduos, a redução nas taxas de incidência, observada para a epidemia em geral, não apenas no Brasil, mas em diversos países ao redor do mundo, ainda é modesta e se mantém alta e relativamente constante. Como tendência, a pandemia no mundo está concentrada predominantemente entre indivíduos de grupos historicamente associados à

infecção, como homens que fazem sexo com homens (HSH), profissionais do sexo (PS), usuários de drogas injetáveis (UDI) e transgêneros femininos (UNAIDS 2013; Beyrer et al. 2014). No Brasil e no restante da América Latina, a epidemia é ainda especialmente concentrada em indivíduos HSH, PS e transgêneros femininos (Ministério da Saúde et al. 2014). Apesar da consistente quantidade de informações a respeito destes grupos e da vulnerabilidade frente às altas taxas de incidência, curiosamente a atenção reportada a esta parcela da população é ainda limitada, incluindo, até mesmo, o baixo direcionamento de investimentos financeiros (Arán-Matero et al. 2011).

A eficaz resposta à epidemia de HIV, enfatizando toda sua heterogeneidade e incluindo programas ativos de prevenção e tratamento de indivíduos infectados requer ampla compreensão de sua magnitude. Ferramentas de análises moleculares se revelaram, recentemente, poderosos métodos para ampliar o entendimento das dinâmicas de transmissão do HIV na epidemia e contribuir para os programas de vigilância já vigentes (Volz et al. 2013; Little et al. 2014; Wertheim et al. 2014). Estes métodos, em conjunto com dados sócio-demográficos, geográficos e/ou clínicos, geram condições adequadas para caracterizar a epidemia através de uma abordagem não clássica. Portanto, a epidemiologia molecular surge como uma importante ferramenta para melhor definir a dinâmica do HIV na população e pode direcionar o foco das abordagens de interferência na epidemia.

Assim, este trabalho buscou entender a dinâmica epidemiológica de transmissão do HIV-1 subtipo B (HIV-1B) no Brasil através, principalmente, da caracterização molecular das transmissões deste subtipo e de suas formas relacionadas. Estas análises resgatam a dinâmica de transmissão e disseminação da epidemia de HIV-1B no Brasil aliando a objetividade dos métodos de análise filogenética e filodinâmica a questões ligadas à saúde pública. O entendimento da dinâmica de disseminação em nível macro

pode beneficiar programas de prevenção, permitindo a identificação dos principais pontos de transmissão no país, comportamento de risco e mutações de resistência circulantes. Estes achados poderão, ainda, influenciar no direcionamento a públicos-alvo específicos em campanhas públicas que visem diminuir a transmissão e, consequentemente, retardem o aumento de novos casos de infectados.

O entendimento da disseminação da epidemia a partir da África e a reconstrução histórica das rotas de transmissão do vírus pelas diferentes populações são fatores importantes para o entendimento da dinâmica atual das infecções. A introdução do HIV-1B em determinados grupos, especialmente nos indivíduos HSH, no início da epidemia, nos Estados Unidos, permitiu a rápida disseminação do vírus na população e repercute no atual cenário de HIV/Aids. O impacto da introdução do HIV-1B neste grupo, como efeito fundador, explica a manutenção da associação entre o subtipo B e a população HSH em determinadas regiões do mundo. Além disso, a relação de determinadas variantes do subtipo B com países específicos sugere a manutenção de uma dinâmica regional nas infecções por HIV.

Os resultados descritos neste trabalho revelam que a epidemia de HIV-1 no Brasil e no restante da América do Sul, ao contrário de outros locais no mundo, é primariamente influenciada por questões locais (Yirrell et al. 1998; Fisher et al. 2010; Wertheim et al. 2014). A análise de disseminação da variante B" corrobora com esta sobreposição das tendências regionais para definição da epidemia de HIV já que revelou um padrão similar de prevalência desta variante no sul do Brasil, estável ao longo de 10 anos e diferente do restante do país. Estudos anteriores em Uganda e Reino Unido sugerem que apenas 30% dos eventos de transmissão ocorrem localmente (Yirrell et al. 1998; Fisher et al. 2010). Da mesma forma, um recente estudo abordando a epidemia do HIV de forma global, utilizando sequências disponíveis de vários países, encontrou

resultados similares (Wertheim et al. 2014). A discrepância destes resultados em relação à epidemia instaurada na América do Sul pode ser um reflexo do grau de conectividade entre cidades, facilitando os processos migratórios humanos. Conforme revelado por Gray *et al*. (2009), a epidemia de HIV em Uganda parece ter estreita relação com países vizinhos devido à extrema conexão entre as cidades, especialmente através de rodovias. O mesmo mecanismo de disseminação que garante as altas prevalências do vírus nesta população, poderia também explicar a relação de sua epidemia com outros locais (Gray et al. 2009). A mesma lógica é válida para a epidemia do Reino Unido, um país com intensa relação social e política com outras regiões (Segal et al. 2009). Na América do Sul, por sua vez, o grau de conectividade entre as diferentes regiões parece envolver baixa nodalidade, tanto por vias terrestres, quanto por vias aéreas (Figura 6.1) (Egler 2012). O isolamento nodal das cidades da América do Sul garantiria, assim, uma reduzida influência de outros locais e uma importante limitação local da epidemia de HIV.

Cabe ressaltar que Brasil e Argentina são os países da América do Sul que apresentam as regiões com o maior grau de nodalidade (Figura 6.1) (Egler 2012). Estes dados, associados aos resultados para transmissão internacional de HIV-1B, onde Brasil e Argentina apresentam o maior número de ligações entre indivíduos nas cadeias de transmissão, sugerem que o grau de conectividade entre os países é um fator que afeta diretamente a epidemia. No Brasil, os estados de São Paulo, Rio de Janeiro e Paraná estão entre os locais melhor conectados com o seu entorno (Figura 6.1). Estas observações são válidas para destacar a importante participação destes estados, juntamente com Goiás e Mato Grosso, nas cadeias de transmissão interestaduais de HIV-1B no Brasil.

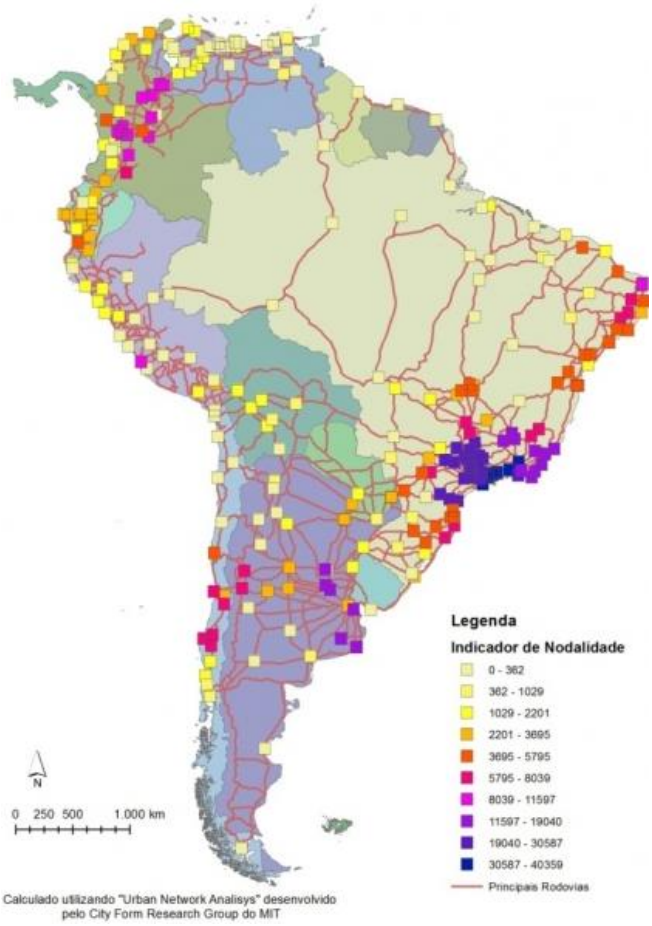


Figura 6.1. Indicadores de nodalidade da rede de cidades na América do Sul e as principais rodovias do continente. Fonte: Egler, 2012.

A heterogeneidade da epidemia, destacada especialmente pelos grupos de exposição, reflete padrões já estabelecidos sobre a prevalência do vírus nas diferentes populações. Entre as cadeias de transmissão encontradas neste estudo observa-se especificamente a relevante proporção de grandes cadeias (abrangendo grande número de indivíduos) envolvendo apenas indivíduos HSH. Como inovação, o extenso banco de dados analisado nos permitiu avaliar o tempo entre a infecção e a transmissão do vírus para um novo indivíduo dentro deste grupo. O intervalo foi estimado em aproximadamente um ano para a população na América do Sul. Dados recentes, especificamente entre a população de indivíduos HSH, corroborando nossos resultados, estima que, durante o primeiro ano, existe uma infecciosidade oito vezes maior que

durante a fase de infecção crônica por HIV (Volz et al. 2013). Neste mesmo estudo, os autores revelam que de 42% a 46% das novas infecções envolvem indivíduos recém-infectados. Estes resultados em conjunto com as análises para a América do Sul revelam uma dinâmica da epidemia do HIV particular para o grupo HSH em que o intervalo de tempo entre as novas infecções é muito estreito e possui ampla participação de indivíduos recém-infectados, que provavelmente desconhecem seu *status* sorológico.

No final de 2013, o Ministério da Saúde anunciou a aprovação de um novo Protocolo Clínico de Tratamento de Adultos com HIV e Aids para brasileiros (http://www.aids.gov.br/pcdt/guia_rapida/). Segundo este, todos os pacientes soropositivos, independente do estágio da doença, deverão receber indicação para uso dos medicamentos antirretrovirais. O ministério afirma que a medida inovadora tem grande impacto na saúde individual, porque garante a melhoria da qualidade de vida dos infectados pelo HIV, e na saúde pública, porque o indivíduo em tratamento com antirretrovirais, ao diminuir sua carga viral, reduz a propagação do vírus. Segundo nossos resultados, é possível inferir que apenas uma pequena parcela das transmissões será evitada pela administração da terapia após o diagnóstico no grupo HSH, já que grande parte das transmissões possivelmente ocorre antes mesmo do próprio diagnóstico (Volz et al. 2013). Estes dados ressaltam a necessidade de uma melhor caracterização da epidemia no Brasil e, além disso, mostram a necessidade de medidas de prevenção específicas para grupos de epidemia concentrada, com maior potencial de transmissão do vírus para outros indivíduos, como a população HSH. Apesar de não representadas em nosso estudo, por falta de informações sobre os pacientes, possivelmente outros grupos onde a epidemia é também concentrada, como profissionais do sexo e transgêneros femininos, o mesmo padrão epidêmico deve ocorrer.

Para a população em geral, encontramos um intervalo de transmissão da infecção de aproximadamente 28 meses. Este período é corroborado por um estudo realizado entre indivíduos heterossexuais, no Reino Unido, infectados por subtipos não-B do HIV-1 (Hughes et al. 2009). Ainda, o agrupamento de sequências em grupos de transmissão nos permitiu avaliar a maior probabilidade de participação de indivíduos *naïve* (indivíduos não usuários da terapia antirretroviral) em grupos de transmissão do que como infecções isoladas. Estes resultados novamente sugerem que a maior parte das transmissões ocorre logo após a infecção e envolve especialmente indivíduos ainda sem acesso à terapia antirretroviral. A concordância de tempo entre diferentes populações e, até mesmo, diferente subtipo sugere uma importante influência do estágio de infecção na transmissão do vírus. Assim, na prática clínica, o reconhecimento de indivíduos recém-infectados teria grande impacto na dinâmica da epidemia de HIV na América do Sul e poderia eficientemente gerar uma resposta significante contra o aumento do número de novos casos.

Os testes de genotipagem, disponibilizados pelo sistema de saúde para pacientes em falha terapêutica, são fundamentais para mapear e identificar as principais mutações, assim como determinar o perfil de resistência aos medicamentos antirretrovirais (http://www.aids.gov.br/aids). Além disso, graças a este sistema, o número de sequências genéticas do vírus, depositadas em banco de dados, aumentou substancialmente nos últimos anos. Nosso estudo permitiu descrever as mutações de resistência circulantes na população de indivíduos infectados na América do Sul e indica uma alta prevalência de mutações associadas à terapia com Inibidores Nucleosídeos da Transcriptase Reversa (INTR) e já associadas com indivíduos em falha terapêutica (M184V e M41L) (van de Vijver et al. 2010). Dentro das cadeias de transmissão identificadas aqui, observou-se ainda uma menor quantidade de mutações

de resistência em comparação ao grupo de sequências não incluso em cadeias. Esse padrão pode estar associado à diminuição da capacidade de replicação das formas que carregam mutações de resistência. No entanto, a maior parte das cadeias de transmissão incluindo indivíduos infectados por vírus que apresentam mutações de resistência foram iniciadas por indivíduos já infectados por vírus mutantes (análise de ancestral comum), demonstrando a importância de contínuos programas de vigilância para mutações de resistência na epidemia.

Por fim, a caracterização global de determinada epidemia, envolve também o entendimento da diversidade molecular do patógeno circulante na população. Na epidemia de HIV-1B no Brasil, particularmente, esta medida é ainda mais importante devido à circulação de variantes virais relacionadas a uma progressão clínica mais lenta e menos agressiva da doença (Santoro-lopes et al. 2000; Casseb et al. 2004; Brito et al. 2006; Araujo et al. 2010). O fato de os indivíduos infectados pela forma B" do HIV-1 apresentarem longo período de infecção pode aumentar a possibilidade de transmissão do vírus para outros indivíduos e, neste caso, seriam importantes alvos na epidemia para medidas de prevenção. Além disso, se detectado um aumento da frequência desta variante na epidemia, a associação com uma progressão lenta e a certeza do potencial de transmissão por longos períodos requereria medidas mais específicas do Ministério da Saúde para diagnóstico da variante. Nossos resultados buscaram entender o comportamento da epidemia desta forma genética ao longo de 10 anos em dois diferentes locais do Brasil, destacando o papel de diferentes grupos de transmissão. Verificou-se certa estabilidade na prevalência da epidemia e, curiosamente, a forma GPGR foi relacionada com transmissão sexual em Florianópolis. Este resultado em conjunto com a avaliação de um recente estudo denota uma epidemia de grande

particularidade em Santa Catarina e reforça nossos achados sobre o impacto da dinâmica local na manutenção da circulação do vírus (Gräf et al. 2011).

Em junho de 2014, através da portaria 1.271, o Ministério da Saúde brasileiro incluiu a infecção por HIV como doença de notificação compulsória. Esta medida é um avanço, pois intensifica a ação dos órgãos junto ao indivíduo infectado, assegura um conhecimento mais amplo a respeito da epidemia e, além disso, permite caracterizar com melhor eficiência as subepidemias locais instauradas no Brasil. Neste sentido, o último boletim epidemiológico publicado pelo Departamento de DST, Aids e Hepatites Virais reitera a necessidade de que a atuação dos órgãos de saúde pública baseiem-se na utilização do conceito de *hotspots*, priorizando estados como Amazonas e Rio Grande do Sul, além de populações-chave (Ministério da Saúde et al. 2014). Estas medidas vão de acordo com as recomendações da UNAIDS e são corroboradas por alguns dos resultados obtidos neste trabalho.

O Brasil deveria descentralizar as ações que impactam na saúde pública e entender a epidemia com um complexo de diversas subepidemias que exigem abordagens diferentes e intervenções específicas. Os dados encontrados aqui demonstram especialmente a divisão da epidemia em subepidemias localizadas com dinâmicas claramente distintas para populações específicas. A concentração da epidemia em determinados grupos de risco cria a necessidade de esforços mais vigorosos nestas populações. A intervenção específica nestes grupos é uma garantia de sucesso na redução de novos casos de HIV e, em última análise, beneficia toda a população. No caso de indivíduos HSH, a epidemia atual no Brasil e América Latina requer um maior esforço dos órgãos de saúde pública para atingir jovens gays, ampliar os programas de testagem e tratamento e implementar novas técnicas de prevenção, como a profilaxia pré-exposição, que poderá ajudar no controle de expansão da

epidemia. As novas respostas para as ações da iniciativa pública deverão envolver abordagens rápidas e no tempo real da dinâmica de cada uma destas microepidemias. Conforme demonstrado aqui, a utilização de técnicas moleculares aliadas a dados clínicos e sociodemográficos pode ser um caminho efetivo e altamente poderoso para projetar o Brasil como um país onde pesquisa e saúde pública caminham juntas.

Capítulo 7: Conclusões

- A extensão atual e o impacto das infecções por HIV-1B no mundo teve grande influência dos processos de migração e do comportamento sexual em humanos e reflete padrões do início da epidemia;
- A circulação de variantes relacionadas ao HIV-1B em regiões específicas do mundo fornece evidências para a influência de fatores locais na dinâmica das infecções pelo HIV;
- A variante B"-GWGR aparentemente circula em todo o Brasil, no entanto, em prevalência distinta para as diferentes regiões do país;
- A epidemia de HIV-1B em Santa Catarina parece refletir uma subepidemia particular, já que variantes virais estão significativamente associadas a grupos de transmissão específicos;
- A epidemia de HIV-1B na América do Sul, especialmente no Brasil, parece ser influenciada primariamente por uma dinâmica local dentro dos países, ou dos estados, no caso do Brasil, provavelmente pelo isolamento nodal dos centros urbanos nestas regiões;
- O tempo médio de transmissão do HIV-1B entre diferentes indivíduos na América do Sul é de aproximadamente 29 meses. Entre o grupo de indivíduos HSH, no entanto, a dinâmica de transmissão é mais acelerada e ocorre em média 12 meses após a infecção. Esses resultados justificam a intervenção dos programas de saúde pública em grupos de epidemia concentrada;
- As taxas de transmissão mais significantes para a epidemia acontecem envolvendo indivíduo *naive*, que provavelmente desconhecem seu estado sorológico;

- Os métodos de filogenética molecular e a ampla amostragem espacial e temporal são capazes de resgatar importantes padrões para a atual epidemia do HIV e podem ajudar significativamente no entendimento das dinâmicas de transmissão.

Capítulo 8: Perspectivas

- Inferir, a partir de reconstruções filogenéticas de sequências coletadas em bancos de dados públicos, o local e ano de origem da variante B”-GWGR;
- Identificar possíveis passos evolutivos para o surgimento da variante B”-GWGR na população humana, bem como relacionar aspectos evolutivos e epidemiológicos com a progressão clínica de infecções causadas por esta forma;
- Entender a dinâmica de dispersão inicial do HIV-1B na América do Sul, especialmente no Brasil.

Referências:

Abecasis AB, Wensing AMJ, Paraskevis D, Vercauteren J, Theys K, Van de Vijver DAMC, Albert J, Asjö B, Balotta C, Beshkov D et al. (2013) HIV-1 subtype distribution and its demographic determinants in newly diagnosed patients in Europe suggest highly compartmentalized epidemics. Retrovirology 10:7.

Almeida SE, de Medeiros RM, Junqueira DM, Gräf T, Passaes CP, Bello G, Morgado MG and L Guimarães M (2012) Temporal dynamics of HIV-1 circulating subtypes in distinct exposure categories in southern brazil. Virol J 9:306.

Arán-Matero D, Amico P, Arán-Fernandez C, Gobet B, Izazola-Licea JA and Avila-Figueroa C (2011) Levels of spending and resource allocation to HIV programs and services in Latin America and the Caribbean. PLoS One 6:e22373.

Araujo AF, Brites C, Monteiro-Cunha J, Santos LA, Galvao-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 Res Hum Retroviruses 26:1249–54.

Arruda L, Romano C, Martinez M, Araújo M, Costa F, Oliveira K, Gonsales C, Duarte A and Casseb J (2011) The HIV-1 Subtype B variant (B'-GWGR motif) was introduced by founder effect among the HIV-1-infected subjects in São Paulo city, Brazil. 6 IAS Conf. HIV Pathog. Treat. Prev.

Audelin AM, Cowan S a, Obel N, Nielsen C, Jørgensen LB and Gerstoft J (2013) Phylogenetics of the Danish HIV epidemic: the role of very late presenters in sustaining the epidemic. J Acquir Immune Defic Syndr 62:102–8.

Aulicino PC, Kopka J, Mangano AM, Rocco C, Iacono M, Bologna R, Sen L and Liebert MA (2005) Sequence analysis of a South American HIV type 1 BC recombinant. AIDS Res Hum Retroviruses 21:158–64.

Auvert B, Sobngwi-Tambekou J, Cutler E, Nieuwoudt M, Lissouba P, Puren A and Taljaard D (2009) Effect of male circumcision on the prevalence of high-risk human papillomavirus in young men: results of a randomized controlled trial conducted in Orange Farm, South Africa. J Infect Dis 199:14–9.

Bailes E, Gao F, Bibollet-Ruche F, Courgnaud V, Peeters M, Marx P a, Hahn BH and Sharp PM (2003) Hybrid origin of SIV in chimpanzees. Science 300:1713.

Barré-Sinoussi F, Chermann JC, Rey F, Nugeyre MC, Chamaret S, Gruest J, Dauget C, Axler-Blin C, Brun-Vezinet F, Rousioux C et al. (1983) Isolation of a T-lymphotropic retrovirus from patient at risk for AIDS. Science 220:868–870.

Bello G, Aulicino PC, Ruchansky D, Guimarães ML, Lopez-Galindez C, Casado C, Chiparelli H, Rocco C, Mangano A, Sen L et al. (2010) Phylodynamics of HIV-1 circulating recombinant forms 12_BF and 38_BF in Argentina and Uruguay. Retrovirology 7:22.

Bello G, Eyer-silva W a, Couto-Fernandez JC, Guimarães ML, Chequer-Fernandez SL, Teixeira SLM, Morgado MG and Guimara ML (2007) Demographic history of HIV-1 subtypes B and F in Brazil. Infect Genet Evol 7:263–70.

Bello G, Guimarães ML, Morgado MG and Guimara ML (2006) Evolutionary history of HIV-1 subtype B and F infections in Brazil. AIDS 20:763–8.

Bello G, Passaes CP, Guimarães ML, Lorete RS, Matos Almeida SE, Medeiros RM, Alencastro PR and Morgado MG (2008) Origin and evolutionary history of HIV-1 subtype C in Brazil. AIDS 22:1993–2000.

Bello G, Simwaka B, Ndhlovu T, Salaniponi F and Hallett TB (2011) Evidence for changes in behaviour leading to reductions in HIV prevalence in urban Malawi. Sex Transm Infect 87:296–300.

Berry N, Davis C, Jenkins A, Wood D, Minor P, Schild G, Bottiger M, Almond N, Rambaut A, Robertson DL et al. (2001) Phylogeny and the origin of HIV-1. Nature 410:1047–1048.

Beyrer C, Baral SD, Weir BW, Curran JW, Chaisson RE and Sullivan PS (2014) A call to action for concentrated HIV epidemics. Curr Opin HIV AIDS 9:95–100.

Bezemer D, Faria NR, Hassan A, Hamers RL, Mutua G, Anzala O, Mandaliya K, Cane P, Berkley JA, Rinke de Wit TF et al. (2014) HIV Type 1 transmission networks among men having sex with men and heterosexuals in Kenya. AIDS Res Hum Retroviruses 30:118–26.

Bibollet-Ruche F, Bailes E, Gao F, Pourrut X, Barlow KL, Clewley JP, Mwenda JM, Langat DK, Chege GK, McClure HM et al. (2004) New simian immunodeficiency virus infecting De Brazza's monkeys (Cercopithecus neglectus): evidence for a cercopithecus monkey virus clade. J Virol 78:7748–62.

Blower S (1991) Behaviour change and stabilization of seroprevalence levels in communities of injecting drug users: correlation or causation? J Acquir Immune Defic Syndr 4:920–3.

Bongertz V, Bou-Habib DC, Brígido LF, Caseiro M, Chequer PJN, Couto-Fernandez JC, Ferreira PC, Galvão-Castro B, Greco D, Guimarães ML et al. (2000) HIV-1 diversity in Brazil: genetic, biologic, and immunologic characterization of HIV-1 strains in three potential HIV vaccine evaluation sites. Brazilian Network for HIV Isolation and Characterization. J Acquir Immune Defic Syndr 23:184–93.

Brenner BG, Roger M, Routy J, Moisi D, Ntemgwa M, Matte C, Baril J, Thomas R, Rouleau D, Bruneau J et al. (2007) High rates of forward transmission events after acute/early HIV-1 infection. J Infect Dis 195:951–9.

Brenner BG and Wainberg M a (2013) Future of phylogeny in HIV prevention. J Acquir Immune Defic Syndr 63 Suppl 2:S248–54.

Brígido LFM, Nunes CC, Oliveira CM, Knoll RK, Ferreira JLP, Freitas CA, Alves MA, Dias C and Rodrigues R (2007) HIV type 1 subtype C and CB Pol recombinants prevail at the cities with the highest AIDS prevalence rate in Brazil. AIDS Res Hum Retroviruses 23:1579–1586.

Brindeiro RM, Diaz RS, Sabino EC, Morgado MG, Pires IL, Brigido L, Dantas MC, Barreira D, Teixeira PR and Tanuri A (2003) Brazilian Network for HIV Drug Resistance Surveillance (HIV-BResNet): a survey of chronically infected individuals. AIDS 17:1063–9.

Brito A De, Komninakis SC V, Oliveira RM De, Fonseca LAM, Duarte AJS and Casseb J (2006) Women Infected with HIV Type 1 Brazilian Variant , Subtype B (B -GWGR Motif) Have Slower Progression to AIDS , Compared with Patients Infected with Subtype B (B-GPGR Motif). Clin Infect Dis 43:0–5.

Cabello M, Junqueira DM and Bello G (2015) Dissemination of nonpandemic Caribbean HIV-1 subtype B clades in Latin America. AIDS 29:483–92.

Cabello M, Mendoza Y and Bello G (2014) Spatiotemporal Dynamics of Dissemination of Non-Pandemic HIV-1 Subtype B Clades in the Caribbean Region. PLoS One 9:e106045.

Callegaro A, Svicher V, Alteri C, Lo Presti A, Valenti D, Goglio A, Salemi M, Cella E, Perno CF, Ciccozzi M et al. (2011) Epidemiological network analysis in HIV-1 B infected patients diagnosed in Italy between 2000 and 2008. Infect Genet Evol 11:624–32.

Carr JK, Foley BT, Leitner T, Salminen M, Korber B and McCutchan F (1998) Reference sequences representing the principal genetic diversity of HIV-1 in the pandemic. Hum retroviruses AIDS 10–19.

Carrion G, Hierholzer J, Montano S, Alava A, Perez J, Guevara A, Laguna-Torres V, Mosquera C, Russell K, Chauca G et al. (2003) Circulating recombinant form CRF02_AG in South America. AIDS Res Hum Retroviruses 19:329–32.

Casseb J, Hong MA, Gonsalez C, Brígido LF, Duarte AJ and Michael-Hendry R (1998) Two variants of HIV-1 B serotype are transmitted heterosexually in São Paulo, Brazil. Braz J Med Biol Res 31:1243–6.

Casseb J, Komninakis S, Abdalla L, Brigido L, Rodrigues R, Araujo F, Rochaveiga a, Almeida a, Flannery B and Michaelhendry R (2002) HIV disease progression: is the Brazilian variant subtype B' (GWGR motif) less pathogenic than US/European subtype B (GPGR)?1, 2. Int J Infect Dis 6:164–169.

Casseb J, Montanheiro P, Komninakis S, Brito A and Duarte AJS (2004) Human immunodeficiency virus type 1 Brazilian subtype B variant showed an increasing avidity of the anti-V3 antibodies over time compared to the subtype B US/European strain in São Paulo, Brazil. Mem Inst Oswaldo Cruz 99:69–71.

Chalmet K, Staelens D, Blot S, Dinakis S, Pelgrom J, Plum J, Vogelaers D, Vandekerckhove L and Verhofstede C (2010) Epidemiological study of phylogenetic transmission clusters in a local HIV-1 epidemic reveals distinct differences between subtype B and non-B infections. BMC Infect Dis 10:262.

Charneau P, Borman AM, Quillent C, Guétard D, Chamaret S, Cohen J, Rémy G, Montagnier L and Clavel F (1994) Isolation and envelope sequence of a highly divergent HIV-1 isolate: definition of a new HIV-1 group. Virology 205:247–53.

Colijn C and Gardy J (2014) Phylogenetic tree shapes resolve disease transmission patterns. Evol Med public Heal 2014:96–108.

Covas DT, Bíscaro TA, Kashima S, Duarte G and Machado AA (1998) High frequency of the GWG (Pro Trp) envelope variant of HIV-1 in Southeast Brazil. J Acquir immune Defic Syndr Hum retrovirology 19:74–9.

Cuevas MT, Ruibal I, Villahermosa ML, Díaz H, Delgado E, Parga EV, Pérez-Alvarez L, de Armas MB, Cuevas L, Medrano L et al. (2002) High HIV-1 genetic diversity in Cuba. AIDS 16:1643–53.

D'arc M, Ayouba A, Esteban A, Learn GH, Boué V, Liegeois F, Etienne L, Tagg N, Leendertz FH, Boesch C et al. (2015) Origin of the HIV-1 group O epidemic in western lowland gorillas. Proc Natl Acad Sci 201502022.

De Medeiros RM, Junqueira DM, Matte MCC, Barcellos NT, Chies JAB and Almeida SEDM (2011) Co-Circulation HIV-1 Subtypes B , C , and CRF31 _ BC in a Drug-Naive Population From Southernmost Brazil : Analysis of Primary Resistance Mutations. J Med Virol 83:1682–1688.

De Oliveira T, Pillay D and Gifford RJ (2010) The HIV-1 subtype C epidemic in South America is linked to the United Kingdom. PLoS One 5:e9311.

Dean G, Pao D, Fisher M, Hue S, Murphy G, Cane PA and Sabin CA (2005) Transmission of HIV-1 during primary infection : relationship to sexual risk and sexually transmitted infections. Aids 85–90.

Delaugerre C, De Oliveira F, Lascoux-Combe C, Plantier J-C and Simon F (2011) HIV-1 group N: travelling beyond Cameroon. Lancet 378:1894.

Dennis AM, Hué S, Hurt CB, Napravnik S, Sebastian J, Pillay D and Eron JJ (2012) Phylogenetic insights into regional HIV transmission. AIDS 26:1813–22.

Diaz RS, Leal E, Sanabani S, Sucupira MC a, Tanuri AAA, Sabino EC, Janini LMM, Leal L, Sobhie R, Leal É et al. (2008) Selective regimes and evolutionary rates of HIV-1 subtype B V3 variants in the Brazilian epidemic. Virology 381:184–193.

Donnell D, Baeten JM, Kiarie J, Thomas KK, Stevens W, Cohen CR, McIntyre J, Lingappa JR and Celum C (2010) Heterosexual HIV-1 transmission after initiation of antiretroviral therapy: a prospective cohort analysis. Lancet 375:2092–8.

Eames KTD and Keeling MJ (2004) Monogamous networks and the spread of sexually transmitted diseases. Math Biosci 189:115–30.

Egler CAG (2012) Nodalidade e rede de cidades na América do Sul. Confins. doi: 10.4000/confins.7878

Faria NR, Rambaut A, Suchard MA, Baele G, Bedford T, Ward MJ, Tatem AJ, Sousa JD, Arinaminpathy N, Pepin J et al. (2014) The early spread and epidemic ignition of HIV-1 in human populations. Science (80-) 346:56–61.

Fisher M, Pao D, Brown AE, Sudarshi D, Gill ON, Cane P, Buckton AJ, Parry J V, Johnson AM, Sabin C et al. (2010) Determinants of HIV-1 transmission in men who have sex with men: a combined clinical, epidemiological and phylogenetic approach. AIDS 24:1739–47.

Fontella R, Soares MA and Schrago CG (2008) On the origin of HIV-1 subtype C in South America. AIDS 22:2001–11.

Franca RFO, Castro-Jorge LA, Neto RJP, Jorge DMM, Lima DM, Colares JKB, Paula SO and da Fonseca BAL (2011) Genotypic characteristics of HIV type 1 based on gp120 hypervariable region 3 of isolates from Southern Brazil. AIDS Res Hum Retroviruses 27:903–9.

Gallo RC, Sarin PS, Gelmann EP, Robert-Guroff M, Richardson E, Kalyanaraman VS, Mann D, Sidhu GD, Stahl RE, Zolla-Pazner S et al. (1983) Isolation of human T-cell leukemia virus in acquired immune deficiency syndrome (AIDS). Science 220:865–867.

Gao F, Bailes E, Robertson DL, Chen Y, Rodenburg CM, Michael SF, Cummins LB, Arthur LO, Peeters M, Shaw GM et al. (1999) Origin of HIV-1 in the chimpanzee Pan troglodytes troglodytes. Nature 397:436–41.

GHO (2013) Global Health Observatory data.

Gilbert MTP, Rambaut A, Wlasiuk G, Spira TJ, Pitchenik AE and Worobey M (2007) The emergence of HIV/AIDS in the Americas and beyond. Proc Natl Acad Sci U S A 104:18566–70.

Gottlieb MS, Schroff R, Schanker HM, Weisman JD, Fan PT, Wolf RA and Saxon A (1981) Pneumocystis carinii pneumonia and mucosal candidiasis in previously healthy homosexual men: evidence of a new acquired cellular immunodeficiency. N Engl J Med 305:1425–31.

Grabowski MK, Lessler J, Redd AD, Kagaayi J, Laeyendecker O, Ndyanabo A, Nelson MI, Cummings D a T, Bwanika JB, Mueller AC et al. (2014) The role of viral introductions in sustaining community-

based HIV epidemics in rural Uganda: evidence from spatial clustering, phylogenetics, and egocentric transmission models. PLoS Med 11:e1001610.

Grabowski MK and Redd AD (2014) Molecular tools for studying HIV transmission in sexual networks. Curr Opin HIV AIDS 9:126–33.

Gräf T, Passaes CPB, Ferreira LGE, Grisard EC, Morgado MG, Bello G and Pinto AR (2011) HIV-1 genetic diversity and drug resistance among treatment naïve patients from Southern Brazil: An association of HIV-1 subtypes with exposure categories. J Clin Virol 51:186–91.

Grant RM, Lama JR, Anderson PL, McMahan V, Liu AY, Vargas L, Goicochea P, Casapía M, Guanira-Carranza JV, Ramirez-Cardich ME et al. (2010) Preexposure Chemoprophylaxis for HIV Prevention in Men Who Have Sex with Men. N Engl J Med 2092–2098.

Gray RHRR, Tatem AJ, Lamers S, Hou W, Laeyendecker O, Serwadda D, Sewankambo N, Wawer M, Quinn TC, Goodenow MM et al. (2009) Spatial phylodynamics of HIV-1 epidemic emergence in east Africa. AIDS 23:F9–F17.

Guimarães M, dos Santos Moreira A, Loureiro R, Galvão-castro B and Morgado MG (2002) High frequency of recombinant genomes in HIV type 1 samples from Brazilian southeastern and southern regions. AIDS Res Hum Retroviruses 18:1261–9.

Guimarães M, Eyer-Silva W, Couto-Fernandez J and Morgado M (2008) Identification of two new CRF_BF in Rio de Janeiro State, Brazil. AIDS 30:433–435.

Guimarães ML, Vicente ACP, Otsuki K, da Silva RFFC, Francisco M, da Silva FG, Serrano D, Morgado MG, Bello G, Ferreira R et al. (2009) Close phylogenetic relationship between Angolan and Romanian HIV-1 subtype F1 isolates. Retrovirology 6:39. doi: 10.1186/1742-4690-6-39

Gürtler LG, Zekeng L, Tsague JM, van Brunn A, Afane Ze E, Eberle J and Kaptue L (1996) HIV-1 subtype O: epidemiology, pathogenesis, diagnosis, and perspectives of the evolution of HIV. Arch Virol Suppl 11:195–202.

Hahn BH, Shaw GM, De Cock KM and Sharp PM (2000) AIDS as a zoonosis: scientific and public health implications. Science 287:607–14.

Hemelaar J (2012) The origin and diversity of the HIV-1 pandemic. Trends Mol Med 18:182–92.

Hemelaar J, Gouws E, Ghys PD and Osmanov S (2011) Global trends in molecular epidemiology of HIV-1 during 2000-2007. AIDS 25:679–89.

Hierholzer J, Montano S, Hoelscher M, Negrete M, Hierholzer M, Avila MM, Carrillo MG, Russi JC, Vinoles J, Alava A et al. (2002) Molecular Epidemiology of HIV Type 1 in Ecuador, Peru, Bolivia, Uruguay, and Argentina. AIDS Res Hum Retroviruses 18:1339–50.

Hué S, Gifford RJ, Dunn D, Fernhill E and Pillay D (2009) Demonstration of sustained drug-resistant human immunodeficiency virus type 1 lineages circulating among treatment-naïve individuals. J Virol 83:2645–54.

Hué S, Pillay D, Clewley JP and Pybus OG (2005) Genetic analysis reveals the complex structure of HIV-1 transmission within defined risk groups. Proc Natl Acad Sci U S A 102:4425–9.

Hughes GJ, Fearnhill E, Dunn D, Lycett SJ, Rambaut A and Leigh Brown AJ (2009) Molecular phylodynamics of the heterosexual HIV epidemic in the United Kingdom. PLoS Pathog 5:e1000590.

Jia Y, Aliyu MH and Jennifer Huang Z (2014) Dynamics of the HIV epidemic in MSM. Biomed Res Int 2014:497543.

Junqueira, de Medeiros RM, Matte MCC, Araújo LAL, Chies JAB, Ashton-Prolla P and Almeida SE de M (2011) Reviewing the History of HIV-1: Spread of Subtype B in the Americas. PLoS One 6:e27489.

Junqueira DM, Medeiros RM De, Leite TCNF, Guimarães ML, Gräf T, Pinto AR and Almeida SEDM (2013) Detection of the B"-GWGR variant in the southernmost region of Brazil: unveiling the complexity of the human immunodeficiency virus-1 subtype B epidemic. Mem Inst Oswaldo Cruz 108:735–40.

Kalish ML, Robbins KE, Pieniazek D, Schaefer A, Nzilambi N, Quinn TC, St Louis ME, Youngpairoj AS, Phillips J, Jaffe HW et al. (2004) Recombinant viruses and early global HIV-1 epidemic. Emerg Infect Dis 10:1227–34.

Kalish ML, Wolfe ND, Ndongmo CB, McNicholl J, Robbins KE, Aidoo M, Fonjungo PN, Alemnji G, Zeh C, Djoko CF et al. (2005) Central African hunters exposed to simian immunodeficiency virus. Emerg Infect Dis 11:1928–30.

Keele BF, Van Heuverswyn F, Li Y, Bailes E, Takehisa J, Santiago ML, Bibollet-Ruche F, Chen Y, Wain L V, Liegeois F et al. (2006) Chimpanzee reservoirs of pandemic and nonpandemic HIV-1. Science 313:523–526.

Kouyos RD, Wyl V Von, Yerly S, Taffe P, Shah C, Bu P, Klimkait T, Weber R, Hirschel B, Cavassini M et al. (2010) Molecular Epidemiology Reveals Long-Term Changes in HIV Type 1 Subtype B Transmission in Switzerland. Epidemiology 2009:8–11.

Kuiken C, Thakallapalli R, Esklid a, de Ronde a, Eskild A and Ronde A De (2000) Genetic analysis reveals epidemiologic patterns in the spread of human immunodeficiency virus. Am J Epidemiol 152:814–22.

Lama JR, Sanchez J, Suarez L, Caballero P, Laguna A, Sanchez JL, Whittington WLH, Celum C and Grant RM (2006) Linking HIV and antiretroviral drug resistance surveillance in Peru: a model for a third-generation HIV sentinel surveillance. J Acquir Immune Defic Syndr 42:501–5.

Leal É, Martins LO, Janini LM and Diaz RS (2007) Spread of HIV-1 BF and CB recombinants in South America. AIDS Res Hum Retroviruses 52841–52841.

Leal E, Silva WP, Sucupira MC, Janini LM and Diaz RS (2008) Molecular and structural characterization of HIV-1 subtype B Brazilian isolates with GWGR tetramer at the tip of the V3-loop. Virology 381:222–9.

Leal É and Villanova FE (2010) Diversity of HIV-1 subtype B: implications to the origin of BF recombinants. PLoS One 5:e11833.

Leigh Brown AJ, Lycett SJ, Weinert L, Hughes GJ, Fearnhill E and Dunn DT (2011) Transmission network parameters estimated from HIV sequences for a nationwide epidemic. J Infect Dis 204:1463–9.

Leitner T (1996) Genetic subtypes of HIV-1. In: Myers G, Foley B, Mellors JW, Korber B, Jeang KT, Wain-Hobson S, editors. Human Retroviruses and AIDS. Theor Biol Biophys Los Alamos Natl Lab Los Alamos, III28–40.

Lemey P, Pybus OGOG, Wang B, Saksena NKNK, Salemi M and Vandamme A-MAM (2003) Tracing the origin and history of the HIV-2 epidemic. Proc Natl Acad Sci U S A 100:6588–92.

Lemey P, Rambaut A and Pybus O (2006) HIV evolutionary dynamics within and among hosts. AIDS Rev 8:125–40.

Levi GC and Vitória MAA (2002) Fighting against AIDS: the Brazilian experience. AIDS 16:2373–83.

Lewis F, Hughes GJ, Rambaut A, Pozniak A, Leigh Brown AJ and Brown AJL (2008) Episodic sexual transmission of HIV revealed by molecular phylodynamics. PLoS Med 5:e50.

Lihana RW, Ssemwanga D, Abimiku A and Ndembi N (2012) Update on HIV-1 diversity in Africa: a decade in review. AIDS Rev 14:83–100.

Little SJ, Kosakovsky Pond SL, Anderson CM, Young JA, Wertheim JO, Mehta SR, May S and Smith DM (2014) Using HIV networks to inform real time prevention interventions. PLoS One 9:e98443.

Louwagie J, Delwart EL, Mullins JI, McCutchan FE, Eddy G and Burke DS (1994) Genetic analysis of HIV-1 isolates from Brazil reveals presence of two distinct genetic subtypes. AIDS Res Hum Retroviruses 10:561–7.

Mehta SR, Delport W, Brouwer KC, Espitia S, Patterson T, Pond SK, Strathdee S a and Smith DM (2010) The relatedness of HIV epidemics in the United States-Mexico border region. AIDS Res Hum Retroviruses 26:1273–7.

Ministério da Saúde, Secretaria de Vigilância em Saúde and Departamento de DST A e HV (2005) Boletim Epidemiológico - Aids e DST.

Ministério da Saúde, Secretaria de Vigilância em Saúde and Departamento de DST A e HV (2014) Boletim Epidemiológico - Aids e DST.

Morgado MG, Guimarães ML, Neves Júnior I, dos Santos VG, Linhares-de-Carvalho MI, Castello-Branco LR, Bastos FI, Castilho E a, Galvão-Castro B and Bongertz V (1998) Molecular epidemiology of HIV in Brazil: polymorphism of the antigenically distinct HIV-1 B subtype strains. The Hospital Evandro Chagas AIDS Clinical Research Group. Mem Inst Oswaldo Cruz 93:383–6.

Morgado MG, Sabino EC, Shpaer EG, Bongertz V, Brigido L, Guimaraes MD, Castilho EA, Galvão-Castro B, Mullins JI and Hendry RM (1994) V3 region polymorphisms in HIV-1 from Brazil: prevalence of subtype B strains divergent from North American/European prototype and detection of subtype F. AIDS Res Hum Retroviruses 10:569–76.

Nadai Y, Eyzaguirre LM, Sill A, Cleghorn F, Nolte C, Charurat M, Collado-Chastel S, Jack N, Bartholomew C, Pape JW et al. (2009) HIV-1 epidemic in the Caribbean is dominated by subtype B. PLoS One 4:e4814.

Nau J-Y, Plantier J-C, Leoz M, Dickerson JE, De Oliveira F, Cordonnier F, Lemée V, Damond F, Robertson DL and Simon F (2009) A new human immunodeficiency virus derived from gorillas. Rev Med Suisse 5:1741.

Ng KT, Ong LY, Lim SH, Takebe Y, Kamarulzaman A and Tee KK (2013) Evolutionary history of HIV-1 subtype B and CRF01_AE transmission clusters among men who have sex with men (MSM) in Kuala Lumpur, Malaysia. PLoS One 8:e67286.

Okie S (2006) Fighting HIV--lessons from Brazil. N Engl J Med 354:1977–81.

Osmanov S, Pattou C, Walker N, Schwardländer B and Esparza J (2002) Estimated global distribution and regional spread of HIV-1 genetic subtypes in the year 2000. J Acquir Immune Defic Syndr 29:184–90.

Pagán I and Holguín Á (2013) Reconstructing the Timing and Dispersion Routes of HIV-1 Subtype B Epidemics in The Caribbean and Central America: A Phylogenetic Story. PLoS One 8:e69218.

Pape JW, Liautaud B, Thomas F, Mathurin JR, St Amand MM, Boncy M, Pean V, Pamphile M, Laroche AC and Johnson WD (1983) Characteristics of the acquired immunodeficiency syndrome (AIDS) in Haiti. N Engl J Med 309:945–50.

Parkin NT and Schapiro JM (2004) Antiretroviral drug resistance in non-subtype B HIV-1, HIV-2 and SIV. Antivir Ther 9:3–12.

Pennings PS, Holmes SP and Shafer RW (2014) HIV-1 transmission networks in a small world. J Infect Dis 209:180–2.

Perrin L, Kaiser L, Yerly S and Ag CRF (2003) Travel and the spread of HIV-1 genetic variants. Lancet Infect Dis 3:22–7.

Pinto ME, Schrago CG, Miranda a B and Russo C a M (2008) A molecular study on the evolution of a subtype B variant frequently found in Brazil. Genet Mol Res 7:1031–44.

Potts KE, Kalish ML, Lott T, Orloff G, Luo CC, Bernard MA, Alves CB, Badaro R, Suleiman J and Ferreira O (1993) Genetic heterogeneity of the V3 region of the HIV-1 envelope glycoprotein in Brazil. Brazilian Collaborative AIDS Research Group. AIDS 7:1191–1197.

Prosperi MCF, Ciccozzi M, Fanti I, Saladini F, Pecorari M, Borghi V, Di Giambenedetto S, Bruzzone B, Capetti A, Vivarelli A et al. (2011) A novel methodology for large-scale phylogeny partition. Nat Commun 2:321.

Prusiner SB (2002) Historical essay. Discovering the cause of AIDS. Science 298:1726.

Pybus OG and Rambaut A (2009) Evolutionary analysis of the dynamics of viral infectious disease. Nat Rev Genet 10:540–50.

Ragonnet-Cronin M, Hodcroft E, Hué S, Fearnhill E, Delpech V, Brown AJL and Lycett S (2013) Automated analysis of phylogenetic clusters. BMC Bioinformatics 14:317.

Ragonnet-cronin M, Ofner-agostini M, Merks H, Pilon R, Rekart M, Archibald CP, Sandstrom PA and Brooks JI (2010) Longitudinal Phylogenetic Surveillance Identifies Distinct Patterns of Cluster Dynamics. J Acquir Immune Defic Syndr 55:102–108.

Rambaut A, Posada D, Crandall KA and Holmes EC (2004) The causes and consequences of HIV evolution. Nat Rev Genet 5:52–61.

Rangel HR, Garzaro D, Gutiérrez CR, Vásquez L, Guillen G, Torres JR and Pujol FH (2009) HIV diversity in Venezuela: predominance of HIV type 1 subtype B and genomic characterization of non-B variants. AIDS Res Hum Retroviruses 25:347–50.

Rieder P, Joos B, von Wyl V, Kuster H, Grube C, Leemann C, Böni J, Yerly S, Klimkait T, Bürgisser P et al. (2010) HIV-1 transmission after cessation of early antiretroviral therapy among men having sex with men. AIDS 24:1177–83.

Rios M, Fernandez J, Jaramillo P, Paredes V, Sanchez JL, Laguna-Torres V a, Carr JK and Ramirez E (2005) Molecular epidemiology of HIV type 1 in Chile: differential geographic and transmission route distribution of B and F subtypes. AIDS Res Hum Retroviruses 21:835–40.

Rothenberg RB, Potterat JJ, Woodhouse DE, Muth SQ, Darrow WW and Klovdahl a S (1998) Social network dynamics and HIV transmission. AIDS 12:1529–36.

Russell KL, Carcamo C, Watts DM, Sanchez J, Gotuzzo E, Euler a, Blanco JC, Galeano a, Alava a, Mullins JI et al. (2000) Emerging genetic diversity of HIV-1 in South America. AIDS 14:1785–91.

Sa Filho D, Sucupira M, Casiero M, Sabino E, Diaz R and Janini L (2006) Identification of two HIV type 1 circulating recombinant forms in Brazil. AIDS Res Hum Retroviruses 1–13.

Salemi M, Gray RR and Goodenow MM (2008) An exploratory algorithm to identify intra-host recombinant viral sequences. Mol Phylogenet Evol 49:618–28.

Sanchez GI, Bautista CT, Eyzaguirre L, Carrion G, Arias S, Sateren WB, Negrete M, Montano SM, Sanchez JL and Carr JK (2006) Molecular epidemiology of human immunodeficiency virus-infected individuals in Medellin, Colombia. Am J Trop Med Hyg 74:674–7.

Santiago ML, Range F, Keele BF, Li Y, Bailes E, Bibollet-ruche F, Fruteau C, Noe R, Peeters M, Brookfield JFY et al. (2005) Simian Immunodeficiency Virus Infection in Free-Ranging Sooty Mangabeys (Cercocebus atys atys) from the Taı ¨ Forest , Co ˆte d ' Ivoire : Implications for the Origin of Epidemic Human Immunodeficiency Virus Type 2. J Virol 79:12515–12527.

Santoro-lopes G, Harrison LEEH, Tavares MD, Xexéo A, Santos ANACEDOS, Schechter M and Dos Santos AC (2000) HIV disease progression and V3 serotypes in Brazil: is B different from B-Br? AIDS Res Hum Retroviruses 16:953–958.

Santos AF a, Lengruber RB, Soares E a, Jere A, Sprinz E, Martinez AMB, Silveira J, Sion FS, Pathak VK and Soares M a (2008) Conservation patterns of HIV-1 RT connection and RNase H domains: identification of new mutations in NRTI-treated patients. PLoS One 3:e1781.

Sauter D, Schindler M, Specht A, Landford WN, Münch J, Kim K, Votteler J, Schubert U, Bibollet-Ruche F, Keele BF et al. (2009) Tetherin-driven adaptation of Vpu and Nef function and the evolution of pandemic and nonpandemic HIV-1 strains. Cell Host Microbe 6:409–21.

Segal UA, Elliott D and Mayadas NS (2009) Immigration Worldwide: Policies, Practices, and Trends.

Sharp PM, Bailes E, Chaudhuri RR, Rodenburg CM, Santiago MO and Hahn BH (2001) The origins of acquired immune deficiency syndrome viruses: where and when? Philos Trans R Soc Lond B Biol Sci 356:867–76.

Sharp PM and Hahn BH (2010) The evolution of HIV-1 and the origin of AIDS. Philos Trans R Soc B Biol Sci 365:2487–2494.

Sharp PM, Hahn BH and B PTRS (2010) The evolution of HIV-1 and the origin of AIDS The evolution of HIV-1 and the origin of AIDS. Phil Trans R Soc B 2487–2494.

Simon F, Mauclère P, Roques P, Loussert-Ajaka I, Müller-Trutwin MC, Saragosti S, Georges-Courbot MC, Barré-Sinoussi F and Brun-Vézinet F (1998) Identification of a new human immunodeficiency virus type 1 distinct from group M and group O. Nat Med 4:1032–7.

Soares E a JM, Santos RP, Pellegrini JA, Sprinz E, Tanuri A and Soares M a (2003) Epidemiologic and molecular characterization of human immunodeficiency virus type 1 in southern Brazil. J Acquir Immune Defic Syndr 34:520–6.

Taylor BS, Hammer SM, Mccutchan FE and D P (2008) The challenge of HIV-1 subtype diversity. N Engl J Med 359:1965–6.

Thompson MA, Aberg JA, Hoy JF, Telenti A, Benson C, Cahn P, Eron JJ, Günthard HF, Hammer SM, Reiss P et al. (2012) Antiretroviral treatment of adult HIV infection: 2012 recommendations of the International Antiviral Society-USA panel. JAMA 308:387–402.

Thomson MM, Delgado E, Herrero I, Villahermosa ML, Vázquez-de Parga E, Cuevas MT, Carmona R, Medrano L, Pérez-Alvarez L, Cuevas L et al. (2002) Diversity of mosaic structures and common

ancestry of human immunodeficiency virus type 1 BF intersubtype recombinant viruses from Argentina revealed by analysis of near full-length genome sequences. J Gen Virol 83:107–19.

Trivedi B (2010) The primate connection. Nature 466:S5.

UNAIDS (2014a) The gap report.

UNAIDS (2013) Report on the global AIDS epidemic. http://www.unaids.org/en/media/unaids/contentassets/documents/epidemiology/2013/gr2013/UNAIDS_Global_Report_2013_en.pdf.

UNAIDS (2011) World AIDS Day Report.

UNAIDS (2012) Report on the global AIDS epidemic.

UNAIDS (2014b) Local Epidemics Issues Brief.

Vallari A, Holzmayer V, Harris B, Yamaguchi J, Ngansop C, Makamche F, Mbanya D, Kaptué L, Ndembi N, Gürtler L et al. (2010) Confirmation of Putative HIV-1 Group P in Cameroon. J Virol 85:1403–7.

Van de Vijver DAMC, Wensing AMJ, Åsjö B, Bruckova M, Jorgensen LB, Camacho R, Horban A, Linka M, Lazanas M, Loveday C et al. (2010) HIV-1 drug-resistance patterns among patients on failing treatment in a large number of European countries. Acta dermatovenerologica Alpina, Pannonica, Adriat 19:3–9.

Vidal N, Peeters M, Mulanga-Kabeya C, Nzilambi N, Robertson D, Ilunga W, Sema H, Tshimanga K, Bongo B and Delaporte E (2000) Unprecedented degree of human immunodeficiency virus type 1 (HIV-1) group M genetic diversity in the Democratic Republic of Congo suggests that the HIV-1 pandemic originated in Central Africa. J Virol 74:10498–507.

Volz EM, Ionides E, Romero-Severson EO, Brandt M-G, Mokotoff E and Koopman JS (2013) HIV-1 Transmission during Early Infection in Men Who Have Sex with Men: A Phylodynamic Analysis. PLoS Med 10:e1001568.

Wertheim JO, Leigh Brown AJ, Hepler NL, Mehta SR, Richman DD, Smith DM and Kosakovsky Pond SL (2014) The global transmission network of HIV-1. J Infect Dis 209:304–13.

Wertheim JO and Worobey M (2009) Dating the age of the SIV lineages that gave rise to HIV-1 and HIV-2. PLoS Comput Biol 5:e1000377.

Wilbe K (2004) Genetic dynamics of HIV-1 : recombination , drug resistance and intrahost evolution.

Worobey M, Gemmel M, Teuwen DE, Haselkorn T, Kunstman K, Bunce M, Muyembe J-J, Kabongo J-MM, Kalengayi RM, Van Marck E et al. (2008) Direct evidence of extensive diversity of HIV-1 in Kinshasa by 1960. Nature 455:661–664.

Yamaguchi J, Devare SG and Brennan CA (2000) Identification of a new HIV-2 subtype based on phylogenetic analysis of full-length genomic sequence. AIDS Res Hum Retroviruses 16:925–930.

Ye J, Lu H, Su X, Xin R, Bai L, Xu K, Yu S, Feng X, Yan H, He X et al. (2014) Phylogenetic and temporal dynamics of human immunodeficiency virus type 1B in China: four types of B strains circulate in China. AIDS Res Hum Retroviruses 30:920–6.

Yerly S, Junier T, Gayet-Ageron A, Amari EB El, von Wyl V, Günthard HF, Hirschel B, Zdobnov E and Kaiser L (2009) The impact of transmission clusters on primary drug resistance in newly diagnosed HIV-1 infection. AIDS 23:1415–23.

Yirrell DL, Pickering H, Palmarini G, Hamilton L, Rutemberwa a, Biryahwaho B, Whitworth J and Brown a J (1998) Molecular epidemiological analysis of HIV in sexual networks in Uganda. AIDS 12:285–90.

Zehender G, Ebranati E, Lai A, Santoro MM, Alteri C, Giuliani M, Palamara G, Perno CF, Galli M, Lo Presti A et al. (2010) Population dynamics of HIV-1 subtype B in a cohort of men-having-sex-with-men in Rome, Italy. J Acquir Immune Defic Syndr 55:156–60.

Apêndices

Trabalhos Científicos realizados em período concomitante e

diretamente relacionados ao

tema central desta tese de Doutorado

Apêndice 01

"New insights into the *In Silico* prediction of HIV protease resistance to nelfinavir"

Dinler A. Antunes, Maurício M. Rigo, Marialva Sinigaglia, Rúbia M. de Medeiros, Dennis M. Junqueira, Sabrina E. M. Almeida, Gustavo F. Vieira

PLoS ONE, 2014

OPEN ACCESS Freely available online

PLOS ONE

New Insights into the *In Silico* Prediction of HIV Protease Resistance to Nelfinavir

Dinler A. Antunes[1,3], Maurício M. Rigo[1,3], Marialva Sinigaglia[1,3], Rúbia M. de Medeiros[2,3], Dennis M. Junqueira[2,3], Sabrina E. M. Almeida[2], Gustavo F. Vieira[1,3]*

1 Núcleo de Bioinformática do Laboratório de Imunogenética (NBLI), Departamento de Genética, Universidade Federal do Rio Grande do Sul. Porto Alegre, Rio Grande do Sul, Brazil, **2** Technological and Scientific Development Center (CDCT), State Foundation in Production and Health Research (FEPPS), Porto Alegre, Rio Grande do Sul, Brazil, **3** Programa de Pós-Graduação em Genética e Biologia Molecular (PPGBM), Universidade Federal do Rio Grande do Sul (UFRGS), Porto Alegre, Rio Grande do Sul, Brazil

Abstract

The Human Immunodeficiency Virus type 1 protease enzyme (HIV-1 PR) is one of the most important targets of antiretroviral therapy used in the treatment of AIDS patients. The success of protease-inhibitors (PIs), however, is often limited by the emergence of protease mutations that can confer resistance to a specific drug, or even to multiple PIs. In the present study, we used bioinformatics tools to evaluate the impact of the unusual mutations D30V and V32E over the dynamics of the PR-Nelfinavir complex, considering that codons involved in these mutations were previously related to major drug resistance to Nelfinavir. Both studied mutations presented structural features that indicate resistance to Nelfinavir, each one with a different impact over the interaction with the drug. The D30V mutation triggered a subtle change in the PR structure, which was also observed for the well-known Nelfinavir resistance mutation D30N, while the V32E exchange presented a much more dramatic impact over the PR flap dynamics. Moreover, our *in silico* approach was also able to describe different binding modes of the drug when bound to different proteases, identifying specific features of HIV-1 subtype B and subtype C proteases.

Citation: Antunes DA, Rigo MM, Sinigaglia M, de Medeiros RM, Junqueira DM, et al. (2014) New Insights into the *In Silico* Prediction of HIV Protease Resistance to Nelfinavir. PLoS ONE 9(1): e87520. doi:10.1371/journal.pone.0087520

Editor: Andrea Cavalli, University of Bologna & Italian Institute of Technology, Italy

Received September 16, 2013; **Accepted** December 22, 2013; **Published** January 31, 2014

Funding: This work was supported by Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPQ), Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES) and Fundação de Amparo à Pesquisa do Estado do Rio Grande do Sul (FAPERGS). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors have declared that no competing interests exist.

* E-mail: fioravanti.vieira@ufrgs.br

Introduction

Human immunodeficiency virus type 1 protease (HIV-1 PR) is a catalytic protein that cleaves the Gag and Gag-Pol viral polyproteins, allowing the virus to efficiently infect new host cells. The HIV-1 PR exists as an aspartyl homodimeric enzyme composed by symmetrical subunits of 99 amino acids each. The access of the substrate to the active site of PR is regulated by two mobile flaps that shift from an open to a closed conformation to bind and cleave the substrate.

The HIV-1 protease is one of the most important targets of antiretroviral therapy used in the treatment of AIDS patients due to its critical role in the viral replication cycle. Protease inhibitors (PI) were developed to inhibit cleavage function of HIV-1 protease by mimicking the reaction intermediates that arises during the hydrolysis of the substrate, disabling the enzyme. The current success of PIs is frequently limited by the emergence of protease gene mutations that confer resistance to this drug class. By changing the structure of the substrate-binding cavity, mutations directly or indirectly interfere with the binding of inhibitors, resulting in viral resistance to PIs.

According to the International AIDS Society, 23 mutations in 16 codons of the protease gene related to major drug-resistance to PIs were identified by phenotypic resistance assays [1]. In addition, it is currently known that polymorphisms in some codons not previously related to major drug-resistance could affect the viral fitness in the presence of the drug. Previous studies demonstrated that the viability to the arising of resistance mutations is generally dependent on the genetic background. Therefore, the genetic context in which the evolutionary variations arise in the protease gene may affect the efficacy of the treatment.

In this context, codons in the protease gene related to major drug resistance to a specific protease inhibitor can provide clues on the important sites to the interaction between drug and target, and it is possible that unusual changes in these same sites can also affect the interaction with the drug. For instance, D30N mutation causes high-level resistance to Nelfinavir (NF) [1,2] and V32I is associated to reduced susceptibility to all PIs, except Saquinavir [1,3]. However, the effect of the presence of alternative amino acids in these same sites is still unclear.

Due to the elevated costs and the extensive time required for *in-vitro* analysis, it is still impractical to use these conventional methods to evaluate the effect of each mutation in view of the genetic background of HIV-1 protease. Thus, computational methods can improve the screening analyzes revealing the role of individual mutations and its impact on the protein function [4–7]. In the present study, we used molecular dynamics and other bioinformatics tools aiming to identify structural features that could indicate the NF-resistance effect of the unusual mutations D30V and V32E, and to evaluate the influence of the HIV-1

genetic background (*i.e.* subtype B and subtype C) over these mutations.

Results

Sequence alignment, homology modeling and molecular docking

Complete identification for the subtype B wild-type (sB-WT) protease sequence, and for all other sequences included in this study, is provided in File S1. Sequence alignment confirmed the presence of mutations at positions 30 and 32, as well as other accessory mutations specific for each protease (Figure S1). All PR models presented 100% of their residues in the most favored regions of Ramachandran Plot (Table S1). Nelfinavir structure was successfully placed in the cavity of all models through molecular docking (Table S1).

Flap opening in a 10 ns MD with NF

Five independent 10 nanoseconds (ns) MD simulations were performed for each one of the four subtype B PR structures studied, sB-WT, sB-D30N, sB-D30V and sB-V32E, totaling 20 MD simulations (or 200 ns). No evident differences were observed in the Root Mean Square Deviation (RMSD) among all five replicated simulations of sB-WT, sB-D30N and sB-D30V (Figure 1 and Figure S2). In these three PR models, the overall conformation of the PR-NF complex (Figure S3) was sustained during the simulations, without significant changes in flaps orientation. On the other hand, the structure of sB-V32E presented a more unstable behavior when simulated with NF (Figure 1). Such model remained with the flaps in a closed conformation in 3 out of 5 simulations, but changed them to an open conformation in the other 2 simulations (Figure 2 and Figure S2).

Extension of selected complexes up to 50 ns

In order to verify if the differences observed among the complexes were not influenced by the short period of simulation, we extended one of each simulation from the proteases that remained in a closed conformation in the first 10 ns (sB-WT, sB-D30N and sB-D30V). For the sB-V32E, however, we extended all the three simulations that ended with a closed conformation. The sB-WT protease remained in a closed conformation bound to NF during 50 ns, while all three simulations of sB-V32E presented a change to an open conformation (Figure 3, Figure S4, Movie S1, Movie S2 and Movie S3). Complexes formed with sB-D30N and sB-D30V presented a change to a semiopen conformation during the simulated time (Figure 4).

Simulation of the V32E exchange without accessory mutations

In order to evaluate the isolated influence of the V32E mutation over the dynamics of the subtype B PR structure, we performed a 50 ns MD simulation of the sB-WT-V32E PR complexed with Nelfinavir. Consistent with all sB-V32E simulations, the sB-WT-V32E protease also changed to an open conformation (Figure 3).

Detailed structural analysis of sB-D30V during a 50 ns MD

Distance measurements between ASP25 (catalytic residue) and ILE50 (tip residue) (Figure S3) were calculated over the sB-D30V 50 ns simulation, and compared with the same measurements from sB-WT (susceptible to NF) and sB-D30N (resistant to NF). Both mutated proteases presented values above 1.58 nm for this distance during the second half of simulation, which is indicative of a semiopen conformation of the PR flaps (Figure 4) [4]. The wild-type PR presented values below 1.4 nm for the same measure (last 15 ns), consistent with a closed conformation. Distance ASP25-NF was also bigger for mutated complexes than for the WT (Figure S5). These distances were also calculated for Chain B (Figure S6), which did not present the same differences. Distances among other key residues were also calculated, highlighting a sequential interaction of PR residues with Nelfinavir (Figures S7, S8 and S9).

Hydrogen bonds between NF and sB-D30V

Analysis of subtype B 50 ns simulations clearly indicated the impact of PR residue 30 mutations over the hydrogen bonds network (Figure 5 and Table S2). After the first half of the simulation only the wild-type was able to sustain stable direct hydrogen bonding between residue 30 and the drug. Moreover, both mutated proteases presented a reduction of hydrogen bonding between the two flaps of the protease, and also between Nelfinavir and the catalytic ASP25 from Chain A (Figures S10, S11 and S12).

Dynamics of D30V and V32E mutations in the subtype C background

Molecular dynamics of subtype C structures complexed with Nelfinavir presented similar results to those observed for subtype B. Both sC-WT and sC-D30V remained in a closed conformation during the whole simulation (50 ns), while sC-V32E PR changed to the open conformation within the first 20 ns of simulation (Figure 6). However, the dynamic behavior of the flaps was different of that observed for sB-V32E (Figure S13, Movie S3 and Movie S4). The sC-V32E structure presented a periodic behavior, changing between open and closed conformation during the simulation.

Detailed structural analysis of sC-D30V during a 50 ns MD

Distance measurements between ASP25 and ILE50 were calculated over the sC-D30V 50ns simulation, and compared with the same measurements from sC-WT. The wild-type PR presented values below the stipulated threshold for the semiopen conformation (1.58 nm) while the mutated PR presented slightly greater values for the most part of the simulation. However, a clear difference between the two simulations is only observed in the first 15 ns (Figure S14A). Distances ASP25-NF presented greater values for subtype C proteases than that observed for sB-WT, with the mutated sC-D30V presenting lower values than the sC-WT (Figure S14B). The same distances were also calculated for Chain B, with similar results (Figure S15). Other distances and hydrogen bonds were also calculated, indicating a different network of residues responsible for the interaction between the drug and the two subtype C proteases (Figures S16 and S17).

Secondary structure analysis over a 50 ns MD

Remarkable conservation of secondary structure was observed for all models during the simulations, even considering the proteases that changed to an open conformation (Figure S18).

Molecular dynamics of apo proteases

Simulations of the unbound (apo) structure of all seven protease variants studied (sB-WT, sB-D30N, sB-D30V, sB-V32E, sC-WT, sC-D30V and sC-V32E) were performed, in order to evaluate the dynamics of each enzyme in the absence of Nelfinavir. As expected, all structures changed to an open conformation before 50 ns of simulation (Figure S19). In order to evaluate the strength

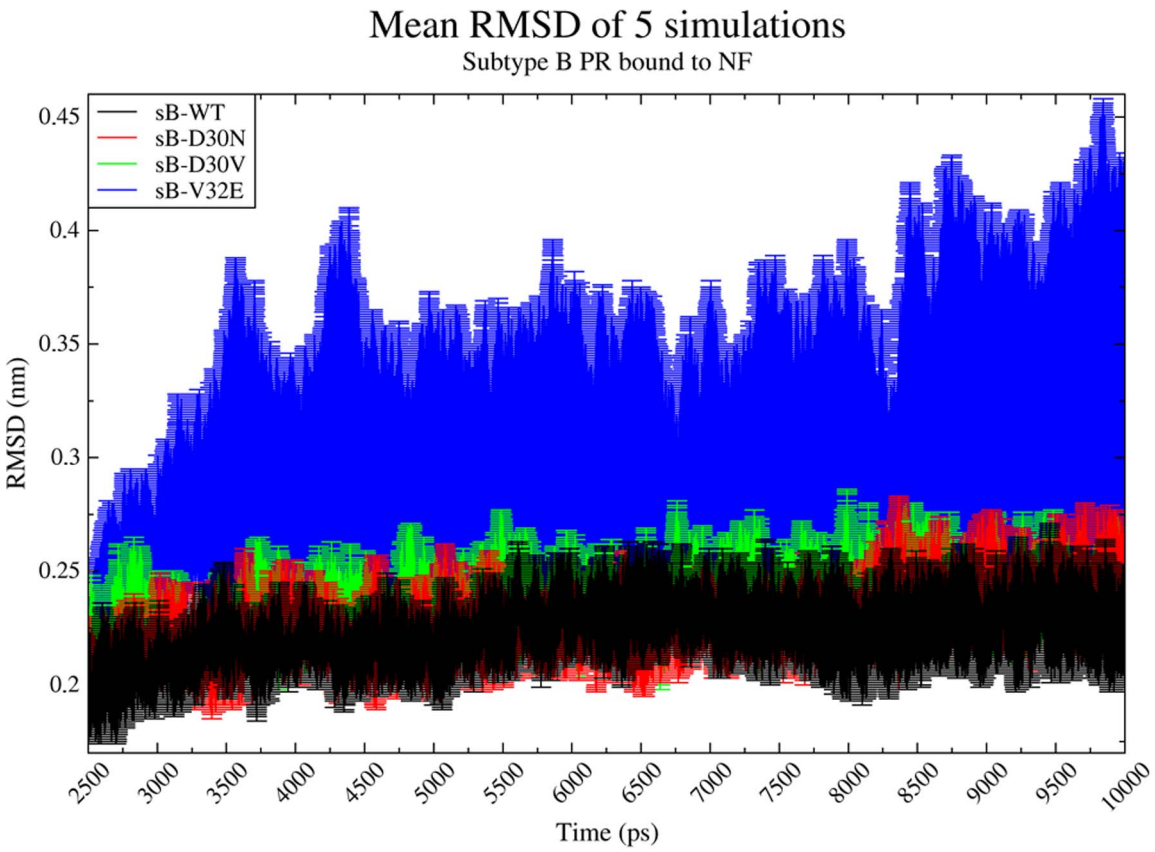


Figure 1. Short replicated simulations of sB-PRs bound to NF. Average and Standard Deviation of the Root Mean Square Deviation (RMSD) for five independent 10 ns simulations of four different subtype B proteases (sB-PRs) bound to Nelfinavir (NF). Greater divergence is observed for sB-V32E, since two of its replicates presented a change to an open conformation of the flaps. Equilibration stages (before 2,500 ps) are not represented. Independent trajectories of each simulation can be observed in Figure S2.
doi:10.1371/journal.pone.0087520.g001

of this tendency a replicate simulation was performed for all proteases, and two of the replicates (one sB-D30N-apo and one sC-WT-apo) did not present the change in the same period (data not shown). Of note, the two "V32E" proteases changed to an open conformation before 5 ns of simulation (and also its replicates), while other proteases presented this conformational shift later on the simulation.

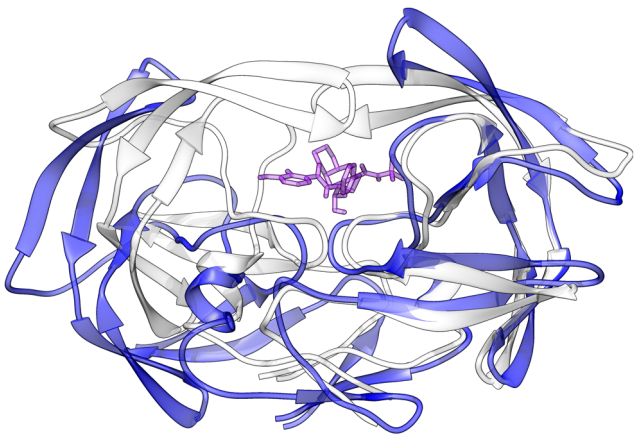


Figure 2. Open conformation of the sB-V32E protease. Comparison between two frames of a molecular dynamics of the sB-V32E PR complexed with NF. Protease structure at 2,500 ps (25 ns) is represented in white (*cartoon*) with Nelfinavir depicted in purple (*sticks*). Protease structure at 50,000 ps (50 ns), in an open conformation, is depicted in blue (*cartoon*).
doi:10.1371/journal.pone.0087520.g002

Conformational variability of Nelfinavir

A longer simulation (100 ns) of unbound Nelfinavir (in solution) was performed in order to see all different conformations adopted and to sample low energy poses of the drug. Using Free Energy Surface (FES) analysis, we were able to identify three different islands of low energy level, each one presenting a different pool of conformations, and all of them different from that conformation observed in crystal structures (Figure 7). The structure with the lowest energy in the entire simulation was observed in the first island (NF-i1). Interestingly, the conformation extracted from the second island (NF-i2) was similar to NF-i1, but with a subtle conformational change which allowed the formation of an internal hydrogen bonding.

Great divergence for the crystal structure was also observed in a simulation of Nelfinavir bound to sB-WT (Figure S20). FES analysis for this simulation presented two low energy islands (Figure S21) and the structures recovered from these islands match with the low energy structures sampled from Nelfinavir unbound 100 ns simulation (NF-i1 and NF-i2). Similar conformations were also observed for sC-WT. Different patterns of FES were observed for Nelfinavir dynamics when bound to each one of different variants, maintaining similar patterns for similar mutations. In the

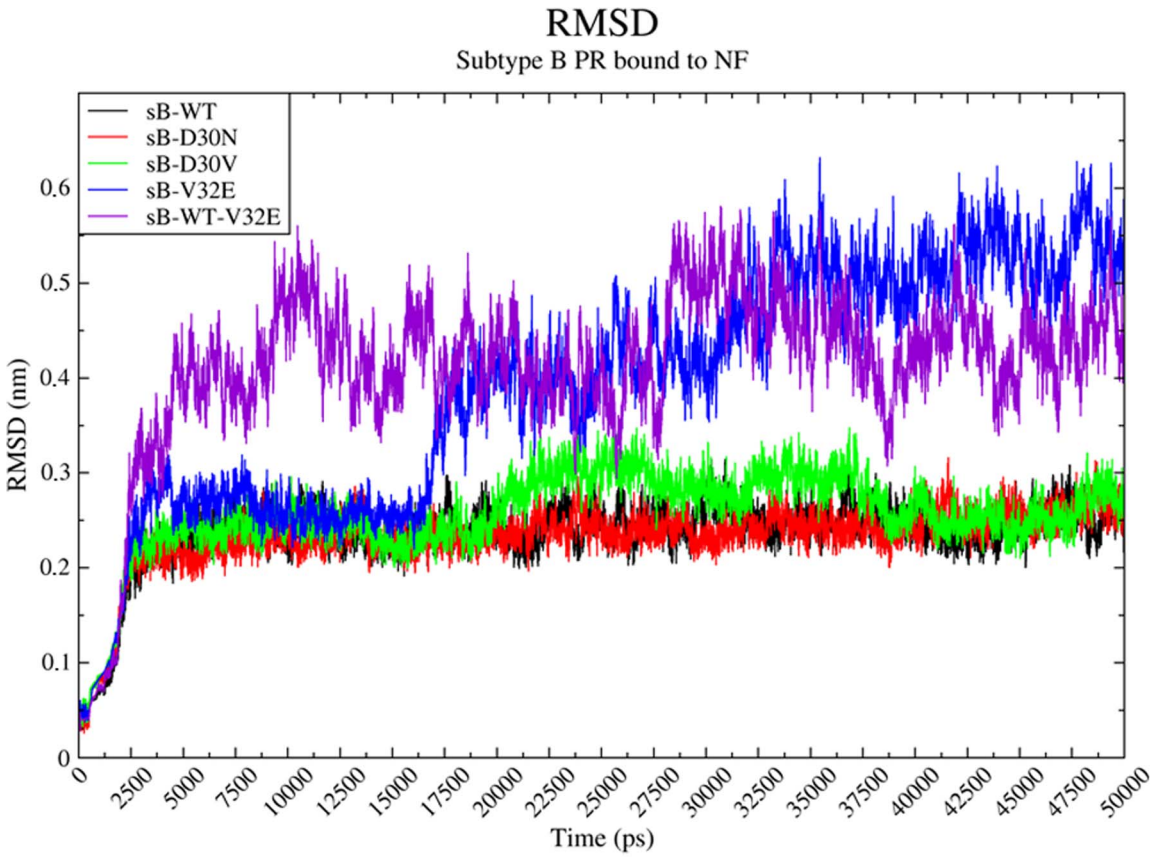


Figure 3. Molecular dynamics of sB-PRs bound to NF. Root Mean Square Deviation (RMSD) of subtype B (sB) proteases bound to Nelfinavir (NF) along 50 ns of molecular dynamics simulation. The colors are given in black, red, green, blue and purple for the wild-type (sB-WT), D30N (sB-D30N), D30V (sB-D30V), V32E (sB-V32E) and wild-type-V32E (sB-WT-V32E), respectively. It is important to note that while sB-WT, sB-D30N and sB-D30V seems to remain in a closed conformation state, sB-WT-V32E and sB-V32E change to an open conformation of the flaps in the first 5 and 15 ns, respectively.
doi:10.1371/journal.pone.0087520.g003

case of proteases with mutations at residue 30, Nelfinavir presented a structure similar to NF-i3 (Figure S20 and Figure S21).

Free Energy Surface (FES) of all simulated systems

FES analysis (File S1) was also performed for all proteases, both bound to Nelfinavir and in the apo form (Figure S22). As expected, all proteases that evolved to a full-open conformation (including apo simulations) presented a similar pattern of FES. Important differences were observed when comparing results among subtypes.

Discussion

In a previous study, de Medeiros *et al.* (2011) [8] evaluated the profile of mutations and polymorphisms in the protease (PR) and reverse transcriptase (RT) genes of HIV-1 from untreated patients living in Porto Alegre, Southernmost Brazil, in order to identify the subtypes and circulating drug resistant genotypes. Two unusual protease mutations – D30V and V32E – were identified, and its effect on drug resistance have not yet been evaluated *in vitro*. Supported by a careful comparison with sB-WT and sB-D30N simulations, we were able to predict the specific impact of each one of these unusual mutations over the structure of subtype B and subtype C proteases, and its effect over the interaction with the protease inhibitor Nelfinavir. These mutated PR structures were not produced by just exchanging these positions in the 1OHR structure (sB-WT), since this procedure would ignore accessory mutations and create PR structures that were not observed *in vivo*. Instead, we choose to model PR structures based on complete sequences previously obtained by de Medeiros *et al.* (2011) [8].

Nelfinavir is a protease inhibitor largely used as part of the treatment given to HIV-1 infected patients [9]. As most of the protease inhibitors, its function is to establish a stable network of hydrogen bonds with different PR residues, keeping the flaps in a closed conformation and blocking the catalytic site to the access of the substrate [10]. As expected, this mechanism was seen in all simulations of wild-type proteases, either performing five independent 10 ns MD replicates of sB-WT (Figure 1), or 50 ns simulations of sB-WT and sC-WT (Figure 3). The PR structure remained in a closed conformation in all these cases.

As an additional control, we also performed simulations of unbound (apo) structures of each protease studied (Figure S19). A change to an open conformation was observed in 86% of apo simulations (12 out of 14), reflecting a natural tendency of opening in the absence of the drug. On the other hand, in the 2 out of 14 simulations, protease stayed in a closed conformation up to 50 ns, suggesting that this opening trend is also strongly influenced by random events such as the influx of water molecules.

The impact of V32E mutation was clearly demonstrated by our simulations, always triggering the change of the PR structure to an open conformation within a period of 50 ns. Due to random features of the system, this important conformational change can happen even earlier in the simulation, as observed in some 10 ns

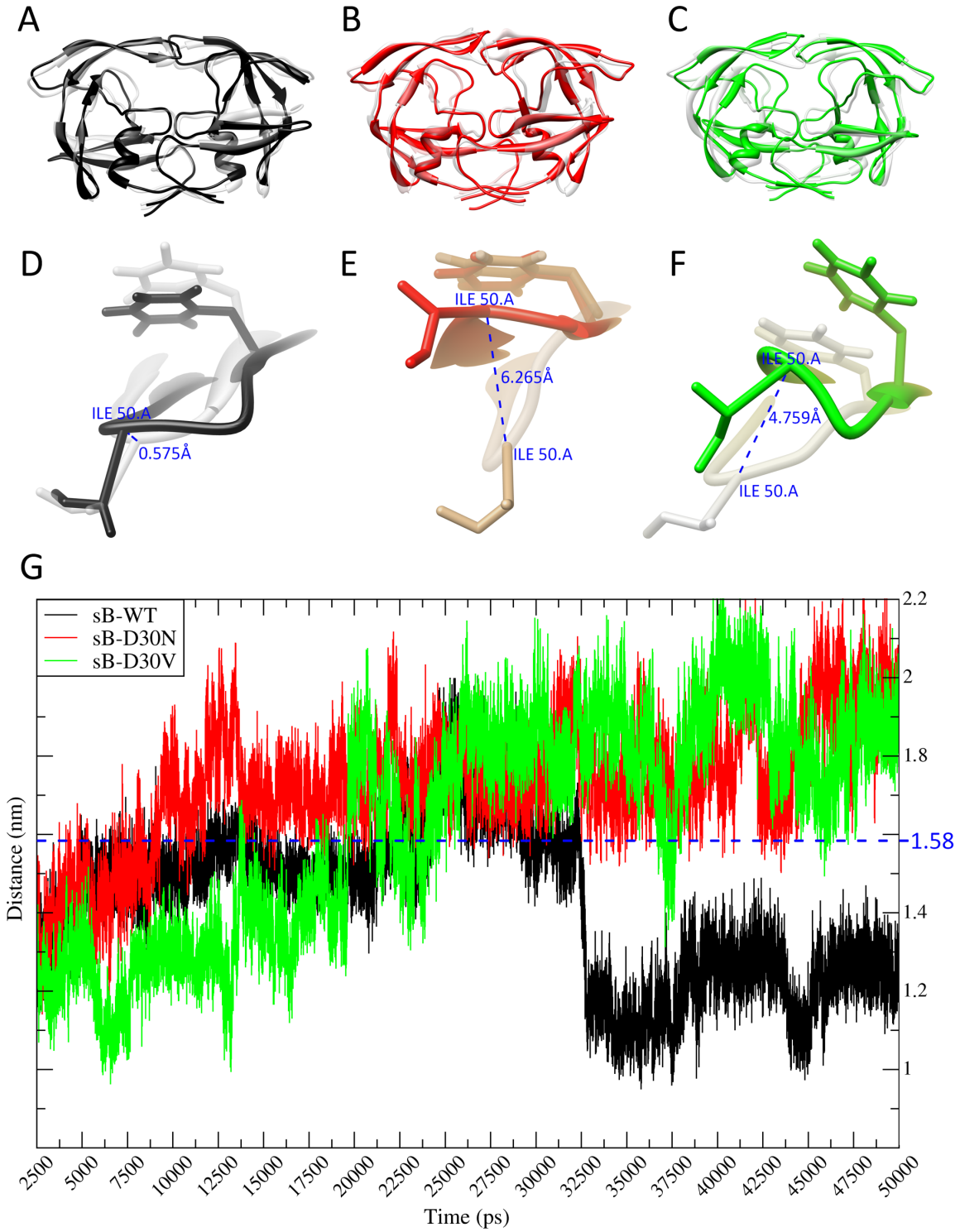


Figure 4. Structural analysis of sB-D30V. (A) Superposition of the structures of sB-WT PR at 2,500 ps of simulation (grey) and at 50,000 ps (black). (B–C) Superposition of sB-D30N (red) and sB-D30V (green) at 50,000 ps over the respective structures at 2,500 ps (grey). (D–F) Measure of the deviation of ILE50 residue from PR Chain A considering the same structures from A, B and C, indicating the extent of Chain A flap movement. (G) Plot of the variation of the ASP25-ILE50 distance (Chain A) along the simulation. The stipulated threshold for semiopen conformation (1.58 nm) is indicated in blue.
doi:10.1371/journal.pone.0087520.g004

simulations (Figure S2). The same "opening behavior" was observed even when this residue was simulated without any accessory mutations (sB-WT-V32E) or in the context of a subtype C protease (sC-V32E) (Figure 2 and Figure 5). However, other residues can certainly modulate the influence of V32E mutation, as observed by the differential behavior between sB-V32E and

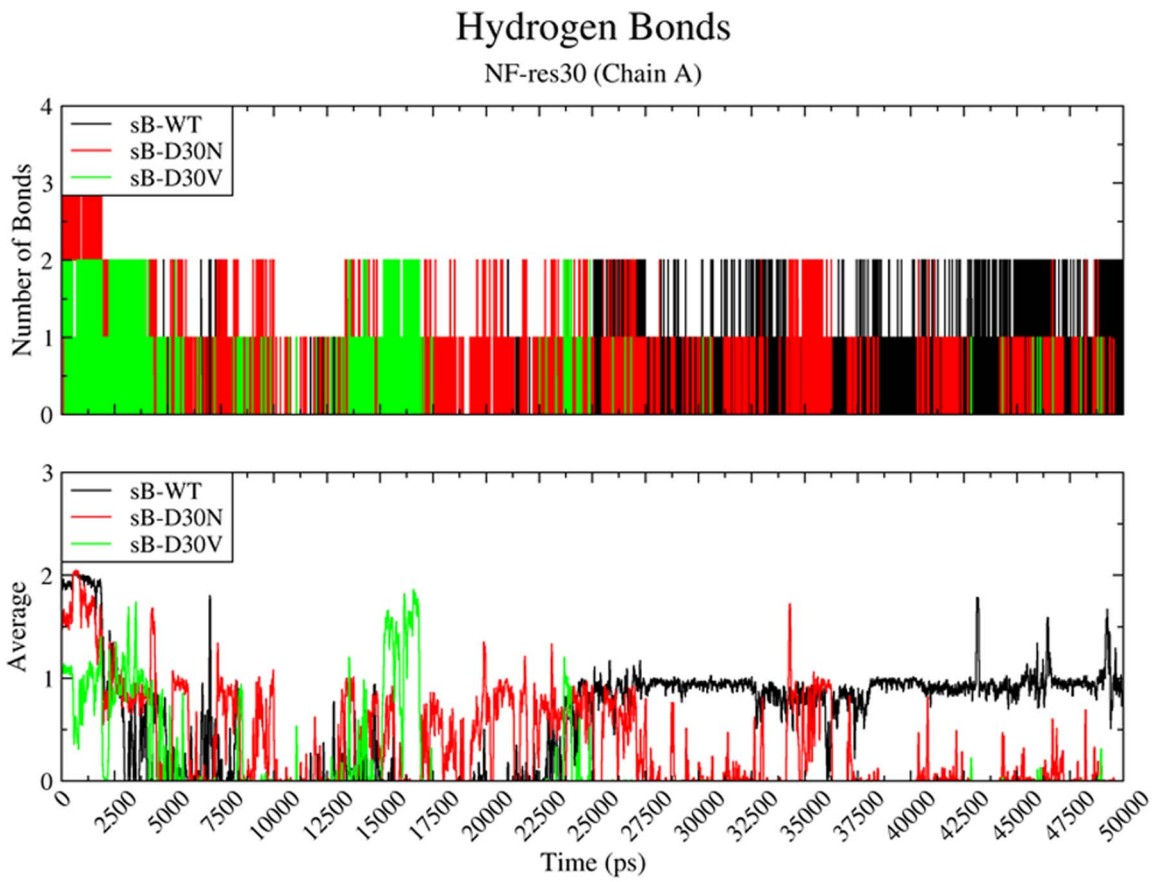


Figure 5. Hydrogen bonds between NF and residue 30 of sB-PRs. Number (above) and average (below) of hydrogen bonds performed between the ligand Nelfinavir and the residue 30 of each subtype B (sB) protease (Chain A) along 50 ns of molecular dynamics simulation. The colors are given in black, red and green for the wild-type (sB-WT), D30N (sB-D30N) and D30V (sB-D30V), respectively.
doi:10.1371/journal.pone.0087520.g005

sC-V32E (Figure S13). This mutation does not prevent direct interactions with the drug (Table S2) and has also no direct impact on the secondary structure of the enzyme (Figure S18). A slightly faster opening behavior was observed for apo proteases bearing this mutation (Figure S19), which could indicate a direct effect on flap stability.

The well-known D30N exchange has been largely studied and described as the primary Nelfinavir resistance mutation [2,6,11–13]. In our 10 ns simulations, this mutation did not affect the flap dynamics of the PR-NF complex, which stayed in a closed conformation. This result is in agreement with previous 10 ns MD data published by Soares, *et. al* 2010 [6], in which the same sB-D30N PR remained in a closed conformation when simulated bound to NF or to the Gag substrate CA/p2. The unusual D30V mutation presented similar RMSD results to those obtained for D30N, both in 10 ns and 50 ns simulations (Figure 1 and Figure 3).

Supported by a series of data [14,15], Perryman *et al.* 2003 [4] discussed that PR flap dynamics is involved with the enzymatic mechanism itself (File S1). These data indicated that the activation free energy barrier of the enzymatic reaction is highly sensitive to the distance between the substrate and the catalytic aspartates, and that the motion of the substrate toward these catalytic residues is tightly coupled to dynamics of the flap tips. They used the ASP25-ILE50 distance (Figure S3) to observe the extent of the flap opening during the MD simulations, defining the ASP25-ILE50 distance from the non-bonded (apo form) semiopen crystal structure 1HHP (1.58 nm) as a threshold to identify snapshots of semiopen conformations (*see* File S1).

In agreement with the discussion from Perryman *et al.* 2003 [4], we also observed differential flap tip motions for the PR variants when compared to the wild-type. Measurements of the distance ASP25-ILE50 (Chain A) presented bigger values for both subtype-B mutated PRs (sB-D30N and sB-D30V), above the threshold of 1.58 nm, which is consistent with the movement toward a semiopen conformation of the flaps (Figure 3G). While the wild-type PR presents almost the same conformation comparing these two snapshots of the simulation (Figures 3A and 3D), both sB-D30N and sB-D30V presented an important opening of the Chain A flap (Figures 3B, 3C, 3E and 3F). Greater motions observed in Chain A when compared to Chain B, for all complexes, are also consistent with the finds from Perryman *et al.* 2003 [4]. The sB-WT complex slowly increased the measured distance during the first half of the simulation and even presented values above the threshold, but suffered a fast accommodation process before 32,500 picoseconds (ps), remaining below the threshold for the rest of the simulation. This accommodation was driven by a sequence of interactions, starting with the hydrogen bonding between ASP30(O)-NF(O46) and ASP25(OD1)-NF(N37), both starting around 22,500 ps (Figures S7 and S10). After that, temporary interaction between ILE50(O) and NF(O8) also seems to help the closing of the Chain A flap. Finally, increase formation of hydrogen bonds between the two flaps help to keep the structure in a closed conformation. This sequence of interactions was not

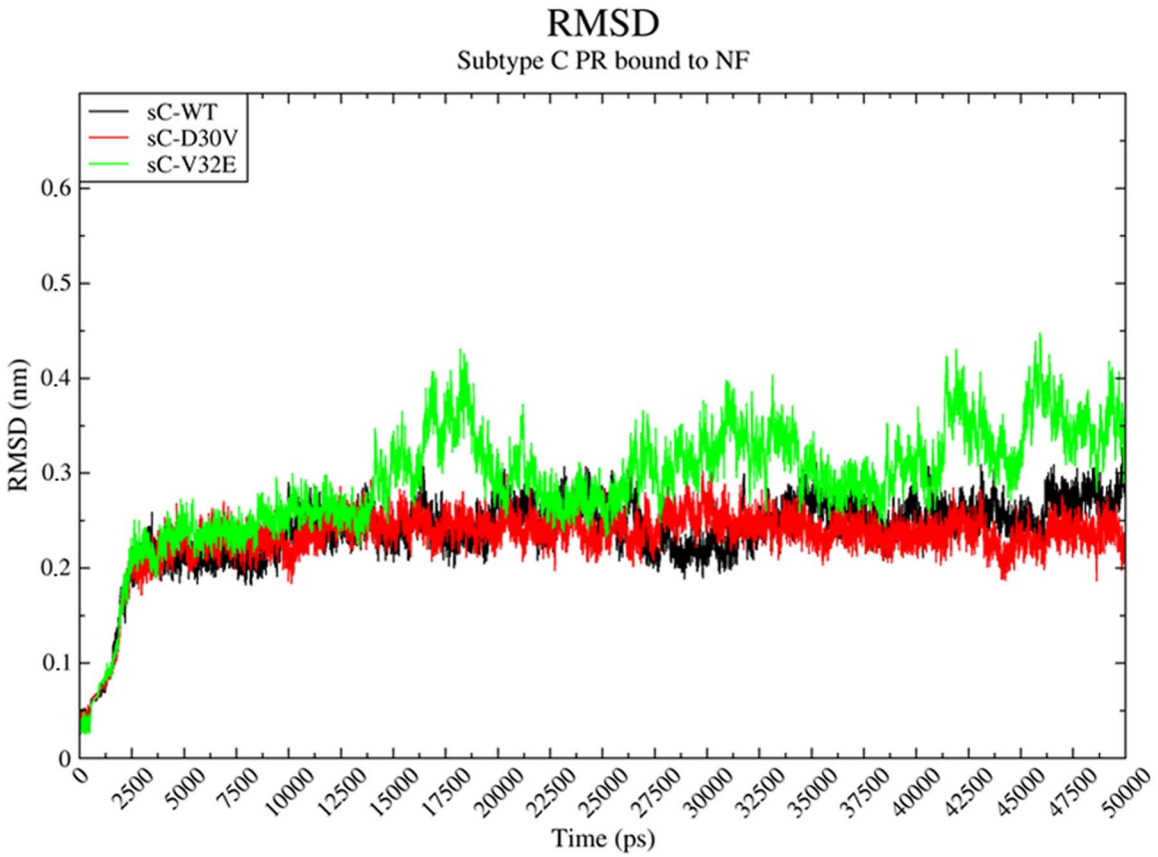


Figure 6. Molecular dynamics of sC-PRs bound to NF. Root Mean Square Deviation (RMSD) of subtype C (sC) proteases bound to Nelfinavir (NF) along 50 ns of molecular dynamics simulation. The colors are given in black, red and green for the wild-type (sC-WT), D30V (sC-D30V) and V32E (sC-V32E), respectively. Note that the sC-V32E RMSD behavior points to an open conformation state of the protease within the first 20 ns, alternating between open and closed conformation along the simulated period. The sC-WT and sC-D30V PRs remained in a closed conformation along the entire simulation.
doi:10.1371/journal.pone.0087520.g006

observed in sB-D30N and sB-D30V, which were not able to sustain the hydrogen bonding between the mutated residue and the drug (Figures S8, S9, S11 and S12).

In order to have some clues about the differential interaction of these variants with the same inhibitor, we evaluated the distance of the catalytic aspartates (ASP25 and ASP124) to the drug (Figures S5 and S6B). What we verified is that indeed Nelfinavir stays much more close to the PR catalytic residues in the cavity of sB-WT than in the cavity of a well-known NF-resistant variant (sB-D30N). The same differential distance was also observed for the unusual sB-D30V variant, corroborating the hypothesis of resistance to Nelfinavir. Moreover, this distance seems to be increasing for the mutants in the second half of the simulation, which combined with semi-open conformations of the flaps might also be a sign of instability.

Interestingly, the subtype C wild-type protease (sC-WT) presented a completely different network of hydrogen bond interactions with Nelfinavir (Figure S16A). Differently from subtype B PRs, in which the side Chain of ASP30 is oriented toward the catalytic site of the enzyme, in sC-WT the side Chain of this residue is oriented in the opposite direction (consistent with PDB crystal structures 2R5P and 2R5Q). Therefore, despite the presence of the same "ASP30", hydrogen bonds between this residue and the drug were observed only in the first 3,750 ps of sC-WT simulation. Shortly after that, Nelfinavir establishes key interactions with other PR residues, such as ALA28 and ILE149, which are able to keep the PR in a closed conformation (Figures S16A and S17A). Little difference was observed between the two sC PRs regarding the ASP25-ILE50 distance (Figures S14A and S15A) and the wild-type presented even higher values for the ASP25-NF interaction (Figure S14B and S15B). However, this mutated complex presented some signs of instability. The key interactions observed in the sC-WT (ALA28-NF and ILE149-NF) were not observed in the sC-D30V (Figure S16B), and there are some previous evidence suggesting low levels of resistance to PIs [16].

Different networks of interaction suggest different binding modes for Nelfinavir. Aiming to explore the dynamic behavior of the drug, we performed a 100 ns simulation of Nelfinavir in solution. Free Energy Surface (FES) analysis of this simulation indicated three different low energy conformations of the drug, all different from the crystal structure (Figure 7). Nelfinavir conformation also diverged from crystal structure in all PR-NF simulated systems (Figure S20). A translation of NF heterocyclic portion (lipophilic dodecahydroisoquinoline ring) [17] was observed early in simulations, adopting a structure more similar to that of NF-i3. Of note, a similar conformation was prevalent for all "D30 mutants", while both wild-type proteases presented a conformation similar to NF-i2 at the end of simulation. It was actually possible to separate Nelfinavir dynamics by the type of protease it is bound to (Figure S20). Proteases sB-V32E and sB-WT-V32E changed to a full-open conformation before 20 ns of simulation

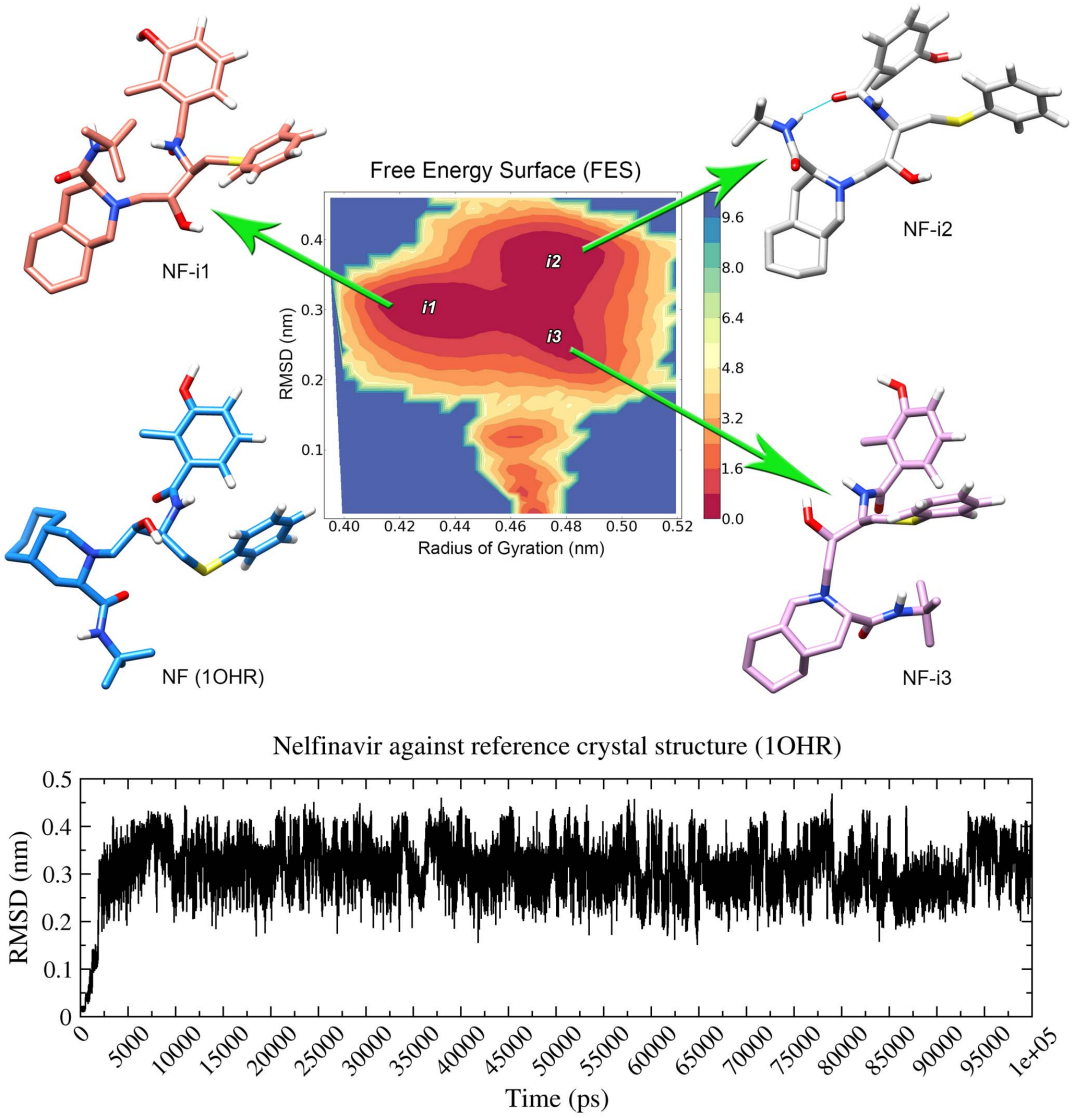


Figure 7. Conformational variability of Nelfinavir. Structural analysis of Nelfinavir in solution along 100 ns of molecular dynamics simulation. Free Energy Surface (FES) of this simulation (top) indicates three "islands" of low energy conformations (i1, i2 and i3), from which different structures were recovered (NF-i1, NF-i2, NF-i3). The crystal structure of Nelfinavir (1OHR) was the input conformation, and Root Mean Square Deviation (RMSD) indicates that all conformations sampled during the simulation differ from original structure by at least 0.2 nm (down).
doi:10.1371/journal.pone.0087520.g007

and their peaks of divergence to NF-i2 are only observed after 30 ns of simulation, being therefore a consequence of flap opening. In the case of sB-WT, the fast conformational change observed for the drug after 30 ns is a consequence of the hydrogen bonds established with ASP25 and ASP30, which formed around 25 ns. These bonds stabilized the drug in the catalytic site and allowed its change to a lower energy conformation, also contributing for protease flaps closing observed after 30 ns.

FES analysis also provided new insights regarding to protein structure and its evolution throughout the simulation. It indicates a progressive increase of RMSD and Radius of Gyration (RoG) in the early stages of sB-WT simulation, which accounts for the diagonal displacement of the tail bellow the island of low energy conformation. Although this island is similar among sB-WT, sB-D30N and sB-D30V, the diagonal displacement of the tail was not observed in the last two structures. This result indicates that subtle structural differences were already in place in early stages of subtype B mutants' simulations, reflecting in a higher value of RoG and leading to a semiopen conformation. Interestingly, distinct patterns were observed for FES of subtype C proteases bound to Nelfinavir, which is probably related to the different network of interactions observed for subtype C proteases. Despite less clear to see, sC-D30V also presented a more vertical displacement of the tail, which is similar with their subtype B counterparts.

Taken together, these PR-NF simulations indicated different patterns of interaction for the same protease inhibitor and, consequently, different impacts of the studied mutations in the context of different HIV-1 subtypes. Moreover, the V32E mutation seems to have a stronger and individual effect over the protease dynamics, inducing flap opening and therefore predisposing to drug dissociation. The previous described V32I mutation has confirmed resistance to multiple PIs [1,3], which could be explained by a similar mechanism of intrinsic propensity to an

open conformation. Future studies will be needed to clarify the shared features between V32E and V32I PRs, and if the unusual V32E also presents resistance to multiple PIs. While the V32E mutation seems to induce Nelfinavir resistance in both studied subtypes, our data only clearly indicated the D30V resistance effect in the subtype B background. Of note, the previous described D30N mutation also has a specific effect on Nelfinavir resistance on subtype B PRs, appearing in lower frequencies in other subtypes [1,9].

Our data suggest a cross talk between Nelfinavir dynamics inside the binding site and protease structure (specific residues, hydrogen bond network, etc). For instance, loss of direct hydrogen bonding with ASP30 may hinder the Nelfinavir's transition to some low energy conformations (similar to NF-i2), driving the drug to alternative lower energy conformations (similar to NF-i3). In turn, this might influence the transition for a semiopen conformation, as observed for sB-D30N, sB-D30V and sC-D30V. Previous works have reported subtle conformational changes in Nelfinavir [13,18–20] and even predicted alternative binding modes to other kinases [17], but a much greater structural variation was observed in our 50 ns and 100 ns simulations. This great conformational flexibility of Nelfinavir might also be involved with additional activities, such as broad antitumor properties [17,21].

The PR sequences studied in the present work were obtained from untreated patients, through pro-viral DNA sequencing [8]. Starting from these sequences, our *in silico* analysis was able not only to identify structural features which differentiate these PR-variants and the wild-type, but also to describe the influence of the HIV-1 genetic background (*I.e.* subtype B and subtype C) over these important NF resistance related codons. Molecular dynamics is a powerful tool which has been largely used to identify features of the PR enzyme and the molecular basis for resistance to PIs [4–7,13,22,23]. The identification of key features involved in the efficacy of different PIs in different HIV-1 subtypes and the fast evolution of computational resources would allow performing extensive *in silico* analysis in a short time and at low cost. Therefore, bioinformatics tools could be applied in association with conventional clinical methods as a virtual screening tool for the impact of new mutations, allowing individualized regimen of antiretrovirals, avoiding treatment failure and promoting durable remission of HIV-1.

Materials and Methods

HIV-1 PR sequences

HIV-1 PR sequences were selected from a set constructed in a previous epidemiologic study conducted in Porto Alegre (Southernmost Brazil) where HIV-1 subtype B and C circulates in equal proportions [8]. This set included 99 partial *pol* sequences of proviral DNA extracted from HIV-1-positive patients not under antiretroviral therapy. In our study, the selection criteria involved the inclusion of sequences harboring mutations not described as resistance mutations but occurring in a codon effectively related to major drug resistance present in both subtype B and subtype C.

Complete information on the selected sequences is provided on File S1. DNA sequences were translated with Expasy translate tool [24] and all protein sequences were aligned with Geneious version 5.1.4 (File S1). An alignment presenting all studied sequences and highlighting the mutations in relation to the wild-type subtype B HIV-1 PR (sB-WT) is presented in Figure S1.

Protease models and ligand parameters

The 3D structure of the sB-WT complexed with NF was obtained from Protein Data Bank (PDB code 1OHR). This structure was used as the initial coordinates for the molecular dynamics (MD) simulation of the sB-WT and also as the template for the molecular homology modeling (including the crystallographic water molecules) of all other PR models studied. Models were generated with Modeller 9.11 [25] (*see* File S1). Models were evaluated with PROCHECK [26] and DOPE score [25]. Therefore, in addition to sB-WT (crystal structure itself), seven PR structures were modeled using 1OHR as template: Subtype B D30V (sB-D30V), Subtype B V32E (sB-V32E), Subtype B D30N (sB-D30N), Subtype C wild-type (sC-WT), Subtype C D30V (sC-D30V), Subtype C V32E (sC-V32E) and sB-WT-V32E (same sequence of 1OHR containing only the V32E mutation).

Atom coordinates of Nelfinavir were obtained from 1OHR. NF parameters were obtained from PRODRG server [27] and from the full NF topology previously calculated and made available by Soares *et al.* 2010 [6] (*see* File S1).

Docking calculations

Modeled PR structures were complexed with Nelfinavir through molecular docking with Autodock Vina 1.1.2 [28] (*see* File S1). The starting coordinates of NF were used for redocking with sB-WT (1OHR) and cross-docking with all generated models. Docking calculations were independently repeated 20 times using the same input, and the best conformation was selected through an automated script developed by our team (File S1). Since the aim of these calculations was just to provide input structures for molecular dynamics simulations, only rigid dockings were performed.

Molecular Dynamics (MD) simulations

All MD simulations were performed with GROMACS v4.5.1 package [29], on Linux platform (Ubuntu 10.10), using GROMOS96 (*53a6*) force field. An appropriate number of sodium (Na+) and chloride (Cl-) counter-ions were added to neutralize the system, with final concentration of 0.15 mol/L. In agreement to the literature, only the catalytic residue ASP124 of PR was protonated [6,30–32]. Further details are provided in File S1. Visual inspection of the MD trajectories was performed with VMD 1.9.1 [33], PyMOL 1.0 [34] and UCSF Chimera [35] (*see* File S1).

Supporting Information

Figure S1 Sequence Alignment. Complete sequences from eight different proteases are depicted with colors indicating biochemical properties of the amino acids (Geneious colors by Polarity). The Identity of the alignment is depicted above the sequences, with regions of 100% identity depicted in green. Both Chains from each protease are depicted sequentially, with Chain A residues ranging from 1 to 99 and Chain B from 100 to 198. Both Chains have the exact same sequence, with residues 124, 129, 131 and 149 being the Chain B equivalents for the Chain A residues 25, 30, 32 and 50, respectively. sB-WT, Subtype B wild-type; sB-WT-V32E, Subtype B wild-type with V32E mutation; sB-D30N, Subtype B D30N; sB-D30V, Subtype B D30V; sB-V32E, Subtype B V32E; sC-WT, Subtype C wild-type; sC-D30V, Subtype C D30V; sC-V32E, Subtype C V32E.
(TIF)

Figure S2 Replication of 10 ns molecular dynamics of sB-PR bound to NF. Root Mean Square Deviation (RMSD) of

the subtype B (sB) protease bound to Nelfinavir (NF) along 10 ns of molecular dynamics simulation. Each graph contains five replicates (represented by the numbers within the brackets in the legend box) of the same protease:Nelfinavir complex depicted in shades of black (sB-WT), red (sB-D30N), green (sB-D30V) and blue (sB-V32E). Note that within 10 ns the sB-V32E changes its conformation from closed to an open state in two out of five replicates (1 and 4).
(TIF)

Figure S3 *Cartoon* representation of the 1OHR crystal structure (sB-WT in a closed conformation). Chains A and B are depicted in different shades of gray. Residues Aspartate 25 (ASP25) and Isoleucine 50 (ILE50) of Chain A are depicted in blue, and the distance between these two residues is also indicated. Residues 129 and 131 (ASP30 and VAL32 from Chain B, respectively) are depicted in green. Protease flaps (residues 43–59) are depicted in dark red, with the tip of the flaps (residues 48–53) represented in light red.
(TIF)

Figure S4 Reproduction of sB-V32E simulation results. Root Mean Square Deviation (RMSD) of three replicas (2, 3 and 5) of the sB-V32E protease bound to Nelfinavir (NF) along 50 ns of molecular dynamics simulation. Each simulation is identified by the same color used in Figure S2. All replicas also changed its conformation to an open state before 50 ns, but each one at a different point of the simulation.
(TIF)

Figure S5 Drug-enzyme distance. Distance variation between Aspartate 25 (ASP25, Chain A) of subtype B proteases and Nelfinavir along 50 ns of molecular dynamics simulation. The colors are given in black, red and green for the wild-type (sB-WT), D30N (sB-D30N) and D30V (sB-D30V), respectively. We could observe an increase in the distance along the simulation for sB-D30N and sB-D30V, pointing to a less effective interaction between the drug and the enzyme.
(TIF)

Figure S6 Distance measurements in Chain B of sB-PRs. Colors are given in black, red and green for the wild-type (sB-WT), D30N (sB-D30N) and D30V (sB-D30V), respectively. (A) Distance variation between Aspartate 124 (ASP25 from Chain B) and Isoleucine 149 (ILE50 from Chain B) along 50 ns of molecular dynamics simulation. No difference is observed among the complexes. (B) Distance variation between Aspartate 124 (ASP25, Chain B) and Nelfinavir in the same period of simulation. Both sB-D30N and sB-D30V have presented slightly bigger distance variation than sB-WT.
(TIF)

Figure S7 Interactions with key residues from sB-WT. Distance variation among the drug (NF) and selected atoms of key residues in the subtype B wild-type (sB-WT) Chain A structure along 50 ns of molecular dynamics simulation. The colors are given in black, gray, light pink and beige for the interaction pairs Isoleucine 50/Aspartate 25 (depicted in black in Figure 4), Aspartate 30(O)/NF(O46), Aspartate 25(OD1)/NF(N37) and Isoleucine 50(O)/NF(O8), respectively. O, Oxygen; O46, Oxygen 46; OD1, Oxygen Delta 1; N37, Nitrogen 37; O8, Oxygen 8; NF, Nelfinavir.
(TIF)

Figure S8 Interactions with key residues from sB-D30N. Distance variation among the drug (NF) and selected atoms of key residues in the sB-D30N Chain A structure along 50 ns of molecular dynamics simulation. The colors are given in black, gray, light pink and beige for the interaction pairs Isoleucine 50/Aspartate 25 (depicted in red in Figure 4), Aspartate 30(O)/NF(O46), Aspartate 25(OD1)/NF(N37) and Isoleucine 50(O)/NF(O8), respectively. O, Oxygen; O46, Oxygen 46; OD1, Oxygen Delta 1; N37, Nitrogen 37; O8, Oxygen 8; NF, Nelfinavir.
(TIF)

Figure S9 Interactions with key residues from sB-D30V. Distance variation among the drug (NF) and selected atoms of key residues in the sB-D30V Chain A structure along 50 ns of molecular dynamics simulation. The colors are given in black, gray, light pink and beige for the interaction pairs Isoleucine 50/Aspartate 25 (depicted in green in Figure 4), Aspartate 30(O)/NF(O46), Aspartate 25(OD1)/NF(N37) and Isoleucine 50(O)/NF(O8), respectively. O, Oxygen; O46, Oxygen 46; OD1, Oxygen Delta 1; N37, Nitrogen 37; O8, Oxygen 8; NF, Nelfinavir.
(TIF)

Figure S10 Key hydrogen bonds between drug sB-WT. Number (above) and average (below) of hydrogen bonds performed among different residues of the sB-WT and the ligand Nelfinavir (NF) along 50 ns of molecular dynamics simulation. The colors are given in green, cyan and blue for the interaction pairs Flap Chain A/Flap Chain B, Aspartate 25/NF and Aspartate 30/NF, respectively.
(TIF)

Figure S11 Key hydrogen bonds between drug sB-D30N. Number (above) and average (below) of hydrogen bonds performed among different residues of the sB-D30N and the ligand Nelfinavir (NF) along 50 ns of molecular dynamics simulation. The colors are given in green, cyan and blue for the interaction pairs Flap Chain A/Flap Chain B, Aspartate 25/NF and Aspartate 30/NF, respectively.
(TIF)

Figure S12 Key hydrogen bonds between drug sB-D30V. Number (above) and average (below) of hydrogen bonds performed among different residues of the sB-D30V and the ligand Nelfinavir (NF) along 50 ns of molecular dynamics simulation. The colors are given in green, cyan and blue for the interaction pairs Flap Chain A/Flap Chain B, Aspartate 25/NF and Aspartate 30/NF, respectively.
(TIF)

Figure S13 Different patterns of conformational change during the simulation. Selected frames from sB-WT (black), sB-D30N (red), sB-V32E (blue) and sC-V32E (green) are depicted in a *cartoon* representation. All structures are presented with Chain A on the left and Chain B on the right, with the drug depicted in *sticks*. All PRs started in a closed conformation (as represented in Figure S3) and presented important conformational changes during the 50 ns of simulation. At 15 ns: Note that all structures presented greater opening movement of Chain A Flap, with exception of SB-V32E that presented a small movement in the opposite direction. It is already possible to observe sB-D30N in a semiopen conformation and sC-V32E in a full open conformation of Chain A Flap. At 25 ns: Note that sB-V32E has already reached the full open conformation while sC-V32E has returned to a situation similar to the semiopen conformation. At 50 ns: Note that sB-WT has returned to a closed conformation (compare to Figure S3) while sB-D30N remains in a semiopen conformation. The sB-V32E has remained in the full open conformation and sC-V32E has also returned to this full open conformation.
(TIF)

Figure S14 Distance measurements in Chain A of sC-PRs. Colors are given in black and red for the wild-type (sC-WT) and D30V (sC-D30V), respectively. (A) Distance variation (above) and average (below) between Aspartate 25 (ASP25) and Isoleucine 50 (ILE50) along 50 ns of molecular dynamics simulation. Differences between the complexes are only clearly observed in the first 15 ns of simulation, although sC-D30V presents slightly higher values in most of the simulated period. The blue line over the averages indicates the threshold for semiopen conformation (1.58 nm). (B) Distance variation between Aspartate 25 (ASP25) and Nelfinavir in the same period of simulation. The wild-type complex presented higher values for this measurement than sC-D30V.
(TIF)

Figure S15 Distance measurements in Chain B of sC-PRs. Colors are given in black and red for the wild-type (sC-WT) and D30V (sC-D30V), respectively. (A) Distance variation between Aspartate 124 (ASP25 from Chain B) and Isoleucine 149 (ILE50 from Chain B) along 50 ns of molecular dynamics simulation. The sC-D30V presented slightly higher values in most of the simulated period, although both complexes presented values below the stipulated threshold for semiopen conformation of Chain A (blue line). (B) Distance variation between Aspartate 124 (ASP25, Chain B) and Nelfinavir in the same period of simulation. The wild-type complex presented higher values for this measurement than sC-D30V.
(TIF)

Figure S16 Interactions with key residues from sC-PRs. Distance variation among the drug (NF) and selected atoms of key residues in both the sC-WT (A) and the sC-D30V (B) structures, along 50 ns of molecular dynamics simulation. The colors are given in gray, brown and pink for the interaction pairs Valine 30(O)/NF(O46), Isoleucine 149(N)/NF(O21) and Alanine 28(N)/NF(O46), respectively. O, Oxygen; O21, Oxygen 21; O46, Oxygen 46; N, Nitrogen.
(TIF)

Figure S17 Key hydrogen bonds between drug and sC-PRs. Number (above) and average (below) of hydrogen bonds performed among the drug (NF) and different residues of both sC-WT (A) and sC-D30V (B), along 50 ns of molecular dynamics simulation. The colors are given in dark green, light green, cyan and blue for the interaction pairs Flap Chain A/Flap Chain B, Isoleucine 149/NF, Aspartate 25/NF and Aspartate 30/NF, respectively. No hydrogen bonds were observed for the pair Isoleucine 149/NF in the sC-D30V simulation.
(TIF)

Figure S18 Secondary structure analysis. Content of secondary structure of each model is compared with the same complex after 25 ns and 50 ns of molecular dynamics simulation. The shades of blue behind each residue indicate the accessibility of that residue. Position of the residues is indicated below the secondary structure maps, for each complex. In each line, Chain A is depicted in the left and Chain B in the right. Observe that in this picture the position of Chain B residues is not represented from 99–198, but starts again from 1–99.
(TIF)

Figure S19 Simulations of unbound sB-PRs (apo form). Root Mean Square Deviation (RMSD) of the unbound proteases along 50 ns of molecular dynamics simulation. All proteases, from both subtypes, changed to an open conformation before 20 ns. Of note, both proteases bearing the V32E mutation (sB-V32E and sC-V32E) presented this change before 5 ns of simulation.
(TIF)

Figure S20 Different conformations of NF bound to each PR. Frames of each PR-NF simulation were recovered each 5 ns and the respective Nelfinavir conformation was used as input to calculate the Root Mean Square Deviation (RMSD) against one of the reference structures. Reference structures included the crystal conformation (from 1OHR) and three low energy conformations recovered from a 100 ns simulation of Nelfinavir in solution (see Figure 7). All dynamic bound conformations of Nelfinavir presented an important divergence from crystal structure. At the second half of simulations, wild-type proteases presented Nelfinavir conformations similar to NF-i1 and NF-i2, while proteases presenting "D30 mutations" presented Nelfinavir conformations similar to NF-i3).
(TIF)

Figure S21 Free Energy Surface for Nelfinavir bound to PRs. Free Energy Surface (FES) representation for Nelfinavir bound to different proteases (PRs) through 50 ns simulations. For each plot, variation on Root Mean Square Deviation (RMSD) and Radius of Gyration (RoG) are indicated in Y and X axis, respectively. The size of dark red "islands" indicates the frequency of low energy conformations with similar values of RMSD and RoG. The "sB-V32E_NF (5)" refers to a replicate of sB-V32E_NF (see Figure S4).
(TIF)

Figure S22 Free Energy Surface for different PRs bound to Nelfinavir. Free Energy Surface (FES) representations for different HIV-1 proteases in 50 ns simulations. For each plot, variation on Root Mean Square Deviation (RMSD) and Radius of Gyration (RoG) are indicated in Y and X axis, respectively. The size of dark red "islands" indicates the frequency of low energy conformations with similar values of RMSD and RoG. The "sB-V32E (5)" refers to a replicate of sB-V32E bound to Nelfinavir (see Figure S4). The suffix "apo" indicates simulations of unbound proteases.
(TIF)

Table S1 Modeling results.
(DOCX)

Table S2 Prevalence (%) of direct hydrogen bond interactions between drug (Nelfinavir) and PRs during 50 ns of simulation.
(DOCX)

Movie S1 Dynamic behavior of the sB-WT bound to Nelfinavir during a 50 ns simulation.
(MP4)

Movie S2 Dynamic behavior of the sB-D30N bound to Nelfinavir during a 50 ns simulation.
(MP4)

Movie S3 Dynamic behavior of the sB-V32E bound to Nelfinavir during a 50 ns simulation.
(MP4)

Movie S4 Dynamic behavior of the sC-V32E bound to Nelfinavir during a 50 ns simulation.
(MP4)

File S1 Supplementary Information on Methods.
(DOCX)

Acknowledgments

We thank the *Centro Nacional de Supercomputação* (CESUP-RS) for allowing access to its computational resources. We also thank the students Jader Peres da Silva and Marcus Fabiano de Almeida Mendes for the help with

the execution of some analyses. Finally, we thank Wesley Júnio Alves da Conceição (*Laboratory for Molecular Modelling and Dynamics - Carlos Chagas Filho Biophysics Institute*), for sharing scripts for Free Energy Surface analysis.

Author Contributions

Conceived and designed the experiments: DAA MMR MS RMM DMJ GFV. Performed the experiments: DAA MMR MS. Analyzed the data: DAA MMR MS RMM DMJ GFV. Wrote the paper: DAA. Reviewed the manuscript: MMR MS RMM DMJ SEMA GFV. Discussed the data: MMR MS RMM DMJ SEMA GFV.

References

1. Rhee SY, Gonzales MJ, Kantor R, Betts BJ, Ravela J, et al. (2003) Human immunodeficiency virus reverse transcriptase and protease sequence database. Nucleic Acids Res 31: 298–303.
2. Santos AF, Soares MA (2011) The impact of the nelfinavir resistance-conferring mutation D30N on the susceptibility of HIV-1 subtype B to other protease inhibitors. Mem Inst Oswaldo Cruz 106: 177–181.
3. Rhee SY, Taylor J, Fessel WJ, Kaufman D, Towner W, et al. (2010) HIV-1 protease mutations and protease inhibitor cross-resistance. Antimicrob Agents Chemother 54: 4253–4261.
4. Perryman AL, Lin J-h, McCammon JA (2004) HIV-1 protease molecular dynamics of a wild-type and of the V82F/I84V mutant: possible contributions to drug resistance and a potential new target site for drugs. Protein Sci 13: 1108–1123.
5. Perryman AL, Lin J-h, McCammon JA (2006) Restrained molecular dynamics simulations of HIV-1 protease: the first step in validating a new target for drug design. Biopolymers 82: 272–284.
6. Soares RO, Batista PR, Costa MGS, Dardenne LE, Pascutti PG, et al. (2010) Understanding the HIV-1 protease nelfinavir resistance mutation D30N in subtypes B and C through molecular dynamics simulations. J Mol Graph Model 29: 137–147.
7. Lu T, Chen Y, Li X-Y (2010) An insight into the opening path to semi-open conformation of HIV-1 protease by molecular dynamics simulation. AIDS 24: 1121–1125.
8. de Medeiros RM, Junqueira DM, Matte MC, Barcellos NT, Chies JA, et al. (2011) Co-circulation HIV-1 subtypes B, C, and CRF31_BC in a drug-naive population from Southernmost Brazil: analysis of primary resistance mutations. J Med Virol 83: 1682–1688.
9. Johnson VA, Calvez V, Gunthard HF, Paredes R, Pillay D, et al. (2013) Update of the drug resistance mutations in HIV-1: March 2013. Top Antivir Med 21: 6–14.
10. Ghosh AK, Chapsal BD, Weber IT, Mitsuya H (2008) Design of HIV protease inhibitors targeting protein backbone: an effective strategy for combating drug resistance. Acc Chem Res 41: 78–86.
11. Chen J, Zhang S, Liu X, Zhang Q (2010) Insights into drug resistance of mutations D30N and I50V to HIV-1 protease inhibitor TMC-114: free energy calculation and molecular dynamic simulation. J Mol Model 16: 459–468.
12. Mitsuya Y, Winters MA, Fessel WJ, Rhee S-y, Hurley L, et al. (2006) N88D facilitates the co-occurrence of D30N and L90M and the development of multidrug resistance in HIV type 1 protease following nelfinavir treatment failure. AIDS Res Hum Retroviruses 22: 1300–1305.
13. Matsuyama S, Aydan A, Ode H, Hata M, Sugiura W, et al. (2010) Structural and energetic analysis on the complexes of clinically isolated subtype C HIV-1 proteases and approved inhibitors by molecular dynamics simulation. J Phys Chem B 114: 521–530.
14. Piana S, Carloni P, Parrinello M (2002) Role of conformational fluctuations in the enzymatic reaction of HIV-1 protease. J Mol Biol 319: 567–583.
15. Piana S, Carloni P, Rothlisberger U (2002) Drug resistance in HIV-1 protease: Flexibility-assisted mechanism of compensatory mutations. Protein Sci 11: 2393–2402.
16. Arora SK, Gupta S, Toor JS, Singla A (2008) Drug resistance-associated genotypic alterations in the pol gene of HIV type 1 isolates in ART-naive individuals in North India. AIDS Res Hum Retroviruses 24: 125–130.
17. Xie L, Evangelidis T, Bourne PE (2011) Drug discovery using chemical systems biology: weak inhibition of multiple kinases may contribute to the anti-cancer effect of nelfinavir. PLoS Comput Biol 7: e1002037.
18. Kozisek M, Bray J, Rezacova P, Saskova K, Brynda J, et al. (2007) Molecular analysis of the HIV-1 resistance development: enzymatic activities, crystal structures, and thermodynamics of nelfinavir-resistant HIV protease mutants. J Mol Biol 374: 1005–1016.
19. Ode H, Matsuyama S, Hata M, Neya S, Kakizawa J, et al. (2007) Computational characterization of structural role of the non-active site mutation M36I of human immunodeficiency virus type 1 protease. J Mol Biol 370: 598–607.
20. Perez MA, Fernandes PA, Ramos MJ (2007) Drug design: new inhibitors for HIV-1 protease based on Nelfinavir as lead. J Mol Graph Model 26: 634–642.
21. Gantt S, Casper C, Ambinder RF (2013) Insights into the broad cellular effects of nelfinavir and the HIV protease inhibitors supporting their role in cancer treatment and prevention. Curr Opin Oncol 25: 495–502.
22. Kar P, Knecht V (2012) Energetic basis for drug resistance of HIV-1 protease mutants against amprenavir. J Comput Aided Mol Des 26: 215–232.
23. Naicker P, Achilonu I, Fanucchi S, Fernandes M, Ibrahim MA, et al. (2012) Structural insights into the South African HIV-1 subtype C protease: impact of hinge region dynamics and flap flexibility in drug resistance. J Biomol Struct Dyn.
24. Artimo P, Jonnalagedda M, Arnold K, Baratin D, Csardi G, et al. (2012) ExPASy: SIB bioinformatics resource portal. Nucleic Acids Res 40: W597–603.
25. Eswar N, Webb B, Marti-Renom MA, Madhusudhan MS, Eramian D, et al. (2006) Comparative protein structure modeling using Modeller. Curr Protoc Bioinformatics Chapter 5: Unit 5 6.
26. Laskowski RA, Rullmannn JA, MacArthur MW, Kaptein R, Thornton JM (1996) AQUA and PROCHECK-NMR: programs for checking the quality of protein structures solved by NMR. J Biomol NMR 8: 477–486.
27. van Aalten DM, Bywater R, Findlay JB, Hendlich M, Hooft RW, et al. (1996) PRODRG, a program for generating molecular topologies and unique molecular descriptors from coordinates of small molecules. J Comput Aided Mol Des 10: 255–262.
28. Trott O, Olson AJ, News S (2010) AutoDock Vina: improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading. J Comput Chem 31: 455–461.
29. Van Der Spoel D, Lindahl E, Hess B, Groenhof G, Mark AE, et al. (2005) GROMACS: fast, flexible, and free. J Comput Chem 26: 1701–1718.
30. Hyland LJ, Tomaszek TA, Jr., Meek TD (1991) Human immunodeficiency virus-1 protease. 2. Use of pH rate studies and solvent kinetic isotope effects to elucidate details of chemical mechanism. Biochemistry 30: 8454–8463.
31. Hyland LJ, Tomaszek TA, Jr., Roberts GD, Carr SA, Magaard VW, et al. (1991) Human immunodeficiency virus-1 protease. 1. Initial velocity studies and kinetic characterization of reaction intermediates by 18O isotope exchange. Biochemistry 30: 8441–8453.
32. Batista PR, Wilter A, Durham EH, Pascutti PG (2006) Molecular dynamics simulations applied to the study of subtypes of HIV-1 protease common to Brazil, Africa, and Asia. Cell Biochem Biophys 44: 395–404.
33. Humphrey W, Dalke A, Schulten K (1996) VMD: visual molecular dynamics. J Mol Graph 14: 33–38, 27–38.
34. DeLano WL, Bromberg S (2004) PyMOL User's Guide. San Francisco: DeLano Scientific LLC.
35. Pettersen EF, Goddard TD, Huang CC, Couch GS, Greenblatt DM, et al. (2004) UCSF Chimera-a visualization system for exploratory research and analysis. J Comput Chem 25: 1605–1612.

Apêndice 02

"Temporal dynamics of HIV-1 circulating subtypes in distinct exposure categories in Southern Brazil"

Sabrina EM Almeida, Rubia M de Medeiros, Dennis M Junqueira, Tiago Gräf, Caroline PB Passaes, Gonzalo Bello, Mariza G Morgado, Monick L Guimarães

Virology Journal, 2012

VIROLOGY JOURNAL

RESEARCH Open Access

Temporal dynamics of HIV-1 circulating subtypes in distinct exposure categories in southern Brazil

Sabrina EM Almeida[1*], Rubia M de Medeiros[1,3], Dennis M Junqueira[1,3], Tiago Gräf[1,4], Caroline PB Passaes[2], Gonzalo Bello[2], Mariza G Morgado[2] and Monick L Guimarães[2]

Abstract

Background: The HIV-1 epidemic in Brazil is predominantly driven by subtype B. However, in Brazilian Southern region subtype C prevails and a relatively high AIDS incidence rate is observed. The aim of the present study was to assess the temporal dynamics of HIV-1 subtypes circulating in patients from distinct exposure categories in Southern Brazil. For this purpose 166 HIV-1 samples collected at the years of 1998 (group I) and 2005–2008 (group II) were analyzed.

Results: Analysis of group I revealed statistically significant ($p < 0.05$) associations between MSM and subtype B as well as between IDU and subtype C; while no statistical significant association between HIV-1 subtypes and exposure category was verified for group II. An overall temporal increase in the prevalence of subtype C and BC recombinants was observed in both HET and MSM populations, accompanied by a proportional decrease in the prevalence of the pure subtype B.

Conclusions: The present study shows an association between HIV subtypes and exposure categories at the middle 1990s in Southern Brazil. Our findings suggest that MSM and IDU populations might have played a major role in the introduction and initial dissemination of subtypes B and C, respectively, in Southern Brazil. This study also suggests a trend towards homogenization of HIV-1 strains across distinct exposure categories as a consequence of an overall increase in the prevalence of subtype C and BC recombinants in both HET and MSM populations.

Keywords: HIV-1, Brazil, Subtypes, Exposure categories, Temporal dynamics

Introduction

The main hallmark of the HIV-1 is an extraordinary evolution rate, which results in high molecular diversity and dynamism of the AIDS epidemic [1]. HIV-1 is classified in four groups and the group M, currently estimated to infect around 33 million people around the world, is subdivided in 9 subtypes (A–D, F–H, J, and K) and 54 circulating recombinant forms (CRFs) [1-3].

Since 1980, Brazil has registered 608,230 cases of AIDS, representing an overall prevalence of 0.6% in adult population [4]. The HIV-1 subtype B is the predominant variant in most of the Brazilian regions followed by subtype F1, subtype C and a large variety of BF1 and BC recombinant forms [4-10]. The distribution of HIV-1 subtypes, however, is not homogeneous across the country and a distinct HIV-1 molecular epidemiologic scenario is observed in the Southern region. Composed by the states of Rio Grande do Sul, Santa Catarina and Paraná, the Brazilian Southern region shows a remarkably high prevalence of subtype C and BC recombinant forms [11-21].

The HIV-1 molecular epidemiology in Porto Alegre, capital of the Rio Grande do Sul state, is characterized by a high prevalence of subtype C (~30-40%), subtype B (~30-45%) and the circulating recombinant form (CRF) 31_BC (~10-25%), and lower prevalence of unique recombinant forms (URFs) and subtype F1 [12,16-18,21]. The southern region of Brazil is not only characterized by a distinct subtype profile, but also by a relative high AIDS incidence rate. This is particularly evident for the city of Porto Alegre, showing the highest AIDS incidence

* Correspondence: sabrina.gene@gmail.com
[1]Centro de Desenvolvimento Técnico e Científico – CDCT, Fundação Estadual de Produção e Pesquisa em Saúde – FEPPS, Av. Ipiranga, 5400, 3° andar, CEP: 90610-000, Porto Alegre, RS, Brazil
Full list of author information is available at the end of the article

rate among all the Brazilian capitals since 1997 [4]. The reported AIDS incidence in Porto Alegre in 2010 (99.8 cases per 100,000 habitants) was more than five times higher than the mean incidence for the whole country (17.9 cases per 100,000 habitants) [4]. In this city, almost 80% of the new AIDS cases registered in 2009 correspond to heterosexual (HET) individuals, approximately 10% to men who have sex with men (MSM) and another 10% to injection drug users (IDU) [22].

Some studies indicate a tendency towards an increasing proportion of subtype C infections over time in some cities from the Southern region. The estimated prevalence of subtype C increased from 36% before 1997 to 53% in 2008 in Rio Grande (Rio Grande do Sul state) [11,23], and from 56% to 78% between 2004 and 2009 in Florianopolis (Santa Catarina state) [14]. Further studies also point to a different spreading of the HIV-1 variants among the exposure categories in southern Brazil. Studies conducted in all the three states of Brazilian Southern region described that subtype C is more frequently seen in the HET population, while subtype B is more common in MSM [14,20,23]. One study also found an association between subtype C infections and the use of intravenous drugs among males in the state of Parana [15]. So far, no significant associations between HIV-1 clades and exposure categories were detected in the city of Porto Alegre, the main capital of Southern Brazil, even though an association of subtype B with anal sex practices and a tendency of subtype C with females have been reported [11,17,21].

Most of the previous studies that analyzed the temporal dynamics of HIV-1 diversity in Southern Brazil were based on the stratification of cross-sectional samples according to time of HIV diagnosis and without taking into account the exposure categories of the individuals. In this study, we described the temporal trends of HIV-1 clades in patients of different exposure categories living in the city of Porto Alegre, through the analysis of two groups of samples collected at 1998 and 2005–2008.

Methods

Study population

Blood samples from 166 HIV-positive patients followed up at different outpatients clinics in the metropolitan region of Porto Alegre, the capital of the Southernmost state of Brazil, were collected at two different time periods: 1998 (n = 83) and 2005–2008 (n = 83). The inclusion criteria for individuals in both studied periods were age over 18-year old, agreement to participate in the study, read and signed an informed consent form. In the second period the individuals had to report no previous antiretroviral therapy. The clinical and demographic data of the patients (age, sex, first positive serology for HIV-1 and CD4+ T-cell counts) are shown in Table 1. This study was approved by the Ethical Research Committee from Fundação Estadual de Produção e Pesquisa em Saúde number18/2005.

HIV-1 amplification, sequencing and subtyping

DNA samples were extracted from 200 μl of whole blood using a QIAamp DNA kit (Qiagen Inc., CA, U.S. A.), according to the manufacturer's protocol. Amplification and sequencing of the PR/RT region was performed as described elsewhere [24]. The sequences generated were ~1,160 nt long and covered the protease (PR) and part of the reverse transcriptase (RT) genes (nucleotides 2253–3413 relative to HXB2).

Nucleotide sequences were aligned using the Clustal X program [25] and three strategies were used to well characterize the HIV-1 sequences as pure subtype, CRF-like or URF using the pol alignment: i) a Neighbor-Joining (NJ) phylogenetic tree was first built under the Tamura-Nei substitution model in 1000 bootstrapped data sets, as implemented in MEGA program [26]; ii) all sequences were subsequently subjected to bootscanning with the Simplot software version 3.5.1 [27], using reference sequences representative of HIV-1 subtypes A1, B, C, and F1 available in Los Alamos database, http://www.hiv.lanl.gov/components/sequence/HIV/search/search.html). Bootstrap values supporting branching with reference sequences were determined in NJ trees constructed using the K2-parameter model [28], based on 100 resamplings, with a 300 nt sliding window moving in steps

Table 1 Clinical and demographic data of the HIV-positive patients from Porto Alegre at two distinct time periods - 1998 (group I) and 2005–2008 (group II)

	Group I (n=83)	Group II (n=83)
Age (years)	33 ± 10	35 ± 10
Gender		
Male	59 (71.0%)	49 (59.3%)
Female	24 (29.0%)	34 (40.7%)
CD4 T cell count		
(cell/mm3)*		
200	23 (29,1%)	16 (28,6%)
200-400	25 (31,6%)	13 (24,1%)
>400	31 (39,2%)	27 (50,0%)
Exposure category		
HET	42 (50.6%)	61 (73.5%)
MSM	32 (38.6%)	22 (26.5%)
IDU	9 (10.8%)	-
First positive serology for HIV (years)	1994 [1989–1998]	2005 [2002–2008]

* information not available for four individuals from group I and twenty seven from group II. HET – heterosexuals; MSM – men who have sex with men; IDU – injection drug users.

of 10 bases; iii) to better characterize the recombination breakpoints suggested in the previous analyses, the putative recombinants were subjected to informative site analyses as described elsewhere [24]. Consensus sequences used in our analyses were generated from a total of 491 subtype B and 164 subtype C pol Brazilian sequences downloaded from the Los Alamos HIV Sequence Database, as implemented in the DAMBE program [29]. Sequences were submitted to GeneBank under the following accession numbers: [JF487830 - JF487913] and [JQ619540 - JQ619621].

Statistical analyzes

Statistical comparisons among subtypes groups and exposure categories were made using Pearson's χ^2-test with adjusted residues and Fisher's exact test when appropriate. Statistical analysis was performed using the SPSS 16.0 statistical package and the significance level was set at $p<0.05$.

Results

Group I was composed by HIV-1 positive samples collected in 1998 from 83 individuals, of which 71.0% were men. Based on medical records, patients were classified into three exposure categories: HET (50.6%), MSM (38.6%) and IDU (10.8%). The mean diagnostic year of patients from group I was 1994 (Table 1). Group II, comprised 83 HIV-1 positive individuals sampled from 2005 to 2008, of which 59.3% were men (Table 1). Regarding the exposure categories, group II showed 73.5% of HET individuals and 26.5% of MSM, the mean diagnostic year being 2005. Most patients of both groups (39,2% and 50%, respectively) presented T CD4 cells counts above 400cell/mm3. All patients included in group II were antiretroviral naïve, while 55% of individuals from group I declared to be under antiretroviral therapy.

According to the phylogenetic, bootscanning, and informative site analyses of the PR/RT region, the 166 HIV-1 samples from Porto Alegre here analyzed were classified as follows: subtype B (n = 65), subtype C (n = 37), CRF31_BC (n = 28), URFs_BC (n = 25), subtype F1 (n = 3), URFs_BF (n = 5), and URF_BCF (n = 3) (Additional file 1: Figure S1). Some of the URFs_BC detected in the present study (n = 13), shared one of the recombination breakpoints with the CRF31_BC and were probably created by the recombination of the CRF31_BC with local subtypes C and B. Since CRF31_BC and the URFs_BC with a CRF31_BC-related structure probably share a common evolutionary history, they were grouped together for further analyses. Other URFs_BC with a mosaic structure not related to the CRF31_BC (n = 12) were allocated in a separate group. The recombination pattern of all URFs and a schematic draw of the CRF31_BC-related structure are provided as Additional file (Additional file 2: Figure S2).

The analyses of HIV diversity and exposure categories for both time periods are described in Table 2. The analysis of HIV-1 clades for group I revealed a significant association ($p < 0.05$) between MSM exposure category and subtype B, as well as between IDU patients and subtype C. The HIV-1 subtype B variant was responsible for more than 70% of infections in MSM individuals, while it represented 50% of infections in HET and only 11% of infections among IDU. By contrast, the HIV-1 subtype C variant was more frequent among IDU patients (44.4%) than among HET (14.3%) and MSM (6.2%) individuals. The frequency of URFs_BC in the IDU population (22.2%) was also higher than in the MSM (3%) and HET (0%) populations ($p<0.05$). The CRF31_BC and other BC recombinants with a related mosaic structure were more frequent in IDU (22.2%) and HET (21.4%) populations than in the MSM (9.4%) group, while subtype F1 and BF1 recombinants displayed a higher frequency in the HET (14.3%) and MSM (9.4%) populations compared to IDU (0%), although those differences were not statistically significant.

Analysis of the group II reveals no significant association between HIV-1 clades and the exposure category, even though some differences in the most prevalent clades circulating in HET and MSM individuals were evident. Most HIV-1 infections in the HET population were associated to subtype C (34.4%) and CRF31_BC and related recombinants (36%), followed by subtype B (21.4%) and URFs_BC (6.6%). By contrast, the majority of HIV-1 infections detected in the MSM population were related to subtype B (31.8%), followed by roughly similar frequencies of CRF31_BC and related recombinants (22.7%), URFs_BC (22.7%), and subtype C (18.2%). In both groups, the frequency of subtype F1 and URFs_BF/URFs_BCF was very low (<5%).

The analysis of the temporal dynamics of HIV-1 clades circulating in the HET and MSM individuals revealed a significant ($p<0.01$) decreasing in prevalence of subtype B in both exposure categories over time (Figure 1). In 1998, 50% of HET and 72% of MSM individuals were infected by subtype B, passing to 21% and 32% in 2005–2008, respectively. At the same time, the prevalence of subtype C, CRF31_BC and related recombinants and URFs_BC increased in both HET and MSM individuals, although only the increase of subtype C in HET (14% to 34%) and of URFs_BC in MSM (3% to 23%) were statistically significant ($p<0.05$) (Figure 1). A decrease in the prevalence of subtype F1 and URFs_BF was also observed and it was more pronounced in the HET population, in which the frequency of those clades decreased from 14% to 2% ($p<0.05$) (Figure 1). When the HET group was analyzed by gender, it was observed a

Table 2 HIV-1 subtype frequencies according to the patient's exposure category

		Exposure Category		
		HET	MSM	IDU
Group I				
	Subtype B	21 (50.0)	23 (72.0)*	1 (11.0)
	Subtype C	6 (14.3)	2 (6.2)	4 (44.6)*
	CRF31_BC and related recombinants	9 (21.4)	3 (9.4)	2 (22.2)
	URFs_BC	-	1 (3.0)	2 (22.2)*
	Subtype F1 and URF_BF	6 (14.3)	3 (9.4)	-
	Total (n=83)	42	32	9
Group II				
	Subtype B	13 (21.4)	7 (31.8)	-
	Subtype C	21 (34.4)	4 (18.2)	-
	CRF31_BC and related recombinants	22 (36.0)	5 (22.7)	-
	URFs_BC	4 (6.6)	5 (22.7)	-
	Subtype F1 and URF_BF/URF_BCF	1 (1.6)	1 (4.5)	-
	Total (n=83)	61	22	-

*significant p value <0.05. Values between brackets are relative percentages of the total of the column. HET – heterosexuals; MSM – men who have sex with men; IDU – injection drug users.

significant ($p<0.05$) increase of subtype C (8.3% to 33.3%) across time in women, significant ($p<0.05$) decrease of subtype B in both women (50% to 24.2%) and men (50% to 17.9%) and significant ($p<0.05$) decrease of subtype F1 and URFs_BF (12.5% to 0%) in women.

Discussion

The current study compared the HIV-1 molecular epidemiologic scenario at two distinct time periods in the city of Porto Alegre, Southern Brazil. Patients from group I, recruited in 1998 were diagnosed between 1986 to 1998 and were likely infected around the early 1990s. Patients from group II, recruited between 2005 and 2008 were diagnosed between 2002 to 2008, and constitute a picture of the HIV-1 epidemic around the early to middle 2000s. Given the broad scope of these two sample groups, the analyses performed here have a temporal feature, shedding some light to the origins of the singular HIV-1 epidemic in Southern Brazil.

Analysis of the group I revealed a significant association between the MSM exposure category and subtype B, as well as between IDU patients and subtype C. The MSM and IDU population have played a major role in

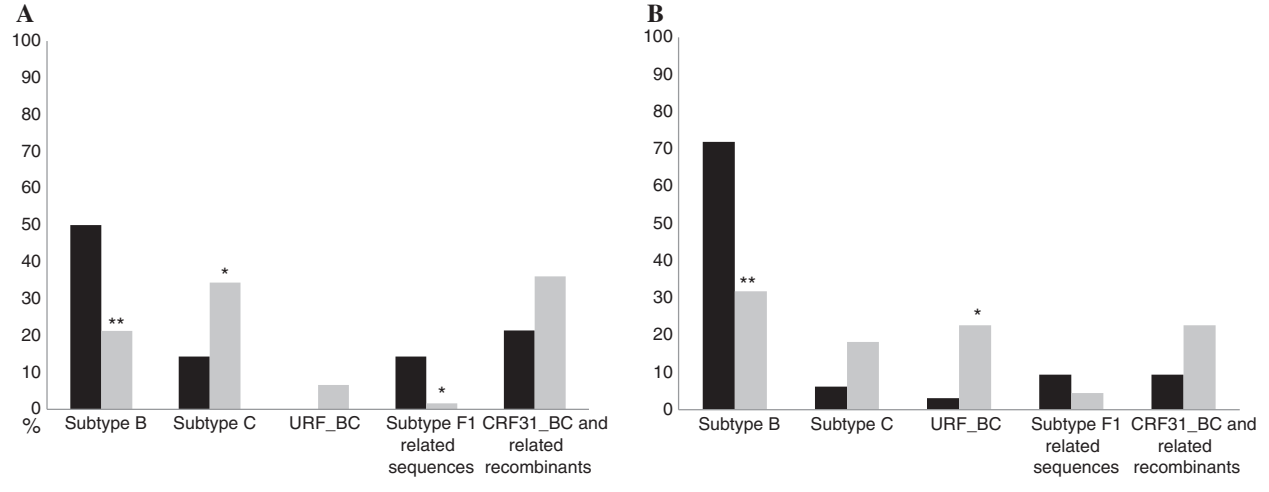


Figure 1 Prevalence of HIV-1 subtypes in Porto Alegre in the two studied populations (1998 and 2005–2008) according to exposure categories (A) Heterosexual and (B) Men who have sex with men. Black bars represent Group I and gray bars represent Group II. Symbols represent statistical significance, * = $p<0.05$; ** = $p<0.01$.

the dynamics of the HIV epidemic in Porto Alegre, particularly up to the middle 90s (Figure 2). Although harm reduction policies promoted by the Brazilian Ministry of Health reduced the proportion of AIDS cases in IDU from more than 25% before 1990 to around 5% in 2010, Porto Alegre continues to figure as the Brazilian city with the highest number of AIDS cases among IDU [4,22]. No temporal analysis of the epidemic in the IDU exposure category was performed as a consequence of the absence of IDU in group II. Selecting a sample of IDU is not a simple task since injecting drug is an illegal and stigmatizing behavior, and these individuals are usually outside the public health services. These results suggest that HIV-1 clades B and C may have been introduced and initially disseminated in Porto Alegre through the MSM and IDU transmission networks, respectively. Similar findings have been recently reported in other Southern Brazilian states. One study reported an association between subtype C infections and the use of intravenous drugs among males in the state of Parana [15], while other studies described an association between subtype B infections and the MSM population in Parana, Santa Catarina and Rio Grande do Sul states [14,20,23].

The proposed scenario is consistent with the proposed evolutionary history of subtype B and C epidemics in Brazil. Some studies suggest that the pandemic subtype B clade was introduced in Brazil around 1965–1970, at least 10–15 years earlier than subtype C [30,31]. Thus, the pandemic subtype B is the most probable HIV-1 clade responsible for the initial AIDS cases described in Brazil, including the city of Porto Alegre, during the early 1980s, that mostly affect the MSM population (Figure 2). The proportion of AIDS cases corresponding to IDU and HET individuals started to growth in Porto Alegre around the middle 1980s (Figure 2), shortly after the estimated introduction of subtype C clade [32]. Of note, the estimated onset date of the CRF31_BC clade (around the late 1980s) [24], coincides with a peak in the proportion of AIDS cases due to IDU infections in Porto Alegre (Figure 2). Since HIV transmission is very efficient through injecting equipment, super-infections with distinct viral forms are constant among IDU [33], which may have allowed the emergence of diverse recombinants forms, including the CRF31_BC.

Analysis of the group II reveals that by the middle 2000s most HIV-1 infections in the HET population were associated to subtype C (34%) and CRF31_BC and related recombinants (36%), while subtype B was the most common genetic variant in the MSM population (32%). However, no significant association between HIV-1 clades and the exposure infection risk were detected at this second time point. This coincides with other studies conducted in Porto Alegre between 2002 and 2009, that also have failed to find a significant association between HIV-1 clades and the exposure category [17,21]. This lack of association may be explained by a progressive intermixing and homogenization of HIV-1 clades

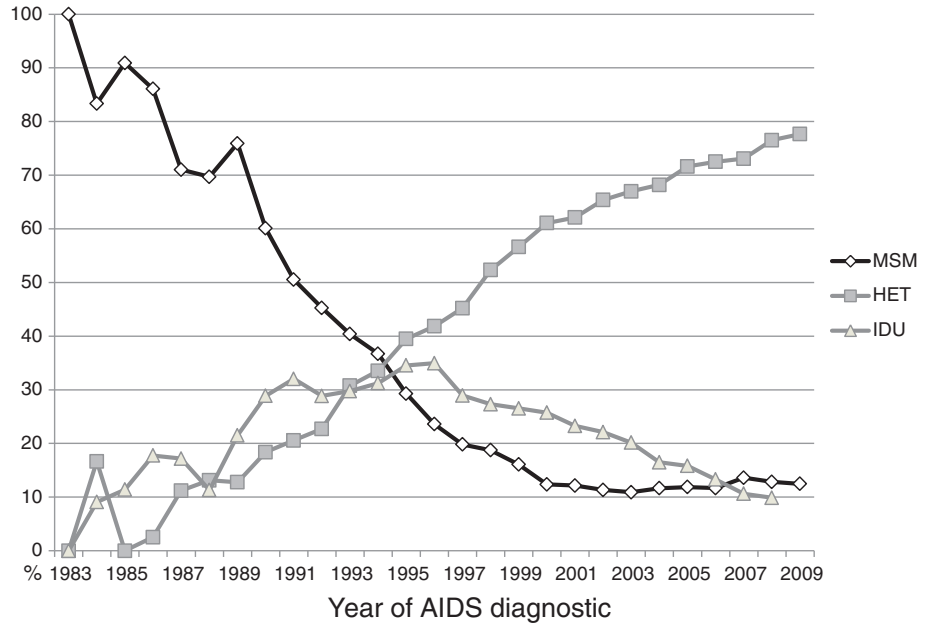


Figure 2 Proportion of AIDS diagnosed individuals according to exposure categories in Porto Alegre – Brazil. Black line with diamonds represents Men who have sex with men, grey line with squares represents heterosexuals and gray line with triangles represents injection drug users. Source: http://www.aids.gov.br/pagina/tabulacao-de-dados - access October 2011.

between different transmission networks in Porto Alegre over time, characteristic of a mature epidemic.

Comparison between groups I and II revealed a temporal increase in the proportion of subtype C (from 14% to 34%) and CRF31_BC and related recombinants (from 21% to 36%) in HET individuals. This is consistent with a previous coalescent analysis that supports an exponential expansion of subtype C and CRF31_BC clades in Porto Alegre during the 1980s and 1990s [34]. This expansion also coincides with the growing proportion of AIDS cases in IDU and, particularly, in the HET population in Porto Alegre (Figure 2), in agreement with the process of feminization of the AIDS epidemic observed throughout the country [4]. Similar to the AIDS Brazilian epidemic, the present study also observed an increase in heterosexual transmissions (50.6%-73.5%) and in female cases (29.0%- 40.7%) between the two groups of patients (Table 1). It is conceivable that shortly after introduction into IDU population, subtype C may have passed to HET. The early introduction of subtype C into the HET group may have promoted the successful spread of this variant in Porto Alegre, even after the decrease of HIV infections among IDU.

An increase in the proportion of subtype C (from 6% to 18%) and CRF31_BC and related recombinants (from 9% to 23%) between the two time periods was also evident among MSM individuals from Porto Alegre, thus providing evidence that dissemination of subtype C and CRF31_BC was not limited to the HET group. These results support the hypothesis that some associations between HIV-1 strains and sexual exposure categories detected in Brazil may be due to founder effects, rather than to a different efficacy of transmission of subtypes B and C through MSM or HET individuals. Analysis of the MSM population further reveals a significant increase in the frequency of URFs_BC (from 3% to 23%), that could be explained by the gradual dissemination of subtype C into this group coupled to high rates of coinfection and/or superinfection. Bisexual men and homosexual IDU might have played an essential role in this process, promoting a linkage between partially isolated transmission networks.

The overall increase in the prevalence of subtype C and BC recombinants in Porto Alegre was accompanied by a proportional decrease in the prevalence of subtype B in both HET (from 50% to 21%) and MSM (from 72% to 32%) populations. One hypothesis to explain this result is that subtype C displays a higher sexual transmissibility than subtype B. Alternatively, the dissimilar outcome of subtypes B and C in the city of Porto Alegre may be a consequence of differences in the transmission networks that promoted the initial dissemination of each subtype in the city. It is possible that subtype C and the CRF31_BC gained access to large networks of IDU and HET groups before subtype B, and this may have conditioned the subsequent dissemination of those clades among all different exposure categories. The great variation in the relative prevalence of clades B and C across different cities in Southern Brazil favors this second hypothesis. Of note, interpretations of the observed patterns should be considered with caution because of the limited sample size and the absence of IDU in the second casuistic. Future studies including larger number of HIV+ patients within different exposure categories will be necessary to confirm the observed trends.

Conclusions

The data presented here points to a possible introduction of subtype B and C through different transmission networks in the city of Porto Alegre. Subtype B was probably introduced through the MSM individuals while subtype C may have been introduced into the IDU network and rapidly disseminated to the HET group. This study also suggests a trend towards homogenization of HIV-1 strains across distinct exposure categories as there has been an overall increase in the prevalence of subtype C and BC recombinants in both HET and MSM populations. Understanding the mechanism responsible for the expansion and contraction of subtypes C and B, respectively, in Porto Alegre, is of paramount importance to understand the HIV-1 dynamics in this country region.

Additional files

Additional file 1: Figure S1. Neighbor-Joining phylogenetic tree with Tamura-Nei substitution model of HIV-1 PR/RT region (2253–3413 relative to HXB2) of samples from group I and II. Only "pure" subtype C, B, F1 and CRF31_BC were included. Bootstrap values above 80% obtained for 1000 replicates are shown in the nodes.

Additional file 2: Figure S2. Schematic drawing showing breakpoint pattern of the URF viruses found in the study. Breakpoint positions were obtained using Simplot 3.5.1 and numbered according to HXB2 reference. Sequences CRF31_BC related are characterized by the presence of a slightly smaller or bigger subtype B fragment.

Abbreviations

AIDS: Acquired immune deficiency syndrome; CRF: Circulating recombinant form; IDU: Injection drug users; HIV: Human immunodeficiency virus; HET: Heterosexual; MSM: Men who have sex with men; NJ: Neighbor-joining; PR: Protease; RT: Reverse transcriptase; URF: Unique recombinant form.

Competing interests

The authors declare that they have no competing interests.

Authors' contribution

SEMA designed the study, helped to draft the manuscript and coordinated the statistical analysis. RMM, DMJ and CPBP collected the samples, carried out the molecular studies and the phylogenetic analysis. TG drafted the manuscript and performed the statistical analysis. GB participated in the study design and helped in the phylogenetic and statistical analysis. MGM helped in the draft of the manuscript. MLG conceived of the study, participated in its design and coordination and edited the manuscript. All authors read and approved the final manuscript.

Acknowledgements
We wish to thank Vera Bongertz for English revision. We also acknowledge Regina Loureiro for patients' recruitment and sample collection.

Author details
[1]Centro de Desenvolvimento Técnico e Científico – CDCT, Fundação Estadual de Produção e Pesquisa em Saúde – FEPPS, Av. Ipiranga, 5400, 3° andar, CEP: 90610-000, Porto Alegre, RS, Brazil. [2]Laboratório de AIDS & Imunologia Molecular. Instituto Oswaldo Cruz – FIOCRUZ. Av. Brasil 4365, 439 - Pavilhão Leonidas Deane, sala 413, CEP: 21040-900, Rio de Janeiro, RJ, Brazil. [3]Programa de Pós-graduação em Genética e Biologia Molecular, Departamento de Genética, Centro de Ciências Biológicas, Universidade Federal do Rio Grande do Sul (UFRGS), 9500 - Prédio 43323M, CEP:91501-970, Porto Alegre, RS, Brazil. [4]Programa de Pós-graduação em Biotecnologia e Biociências, Departamento de Microbiologia, Imunologia e Parasitologia, Centro de Ciências Biológicas, Universidade Federal de Santa Catarina (UFSC), Campus Universitário, CEP: 88040-970, Florianópolis, SC, Brazil.

Received: 2 March 2012 Accepted: 7 December 2012
Published: 12 December 2012

References

1. Tebit DM, Arts EJ: **Tracking a century of global expansion and evolution of HIV to drive understanding and to combat disease.** *Lancet Infect Dis* 2011, **11:**45–56.
2. World Health Organization, Global report: *UNAIDS report on the global AIDS epidemic*. 2010. http://www.unaids.org/globalreport/Global_report.htm.
3. Ng KT, Ong LY, Takebe Y, Kamarulzaman A, Tee KK: **Genome Sequence of a Novel HIV-1 Circulating Recombinant Form 54_01B from Malaysia.** *J Virol* 2012, **86:**11405–11406.
4. Brazilian Ministry of health: *AIDS epidemiological bulletin (in portuguese). January-june 2011; ano viii, n°01*. http://www.aids.gov.br/publicacao/2011/boletim_epidemiologico_2011.
5. Guimarães ML, Dos Santos Moreira A, Loureiro R, Galvão-Castro B, Morgado MG: **High frequency of recombinant genomes in HIV type 1 samples from Brazilian southeastern and southern regions.** *AIDS Res Hum Retroviruses* 2002, **18:**1261–1269.
6. Brígido LF, Franco HM, Custódio RM, Oliveira CA, P Ferreira JL, Eira M, Bergel F, Araújo F, Carvalheiro JR, Rodrigues R: **Molecular characteristics of HIV type 1 circulating in São Paulo.** *Brazil. AIDS Res Hum Retroviruses* 2005, **21:**673–682.
7. Pedroso C, Queiroz AT, Alcântara LC, Drexler JF, Diaz RS, Weyll N, Brites C: **High prevalence of primary antiretroviral resistance among HIV-1-infected adults and children in Bahia, a northeast state of Brazil.** *J Acquir Immune Defic Syndr* 2007, **45:**251–253.
8. Cardoso LP, Queiroz BB, Stefani MM: **HIV-1 pol phylogenetic diversity and antiretroviral resistance mutations in treatment naïve patients from Central West Brazil.** *J Clin Virol* 2009, **46:**134–139.
9. Machado LF, Ishak MO, Vallinoto AC, Lemos JA, Azevedo VN, Moreira MR, Souza MI, Fernandes LM, Souza LL, Ishak R: **Molecular epidemiology of HIV type 1 in northern Brazil: identification of subtypes C and D and the introduction of CRF02_AG in the Amazon region of Brazil.** *AIDS Res Hum Retroviruses* 2009, **25:**961–966.
10. Monteiro-Cunha JP, Araujo AF, Santos E, Galvao-Castro B, Alcantara LC: **Lack of high-level resistance mutations in HIV type 1 BF recombinant strains circulating in northeast Brazil.** *AIDS Res Hum Retroviruses* 2011, **27:**623–631.
11. Soares EA, Martínez AM, Souza TM, Santos AF, Da Hora V, Silveira J, Bastos FI, Tanuri A, Soares MA: **HIV-1 subtype C dissemination in southern Brazil.** *AIDS* 2005, **19**(Suppl 4):S81–S86.
12. Brígido LF, Nunes CC, Oliveira CM, Knoll RK, Ferreira JL, Freitas CA, Alves MA, Dias C, Rodrigues R, Program RC: **HIV type 1 subtype C and CB Pol recombinants prevail at the cities with the highest AIDS prevalence rate in Brazil.** *AIDS Res Hum Retroviruses* 2007, **23:**1579–1586.
13. Rodrigues R, Scherer LC, Oliveira CM, Franco HM, Sperhacke RD, Ferreira JL, Castro SM, Stella IM, Brigido LF: **Low prevalence of primary antiretroviral resistance mutations and predominance of HIV-1 clade C at polymerase gene in newly diagnosed individuals from south Brazil.** *Virus Res* 2006, **116:**201–207.
14. Gräf T, Passaes CP, Ferreira LG, Grisard EC, Morgado MG, Bello G, Pinto AR: **HIV-1 genetic diversity and drug resistance among treatment naïve patients from Southern Brazil: an association of HIV-1 subtypes with exposure categories.** *J Clin Virol* 2011, **51:**186–191.
15. Raboni SM, Almeida SM, Rotta I, Ribeiro CE, Rosario D, Vidal LR, Nogueira MB, Riedel M, Winhescki MG, Ferreira KA, Ellis R: **Molecular epidemiology of HIV-1 clades in Southern Brazil.** *Mem Inst Oswaldo Cruz* 2010, **105:**1044–1049.
16. Santos AF, Schrago CG, Martinez AM, Mendoza-Sassi R, Silveira J, Sousa TM, Lengruber RB, Soares EA, Sprinz E, Soares MA: **Epidemiologic and evolutionary trends of HIV-1 CRF31_BC-related strains in southern Brazil.** *J Acquir Immune Defic Syndr* 2007, **45:**328–333.
17. Dias CF, Nunes CC, Freitas IO, Lamego IS, Oliveira IM, Gilli S, Rodrigues R, Brigido LF: **High prevalence and association of HIV-1 non-B subtype with specific sexual transmission risk among antiretroviral naïve patients in Porto Alegre, RS, Brazil.** *Rev Inst Med Trop Sao Paulo* 2009, **51:**191–196.
18. de Medeiros RM, Junqueira DM, Matte MC, Barcellos NT, Chies JA, Matos Almeida SE: **Co-circulation HIV-1 subtypes B, C, and CRF31_BC in a drug-naïve population from Southernmost Brazil: analysis of primary resistance mutations.** *J Med Virol* 2011, **83:**1682–1688.
19. Toledo PV, Carvalho DS, Rossi SG, Brindeiro R, de Queiroz-Telles F: **Genetic diversity of human immunodeficiency virus-1 isolates in Paraná.** *Brazil. Braz J Infect Dis* 2010, **14:**230–236.
20. Silva MM, Telles FQ, da Cunha CA, Rhame FS: **HIV subtype, epidemiological and mutational correlations in patients from Paraná.** *Brazil. Braz J Infect Dis* 2010, **14:**495–501.
21. Simon D, Béria JU, Tietzmann DC, Carli R, Stein AT, Lunge VR: **Prevalence of HIV-1 subtypes in patients of an urban center in Southern Brazil.** *Rev Saude Publica* 2010, **44:**1094–1101.
22. Brazilian Ministry of health: *Coordenação DST/AIDS, Brasil*. http://www.aids.gov.br/pagina/tabulacao-de-dados.
23. Silveira J, Santos AF, Martínez AM, Góes LR, Mendoza-Sassi R, Muniz CP, Tupinambás U, Soares MA, Greco DB: **Heterosexual transmission of human immunodeficiency virus type 1 subtype C in southern Brazil.** *J Clin Virol* 2012, **54:**36–41.
24. Passaes CP, Bello G, Lorete RS, Matos Almeida SE, Junqueira DM, Veloso VG, Morgado MG, Guimarães ML: **Genetic characterization of HIV-1 BC recombinants and evolutionary history of the CRF31_BC in Southern Brazil.** *Infect Genet Evol* 2009, **9:**474–482.
25. Thompson JD, Gibson TJ, Plewniak F, Jeanmougin F, Higgins DG: **The CLUSTAL_X windows interface: flexible strategies for multiple sequence alignment aided by quality analysis tools.** *Nucleic Acids Res* 1997, **25:**4876–4882.
26. Tamura K, Dudley J, Nei M, Kumar S: **MEGA4: Molecular Evolutionary Genetics Analysis (MEGA) software version 4.0.** *Mol Biol Evol* 2007, **24:**1596–1599.
27. Ray S: *Simplot [computer program] version 3.5.1*. http://sray.med.som.jhmi.edu/scroftware/.
28. Kimura M: **A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences.** *J Mol Evol* 1980, **16:**111–120.
29. Xia X, Xie Z: **DAMBE: software package for data analysis in molecular biology and evolution.** *J Hered* 2001, **92:**371–373.
30. Bello G, Guimarães ML, Morgado MG: **Evolutionary history of HIV-1 subtype B and F infections in Brazil.** *AIDS* 2006, **20:**763–768.
31. Bello G, Eyer-Silva WA, Couto-Fernandez JC, Guimarães ML, Chequer-Fernandez SL, Teixeira SL, Morgado MG: **Demographic history of HIV-1 subtypes B and F in Brazil.** *Infect Genet Evol* 2007, **7:**263–270.
32. Bello G, Passaes CP, Guimarães ML, Lorete RS, Matos Almeida SE, Medeiros RM, Alencastro PR, Morgado MG: **Origin and evolutionary history of HIV-1 subtype C in Brazil.** *AIDS* 2008, **22:**1993–2000.
33. Bergenstrom AM, Abdul-Quader AS: **Injection drug use, HIV and the current response in selected low-income and middle-income countries.** *AIDS* 2010, **24**(Suppl 3):S20–S29.
34. Bello G, Guimarães ML, Passaes CP, Matos Almeida SE, Veloso VG, Morgado MG: **Short communication: Evidences of recent decline in the expansion rate of the HIV type 1 subtype C and CRF31_BC epidemics in southern Brazil.** *AIDS Res Hum Retroviruses* 2009, **25:**1065–1069.

doi:10.1186/1743-422X-9-306
Cite this article as: Almeida *et al.*: Temporal dynamics of HIV-1 circulating subtypes in distinct exposure categories in southern Brazil. *Virology Journal* 2012 **9**:306.

Apêndice 03

"Naturally occurring resistance mutations to HIV-1 entry inhibitors in subtypes B, C, and CRF31_BC"

Araújo LA, Junqueira DM, de Medeiros RM, Matte MC, Almeida SEM

Journal of Clinical Virology, 2012

Journal of Clinical Virology 54 (2012) 6–10

Contents lists available at SciVerse ScienceDirect

Journal of Clinical Virology

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

Naturally occurring resistance mutations to HIV-1 entry inhibitors in subtypes B, C, and CRF31_BC

Leonardo Augusto Luvison Araújo[a,*], Dennis Maletich Junqueira[a,b], Rubia Marília de Medeiros[a,b], Maria Cristina Cotta Matte[a,b], Sabrina Esteves de Matos Almeida[a]

[a] Centro de Desenvolvimento Científico e Tecnológico (CDCT), Fundação Estadual de Produção e Pesquisa em Saúde (FEPPS), Porto Alegre, Brazil
[b] Programa de Pós-Graduação em Genética e Biologia Molecular, Universidade Federal do Rio Grande do Sul (UFRGS), Porto Alegre, Brazil

ARTICLE INFO

Article history:
Received 31 August 2011
Received in revised form
30 December 2011
Accepted 9 January 2012

Keywords:
HIV-1
Subtypes
Natural resistance
Entry inhibitors

ABSTRACT

Background: Entry inhibitors are a class of antiretroviral (ARV) drugs that prevent HIV replication by blocking viral entry into the host cell. The investigation of naturally occurring mutations associated with entry inhibitors across subtypes is required because genetic differences between HIV-1 variants may influence the emergence of drug resistance. Despite the importance of subtype C, which predominates globally, the majority of studies include only subtype B strains.
Objectives: To investigate the presence of natural resistance mutations to entry inhibitors in HIV-1 subtypes B, C, and CRF31_BC strains.
Study design: Eighty samples were collected from antiretroviral-naïve patients. The gp41 gene from 67 patients and the gp120 gene from 65 patients were partially sequenced. Resistance mutations to entry inhibitors Enfuvirtide, Maraviroc, and Vicriviroc were screened.
Results: ENF resistance-associated mutations of HR1 and HR2 on gp41 were not associated with any subtype. However, the major polymorphisms detected in HR1: N42S, L54M, and A67T were most prevalent in subtype C ($p < 0.001$). Mutations A316T and R315Q in gp120, which are related to MVC and VCV reduced susceptibility respectively, were predominant in subtype C ($p < 0.05$).
Conclusions: This study shows that many more resistance-associated mutations to entry inhibitors in ARV-naïve patients occur in subtype C compared with subtype B strains. However, further studies will be necessary to elucidate if the differential genetic background of HIV subtypes can affect the efficacy of treatment with entry inhibitors.

1. Background

Classical therapies for HIV infection inhibit the viral enzymes reverse transcriptase and protease within the host cell.[1] Recently, therapies that target other steps of the HIV replicative cycle have been developed. Research has focused on the inhibition of the complex machinery of the entry process into the host cell, which has important implications in the overall viral replicative fitness.[2] Only two entry inhibitors have been approved by the Food and Drug Administration (FDA),[3] Enfuvirtide (ENF) and Maraviroc (MVC), which have been used in addition to other antiretroviral drugs, especially in cases of therapeutic failure.[4] The entry inhibitor in the most advanced stage of clinical development until then was Vicriviroc (VCV),[5] however due to the recent results obtained in the phase III clinical trial the development of the drug will be discontinued for the treatment of HIV infection.[6]

ENF targets the HIV protein gp41 by blocking the conformational changes necessary for the fusion process, while MVC and VCV are negative allosteric modulators of the CCR5 cell receptor, leading to a conformational change that blocks the interaction with gp120.[4]

Resistance to ENF has been shown to be associated with changes at amino acids 36–45 in the first heptad repeat region (HR1) of HIV-1 gp41, mainly in the highly conserved GIV motif at positions 36–38.[7,8] Likewise, others regions mapped within HR1, and in the second heptad repeat region (HR2), could contribute to ENF susceptibility.[8,9] On the other hand, *in vitro* studies have associated MVC and VCV reduced susceptibility with changes in gp120, mainly in the V3 loop motif.[10,11]

The inherent capability of HIV-1 to generate virological failure is due to HIV high rate of replication and error rate of reverse

Abbreviations: ARV, antiretroviral; FDA, Food and Drug Administration; ENF, Enfuvirtide; MVC, Maraviroc; VCV, Vicriviroc; HR1, first heptad repeat region; HR2, second heptad repeat region; pol, polimerase; env, envelope; CRF, circulating recombinant form; URF, unique recombinant form.

* Corresponding author at: Centro de Desenvolvimento Científico e Tecnológico (CDCT) – Fundação Estadual de Produção e Pesquisa em Saúde (FEPPS), Av. Ipiranga 5400, Porto Alegre, CEP: 90610-000, Brazil. Tel.: +55 51 3339 2386; fax: +55 51 3339 3654.
E-mail address: leonardo_luvison@hotmail.com (L.A.L. Araújo).

1386-6532/$ – see front matter © 2012 Elsevier B.V. All rights reserved.
doi:10.1016/j.jcv.2012.01.005

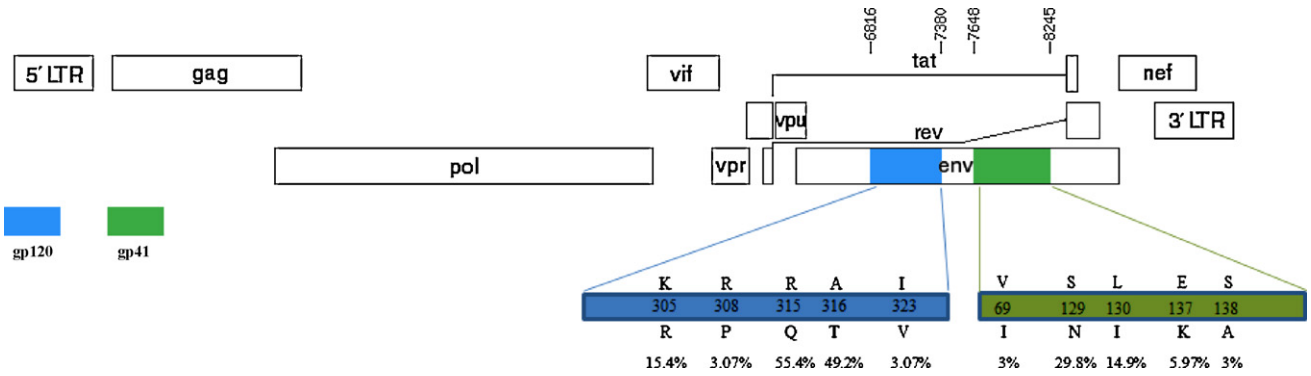


Fig. 1. Location in the HIV reference genome HXB2 of mutations to ENF (gp41) and CCR5 inhibitors (gp120). Blue and green bars highlight the main amino acid sites of change in gp120 and gp41, respectively, from wild type (amino acids above the bars) to resistant forms (amino acids under the bars). The overall mutation frequency in our population study are demonstrated. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of the article.)

transcriptase.[12] Genetic differences among HIV-1 variants can influence the emergence of drug resistance, as viral subtypes differ in their nucleotide composition.[13] Subtypes B and C are disseminated globally, with subtype C being responsible for more than half of all global infections.[14] Despite the predominance of HIV-1 subtype B in the Americas, in the southernmost region of Brazil subtypes B, C, and CRF31_BC co-circulate.[15] This region, in the face of the worldwide importance of subtypes B and C, represents an important epidemiological site to compare the mutation profiles of these subtypes.

2. Objective

The present study aims to investigate the presence of entry inhibitors resistance-associated mutations and to characterize the envelope gene polymorphisms in HIV-1 subtypes B, C, and CRF31_BC from antiretroviral-naïve patients.

3. Study design

This study included blood samples from 80 HIV-1 positive and antiretroviral (ARV) therapy naïve patients collected between 2006 and 2007 in Porto Alegre, in southernmost Brazil. The genotypes of the samples were determined using the pol gene[15] according to the REGA HIV subtyping tool (http://www.bioafrica.net/rega-genotype/html/subtypinghiv.html). The gp41 and gp120 genes were partially amplified by nested PCR, as described elsewhere.[16–18] The PCR products were sequenced using Big Dye Terminator Cycle Sequencing v3.1 (Applied Biosystems, Foster City, USA). Sequences were manually edited (Bioedit) and compared with the HXB2 reference sequence retrieved from the Los Alamos HIV Sequence Database (http://www.hiv.lanl.gov). To examine envelope (*env*) subtypes, we reconstructed two phylogenetic trees by the neighbor-joining method based on the DNA sequences of gp120 and gp41. The analysis of resistance-associated mutations to ENF included: (i) mutations in HR1 at codons 36–45 of gp41, described in the International AIDS Society (IAS): G36D/S, I37V, V38A/E/M, Q39R, Q40H, N42T, and N43D,[19] (ii) mutations in HR1 related to ENF monotherapy: Q32H/R, Q39R, R46M, and V69I,[20] and (iii) HR2 mutations: S129N, L130I, E137K, and S138A.[9,21] The gp120 following mutations related to reduced susceptibility were considered for MVC and VCV, respectively: (i) A316T, and I323V,[10] and (ii) K305R, R308P, and R315Q.[11] Polymorphisms were arbitrarily defined as mutations that occurred in more than 5% of sequences. In order to better understand the relationship of the subtypes and the primary resistance to entry inhibitors, all envelope sequences of ARV-therapy naïve individuals with HIV-1 subtypes B and C from the Los Alamos National Laboratory (LANL) HIV database were downloaded (accessed November 2011). The dataset contained sequences isolated from different countries around the world. Statistical comparisons of the subtype groups were made using Pearson's χ^2 test with adjusted residues (indicate the importance of each categories to the ultimate chi-square value) and Fisher's exact test using the SPSS 16.0 statistical package; the significance level was set at $p < 0.05$.

4. Results

Of the 80 samples evaluated, 67 provided a PCR product that could be sequenced for the partial gp41 gene (HXB2 7648–8245) and 65 for the partial gp120 gene (HXB2 6816–7380) – Fig. 1. As expected, 96.25% of the samples were concordant in the *pol* and *env* (gp41 and/or gp120) subtypes. The 80 samples comprised subtype C (46.26%), B (33.74%), CRF31_BC (10%), or URF (10%). The average period of positive serology of samples was 33 months; the mean $CD4^+$ T-cell and viral loads were approximately 380 cells/mm^3 and 4 log copies/mL, respectively.

ENF resistance-associated mutations and polymorphisms in the HR1 and HR2 regions of gp41 are detailed in Table 1. No resistance mutation was found in codons 36–45 of the HR1 region. The V69I mutation in HR1, associated with resistance to ENF monotherapy, was detected in 3% (2/67) of the samples. N42S, L54M, and A67T were the major polymorphisms detected in the HR1 region, and were most prevalent in subtype C ($p < 0.001$). HR2 mutations S129N, L130I, E137K, and S138A were found in 43.2% (29/67) of samples. In the HR2 region, we found a high prevalence of polymorphisms in all subtypes.

MVC and VCV mutations related to reduced susceptibility and polymorphisms in gp120 across the subtypes are summarized in Table 2. 52.3% (34/65) of samples harbor at least one mutation related to reduced susceptibility to MVC, corresponding 18.2% (4/22) of subtype B and 75% (21/28) of subtype C samples. Mutation A316T (19/28 – 67.8%) was most prevalent in subtype C ($p < 0.05$), and I323V was exclusively found in subtype C strains. 69.2% (45/65) of individuals had HIV strains harboring at least one mutation related to reduced susceptibility to VCV, corresponding to 27.3% (6/22) of subtype B and 96.4% (27/28) of subtype C samples. Mutation R315Q (26/28 – 92.85%) was predominant in subtype C ($p < 0.001$).

Analysis in Los Alamos database (Table 3) emphasizes the well described highly conserved amino acid sequence at residues 36–45 of gp41 from ARV-naïve patients.[22] Moreover a similar profile of mutations related to ENF was found in our population study and in sequences derived from the LANL database, except for substitutions E137K and S138A. This difference may be due to the emergence and evolutionary history of the subtypes B and C infections in Brazil[23,24]

Table 1
Enfuvirtide mutations resistance and subtype-specific polymorphisms of gp41 defined by analyzing adjusted Pearson's chi-square residues.

	Subtype					
Mutation ENF	B ($n=21$)	C ($n=31$)	CRF_31 ($n=7$)	URF ($n=8$)	p value	Overall ($n=67$)
V69I	9.5%	0%	0%	0%	NS	3%
S129N	23.8%	35.5%	42.85%	12.5%	NS	29.8%
L130I	19%	16.12%	14.3%	0%	NS	14.9%
E137K	4.7%	6.45%	0%	12.5%	NS	5.97%
S138A	4.7%	3.22%	0%	0%	NS	.3%
Polymorphisms gp41						
Q32L/K	14.3%	0%	0%	25%	**<0.05**	7.46%
Adjusted residuals	1.44	−2.16	−0.79	2.01		
residuals p value	0.151	**<0.05**	0.427	**<0.05**		
N42S	33.3%	90.3%	100%	37.5%	**<0.001**	67.16%
Adjusted residuals	−3.98	3.75	1.95	−1.9		
residuals p value	**<0.001**	**<0.001**	0.051	0.057		
R46K	23.8%	12.9%	0%	25%	NS	16.41%
L54M	47.6%	90.3%	100%	25%	**<0.001**	70.1%
Adjusted residuals	−2.72	3.35	1.82	−2.97		
residuals p value	**<0.05**	**<0.001**	0.068	**<0.05**		
A67T	33.3%	80.64%	100%	0%	**<0.001**	58.2%
Adjusted residuals	−2.79	3.46	2.37	−3.56		
residuals p value	**<0.05**	**<0.001**	**<0.05**	**<0.001**		
S129K/G/D/Q/V/Y/H	71.4%	54.8%	42.85%	50%	NS	61.2%
L130T/Q/E/X	57.1%	80.64%	85.71%	50%	NS	70.1%
E137D/Q/E/X	42.85%	74.2%	71.43%	50%	NS	61.2%
S138R	9.5%	6.45%	0%	12.5%	NS	7.46%

when compared to worldwide molecular epidemiology of these subtypes. The frequency of mutations related to reduced susceptibility to MVC and VCV derived from the LANL database is correlated with the find in the present study.

5. Discussion

To the best of our knowledge, this is the first study that evaluates the prevalence of natural resistance mutations to entry inhibitors among subtypes B, C, and CRF31_BC. All CRF31_BC samples showed subtype C envelope patterns, as previously described.[25] The URFs showed different BC recombination patterns (URF_BC).

In Brazil, ENF has been used since 2005 and MVC since 2007 in therapeutic rescue strategies for patients failing to previous ARV – regimens.[26] Due to the dates of sample collection viral mutations described here, in the ARV-naïve population, were unlikely selected and transmitted from patients under use of entry inhibitors antiretroviral therapy. Therefore, these

Table 2
CCR5 inhibitors mutations related to reduced suscetibility and subtype-specific polymorphisms of gp120 defined by analyzing adjusted Pearson's chi-square residues.

	Subtype					
Mutations CCR5 inhibitors	B ($n=22$)	C ($n=28$)	CRF_31 ($n=8$)	URF ($n=7$)	p value	Overall ($n=65$)
Vicriviroc						
K305R	18.2%	10.7%	0%	42.85%	NS	15.4%
R308P	4.5%	0%	0%	14.3%	NS	3.07%
R315Q	4.5%	92.85%	75%	42.85%	**<0.001**	55.4%
Adjusted residuals	−5,9	5.29	1.19	−0.71		
residuals p value	**<0.001**	**<0.001**	0.233	0.48		
Maraviroc						
A316T	18.2%	67.8%	75%	42.85	**<0.05**	49.2%
Adjusted residuals	−3.58	2.61	1.56	−0.36		
residuals p value	**<0.001**	**<0.05**	0.129	0.721		
I323V	0%	7.14%	0%	0%	NS	3.07%
Polymorphisms gp120						
K305E/Q/N/A	9.1%	53.5%	62.5%	0%	**<0.05**	33.84%
Adjusted residuals	−3.02	2.92	1.83	−2.00		
residuals p value	**<0.05**	**<0.05**	0.067	**<0.05**		
S306G	54.5%	10.7%	0%	0%	**<0.001**	23.1%
Adjusted residuals	4.31	−2.06	−1.65	−1.53		
residuals p value	**<0.001**	**<0.05**	0.098	0.125		
R308H/G/N	63.6%	14.3%	25%	42.85%	**<0.05**	35.4%
Adjusted residuals	3.41	−3.09	−0.66	0.44		
residuals p value	**<0.001**	**<0.05**	0.512	0.662		
G312A	9%	0%	0%	14.3%	NS	4.6%
P313G/W/L/M/F	27.3%	0%	0%	28.5%	**<0.05**	12.3%
Adjusted residuals	2.63	−2.63	−1.13	1.39		
residuals p value	**<0.05**	**<0.05**	0.258	0.166		
R315H/K/G	13.6%	3.57%	12.5%	0%	NS	7.7%
F317L/Y/W/V	18.2%	7.14%	0%	28.6%	NS	12.3%

Table 3
Subtype-specific primary mutations to Enfuvirtide and CCR5 inhibitors in worldwide samples downloaded from LANL database; defined by adjusted Pearson's chi-square residues.

Mutation ENF	Subtype		
	B (n = 406)	C (n = 416)	p value
Q32H/R	1.5%	1.2%	NS
R46M	2.4%	1.2%	NS
V69I	5.6%	2.8%	<0.05
S129N	11.6%	34.1%	<0.001
L130I	12.5%	20.7%	<0.05
E137K	17%	3.4%	<0.001
S138A	7.4%	0.7%	<0.001
Mutation CCR5 inhibitors	**Subtype**		
	B (n = 621)	C (n = 460)	p value
Vicriviroc			
K305R	17%	10.2%	<0.05
R308P	11.9%	0%	<0.001
R315Q	4%	96.1%	<0.001
Maraviroc			
A316T	5.9%	66.1%	<0.001
I323V	5%	5.6%	NS

results are related to naturally occurring mutations to entry inhibitors.

Resistance to ENF is characterized by a low genetic barrier, as phenotypic resistance arises with mutations in the codons 36–45.[27] Polymorphisms, in parallel with the resistance mutations in different portions of the gp41 HR domains, may be associated with the recovery of gp41 function.[27] Notably, subtype C exhibited higher polymorphism rates than subtype B along the gp41 sequence (Table 1). Under selective pressure of ENF, this variation may compensate for the conformational changes induced by mutations in HR1, as the envelope context in which the HR1 mutations occurred had a strong impact on phenotypic resistance.[2]

We found a high prevalence of polymorphisms 318Y (95.4%) and 319A (92.3%) in the V3 crown region of HIV-1 gp120, which were associated with increased replicative fitness and reduced sensitivity to ENF.[28] The natural polymorphism N42S occurred with the highest frequency in subtype C ($p > 0.001$) and did not reduce sensitivity to ENF.[7]

Different mutational patterns in the V3 region across subtypes may have clinical significance by influencing the effectiveness of CCR5 inhibitors. In Brazil, the V3 linage GWGR is extremely common in subtype B and is rare elsewhere.[29] We found that motif only in subtype B and the frequency of MVC resistance and polymorphisms in subtype B was similar to that found by Alencar et al. (Table 2). However, subtype C, which had not been analyzed for MVC mutations related to reduced susceptibility in Brazil, showed high proportion of mutation A316T (67.8% – $p < 0.05$). Mutations A316T and I323V have the greatest impact on MVC susceptibility, although both mutations are required for high-level of reduced susceptibility.[10]

Subtype C usually contains a highly conserved GPGQ amino acid motif at position 312–315, while GPGR is predominant in subtype B worldwide.29 In this study, 92.85% of subtype C samples harbor the R315Q ($p < 0.001$) mutation, a replacement reported to be associated with reduced susceptibility to VCV.[11] Furthermore, we found in the sequences downloaded from Los Alamos HIV database a higher rate of mutations related to reduced susceptibility to allosteric modulators of CCR5 (MVC and VCV) in subtype C, agreeing with Gonzalez et al.[30] However, despite the in vitro and in vivo studies show that changes in V3 region are an important resistance pathway to CCR5 inhibitors, these mutations conferred resistance only when present in certain genetic contexts since the envelope seems to modulate the overall effect of these changes.[31]

Despite the relatively small sample size, our data corroborates previous findings, and indicates that subtype C has a natural accumulation of polymorphisms that can generate resistance to entry inhibitors compared to subtype B. Further *in vitro* and *in vivo* studies are required to elucidate whether the differential genetic background of the subtypes analyzed affects the efficacy of antiretroviral treatment with this ARV-class.

Funding

Grant sponsor: State Foundation of Production and Research in Health of Rio Grande do Sul; Grant sponsor: Brazilian Ministry of Health (Programa DST-AIDS).

Conflict of interest

All authors declare to have no conflict of interest.

Ethical Approval

This study was approved by Fundação Estadual de Produção e Pesquisa em Saúde Ethical Research committee (process 18/2005).

Acknowledgment

We thank Maria Lucia Rosa Rossetti for encouraging us.

References

1. Pomerantz RJ, Horn DL. Twenty years of therapy for HIV-1 infection. *Nat Med* 2003;**9**:867–73.
2. Rangel HR, Weber J, Chakraborty B, Gutierrez A, Marotta ML, Mirza M, et al. Role of the human immunodeficiency virus type 1 envelope gene in viral fitness. *J Virol* 2003;**77**:9069–73.
3. U.S. Food and Drug Administration. Approved HIV-1 antiretrovirals, http://www.fda.gov [accessed 15.06.11].
4. Lobritz MA, Ratcliff AN, Arts EJ. HIV-1 entry inhibitors, and resistance. *Viruses-Basel* 2010;**2**:1069–105.
5. Gilliam BL, Riedel DJ, Redfield RR. Clinical use of CCR5 inhibitors in HIV and beyond. *J Transl Med* 2011;**9**(Suppl. 1):S9.
6. Gathe J, Diaz R, Fatkenheuer G, Zeinecker J, Mak C, Vilchez R, et al. Phase 3 trials of vicriviroc in treatment-experienced subjects demonstrate safety but not significantly superior efficacy over potent background regimens alone. In: *Program and abstracts of the 17th conference on retroviruses & opportunistic infections*. 2010. Abstract 54LB.
7. Sista PR, Melby T, Davison D, Jin L, Mosier S, Mink M, et al. Characterization of determinants of genotypic and phenotypic resistance to enfuvirtide in baseline and on-treatment HIV-1 isolates. *AIDS* 2004;**18**:1787–94.
8. Carmona R, Perez-Alvarez L, Munoz M, Casado G, Delgado E, Sierra M, et al. Natural resistance-associated mutations to Enfuvirtide (T20) and polymorphisms in the gp41 region of different HIV-1 genetic forms from T20 naive patients. *J Clin Virol* 2005;**32**:248–53.
9. Xu L, Pozniak A, Wildfire A, Stanfield-Oakley SA, Mosier SM, Ratcliffe D, et al. Emergence and evolution of enfuvirtide following long-term therapy involves heptad repeat 2 mutations within gp41. *Antimicrob Agents Chemother* 2005;**49**:1113–9.
10. Westby M, Smith-Burchnell C, Mori J, Lewis M, Mosley M, Stockdale M, et al. Reduced maximal inhibition in phenotypic susceptibility assays indicates that viral strains resistant to the CCR5 antagonist maraviroc utilize inhibitor-bound receptor for entry. *J Virol* 2007;**81**:2359–71.
11. Ogert RA, Wojcik L, Buontempo C, Ba L, Buontempo P, Ralston R, et al. Mapping resistance to the CCR5 co-receptor antagonist vicriviroc using heterologous chimeric HIV-1 envelope genes reveals key determinants in the C2-V5 domain of gp120. *Virology* 2008;**373**:387–99.
12. Mansky LM, Temin HM. Lower in vivo mutation rate of human immunodeficiency virus type 1 than that predicted from the fidelity of purified reverse transcriptase. *J Virol* 1995;**69**:5087–94.
13. Geretti AM. HIV-1 subtypes: epidemiology and significance for HIV management. *Curr Opin Infect Dis* 2006;**19**:1–7.
14. Koh WW, Forsman A, Hué S, Van der Velden GJ, Yirrell DL, McKnight A, et al. Novel subtype C human immunodeficiency virus type 1 envelopes cloned directly from plasma: coreceptor usage and neutralization phenotypes. *J Gen Virol* 2010;**91**:2374–80.
15. De medeiros RM, Junqueira DM, Matte MCC, Barcellos NT, Chies JAB, Almeida SEM. Co-circulation HIV-1 subtypes B C, and CRF31_BC in a drug-naive population from southernmost Brazil: analysis of primary resistance mutations. *J Med Virol* 2011;**83**:1682–8.

16. Guimarães ML, dos Santos Moreira A, Loureiro R, Galvão-Castro B, Morgado MG. High frequency of recombinant genomes in HIV type 1 samples from Brazilian southeastern and southern regions. *AIDS Res Hum Retroviruses* 2002;**18**: 1261–9.
17. Fang G, Weiser B, Kuiken C, Philpott SM, Rowland-Jones S, Plumer F, et al. Recombination following superinfection by HIV-1. *AIDS* 2004;**18**:153–9.
18. Qu S, Ma L, Yuan L, Xu W, Hong K, Xing H, et al. Co-receptor usage and prediction of V3 genotyping algorithms in HIV-1 subtype B' from paid blood donors experienced anti-retroviral therapy in Chinese central province. *J Virol* 2010;**22**(7):280.
19. Johnson VA, Brun-Vezinet F, Clotet B, Gunthard HF, Kuritzkes DR, Pillay D, et al. Update of the drug resistance mutations in HIV-1: December 2010. *Top HIV Med* 2010;**18**:156–63.
20. Wei X, Decker JM, Liu H, Zhang Z, Arani RB, Kilby JM, et al. Emergence of resistant human immunodeficiency virus type 1 in patients receiving fusion inhibitor (T20) monotherapy. *Antimicrob Agents Chemother* 2002;**46**: 1896–905.
21. Teixeira C, De Sá-Filho D, Alkmim W, Janini LM, Diaz RS, Komninakis S. Short communication: high polymorphism rates in the HR1 and HR2 gp41 and presence of primary resistance-related mutations in HIV type 1 circulating in Brazil: possible impact on enfuvirtide efficacy. *AIDS Res Hum Retroviruses* 2010;**26**:307–11.
22. Greenberg ML, Greenberg N. Resistance to enfuvirtide, the first HIV fusion inhibitor. *J Antimicrob Chemother* 2004;**54**:333–40.
23. Junqueira DM, de Medeiros RM, Matte MCC, Araújo LAL, Chies JAB, et al. Reviewing the history of HIV-1: spread of subtype B in the Americas. *PLoS ONE* 2011;**6**(11):e27489.
24. Bello G, Passaes CPB, Guimarães ML, Lorete RS, Almeida SEM, de Medeiros RM, et al. Origin and evolutionary history of HIV-1 subtype C in Brazil. *AIDS* 2008;**22**:1993–2000.
25. Passaes CP, Bello G, Lorete RS, Almeida SEM, Junqueira DM, Veloso VG, et al. Genetic characterization of HIV-1 BC recombinants and evolutionary history of the CRF31_BC in Southern Brazil. *Infect Genet Evol* 2009;**9**:474–82.
26. Alencar CS, Nishiya AS, Ferreira S, Giret MTM, Diaz RS, Sabino EC. Evaluation of primary resistance to HIV entry inhibitors among Brazilian patients failing reverse transcriptase/protease inhibitors treatment reveal high prevalence of maraviroc resistance-related mutations. *AIDS Res Hum Retroviruses* 2010;**26**:1267–71.
27. Menzo S, Castagna A, Monachetti A, Hasson H, Danise A, et al. Genotype and phenotype patterns of human immunodeficiency virus type 1 resistance to enfuvirtide during long-term treatment. *Antimicrob Agents Chemother* 2004;**48**:3253–9.
28. Lobritz MA, Marozsan AJ, Troyer RM, Arts EJ. Natural variation in the V3 crown of human immunodeficiency virus type 1 affects replicative fitness and entry inhibitor sensitivity. *J Virol* 2007;**81**:8258–69.
29. Leal E, Villanova FE. Diversity of HIV-1 subtype B: implications to the origin of bf recombinants. *PLoS ONE* 2010;**28**:e11833.
30. Gonzalez S, Gondwe C, Tully D, Minhas V, Shea D, Kankasa C, et al. Short communication: antiretroviral therapy resistance mutations present in the HIV type 1 subtype C pol and env regions from therapy-naive patients in Zambia. *AIDS Res Hum Retroviruses* 2010;**26**:795–803.
31. Moore JP, Kuritzkes DR. A pièce de resistance: how HIV-1 escapes small molecule CCR5 inhibitors. *Curr Opin HIV AIDS* 2009;**4**:118–24.

Apêndice 04

"Dissemination of nonpandemic Caribbean HIV-1 subtype B clades in Latin America"

Cabello M, Junqueira DM, Bello G

AIDS, 2015

Dissemination of nonpandemic Caribbean HIV-1 subtype B clades in Latin America

Marina Cabello[a], Dennis Maletich Junqueira[b,c,d] and Gonzalo Bello[a]

Objective: To estimate the prevalence of the HIV-1 subtype B pandemic ($B_{PANDEMIC}$) and Caribbean (B_{CAR}) clades in Latin America and to reconstruct the spatiotemporal dynamics of dissemination of the B_{CAR} clades in the region.

Design: A total of 7654 HIV-1 subtype B *pol* sequences collected from 18 different Latin American countries between 1989 and 2011 were analyzed together with subtype B reference sequences representative of the $B_{PANDEMIC}$ (US/France = 300) and the B_{CAR} (Caribbean = 279, Panama = 37) clades.

Methods: Phylogeographic and evolutionary parameters were estimated from sequence data using maximum likelihood and Bayesian coalescent-based methods.

Results: Nonpandemic B_{CAR} strains were probably disseminated from the Caribbean islands of Hispaniola and Trinidad and Tobago into Latin America since the early 1970s. The B_{CAR} strains reached nearly all countries from Latin America here analyzed and in some of them were spread locally, although their overall prevalence in the region is low. The $B_{PANDEMIC}$ clade comprises more than 90% of subtype B infections in most countries analyzed, with exception of Suriname, French Guyana and probably Guyana, where both $B_{PANDEMIC}$ and B_{CAR} clades seem to circulate at a similar prevalence.

Conclusion: This study demonstrates that nonpandemic subtype B lineages of Caribbean origin have been disseminated into Latin America shortly after the estimated introduction of subtype B in the continent. Despite their early dissemination, the B_{CAR} strains account for a minor fraction of current HIV-1 subtype B infections in the region that are mainly driven by spreading of the globally disseminated $B_{PANDEMIC}$ clade.

AIDS 2015, **29**:483–492

Keywords: HIV-1, Latin America, nonpandemic, phylogeography, subtype B

Introduction

An estimated 1.5 million people were living with the HIV type 1 (HIV-1) in Latin America in 2012, most of them concentrated in Brazil (40%), Mexico (11%), Colombia (10%), Venezuela (7%) and Argentina (6.5%) [1]. The HIV prevalence in the adult population (15–49 years) ranges from 0.2% in Mexico to more than 1.0% in Belize, Guyana and Suriname [1]. Most of the HIV epidemics in this region are concentrated in and around networks of MSM, although heterosexual HIV transmission is increasing in the older epidemics in South America and injecting drug use is another significant route of HIV transmission, especially in the southern cone of South America and in Mexico [2].

The HIV-1 group M subtype B is the most prevalent clade in Latin America, accounting for about 70% of infections in the region [3]. The spread of HIV-1 subtype B in the Americas probably occurred via a single introduction event from Central Africa into Haiti around the middle 1960s and later dissemination of the virus from Haiti to

[a]Laboratório de AIDS e Imunologia Molecular, Instituto Oswaldo Cruz, FIOCRUZ, Rio de Janeiro, [b]Centro de Desenvolvimento Científico e Tecnológico, Fundação Estadual de Produção e Pesquisa em Saúde, [c]Uniritter Laureate International Universities, Departamento de Ciências da Saúde, and [d]Programa de Pós-Graduação em Genética e Biologia Molecular (PPGBM), Universidade Federal do Rio Grande do Sul (UFGRS), Porto Alegre, Rio Grande do Sul, Brazil.

Correspondence to Gonzalo Bello, Laboratório de AIDS e Imunologia Molecular, Instituto Oswaldo Cruz, FIOCRUZ, Av Brasil 4365, 21045-900 Rio de Janeiro, RJ, Brazil.

Tel +55 21 3865 8227; fax: +55 21 3865 8173; e-mail: gbellobr@gmail.com/gbello@ioc.fiocruz.br.
Received: 5 October 2014; revised: 14 November 2014; accepted: 18 November 2014.

DOI:10.1097/QAD.0000000000000552

ISSN 0269-9370

other Caribbean islands and to the United States [4]. The virus that entered the United States was further disseminated from this country to other countries around the world, establishing a 'subtype B pandemic' ($B_{PANDEMIC}$) clade, whereas other subtype B lineages seem to have remained mostly restricted to the Caribbean [subtype B Caribbean (B_{CAR}) clades] [4]. A recent study conducted by our group analyzed 1042 HIV-1 subtype B *pol* gene sequences from 14 different Caribbean countries and revealed that nonpandemic B_{CAR} lineages have been widely disseminated through the Caribbean region since the late 1960s, accounting for an important fraction of current HIV-1 infections in several countries including Haiti and the Dominican Republic (~75%), Jamaica (~50%) Trinidad and Tobago (~95%), and other Lesser Antilles (~40–75%) [5].

Two previous studies suggest that nonpandemic B_{CAR} lineages may have been also directly disseminated from the Caribbean islands into South [6] and Central [7] American countries. The study of Junqueira *et al.* [6] identified a few HIV-1 subtype B *pol* sequences from Brazil, Colombia, Guyana, Suriname and Venezuela that were phylogenetically intermixed among basal nonpandemic Caribbean sequences, suggesting a direct epidemiological link between the Caribbean and South American epidemics. Another recent study showed that a minor fraction (5.5%) of Panamanian subtype B *pol* sequences were also intermixed among nonpandemic B_{CAR} strains and further suggests that some of those B_{CAR} clades were mainly disseminated in Panama by heterosexual transmission [7]. Overall, these results suggest that the B_{CAR} clades have not remained confined to the Caribbean region, but have been also disseminated to continental regions of the Americas. The relative prevalence of the $B_{PANDEMIC}$ and B_{CAR} clades across different Latin American countries, however, remains largely unknown.

The objective of this study was to estimate the current prevalence of the $B_{PANDEMIC}$ and B_{CAR} clades in Latin America and to reconstruct the spatiotemporal dynamics of dissemination of the HIV-1 B_{CAR} clades in the region. For this, we used a comprehensive dataset of HIV-1 subtype B *pol* sequences ($n = 7654$) isolated from 18 different Latin American countries between 1989 and 2011. These Latin American sequences were combined with subtype B reference sequences representative of the $B_{PANDEMIC}$ (US/France = 300) and the B_{CAR} (Caribbean/Panama = 316) clades and then subjected to maximum likelihood and Bayesian phylogeographic analyses.

Methods

HIV-1 subtype B *pol* sequence dataset

We downloaded all HIV-1 subtype B *pol* sequences from Latin America that covered the entire protease and partial reverse transcriptase (PR/RT) regions (nucleotides 2253–3260 relative to HXB2 clone) and were available at the Los Alamos HIV Database (http://www.hiv.lanl.gov) by December 2013. Additional HIV-1 subtype B *pol* sequences from Latin America covering only part of the reverse transcriptase (nucleotides 2673–3203 relative to the HXB2 clone) were also downloaded for some countries with few PR/RT sequences available (Bolivia, Suriname and French Guyana). The subtype assignment of all sequences included here was confirmed using the REGA HIV subtyping tool v.2 [8] and by performing phylogenetic analyses (see below) with HIV-1 group M subtype reference sequences. Only one sequence per individual was selected and those sequences containing frameshift mutations or with incorrect subtype assignment were removed. This resulted in a final dataset of 7654 subtype B *pol* sequences isolated from 18 Latin American countries between 1989 and 2011. These sequences were aligned with subtype B *pol* (PR/RT) sequences from the United States ($n = 165$), France ($n = 135$), the Caribbean ($n = 279$) and Panama ($n = 37$), representative of the $B_{PANDEMIC}$ and the B_{CAR} clades as described previously [5,7]. Sequences were aligned using the Clustal W program [9] and all sites associated with major antiretroviral drug resistance in protease (30, 32, 46, 47, 48, 50, 54, 76, 82, 84, 88 and 90) and reverse transcriptase (41, 65, 67, 69, 70, 74, 100, 101, 103, 106, 115, 138, 151, 181, 184, 188, 190, 210, 215, 219 and 230) were excluded. All alignments are available from the authors upon request.

Phylogenetic analysis

Maximum likelihood phylogenetic trees were inferred under the generalized time reversible (GTR)+I+Γ nucleotide substitution model selected using the jModeltest program [10]. The maximum likelihood trees were reconstructed with the PhyML program [11] using an online web server [12]. Heuristic tree search was performed using the subtree-pruning-regrafting (SPR) branch-swapping algorithm and the reliability of the obtained topology was estimated with the approximate likelihood-ratio test [13] based on the Shimodaira–Hasegawa-like procedure. The maximum likelihood trees were visualized using the FigTree v1.4.0 program [14].

Analysis of the spatiotemporal dispersion pattern

The evolutionary rate, the age of the most recent common ancestor (T_{MRCA}) and the spatial diffusion pattern of nonpandemic HIV-1 subtype B clades circulating in South America were jointly estimated using the Bayesian Markov Chain Monte Carlo approach as implemented in BEAST v1.8 [15,16] with BEAGLE to improve run-time [17]. Analyses were performed using the GTR+I+Γ_4 nucleotide substitution model, a relaxed uncorrelated lognormal molecular clock model [18], and a Bayesian skyline coalescent tree prior [19]. The mean evolutionary rates previously estimated for the subtype B *pol* gene ($2.0–2.5 \times 10^{-3}$ substitutions/site per year)

[7,20–22] were incorporated as an informative prior interval. Migration events throughout the phylogenetic history and the most relevant migration pathways were reconstructed using a reversible discrete phylogeography model and the Bayesian stochastic search variable selection approach [23], with a continuous-time Markov chain (CTMC) rate reference prior [24]. Three Markov Chain Monte Carlo chains were run for 500×10^6 generations and then combined using LogCombiner v1.8. Convergence and uncertainty of parameter estimates were assessed by calculating the effective sample size and 95% highest probability density (HPD) values, respectively, after excluding the initial 10% of each run with Tracer v1.6 [25]. The maximum clade credibility tree was summarized with TreeAnnotator v1.8 and visualized with FigTree v1.4.0. Migratory events were summarized using the cross-platform SPREAD application [26].

Results

Detection of HIV-1 subtype B Caribbean clades in the majority of Latin American countries

In order to estimate the relative prevalence of pandemic ($B_{PANDEMIC}$) and nonpandemic (B_{CAR}) subtype B lineages in Latin America, *pol* (PR/RT) sequences from different Latin American countries were divided into six subsets: Central America ($n = 688$), Mexico ($n = 1677$), Argentina ($n = 1548$), Brazil-I ($n = 1329$), Brazil-II ($n = 1329$) and other South American countries ($n = 909$). A seventh subset containing shorter subtype B *pol* (reverse transcriptase) sequences from some Latin American countries poorly represented in the PR/RT dataset (Bolivia = 45, French Guyana = 108, Suriname = 21) was also constructed. Each of the seven Latin American subsets was combined with a reference subtype B dataset selected from a previous study [5] containing 500 sequences representative of the $B_{PANDEMIC}$ (US/France = 300) and the BCAR (Caribbean = 200) clades (Table S1, http://links.lww.com/QAD/A619). The maximum likelihood analyses of all PR/RT (Fig. 1a and Fig. S1, http://links.lww.com/QAD/A619) and reverse transcriptase (Fig. 1b) subsets confirmed the complete segregation of the $B_{PANDEMIC}$ reference sequences in a highly supported (approximate likelihood-ratio test >0.90) monophyletic clade nested within basal B_{CAR} reference sequences. The maximum likelihood analyses also confirmed the circulation of B_{CAR} sequences in most Latin American countries, although with highly variable prevalence (Fig. 2 and Table S2, http://links.lww.com/QAD/A619). The B_{CAR} sequences reach a high prevalence (40–50%) in French Guyana and Suriname; low prevalence (1–10%) in Brazil, Colombia, Ecuador, Mexico, Panama and Venezuela; and very low prevalence (<1%) in Argentina, El Salvador, Honduras and Peru. We found no evidence of circulation of B_{CAR} clades in Bolivia and Chile. The number of PR/RT or reverse transcriptase sequences from Belize, Costa Rica, Guatemala, Guyana, Nicaragua, Paraguay and Uruguay was too small ($n < 10$) to allow any conclusion about the relative prevalence of different subtype B clades circulating in those Latin America countries.

Spatiotemporal dispersal pattern of the HIV-1 subtype B Caribbean clades in Latin America

To reconstruct the origin and spatiotemporal dynamics of nonpandemic subtype B Latin American lineages, the HIV-1 B_{CAR} PR/RT sequences with known sampling date from Latin America here identified ($n = 103$) were combined with B_{CAR} PR/RT sequences from the most widely sampled ($n > 10$) Caribbean islands [Dominican Republic ($n = 123$), Jamaica ($n = 73$), Trinidad and Tobago ($n = 50$), and Haiti ($n = 12$)] and from Panama ($n = 37$), previously identified [5,7]. The B_{CAR} sequences were further aligned with subtype D PR/RT sequences ($n = 10$) from the Democratic Republic of Congo that was pointed as the most probable source of subtype B strain introduced in the Americas [4]. HIV-1 subtypes B and D sequences were classified into 14 discrete geographic locations (Table S3, http://links.lww.com/QAD/A619) and subjected to Bayesian phylogeographic analysis.

The mean estimated evolutionary rate of the HIV-1 B_{CAR}/D *pol* dataset was 2.1×10^{-3} substitutions/site per year (95% HPD 2.0×10^{-3} – 2.2×10^{-3} substitutions/site per year), whereas the corresponding median coefficient of rate variation was 0.31 (95% HPD: 0.27–0.35), supporting the selection of a relaxed molecular clock model. The root location of the HIV-1 subtype B ancestor was most probably placed in the island of Hispaniola (Dominican Republic/Haiti) (posterior state probability = 0.92) (Fig. 3), consistent with previous findings [4,5]. The median estimated T_{MRCA} of subtypes B/D (1956), subtype D (1968) and subtype B (1968) were also very similar to that previously obtained using different *pol* and *env* datasets [4,5] (Table 1). The close match of major spatiotemporal calibration points across different studies validates the time scale inferred from this analysis and indicates that the overall phylogeographic reconstruction was quite robust to the inclusion of new B_{CAR} sequences from Latin America.

After the introduction of HIV-1 subtype B into Hispaniola around the middle 1960s, nonpandemic B_{CAR} lineages were independently disseminated to other countries from the Caribbean and Latin America from the early 1970s onwards. Some of those viral migrations seeded secondary outbreaks that resulted in the origin of several country-specific B_{CAR} subclades including those previously identified in Trinidad and Tobago ($B_{CAR\text{-}TT}$) [4,5], Jamaica ($B_{CAR\text{-}JM\text{-}I}$) [5] and Panama ($B_{CAR\text{-}PA\text{-}I}$, $B_{CAR\text{-}PA\text{-}II}$ and $B_{CAR\text{-}PA\text{-}III}$) [7], and others here identified in Argentina ($B_{CAR\text{-}AR}$), Brazil ($B_{CAR\text{-}BR\text{-}I}$, $B_{CAR\text{-}BR\text{-}II}$ and $B_{CAR\text{-}BR\text{-}III}$), Guyana ($B_{CAR\text{-}GY}$), Mexico ($B_{CAR\text{-}MX\text{-}I}$ and $B_{CAR\text{-}MX\text{-}II}$) and Venezuela

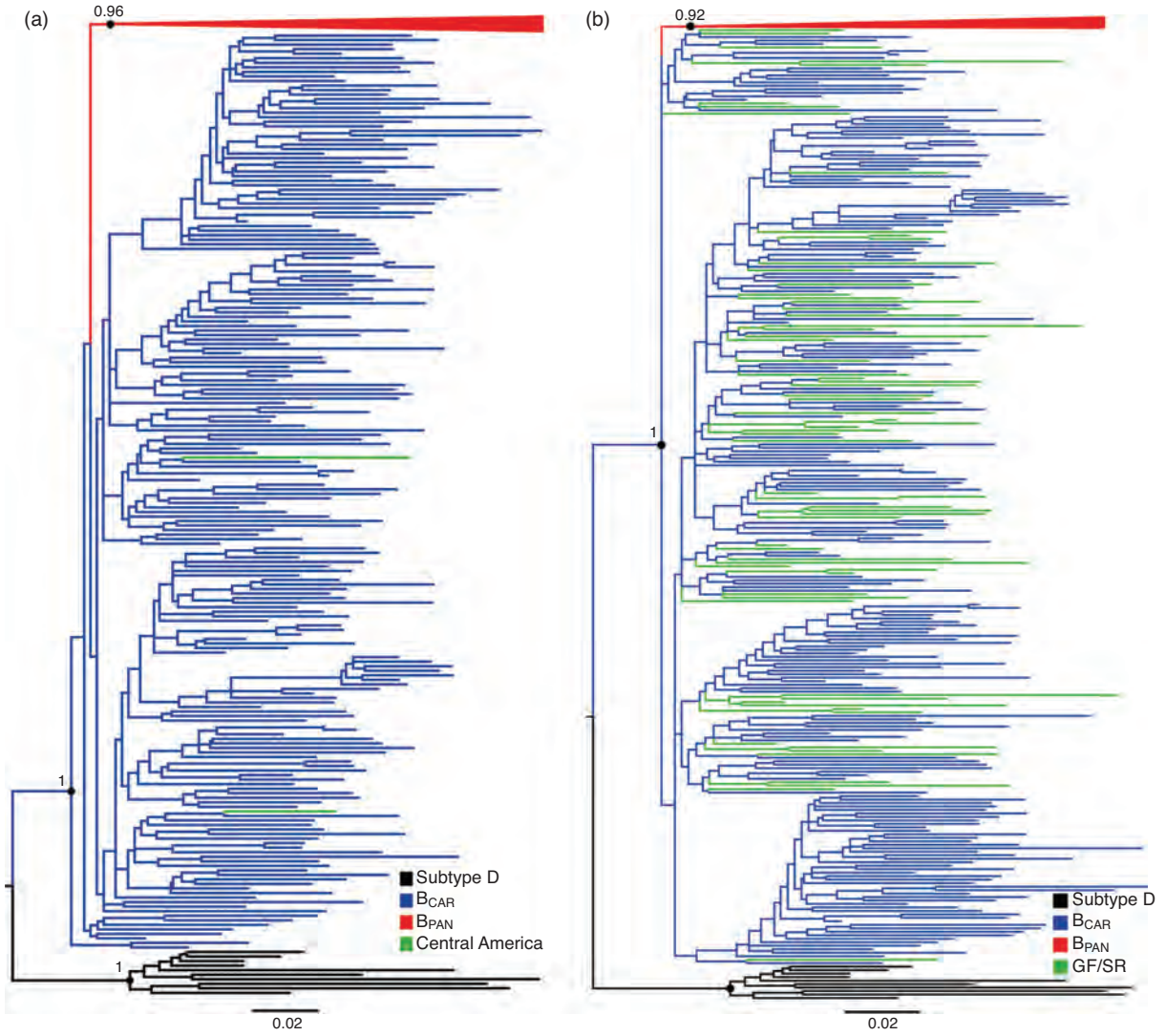


Fig. 1. Maximum likelihood phylogenetic tree of (a) HIV-1 subtype B *pol* protease and partial reverse transcriptase sequences (∼1000 nucleotides) circulating in Central America ($n = 688$) and representative sequences of the $B_{PANDEMIC}$ (US = 165, France = 135) and the B_{CAR} (Caribbean = 200) clades; (b) HIV-1 subtype B *pol* reverse transcriptase (∼600 nucleotides) sequences from Bolivia ($n = 45$), French Guyana ($n = 108$), Suriname ($n = 21$) and the representative sequences of the $B_{PANDEMIC}$ and the B_{CAR} clades. Branches are colored according to the geographic origin/clade classification of each sequence as indicated at the legend (bottom right). The $B_{PANDEMIC}$ clade was collapsed for visual clarity. The approximate likelihood-ratio test support values are indicated at key nodes. Trees were rooted using HIV-1 subtype D reference sequences. The branch lengths are drawn to scale with the bar at the bottom indicating nucleotide substitutions per site. B_{CAR}, subtype B Caribbean; $B_{PANDEMIC}$, subtype B pandemic.

(B_{CAR-VE}) (Fig. 3). The nonpandemic clades B_{CAR-TT}, $B_{CAR-JM-I}$ and $B_{CAR-BR-I}$ seem to have originated around the early 1970s, whereas most of the remaining country-specific B_{CAR} clades probably arose between the late 1970s and the middle 1980s (Fig. 3 and Table 1).

Reconstruction of viral migrations across time suggests that Hispaniola was the major hub of dissemination of nonpandemic subtype B clades in the region and further identified a few secondary hubs in the Caribbean (Trinidad and Tobago) and South America (Brazil and Guyana) (Fig. 4a and 4b). The B_{CAR-TT} clade was independently disseminated from Trinidad and Tobago to other Caribbean islands and to several South American countries including Brazil, Guyana (originating the B_{CAR-GY} clade), Suriname and Venezuela. The B_{CAR-GY} clade was disseminated from Guyana to Suriname and the $B_{CAR-BR-I}$ clade was disseminated from Brazil to Argentina at multiple times (originating the B_{CAR-AR} clade). The Bayes factor tests for significant nonzero rates supports epidemiological linkage between Hispaniola and most other Caribbean and Latin American countries included in the study (with exception of Argentina and Guyana) as well as between Trinidad and Tobago and

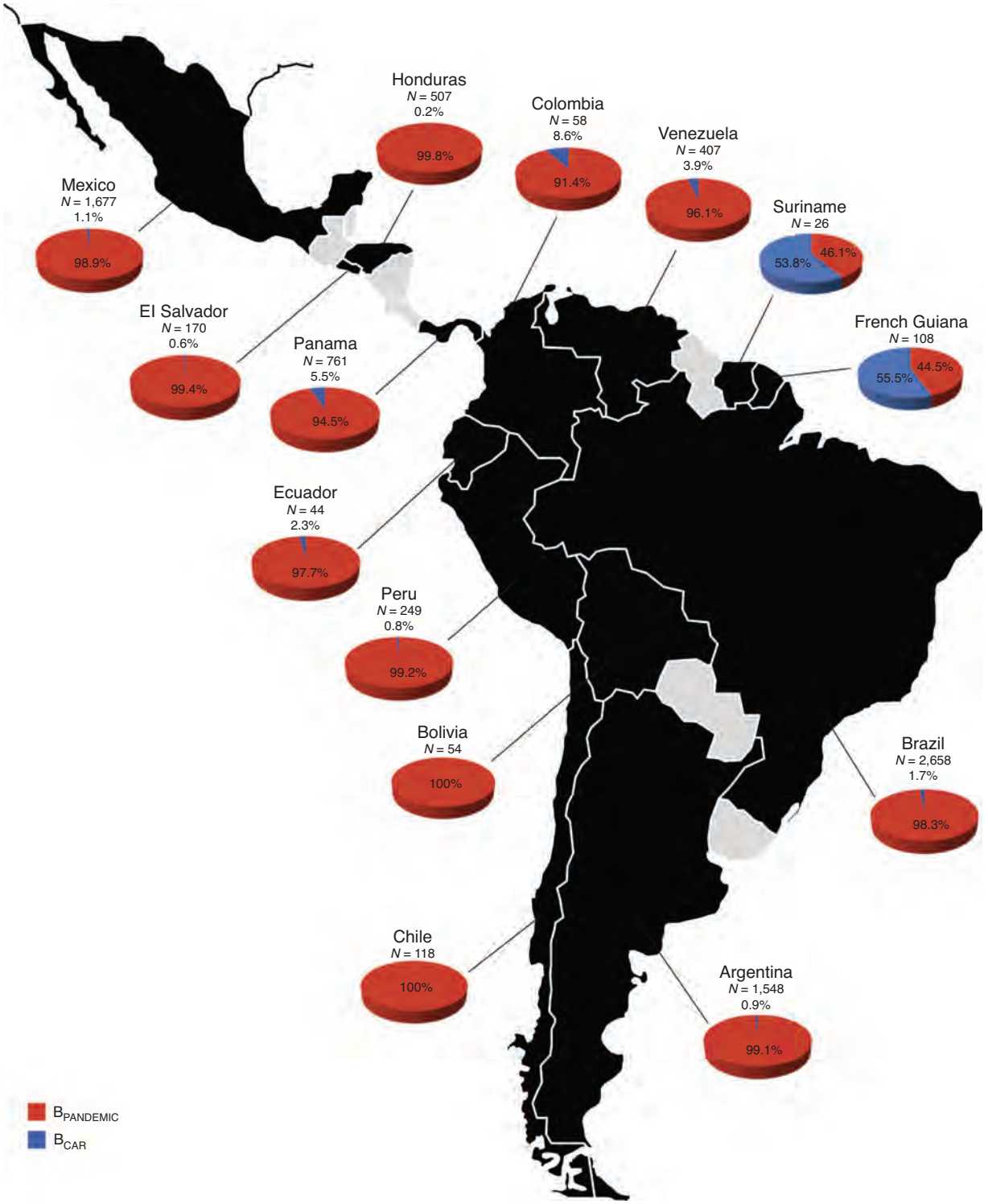


Fig. 2. Estimated proportion of B_{CAR} and $B_{PANDEMIC}$ clades among HIV-1 subtype B-infected individuals from different Latin American countries according to the maximum likelihood analyses. The total number of sequences analyzed in each locality is indicated. Proportions in Panama were estimated in a previous study [7]. Proportions in Latin American countries poorly sampled (n <10) were not estimated. B_{CAR}, subtype B Caribbean; $B_{PANDEMIC}$, subtype B pandemic.

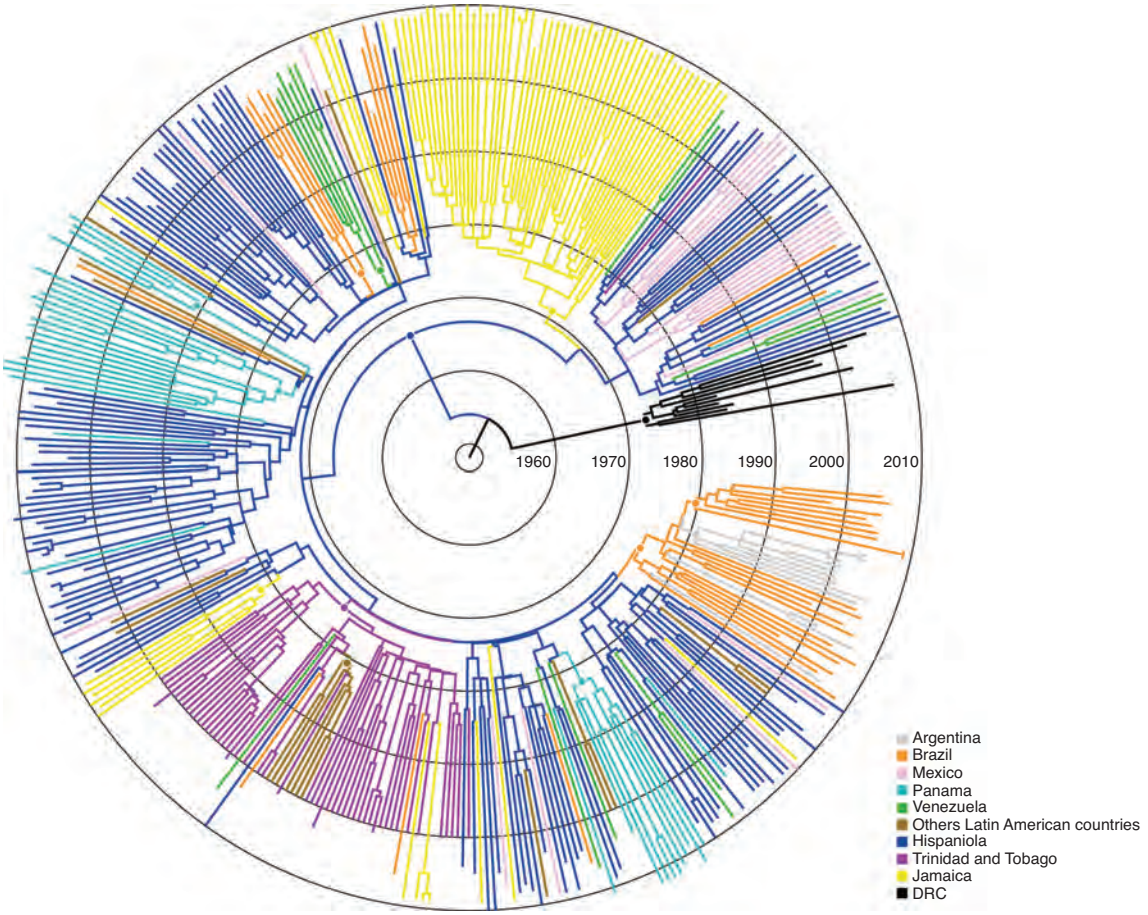


Fig. 3. Time-scaled Bayesian Markov Chain Monte Carlo tree of *pol* protease and partial reverse transcriptase sequences of HIV-1 B_{CAR} lineages from Latin America and the Caribbean, and subtype D reference sequences from the Democratic Republic of Congo (DRC). Branches are colored according to the most probable location state of their descendent nodes as indicated in the legend (bottom right). Colored circles indicate the positions of nodes corresponding to the most recent common ancestors of major country-specific clades (clade size ≥4). Branch lengths are depicted in units of time (years). The tree was automatically rooted under the assumption of a relaxed molecular clock.

Table 1. Bayesian time-scale estimates of most recent common ancestor of HIV-1 subtypes B and D and major subtype B Caribbean clades from Latin America and the Caribbean.

Clade	T_{MRCA} current study	T_{MRCA} Cabello *et al.* [5]	T_{MRCA} Gilbert *et al.* [4]
Subtypes B/D	1956 (1946–1963)	1952 (1943–1960)	1954 (1946–1961)
Subtype D	1968 (1961–1973)	1965 (1958–1971)	1966 (1961–1971)
Subtype B	1968 (1963–1972)	1964 (1959–1969)	1966 (1962–1970)
$B_{CAR\text{-}TT}$	1973 (1969–1976)	1969 (1966–1973)	1973 (1970–1976)
$B_{CAR\text{-}JM\text{-}I}$	1973 (1969–1976)	1971 (1967–1975)	–
$B_{CAR\text{-}JM\text{-}II}$	1982 (1977–1987)	–	–
$B_{CAR\text{-}BR\text{-}I}$	1973 (1971–1977)	–	–
$B_{CAR\text{-}BR\text{-}II}$	1979 (1975–1983)	–	–
$B_{CAR\text{-}BR\text{-}III}$	1983 (1977–1987)	–	–
$B_{CAR\text{-}MX\text{-}I}$	1980 (1974–1986)	–	–
$B_{CAR\text{-}MX\text{-}II}$	1981 (1976–1987)	–	–
$B_{CAR\text{-}PA\text{-}I}$	1977 (1973–1981)	–	–
$B_{CAR\text{-}PA\text{-}II}$	1980 (1976–1984)	–	–
$B_{CAR\text{-}PA\text{-}III}$	1989 (1983–1995)	–	–
$B_{CAR\text{-}AR}$	1979 (1975–1982)	–	–
$B_{CAR\text{-}VE}$	1978 (1974–1983)	–	–
$B_{CAR\text{-}GY}$	1980 (1977–1984)	–	–

$B_{CAR\text{-}AR}$, subtype B Caribbean subclades in Argentina; $B_{CAR\text{-}BR}$, subtype B Caribbean subclades in Brazil; $B_{CAR\text{-}GY}$, subtype B Caribbean subclades in Guyana; $B_{CAR\text{-}JM}$, subtype B Caribbean subclades in Jamaica; $B_{CAR\text{-}MX}$, subtype B Caribbean subclades in Mexico; $B_{CAR\text{-}PA}$, subtype B Caribbean subclades in Panama; $B_{CAR\text{-}TT}$, subtype B Caribbean subclades in Trinidad and Tobago; $B_{CAR\text{-}VE}$, subtype B Caribbean subclades in Venezuela; T_{MRCA}, age of the most recent common ancestor.

Fig. 4. Spatiotemporal dynamics of dissemination of nonpandemic HIV-1 B_{CAR} clades in Latin America. (a and b) Viral migration events occurred between 1970 and 2013 are indicated. Lines between locations represent branches in the Bayesian maximum clade credibility tree along which location transitions occurred. The line's color informs the estimated years of the viral migrations and only the earliest transitions between each location pair were represented. (c and d) Most significant epidemiological links of the dissemination process of B_{CAR} clades. Only epidemiological links supported by Bayes factor rates more than 3 are displayed. Viral migrations and most significant epidemiological links connecting the Hispaniola (a and c) and Trinidad and Tobago (b and d) with Latin American countries were separated in independent panels only for visual clarity. B_{CAR}, subtype B Caribbean.

Jamaica/Guyana/Brazil, between Brazil and Argentina, and between Guyana and Suriname (Fig. 4c and 4d and Table S4, http://links.lww.com/QAD/A619).

Discussion

The HIV-1 subtype B virus was probably originally introduced into Haiti seeded by the epidemic from the Democratic Republic of Congo around the middle 1960s [4]. After a short period of local expansion within the island of Hispaniola (shared by Haiti and the Dominican Republic), the virus seems to have moved out on several independent occasions. The introduction of the virus into the United States around the late 1960s explosively amplified the number of new cases of HIV-1 subtype B infection and originates a $B_{PANDEMIC}$ strain that was disseminated across the world [4]. Other secondary outbreaks simultaneously emerged in the Caribbean [5] and Latin America [6,7] as the result of short-distance disseminations of nonpandemic B_{CAR} strains out of Hispaniola. This study demonstrates that B_{CAR} strains reached nearly all countries in Latin America, although their prevalence is usually much lower than that estimated for the $B_{PANDEMIC}$ clade (Fig. 2). The only exceptions in the region were Suriname, French Guyana and probably Guyana, where both $B_{PANDEMIC}$ and B_{CAR} clades seem to circulate at roughly similar prevalence.

Our results indicate that Haiti and Dominican Republic, which together are home to about 75% of people living with HIV in the Caribbean [27], were probably the major sources of B_{CAR} lineages disseminated into the region. Nonpandemic B_{CAR} strains started to spread from Hispaniola in the beginning of the 1970s and would have reached Trinidad and Tobago, Jamaica, Brazil, Colombia, Ecuador, El Salvador, Honduras, Mexico, Panama, Suriname and Venezuela in the following years. Trinidad and Tobago can be viewed as a secondary hub, seeding tertiary B_{CAR} outbreaks in short-distanced countries such as Jamaica, Venezuela, Guyana and Brazil. Jamaica, by contrast, seems to have played a minor role in the regional dispersion of B_{CAR} strains. We also identified short-distance spreading of B_{CAR} lineages from Brazil to Argentina and from Guyana to Suriname, indicating that some South American countries also acted as secondary hubs of dissemination of nonpandemic subtype B lineages in the region.

Although Dominican Republic, Haiti and Trinidad and Tobago were pointed as the most important sources of B_{CAR} lineages disseminated to Latin America, we cannot rule out the possible role of other Caribbean islands with high prevalence of B_{CAR} strains such as Martinique, Guadeloupe and other Lesser Antilles [5] not included in our phylogeographic analysis because of the very low numbers ($n < 10$) of PR/RT sequences available. This geographical sampling bias may have resulted in an overestimation of the role of Dominican Republic, Haiti and Trinidad and Tobago as source of B_{CAR} lineages in the region. The use of more geographically balanced HIV-1 subtype B Caribbean datasets will be of paramount importance to obtain more precise estimates of the contribution of each Caribbean island in the regional dissemination of nonpandemic subtype B strains.

Several country-specific B_{CAR} clades were detected in Argentina, Brazil, Guyana, Mexico, Panama and Venezuela, suggesting that despite their overall low prevalence, nonpandemic subtype B lineages have been disseminated locally in several Latin American countries. Estimation of the T_{MRCA} of those country-specific B_{CAR} clades further suggests that B_{CAR} lineages started to be disseminated from the Caribbean into Latin America between the early 1970s and the early 1980s. This time scale coincides with the global dissemination of the $B_{PANDEMIC}$ clade from the United States [4] and with the estimated origin of several $B_{PANDEMIC}$ lineages in Latin America [7,28]. Although the $B_{PANDEMIC}$ and the B_{CAR} clades probably arrived at the same time in Latin America, the $B_{PANDEMIC}$ strain was able to ignite much larger outbreaks and infected a much larger number of individuals than any B_{CAR} strain in most of the countries analyzed.

The different epidemic outcomes of the $B_{PANDEMIC}$ and B_{CAR} lineages in Latin America could be related to virological and/or sociological factors. Notably, the highest HIV prevalence rates (>1%) in Latin America and the Caribbean were detected among countries with a high proportion (≥50%) of B_{CAR} clades like Haiti, Bahamas, Guyana, Jamaica and Trinidad and Tobago [5], thus arguing against the hypothesis of a low epidemic potential of B_{CAR} lineages. Transmission route is clearly an important factor shaping the HIV dissemination dynamics and major differences in the epidemic outcome of distinct subtype B clades may have appeared as a consequence of differences in the underlying transmission networks. We suggest that in most Latin American countries the $B_{PANDEMIC}$ strain was introduced and initially disseminated within highly connected networks of MSM and injecting drug users, whereas the B_{CAR} clades were mainly disseminated through heterosexual networks with lower rates of partner exchanges, which may explain the more successful dissemination of the $B_{PANDEMIC}$ lineage.

The remarkably successful dissemination of B_{CAR} clades in some northern countries of South America including French Guyana, Suriname and Guyana, probably reflects the high mobility of people between these countries and the Caribbean islands [29]. This is facilitated not only by the geographical proximity of those South American countries to the Caribbean islands, but also by cultural, linguistic and socioeconomic ties. Suriname and Guyana are members of the Caribbean common market, an

organization of 15 Caribbean nations and dependencies that also includes Bahamas, Belize, Haiti, Jamaica, Trinidad and Tobago and several other Lesser Antilles islands. The Caribbean common market not only promotes economic integration, but also facilitates the free movement of individuals for tourism or labor among countries. It is noted that a significant proportion (10%) of immigrants residing in Trinidad and Tobago are from Guyana [29], which may explain the epidemiological link observed between nonpandemic $B_{CAR\text{-}TT}$ and $B_{CAR\text{-}GY}$ clades circulating in Trinidad and Tobago and Guyana, respectively.

The higher frequency of B_{CAR} clades in Colombia, Panama and Venezuela (4–9% of subtype B infections) when compared with other Latin American countries (<2% of subtype B infections) also probably reflects a more frequent population mobility as a consequence of greater geographical proximity and historical links. It is interesting to note that the first reported Panamanian AIDS case was a Haitian woman diagnosed in September 1984 [30], which supports a longstanding presence of viruses of Caribbean origin in Panama. This country is also an important commercial hub because of the presence of the Panama Canal that promotes transit of people and goods. Junqueira *et al*. previously noted that a boom in oil production in Venezuela attracted immigrants from several countries in the region between 1970 and 1980, including people from Trinidad and Tobago and the Dominican Republic, which may have promoted the introduction of B_{CAR} strains into Venezuela during that time. Furthermore, Colombia and Venezuela have been pointed out as the most important source countries in South America for tourists and labor migrants (including female sex workers) to many Caribbean islands (particularly in the Netherlands Antilles) [29].

In summary, this study demonstrates that several nonpandemic HIV-1 B_{CAR} strains have been disseminated from the Caribbean into Latin America since the early 1970s. The B_{CAR} strains reached nearly all countries from Latin America here analyzed and in some of them were spread locally, establishing secondary outbreaks. Despite the early and widespread dissemination of B_{CAR} strains in the continent, HIV-1 subtype B epidemics in most Latin American countries were mainly driven by the $B_{PANDEMIC}$ clade that accounts for most (>90%) of current HIV-1 subtype B infections in the region. The only exceptions were Suriname, French Guyana and probably Guyana, where both $B_{PANDEMIC}$ and B_{CAR} clades seem to circulate at roughly similar prevalence as observed in many Caribbean islands. Intraregional population mobility combined with chance founder events in populations with high rates of partner exchange were probably the major forces driving the actual distribution of the different subtype B strains in the Americas.

Acknowledgements

The authors wish to thank Dr Vera Bongertz for critical review of the manuscript.

Contributions: The study was conceived and designed by G.B. Data acquisition was performed by M.C. All authors contributed to the data analysis and final version of the review.

Source of funding: This work was supported by Public Health Service grants E-26/110.439/2014 from the FAPERJ and 472896/2012-1 from the CNPq. M.C. was funded by a fellowship from Instituto Oswaldo Cruz-FIOCRUZ.

Conflicts of interest

There are no conflicts of interest.

References

1. UNAIDS. Report on the global AIDS epidemic. http://www.unaids.org/en/media/unaids/contentassets/documents/epidemiology/2013/gr2013/UNAIDS_Global_Report_2013_en.pdf). 2013 [accessed 5 October 2014].
2. UNAIDS. Global HIV/AIDS response. Progress Report 2011. http://www.unaids.org/en/media/unaids/contentassets/documents/unaidspublication/2011/20111130_UA_Report_en.pdf). 2011 [accessed 5 October 2014].
3. Hemelaar J, Gouws E, Ghys PD, Osmanov S. **Global trends in molecular epidemiology of HIV-1 during 2000–2007.** *AIDS* 2011; **25**:679–689.
4. Gilbert MT, Rambaut A, Wlasiuk G, Spira TJ, Pitchenik AE, Worobey M. **The emergence of HIV/AIDS in the Americas and beyond.** *Proc Natl Acad Sci U S A* 2007; **104**:18566–18570.
5. Cabello M, Mendoza Y, Bello G. **Spatiotemporal dynamics of dissemination of non-pandemic HIV-1 subtype B clades in the Caribbean region.** *PLoS One* 2014; **9**:e106045.
6. Junqueira DM, de Medeiros RM, Matte MC, Araujo LA, Chies JA, Ashton-Prolla P, *et al*. **Reviewing the history of HIV-1: spread of subtype B in the Americas.** *PLoS One* 2011; **6**:e27489.
7. Mendoza Y, Martinez AA, Castillo Mewa J, Gonzalez C, Garcia-Morales C, Avila-Rios S, *et al*. **Human immunodeficiency virus type 1 (HIV-1) subtype B epidemic in Panama is mainly driven by dissemination of country-specific clades.** *PLoS One* 2014; **9**:e95360.
8. de Oliveira T, Deforche K, Cassol S, Salminen M, Paraskevis D, Seebregts C, *et al*. **An automated genotyping system for analysis of HIV-1 and other microbial sequences.** *Bioinformatics* 2005; **21**:3797–3800.
9. Thompson JD, Gibson TJ, Plewniak F, Jeanmougin F, Higgins DG. **The CLUSTAL_X windows interface: flexible strategies for multiple sequence alignment aided by quality analysis tools.** *Nucleic Acids Res* 1997; **25**:4876–4882.
10. Posada D. **jModelTest: phylogenetic model averaging.** *Mol Biol Evol* 2008; **25**:1253–1256.
11. Guindon S, Dufayard JF, Lefort V, Anisimova M, Hordijk W, Gascuel O. **New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0.** *Syst Biol* 2010; **59**:307–321.
12. Guindon S, Lethiec F, Duroux P, Gascuel O. **PHYML online: a web server for fast maximum likelihood-based phylogenetic inference.** *Nucleic Acids Res* 2005; **33**:W557–W559.
13. Anisimova M, Gascuel O. **Approximate likelihood-ratio test for branches: a fast, accurate, and powerful alternative.** *Syst Biol* 2006; **55**:539–552.
14. Rambaut A. FigTree v1.4: Tree Figure Drawing Tool. http://tree.bio.ed.ac.uk/software/figtree/2009 [accessed 5 October 2014].

15. Drummond AJ, Nicholls GK, Rodrigo AG, Solomon W. **Estimating mutation parameters, population history and genealogy simultaneously from temporally spaced sequence data.** *Genetics* 2002; **161**:1307–1320.
16. Drummond AJ, Rambaut A. **BEAST: Bayesian evolutionary analysis by sampling trees.** *BMC Evol Biol* 2007; **7**:214.
17. Suchard MA, Rambaut A. **Many-core algorithms for statistical phylogenetics.** *Bioinformatics* 2009; **25**:1370–1376.
18. Drummond AJ, Ho SY, Phillips MJ, Rambaut A. **Relaxed phylogenetics and dating with confidence.** *PLoS Biol* 2006; **4**:e88.
19. Drummond AJ, Rambaut A, Shapiro B, Pybus OG. **Bayesian coalescent inference of past population dynamics from molecular sequences.** *Mol Biol Evol* 2005; **22**:1185–1192.
20. Hue S, Pillay D, Clewley JP, Pybus OG. **Genetic analysis reveals the complex structure of HIV-1 transmission within defined risk groups.** *Proc Natl Acad Sci U S A* 2005; **102**:4425–4429.
21. Zehender G, Ebranati E, Lai A, Santoro MM, Alteri C, Giuliani M, *et al*. **Population dynamics of HIV-1 subtype B in a cohort of men-having-sex-with-men in Rome, Italy.** *J Acquir Immune Defic Syndr* 2010; **55**:156–160.
22. Chen JH, Wong KH, Chan KC, To SW, Chen Z, Yam WC. **Phylodynamics of HIV-1 subtype B among the men-having-sex-with-men (MSM) population in Hong Kong.** *PLoS One* 2011; **6**:e25286.
23. Lemey P, Rambaut A, Drummond AJ, Suchard MA. **Bayesian phylogeography finds its roots.** *PLoS Comput Biol* 2009; **5**:e1000520.
24. Ferreira MAR, Suchard MA. **Bayesian analysis of elapsed times in continuous-time Markov chains.** *Can J Stat* 2008; **26**:355–368.
25. Rambaut A, Drummond A. Tracer v1.6. http://tree.bio.ed.ac.uk/software/tracer/2007 [accessed 5 October 2014].
26. Bielejec F, Rambaut A, Suchard MA, Lemey P. **SPREAD: spatial phylogenetic reconstruction of evolutionary dynamics.** *Bioinformatics* 2011; **27**:2910–2912.
27. UNAIDS. AIDS Info Database. http://www.unaids.org/en/dataanalysis/datatools/aidsinfo/2013 [accessed 5 October 2014].
28. Murillo W, Veras N, Prosperi M, de Rivera IL, Paz-Bailey G, Morales-Miranda S, *et al*. **A single early introduction of HIV-1 subtype B into Central America accounts for most current cases.** *J Virol* 2013; **87**:7463–7470.
29. Borland R, Faas L, Marshall D, McLean R, Schroen M, Smit M, *et al*. *HIV/AIDS and mobile populations in the Caribbean: a baseline assessment*. International Organization for Migration; 2004 Available from: http://www.iom.int/jahia/webdav/site/myjahiasite/shared/shared/mainsite/published_docs/books/hiv_mobile_carribbean.pdf. [accessed 5 October 2014].
30. de Ycaza MM, Rios V, Miranda E, Narvaez E, Sanchez G. **[Acquired immune deficiency syndrome. First confirmed case in Panama].** *Rev Med Panama* 1985; **10**:66–73.

Apêndice 05

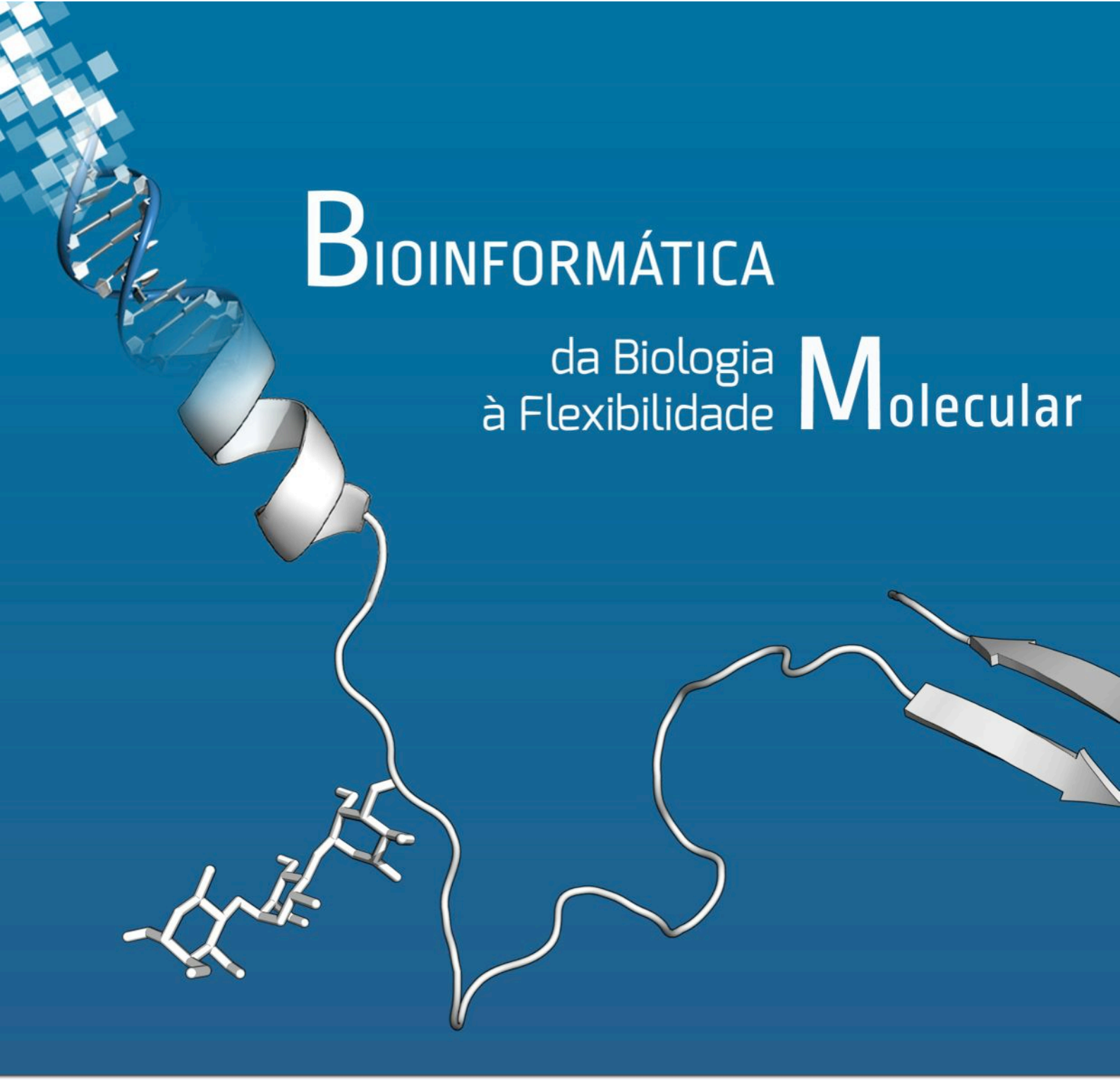
"Alinhamentos"

Junqueira DM, Braun RL, Verli H

Bioinformática: da Biologia à Flexibilidade Molecular, 2014

Capítulo 03

E-book disponível em: http://www.ufrgs.br/bioinfo/ebook/



Bioinformática

da Biologia à Flexibilidade Molecular

Hugo Verli (org)

Apoio:

CNPq CAPES FAPERGS UFRGS Universidade Federal do Rio Grande do Sul

Conteúdos

Autores

Bruno César Feltes
Centro de Biotecnologia, UFRGS

Camila S. de Magalhães
Pólo de Xerém, UFRJ

Charley Christian Staats
Centro de Biotecnologia, UFRGS

Dennis Maletich Junqueira
Depto Genética, UFRGS

Diego Bonatto
Centro de Biotecnologia, UFRGS

Edwin A. Yates
Instituto de Biologia Integrativa, Universidade de Liverpool

Fabio Lima Custódio
Laboratório Nacional de Computação Científica

Fernanda Rabaioli da Silva
Centro de Biotecnologia, UFRGS

Fernando V. Maluf
Centro de Inovação em Biodiversidade e Fármacos, IFSC - USP

Glaucius Oliva
Centro de Inovação em Biodiversidade e Fármacos, IFSC - USP

Gregório K. Rocha
Laboratório Nacional de Computação Científica

Guilherme Loss de Morais
Laboratório Nacional de Computação Científica

Helena B. Nader
Departamento de Bioquímica, Unifesp

Hugo Verli
Centro de Biotecnologia, UFRGS

Isabella A. Guedes
Laboratório Nacional de Computação Científica

Ivarne L. S. Tersariol
Departamento de Bioquímica, Unifesp

João Renato C. Muniz
Grupo de Biotecnologia Molecular, IFSC - USP

Joice de Faria Poloni
Centro de Biotecnologia, UFRGS

Laurent E. Dardenne
Laboratório Nacional de Computação Científica

Luís Maurício T. R. Lima
Faculdade de Farmácia, UFRJ

Marcelo A. Lima
Departamento de Bioquímica, Unifesp

Marcius da Silva Almeida
Instituto de Bioquímica Médica, UFRJ

Priscila V. S. Z. Capriles
PPG Modelagem Computacional, UFJF

Raphael Trevizani
Laboratório Nacional de Computação Científica

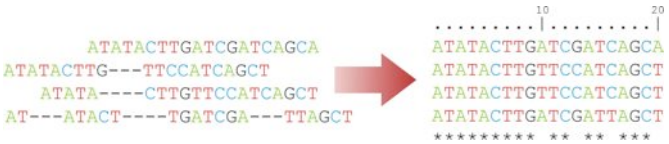
Rafael V. C. Guido
Centro de Inovação em Biodiversidade e Fármacos, IFSC - USP

Rodrigo Ligabue Braun
Centro de Biotecnologia, UFRGS

Rogério Margis
Centro de Biotecnologia, UFRGS

Yraima Cordeiro
Faculdade de Farmácia, UFRJ

3. Alinhamentos



Alinhamento de múltiplas sequências.

Dennis Maletich Junqueira
Rodrigo Ligabue Braun
Hugo Verli

3.1. Introdução

O avanço nas técnicas de sequenciamento do DNA tem permitido um crescente aumento no número de genomas disponíveis em bancos de dados públicos. Esta maior disponibilidade exigiu um grande aumento na capacidade computacional de armazenamento e no investimento em desenvolvimento de técnicas de processamento adequadas para a análise destes dados. Algoritmos de análise tiveram de ser criados e aperfeiçoados e, dentre estes, as técnicas de alinhamento de sequências tornaram-se ferramentas essenciais e primordiais na análise de sequências biológicas. Atualmente, diversos programas *online*, ou mesmo de instalação local, são capazes de alinhar centenas de sequências em poucos minutos.

Devido à extensão de suas aplicações, o alinhamento de sequências biológicas é um processo de fundamental importância para a bioinformática. Conceitualmente, os alinhamentos são técnicas de comparação entre duas ou mais sequências biológicas, que buscam séries de caracteres individuais que se encontram na mesma ordem nas sequências analisadas.

Em geral, as moléculas consideradas por estes programas, sejam elas formadas por nucleotídeos (DNA ou RNA) ou aminoácidos (peptídeos e proteínas), são polímeros representados por uma série de caracteres, e a comparação entre as moléculas depende apenas da comparação entre as respectivas letras. Apesar da facilidade e da aparente simplicidade do processo, a análise de similaridade das sequências é uma tarefa complexa e uma etapa decisiva para grande parte dos métodos de bioinformática que fazem uso de sequências biológicas.

Durante o alinhamento, as sequências são organizadas em linhas e os caracteres biológicos integram as colunas do alinhamento (Figura 1-3). Seguido à organização inicial, algoritmos específicos buscarão a melhor correspondência para as sequências em questão, permitindo a criação de espaços entre estes caracteres para que, ao final, todas as sequências tenham o mesmo comprimento. Isto possibilita uma fácil visualização da similaridade, permitindo que caracteres

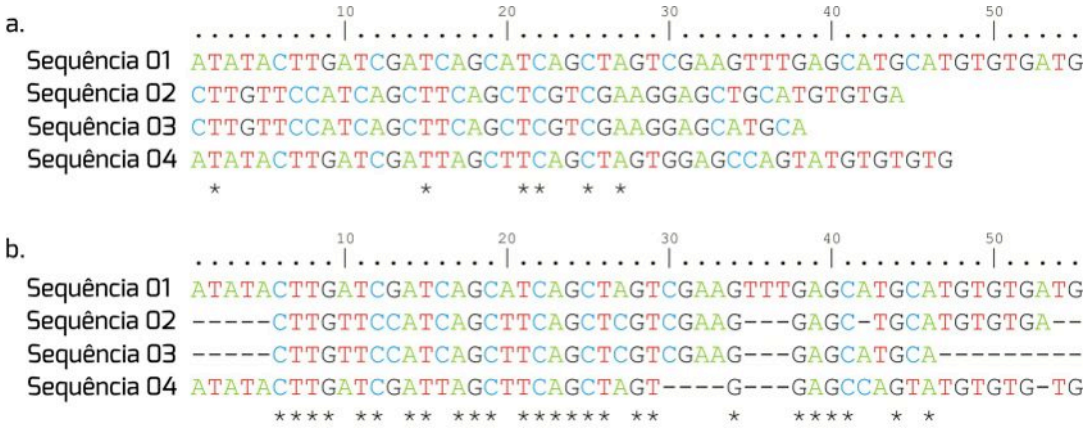


Figura 1-3: Alinhamento de quatro sequências de nucleotídeos envolvendo 55 caracteres. *a*) Grupo de sequências não alinhadas, cada sequência ocupando uma linha individual. *b*) Grupo de sequências alinhadas, onde caracteres idênticos são dispostos em uma mesma coluna e estas são identificadas por asteriscos (dispostos na parte inferior do alinhamento). Nucleotídeos ausentes em determinadas sequências são substituídos por hifens para identificar eventos de inserção/deleção.

idênticos ou similares em cada uma das sequências integrem a mesma coluna. A ideia central destes algoritmos é minimizar as diferenças entre as sequências, buscando um alinhamento ótimo. Comumente, a similaridade entre as sequências envolvidas é expressa pelo termo identidade, que quantifica a porcentagem de caracteres idênticos entre duas sequências.

A relevância e abrangência do uso do método tornam os procedimentos de alinhamento o cerne para diferentes campos dentro da grande área da bioinformática. Além de fundamentais em pesquisas de filogenética e análise evolutiva, os alinhamentos são exigidos em estudos de inferência estrutural e funcional de proteínas, análises de similaridade e identificação de sequências e em estudos aplicados ao campo da genômica.

Através dos métodos de alinhamento, é possível obter informações a respeito da relação evolutiva entre organismos, indivíduos, genes ou entre sequências diversas (Figura 2a-3). Se duas sequências distintas podem ser alinhadas com certo grau de similaridade, é possível inicialmente assumir que elas compartilharam, em algum momento do tempo passado, um ancestral comum e, por isso, são evolutivamente relacionadas. A partir da separação destas sequências de seu ancestral comum, individualmente cada uma delas acumulou diferentes variações ao longo do processo evolutivo. O termo homologia é utilizado frequentemente para definir estes eventos onde, através da relação de ancestralidade, dois indivíduos distintos possuem regiões em seu DNA (incluindo regiões codificantes) herdadas de um ancestral comum. Neste caso, a similaridade deve-se à descendência comum e, portanto, as sequências envolvidas na análise são ditas homólogas.

Cabe ressaltar que a homologia não requer necessariamente alta identidade de caracteres entre as sequências, uma vez que a maior ou menor identidade entre elas dependerá da taxa de evolução do organismo ou da espécie (consultar capítulo 5). Ainda, a similaridade entre sequências pode ser gerada não somente por descendência, mas por pressão seletiva de um determinado ambiente. Nestes casos, teremos regiões similares na sequência de nucleotídeos (ou aminoácidos) que surgiram de maneira independente, sem qualquer relação de descendência, e evoluíram por convergência, não sendo portanto homólogas. Assim, não é possível quantificar a homologia entre as sequências envolvidas, somente dizer se há ou não. Quando identificamos quantos caracteres se repetem nas mesmas posições entre duas ou mais sequências estamos, de fato, verificando a identidade entre estas, e não a homologia.

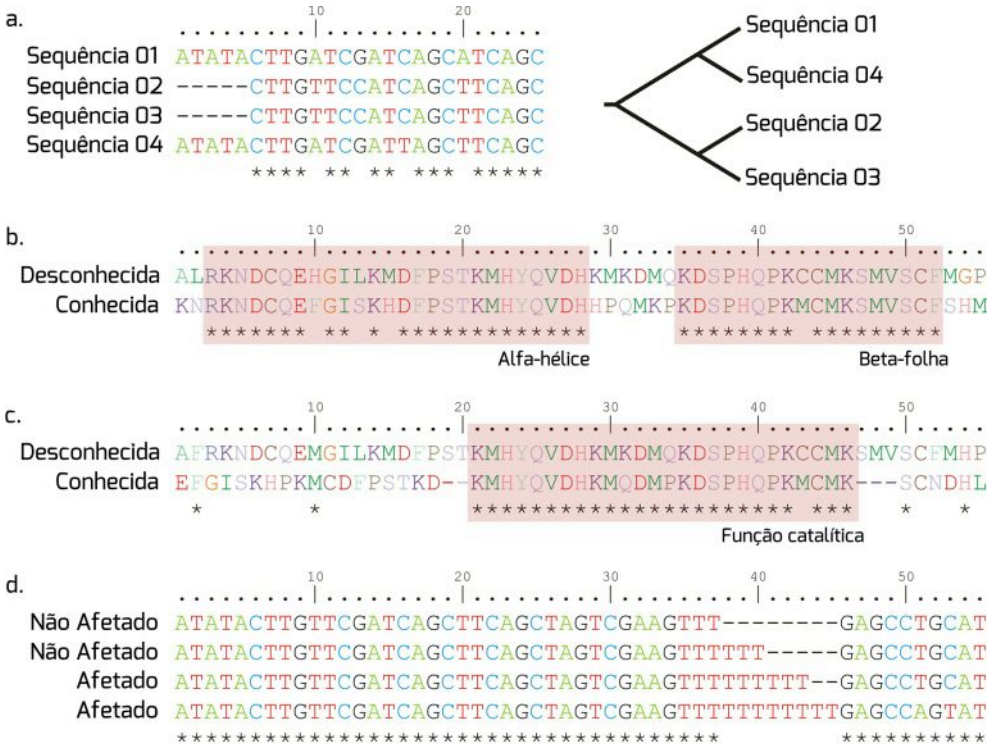


Figura 2-3: Aplicações dos métodos de alinhamento de sequências biológicas. *a*) Inferência filogenética a partir do alinhamento de quatro sequências de nucleotídeos. *b*) Inferência da estrutura de uma proteína alvo (Desconhecida) a partir do alinhamento com uma sequência de aminoácidos cuja estrutura tridimensional é conhecida (Conhecida). *c*) Inferência da função de um domínio proteico a partir da comparação de sequências de aminoácidos. *d*) Comparação de sequências de uma porção de determinado gene de indivíduos afetados e não afetados por uma doença genética. Os asteriscos identificam colunas com total similaridade dos caracteres.

As técnicas de alinhamento vêm se mostrando fundamentais na construção de algoritmos que visam comparar a informação de diversas sequências biológicas. À exemplo do programa BLAST, estes algoritmos permitem comparar uma sequência alvo com milhares de dados disponíveis em grandes bancos de armazenamento, fornecendo um valor de significância estatística associada a esta comparação de similaridade. Devido à facilidade de acesso e rapidez no processamento de dados, estes programas vêm cada vez mais ampliando as possibilidades e opções para o tipo de comparação ou pesquisa a ser realizada.

Os métodos de alinhamento podem ainda ser necessários para fornecer informações a respeito da função e da estrutura de sequências biológicas, particularmente nos alinhamentos de ribonucleotídeos e aminoácidos (Figura 2-3). Nestes casos, a similaridade entre duas ou mais sequências (dada em porcentagem) revela padrões referentes à composição química e podem fornecer embasamento para a definição de um arranjo tridimensional semelhante, principalmente no caso de proteínas (Figura 2b-3). A mesma relação é feita para inferir a função de domínios de uma proteína recém-descoberta, ainda sem função definida. Sabendo que sua forma está diretamente relacionada à sua função, através da comparação com outras proteínas com estrutura e função já estabelecidas, é possível inferir a função realizada por determinado domínio da proteína sob investigação (Figura 2c-3). Nestes casos, as sequências envolvidas no alinhamento não são necessariamente homólogas. Através do fenômeno da evolução convergente, diferentes regiões codificantes do DNA podem gerar produtos proteicos com funções similares, sem obrigatoriamente compartilharem um ancestral comum.

Finalmente, as técnicas de alinhamento

têm grande importância para a análise de genes e genomas. Com o aumento da disponibilidade de sequências nucleotídicas de genomas completos, e mesmo com o surgimento de modernas técnicas de biologia molecular, como o *microarray* e *deep sequencing*, os métodos de comparação permitiram o entendimento a respeito da variabilidade genética de indivíduos e populações.

A comparação entre genomas de diferentes espécies, ou até mesmo de indivíduos da mesma espécie, possibilita a análise de variações (mutações ou polimorfismos) nas sequências e, em alguns casos, permite a identificação de relações entre variações no DNA e susceptibilidade a determinadas doenças, beneficiando o campo da genética e áreas relacionadas. Adicionalmente, como um recurso para a caracterização de eventos evolutivos, os alinhamentos permitem análises comparativas entre genomas. A abrangência e importância evolutiva dos eventos de quebra e reparo de DNA, ou mesmo dos eventos de recombinação, inversões e translocações, tem sido desvendados, primariamente, através dos métodos de alinhamento.

Além do alinhamento de sequências, o alinhamento de estruturas constitui outra importante ferramenta em estudos de bioinformática. A metodologia é bastante diferente daquela empregada em alinhamentos de sequências, pois passamos de um problema unidimensional para um problema tridimensional. Sua utilização passou a ser difundida a partir de 1978, com o trabalho de Rossmann e Argos, comparando os sítios ativos de enzimas cujas estruturas eram conhecidas até aquele momento. Os métodos de sobreposição simples de estruturas estão disponíveis há mais tempo, tendo sido propostos a partir da década de 1970, enquanto os métodos de comparação e alinhamento se desenvolveram posteriormente, principalmente a partir da década de 1990.

A comparação de estruturas se refere à análise de similaridades e diferenças entre duas ou mais estruturas, enquanto o alinhamento de estruturas se refere à determinação de quais aminoácidos seriam equivalentes entre tais estruturas. É importante destacar também a diferença entre alinhamento e sobreposição de estruturas. Apesar desses termos ainda serem empregados na literatura como sinônimos, eles se referem a procedimentos diferentes. Conforme mencionado acima, enquanto o alinhamento de estruturas busca identificar equivalências entre pares de aminoácidos nas estruturas a serem sobrepostas, a sobreposição necessita desse conhecimento prévio sobre as equivalências.

Sendo assim, a sobreposição estrutural busca solucionar um problema muito mais simples, ou seja, minimizar a distância entre dois resíduos já reconhecidos como equivalentes. Isso se dá por encontrar transformações que satisfazem o menor desvio médio quadrático (RMSD) ou as equivalências máximas dentro de um valor limite para o RMSD.

Considerando que a estrutura das proteínas é mais conservada que a sequência, o alinhamento de estruturas confere maior especificidade ao alinhamento de sequências quando comparado ao alinhamento de sequências independente de estrutura. A maioria dos métodos de sobreposição de estruturas é adequado para identificar similaridades entre estruturas proteicas. O alinhamento de duas ou mais estruturas, porém, constitui uma tarefa mais difícil, e sua precisão depende tanto do método usado quanto do objetivo do usuário.

3.2. Alinhando sequências

À primeira vista, o processo de alinhamento entre diferentes sequências parece simples e não sujeito a qualquer tipo de erro. No entanto, esta afirmativa só é verdadeira em casos onde os organismos envolvidos possuem uma baixa taxa evolutiva (Figura 3a-3). Quando consideramos sequências homólogas amostradas de organismos com alta taxa evolutiva, ou até mesmo sequências similares, porém não homólogas, nos deparamos com casos particulares que tornam o processo de alinhamento complexo e, muitas vezes, sujeito a uma interpretação especialmente subjetiva por parte do usuário (Figura 3b-3).

A comparação de sequências homólogas de organismos evolutivamente distantes é um desafio para os programas de alinhamento. As diferentes pressões seletivas moldam os genomas de maneira imprevisível e, muitas vezes, acarretam a perda ou ganho de nucleotídeos ao longo do processo evolutivo. Para estes casos, a adição de lacunas (*gaps*) em matrizes de alinhamento, representadas por "-", é possível e muitas vezes necessária. As lacunas representam um ou mais eventos de inserção ou deleção de nucleotídeos. Estes eventos, comumente chamados de "indels" (*in* para inserção, e *del* para deleção), são fruto de processos mutagênicos (espontâneos ou induzidos) e, dependendo da região atingida, podem ser expressos nas moléculas de RNA e nas proteínas, onde poderão gerar consequências moleculares. Erros de replicação gerados pela DNA-polimerase durante a replicação do DNA, ou mesmo os eventos de recombinação, são os principais fatores atrelados à geração destes *indels* nos genomas. Em regiões codificadoras, estes eventos podem acarretar mudanças no quadro de leitura da proteína e torná-la não funcional.

Em termos analíticos, a inserção de lacunas dificulta o processo de alinhamento e exige interpretações cautelosas. Para determinados casos, especialmente em análises evolutivas e filogeográficas, é comum que regiões do alinhamento com determinado nível de incerteza, especialmente regiões com grande número de lacunas, sejam eliminadas

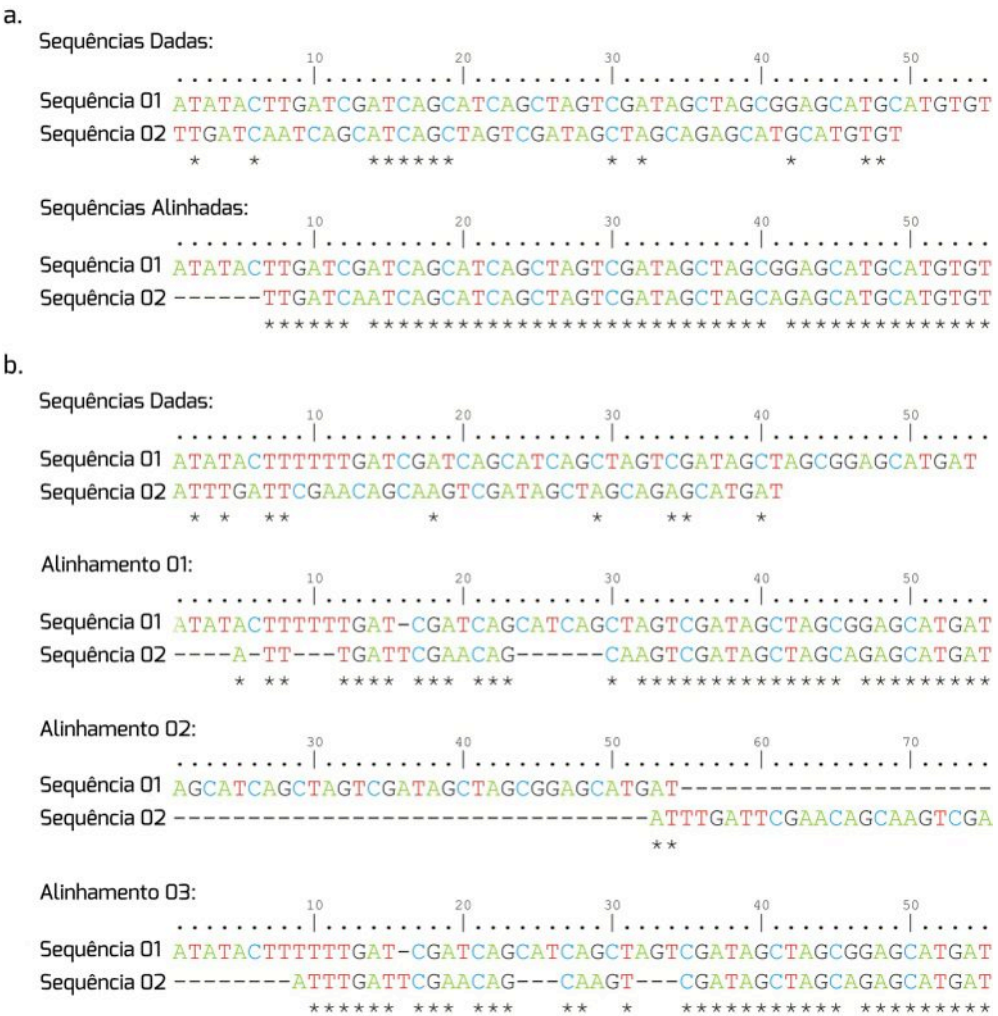


Figura 3-3: Alinhamentos de nucleotídeos. *a*) Duas sequências homólogas originadas de organismos com baixa taxa de evolução são dadas e seu alinhamento é proposto. *b*) Duas sequências homólogas amostradas de organismos com alta taxa de evolução são dadas e diferentes alinhamentos são propostos. Os hifens representam eventos de inserção ou deleção únicos na sequência. Os asteriscos identificam colunas com total similaridade dos caracteres.

da análise. Contudo, até o momento não existem programas capazes de lidar com as lacunas de forma coerentemente biológica. Apesar de sabermos que se tratam de eventos evolutivos comuns e bem caracterizados, as incertezas sobre o número de eventos e sua intensidade tornam as lacunas, em grande parte dos casos, um fator de confusão para análises de alinhamento.

Conforme mostrado na Figura 3-3, diferentes alinhamentos são possíveis para um mesmo grupo de sequências. A pergunta que se segue é: como reconhecer o melhor resultado quando nos deparamos com diversos alinhamentos possíveis para um mesmo conjunto de dados? Buscou-se resolver este problema através da criação de um sistema de pontuação para comparar os resultados de diferentes alinhamentos. Caracteres idênticos em sequências diferentes representam igualdades ou correspondências (*matches*) e, por serem resultados preferenciais durante o processo de alinhamento, são pontuados positivamente. Pelo contrário, caracteres não idênticos que ocupam a mesma coluna são chamados de desigualdades, ou *mismatches*, e recebem atribuições negativas. Como resultado, o melhor alinhamento possível para duas sequências é aquele que maximiza a pontuação total, somando os valores de *matches* e debitando os valores de *mismatches*.

Do ponto de vista biológico, as mudanças entre as bases nitrogenadas nas sequências de nucleotídeos não ocorrem com a mesma probabilidade (Figura 4a-3). Sendo assim, podemos atribuir valores de *mismatches* diferentes às transições (trocas de purinas por purinas ou pirimidinas por pirimidinas) e às transversões (trocas de purinas por pirimidinas ou pirimidinas por purinas). Para sequências de aminoácidos, é necessário escolher ativamente uma matriz de pontuação específica. Essas matrizes são resultados diretos de estudos de variação proteica e estão diretamente relacionadas à probabilidade de substituição de um aminoácido por outro (matrizes BLOSUM e PAM). Atualmente, as matrizes BLOSUM são as mais disseminadas e aplicadas para os mais diversos casos de comparação entre sequências de aminoácidos (Figura 4b-3).

a.

	A	C	G	T
A	1	-2	-2	-2
C		1	-2	-2
G			1	-2
T				1

b.

	A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V
A	5	-2	-2	-2	0	0	0	0	-2	-2	-3	-2	-1	-2	0	0	0	-2	-3	0
R		5	-2	-3	-3	0	-1	-2	0	-3	-4	1	-3	-3	-2	-2	0	0	-3	-4
N			5	0	0	0	-2	0	0	-4	-5	-2	-3	-3	-2	0	0	-2	-2	-5
D				5	-4	0	1	-1	0	-5	-6	-3	-4	-4	0	-2	-2	-2	-2	-5
C					8	-2	-3	-1	-1	0	-2	-3	0	-1	-1	1	0	0	-2	0
Q						5	2	0	0	-2	-4	0	-2	-3	0	0	0	0	-2	-3
E							5	0	0	-3	-4	0	-3	-3	0	0	0	-2	-3	-3
G								6	0	-4	-5	-2	-3	-2	-2	0	0	0	-2	-3
H									6	-3	-4	0	-2	0	0	0	0	0	2	-2
I										4	0	-3	2	0	-2	-3	0	0	-3	2
L											4	-4	0	0	-3	-4	-3	0	-4	0
K												4	-2	-4	-1	-2	0	0	-3	-4
M													6	0	-3	-3	-2	0	-3	2
F														6	-3	-2	-2	2	2	0
P															7	0	0	-2	-3	0
S																4	2	-2	-2	-3
T																	5	-1	-3	0
W																		9	2	-1
Y																			7	-3
V																				4

Figura 4-3: Matrizes de custo utilizadas no cálculo de pontuação dos alinhamentos. *a*) Matriz de custo exemplo utilizada para cálculos de pontuação em alinhamentos de nucleotídeos. *b*) Matriz de custo BLOSUM62 utilizada para cálculo da pontuação em alinhamentos de aminoácidos.

Ainda, é necessário que as lacunas de alinhamentos recebam determinadas pontuações, pois são frequentemente encontradas em alinhamentos de dados biológicos. Se lacunas podem ser adicionadas em qualquer posição sem qualquer restrição, tanto nas extremidades quanto no interior das sequências, é possível gerar alinhamentos com mais lacunas do que propriamente caracteres a serem comparados (Figura 3b-3, alinhamento 2). Com o intuito de prevenir inserção excessiva, a adição de lacunas é penalizada durante a atribuição da pontuação de uma sequência, conforme um conjunto de parâmetros, chamado de penalidades por lacuna (*gap penalties*, PL). A abrangência da lacuna é pontuada pelo respectivo número de *indels* presentes no alinhamento. A fórmula mais comum para cálculo destas penalizações segue abaixo:

$$PL = g + e\,(L - 1)$$

onde L é o tamanho da lacuna (número de *indel*s presentes na lacuna), g é a penalidade pela abertura da lacuna (necessária para evitar que os alinhamentos contenham lacunas desnecessárias) e e é a penalidade atribuída a

cada *indel* (novamente para evitar grandes lacunas sem necessidade). Os valores de penalidade por lacuna são desenhados para reduzir a pontuação de um alinhamento quando este possui uma quantidade de *indels* desnecessária. Apesar da disseminação deste conceito, não há qualquer relação matemática ou biológica sustentando este cálculo. É importante destacar que, através da propriedade de "alinhamento livre de colunas em branco" (ou seja, *gaps* não são alinhados), as penalizações ainda impedem o alinhamento de *indels* entre as sequências envolvidas na análise. Assim, o melhor alinhamento entre as sequências será dado por um valor que resulta da soma dos valores associados a cada um dos *matches*, *mismatches* e lacunas, de acordo com um critério pré-definido (Figura 5-3).

O método de pontuação foi a solução encontrada para avaliar e classificar diferentes alinhamentos em busca da melhor explicação para a relação evolutiva entre as sequências. O próximo problema encontrado foi enumerar todas as possibilidades de alinhamentos para um grupo de dados. Assumindo-se duas sequências com tamanho de 100 caracteres cada, poderíamos enumerar até 10^{77} possíveis alinhamentos, diferentes entre si. A extensão de possibilidades inviabiliza a enumeração de todos os casos devido ao tempo e ao requerimento de enorme processamento destes dados. Apesar da exigência computacional, alguns algoritmos são capazes de realizar tal tarefa e ainda aplicar o método de pontuação para cada um dos casos, em busca do melhor resultado. No entanto, estes algoritmos não são capazes de lidar com sequências que contenham mais que algumas dezenas de caracteres. Em virtude da capacidade de explorar todas as soluções do problema, o processo realizado por estes algoritmos é chamado de "alinhamento ótimo".

Contudo, em virtude da inerente demora do processo, foi necessário desenvolver algoritmos que acelerassem a busca de um alinhamento capaz de explicar de maneira ótima os processos evolutivos para um determinado grupo de sequências sem, no entanto, enumerar todas as possibilidades. Os alinhamentos gerados por estes programas são chamados heurísticos, e compreendem métodos aproximados de busca pelo resultado ótimo. Diferentes métodos foram criados para diferentes tipos de alinhamento (Figura 6-3). Entre estes, devido à eficiência e à rapidez de processamento das informações de um alinhamento, incluindo o cálculo de pontuação, os algoritmos de programação dinâmica são, atualmente, os mais utilizados para este fim, tanto em alinhamentos simples como integrado aos algoritmos de alinhamentos múltiplos.

É fundamental assumirmos, para a maior parte dos problemas em bioinformática, o alinhamento como um modelo de relação evolutiva entre as sequências envolvidas. E como modelo, está sujeito à presença de certos problemas na explicação dos eventos evolutivos reais. Portanto, os alinhamentos devem ser avaliados com extrema cautela. A facilidade e a aparente simplicidade na análise dos programas tornam o processo mecânico e desvinculado de análises críticas pela maior parte dos usuários. A associação dos métodos de alinhamento a outras análises de bioinformática tende a desvincular a real importância desta técnica e a coloca apenas como um procedimento, e não formalmente como uma técnica sujeita à análise crítica. Isto pode ocasionar na obtenção de modelos incorretos ou mesmo de falsos positivos.

3.3. Tipos de alinhamento

Em estudos de bioinformática, é comum compararmos moléculas de dois ou mais indivíduos, sejam eles da mesma espécie ou de espécies diferentes. Quanto maior o número de sequências comparadas, maior o tempo exigido para conclusão do alinhamento e, dependendo das sequências envolvidas, maior a dificuldade dos algoritmos em encontrar o melhor resultado. Conforme a quantidade de sequências envolvidas, podemos dividir os alinhamentos em dois tipos: alinhamentos simples, ou par-a-par, e alinhamentos múltiplos, ou de múltiplas sequências (Figura 7-3).

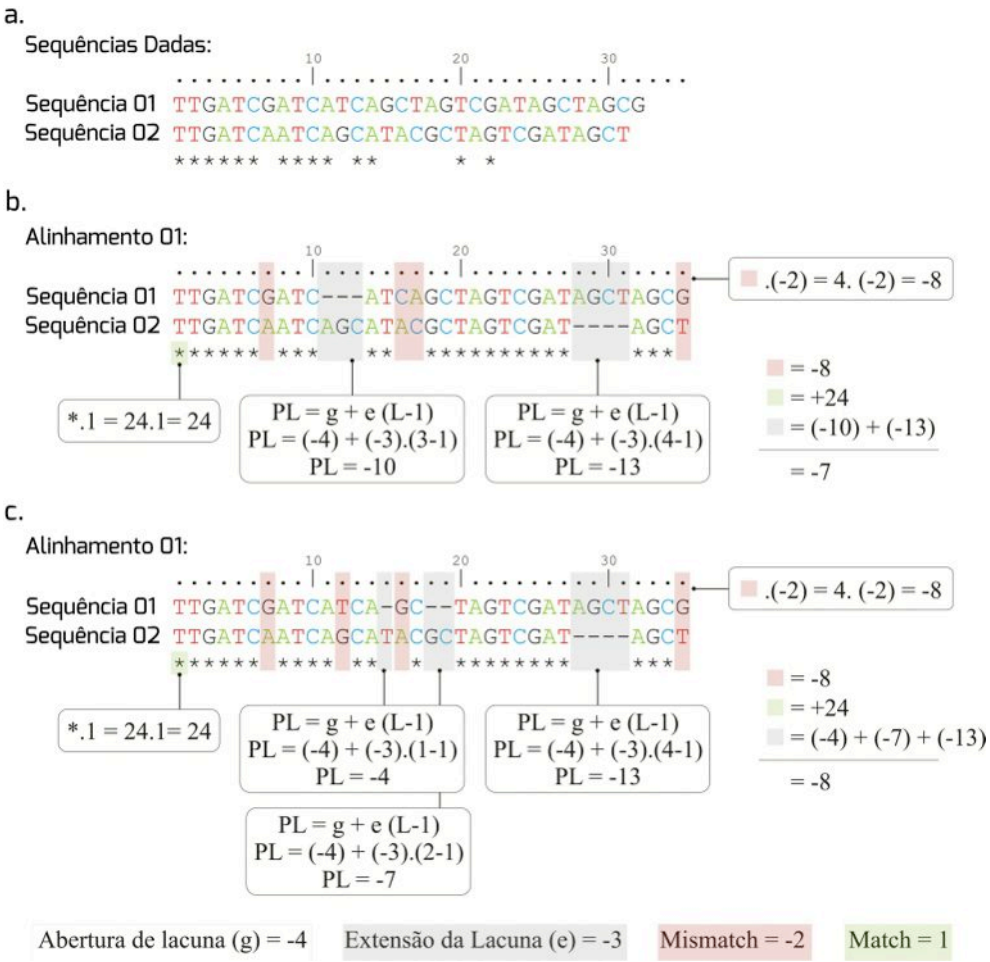


Figura 5-3: Esquema de pontuação para avaliação de alinhamentos. *a*) Duas sequências de desoxirribonucleotídeos não alinhadas. *b*) Proposição de um alinhamento para as sequências dadas em *a*. O alinhamento possui 24 colunas de *matches*, 4 colunas de *mismatches* e duas lacunas com 3 e 4 *indels*. A pontuação total para o alinhamento desta sequência é -7. *c*) Proposição de um segundo alinhamento para as sequências dadas em *a*. O alinhamento possui 24 colunas de *matches*, 4 colunas de *mismatches* e três lacunas com 1, 2 e 4 *indels*. A pontuação total para o alinhamento desta sequência é -8. A partir deste exemplo, o alinhamento com a maior pontuação é o mostrado em *b*. Os valores de pontuação utilizados neste exemplo são especificados na parte inferior da figura.

Os alinhamentos simples descrevem especificamente a relação de similaridade entre duas sequências quaisquer. Já os alinhamentos múltiplos incluem três ou mais sequências na análise de similaridade e, dependendo do objetivo do usuário, podem envolver até centenas de sequências.

Conceitualmente, ainda podemos dividir os alinhamentos, tanto simples, como múltiplos, em dois grandes tipos. Os alinhamentos que levam em consideração toda a extensão das sequências são conhecidos como globais, enquanto aqueles que buscam pequenas regiões de similaridade são chamados de locais (Figura 7-3). Em algoritmos que buscam o alinhamento global de duas sequências, reforça-se a busca do alinhamento completo das sequências envolvidas, procurando incluir o maior número de *matches* do início ao final das sequências. Quando necessário, estes algoritmos permitem a inserção de lacunas para que as sequências tenham o mesmo tamanho no resultado do alinhamento (Figura 7b-3).

Graficamente, os sítios com caracteres idênticos são representados ligados por barras verticais, enquanto os sítios que possuem caracteres diferentes nas duas sequências, ou

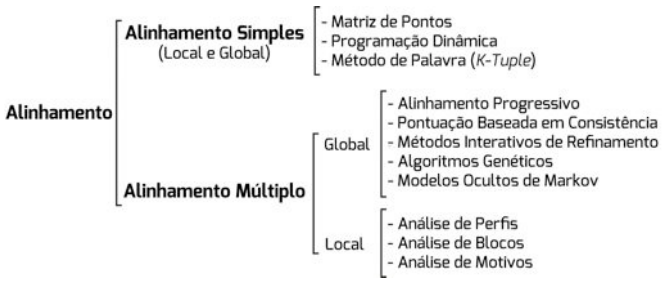


Figura 6-3: Tipos de alinhamento e os algoritmos aplicados à bioinformática.

mesmo a presença de uma lacuna em uma delas, permanecem sem qualquer notação (Figura 7-3). O principal algoritmo envolvido no processamento de alinhamentos globais é aquele desenvolvido por Needleman e Wunsch durante a década de 1970. Além de ter uma notável importância metodológica, este algoritmo tem grande importância na história do alinhamento, pois foi o primeiro algoritmo a aplicar o método de programação dinâmica para a comparação de sequências biológicas.

Em seu início, os métodos de alinhamento eram utilizados especialmente para a comparação par-a-par de sequências de proteínas inteiras. No entanto, com a ampliação da disponibilidade de sequências completas de proteínas, foi necessário buscar métodos de alinhamento que privilegiassem a busca de similaridade, não entre sequências completas, mas apenas entre porções isoladas destas sequências. Durante a década de 1980 iniciou-se o desenvolvimento de novos algoritmos de alinhamento, já que os desenvolvidos até aquele momento não eram aplicáveis para esta particularidade. Entre estes novos algoritmos, o desenvolvido por Smith e Waterman, em 1981, ganhou maior destaque e atualmente é o principal algoritmo utilizado por programas para realização de alinhamentos locais. Nestes casos, privilegia-se o alinhamento de partes da sequência, buscando apenas as regiões com a maior similaridade (Figura 7c-3). Em algoritmos para busca local, o alinhamento pára no final das regiões de alta similaridade e substitui as regiões excluídas por hifens (lacunas) no resultado final (Figura 7c-3).

3.4. Alinhamento simples

Para entender como se processa um alinhamento par-a-par e como o grau de si-

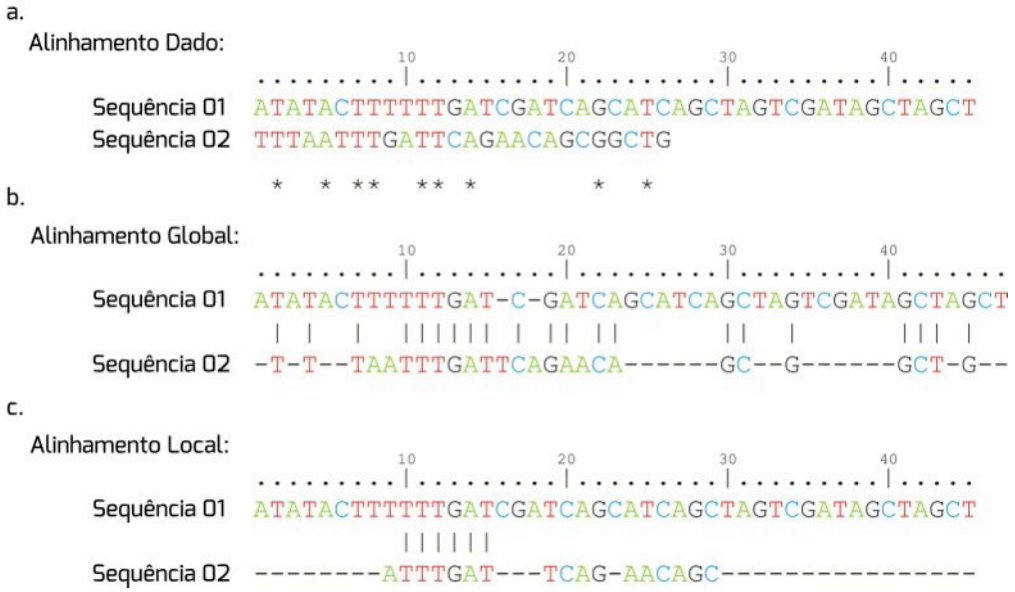


Figura 7-3: Diferenças entre alinhamento local e global. *a*) Duas sequências de nucleotídeos de tamanhos diversos são amostradas e alinhadas por algoritmos diferentes. *b*) No alinhamento local, a prioridade é encontrar as regiões altamente similares, independentemente do tamanho desta região. Neste caso, porções da sequência que não foram alinhadas com alta similaridade foram excluídas do resultado final. *c*) No alinhamento global, as duas sequências são alinhadas por completo, independentemente do número de lacunas que tenham que ser inseridas.

milaridade entre elas pode ser computado, apresentamos três dos principais algoritmos desenvolvidos para este fim: algoritmos de programação dinâmica, análise de matriz de pontos (*dot matrix*) e método de palavra ou *k-tuple*.

A programação dinâmica é, atualmente, o método mais utilizado por programas para realizar o alinhamento de sequências. Em casos simples (par-a-par), é capaz de encontrar o melhor alinhamento para duas sequências através da aplicação da pontuação de similaridades. É, portanto, um método de execução relativamente rápida nos computadores modernos, requerendo um tempo e memória de processamento proporcional ao produto do tamanho das duas sequências envolvidas.

O método é baseado no princípio de otimização de Bellmann, e propõe a solução de problemas complexos através da resolução dos seus diversos subproblemas. Os subproblemas são resolvidos e seus resultados são armazenados pelo algoritmo. A vantagem funcional da resolução em partes é que, geralmente, problemas complexos combinam uma série de subproblemas. Como o algoritmo acumula os resultados dos diferentes subproblemas, acelera a resolução do problema complexo. Assim, a designação "programação" nada tem a ver com programação de computadores, mas com a organização dos resultados já solucionados para resolução de um problema maior.

Conforme discutimos anteriormente, em determinados casos, duas sequências podem apresentar diferentes alinhamentos. Se não há *indels* e as sequências são similares, o alinhamento é rápido e não deixa dúvidas. No entanto, quando existe certa diversidade entre as sequências envolvidas e uma quantidade de suficiente de *indels*, a solução para o alinhamento é menos óbvia visualmente. Nestes casos, os algoritmos de programação dinâmica buscarão solucionar os subproblemas envolvidos e fornecerão o melhor resultado.

Para cálculo do melhor alinhamento entre duas sequências, o algoritmo de programação dinâmica necessita da especificação de um esquema de pontuação, seja ele referente a nucleotídeos ou aminoácidos. Da mesma forma, é necessário fornecer um valor de penalidade para a abertura e extensão das lacunas. A partir destas informações, o algoritmo calculará uma relação entre todos os caracteres das sequências e fornecerá o melhor alinhamento como resultado final.

Como exemplo, consideraremos a Figura 8-3. São dadas duas sequências, sequência 1 e sequência 2, um esquema de pontuação e, para facilitar o entendimento do cálculo, um valor único de penalidade por lacuna de -8. O algoritmo toma as sequências e transforma a relação entre elas em uma tabela, onde as linhas são definidas pelos caracteres da sequência 01, e as colunas pelos caracteres da sequência 02. A fim de permitir lacunas no início do alinhamento, o algoritmo impõe a inserção de uma coluna e de uma linha iniciais contendo o símbolo de *indel*. A partir deste ponto, para cada um dos elementos da matriz, o algoritmo calculará a melhor pontuação dos subcaminhos associados ao alinhamento: uma substituição, uma inserção na sequência 01 ou uma inserção na sequência 2. Assim, o melhor subcaminho será calculado segundo uma função de pontuação, conforme abaixo:

$$F(i,j) = max \begin{cases} valor\ da\ célula\ na\ diagonal\ superior\ esquerda + pontuação\ da\ similaridade \\ valor\ da\ célula\ acima + valor\ da\ penalidade\ por\ lacuna \\ valor\ da\ célula\ à\ esquerda + valor\ da\ penalidade\ por\ lacuna \end{cases}$$

A partir do elemento (1,1) da matriz e ao longo da primeira linha, apenas a terceira condição é satisfeita (valor da célula à esquerda + valor da penalidade por lacuna). Na primeira coluna, apenas a segunda condição é satisfeita. Para outros elementos, as três condições devem ser calculadas e aquela que resultar no maior valor é escolhida para formar a matriz. Além disso, os procedimentos dos algoritmos de programação dinâmica podem ser representados por pequenas setas para indicar qual subcaminho obteve o melhor valor (Figura 8-3).

Outro método importante na área de alinhamento de sequências é a análise de matriz de pontos ou matriz *dot*. É um método simples e bastante eficiente em análises de

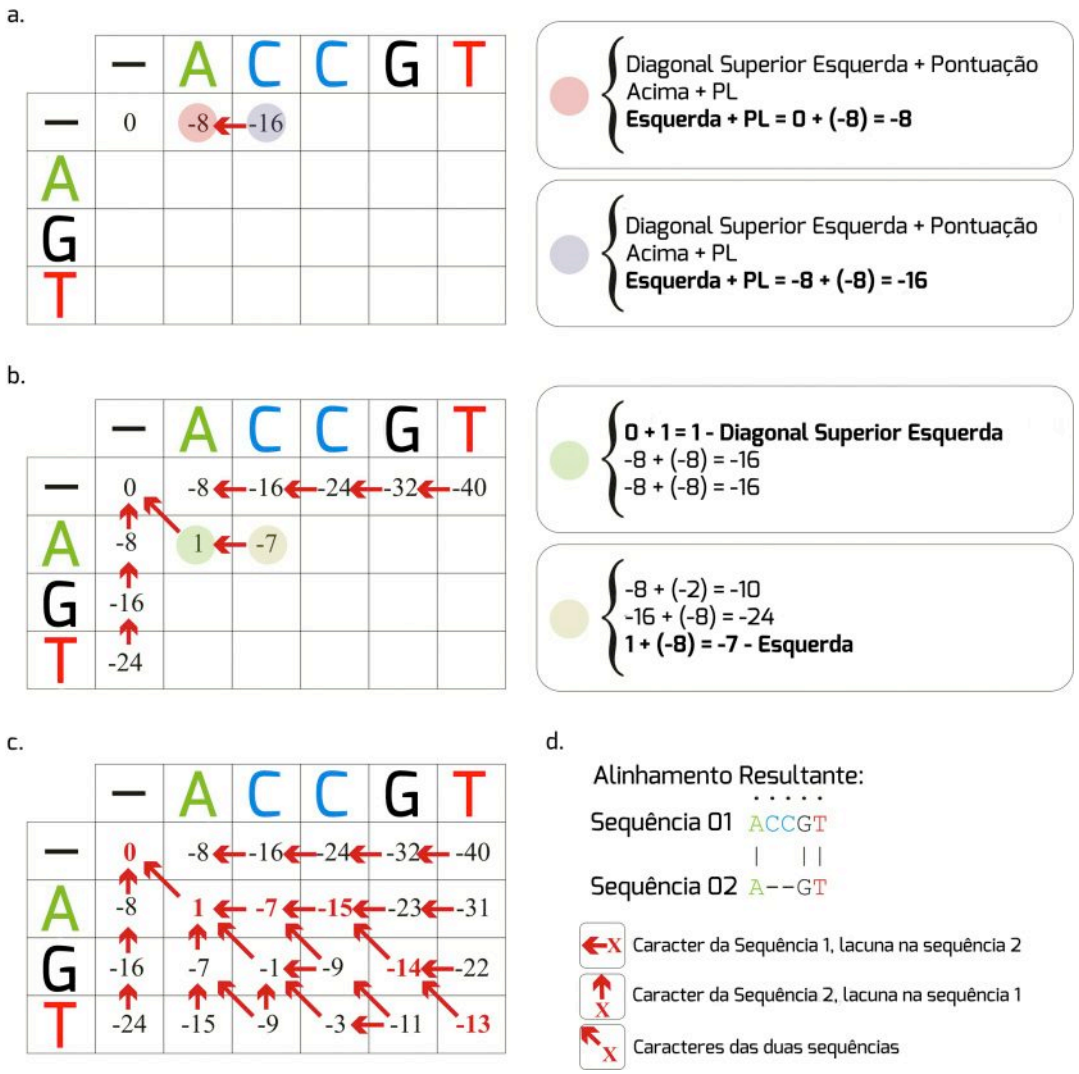


Figura 8-3: Alinhamento de duas sequências de nucleotídeos através do método de programação dinâmica. *a*) As sequências a serem alinhadas são dispostas em uma tabela onde o número de colunas corresponde ao número de caracteres da sequência 1 mais um (devido à adição de uma coluna para uma lacuna) e o número de linhas corresponde ao número de caracteres da sequência 2 mais um. O caractere atribuído à primeira linha e à primeira coluna é, por definição, o símbolo "-", atribuído a uma lacuna. Através da matriz de penalidades calculam-se os valores para as três possibilidades $F(i,j)$, buscando a equação que resulte no maior valor. O valor arbitrário de penalidade por lacuna (PL) é de -8. Em virtude de a primeira linha não possuir valores de comparação na diagonal superior esquerda e acima, considera-se apenas a terceira equação. *b*) O valor demarcado em verde é o primeiro a ser calculado após o preenchimento da primeira linha e primeira coluna, representando o menor valor encontrado no cálculo para $F(i,j)$. Além do cálculo, o algoritmo de programação dinâmica insere informações a respeito da direção da informação. Como o valor "1" foi o maior valor encontrado e representa o cálculo utilizando a informação situada na diagonal superior esquerda, demarcada em verde, insere-se uma seta nesta direção. *c*) O preenchimento completo da tabela e as respectivas setas ilustrando a direção da informação. Algumas casas estão demarcadas com duas setas, pois apresentaram dois valores máximos idênticos na resolução das equações. Ao final dos cálculos, iniciando pelo canto inferior direito, seguem-se as setas em busca dos maiores valores. *d*) Relacionando os dados da tabela com a simbologia apresentada, chega-se ao alinhamento final entre as sequências 1 e 2.

deleções/inserções e para detectar repetições diretas ou inversas, especialmente em sequências de nucleotídeos. Além disso, vem sendo utilizado para buscar regiões de pareamentos intra-cadeia capazes de formar estruturas 2árias em moléculas de RNA. Este método permite a visualização gráfica das regiões de similaridade entre sequências através da construção de uma matriz de identidade. O número de linhas desta matriz é definido pelo número de caracteres de uma das sequências, e o número de colunas é definido pelo número de caracteres da outra sequência a ser comparada (Figura 9-3). É primariamente um método visual, e não fornece o alinhamento propriamente dito como resultado final, embora seja frequentemente utilizado quando se deseja visualizar as regiões de similaridade entre duas sequências.

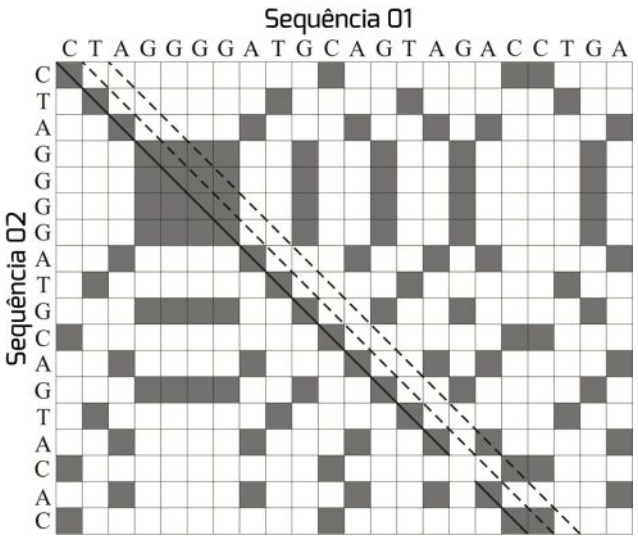


Figura 9-3: Análise de matriz de pontos de duas sequências de DNA. Os pontos assinalados em cinza representam a concordância de caracteres entre a sequência 1 e a sequência 2. A partir da diagonal direita inferior, são traçadas diferentes retas. Aquela que atingir o maior número de pontos assinalados deve ser escolhida como resultado para o alinhamento entre as duas sequências. A linha contínua representa a possibilidade mais adequada a esta análise e as linhas tracejadas representam possibilidades de insucesso.

Neste método, inicialmente, uma das sequências é disposta na vertical e a outra na horizontal (Figura 9-3). Regiões do gráfico que possuam o mesmo caractere tanto na sequência disposta na horizontal, quanto na sequência disposta na vertical, serão assinalados. Esta marcação representa os possíveis correspondências (*matches*) entre uma sequência e outra.

Qualquer região de similaridade entre as duas sequências será evidenciada por uma linha diagonal de assinalações. Pontos não dispostos na diagonal representam correspondências aleatórias que não estão relacionadas com a similaridade entre as sequências. A detecção de regiões de alta similaridade pode ser beneficiada, em alguns casos, através da comparação de dois ou mais caracteres ao mesmo tempo. Nestes casos, é necessário escolher um número de caracteres como janela.

Além disso, arbitrariamente, um número de correspondências deve ser escolhido. Por exemplo, para comparar duas sequências com 100.000 caracteres, podemos escolher uma janela de 15 caracteres e 10 correspondências requeridas. O algoritmo varrerá a matriz de 15 em 15 caracteres e, quando, entre estes quinze caracteres, existirem 10 formando correspondências entre as duas sequências, o algoritmo inserirá uma marcação de similaridade. Geralmente, esta variação do método é utilizada para a comparação de longas sequências de DNA.

Por último, outro algoritmo bastante comum no alinhamento par-a-par de dados biológicos é o *k-tuple*, ou método de palavras. Este método é geralmente mais rápido que o método de programação dinâmica, embora não garanta o melhor alinhamento como resultado. Este tipo de algoritmo é especialmente útil em casos onde se busca similaridade de uma única sequência contra um grande conjunto de dados. Para isso, o algoritmo dividirá uma sequência alvo em pequenas sequências, geralmente conjuntos de dois a seis caracteres, chamados de palavras. Da mesma forma, o conjunto total de sequências do banco de dados terá cada uma das sequências subdivida em pequenas pala-

vras. As palavras da sequência alvo serão comparadas às palavras oriundas do banco de dados. Após a busca de identidade, o algoritmo alinhará as duas sequências completas (sequência oriunda do banco de dados que teve uma palavra similar com umas das palavras da sequência alvo e a própria sequência alvo) a partir das palavras similares e estenderá a análise de similaridade para as regiões vizinhas, antes e depois da palavra similar. Através de uma matriz de penalidade, o algoritmo calculará o alinhamento que teve o maior valor de pontuação. É comum, para esta segunda etapa dos cálculos de similaridade, a utilização de algoritmos de programação dinâmica.

3.5. Alinhamento múltiplo global

Da mesma forma que no caso dos alinhamentos simples, o método de programação dinâmica é usualmente utilizado para lidar com múltiplas sequências. Nestes casos, utiliza-se o conceito de soma ponderada dos pares (*weighted sum of pairs*, WSP). Através deste conceito, para qualquer alinhamento múltiplo de sequências, uma pontuação para cada par possível formado por estas sequências será calculada (Figura 8-3) e, ao final, os valores de similaridade para cada um dos pares serão somados. Apesar de conceitualmente simples, este método exige grande capacidade computacional e, dependendo da quantidade de sequências envolvidas, pode requerer longo tempo para processamento.

Métodos alternativos tiveram que ser criados para acelerar os cálculos para alinhamento de sequências, incluindo-se: alinhamento progressivo, pontuação baseada em consistência (*consistency-based scoring*), métodos iterativos de refinamento, algoritmos genéticos e modelos ocultos de Markov. Cabe ressaltar que todos estes métodos realizam buscas aproximadas pelo resultado ótimo e, portanto, se tratam de métodos heurísticos.

Alinhamento progressivo

Leva em consideração a relação evolutiva entre as sequências. Os algoritmos utilizam as relações filogenéticas para gerar o resultado de alinhamento. Inicialmente, são realizados alinhamentos par-a-par de todos os possíveis pares. Nesta comparação, verifica-se apenas o número de caracteres diferentes entre as duas sequências (verificar o conceito de distância evolutiva observada no capítulo 6). Estas distâncias serão utilizadas para a construção de uma filogenia (geralmente através do método de *neighbor-joining*). A partir desta filogenia o alinhamento será construído progressivamente, dependendo da relação entre as sequências sendo, por isso, chamado de alinhamento progressivo.

Tomemos como exemplo um ramo de uma dada filogenia que inclui duas sequências. O algoritmo construirá um alinhamento através de programação dinâmica para estas duas sequências. A partir deste primeiro alinhamento, estas duas sequências serão agora tratadas como uma, e serão alinhadas à próxima sequência filogeneticamente relacionada. Devemos notar que todo o restante das sequências será alinhado baseando-se neste primeiro par. É um método rápido e amplamente utilizado para alinhar um grande número de sequências. Atualmente, os programas mais populares de alinhamento progressivo são o CLUSTALW e CLUSTALX.

Pontuação baseada em consistência

Baseado no algoritmo de alinhamento progressivo, não leva em consideração apenas o primeiro par de sequências alinhadas. Durante a realização do cálculo, realiza outros alinhamentos par-a-par para aperfeiçoar as comparações entre as sequências. O principal programa a utilizar este algoritmo é o T-COFFEE.

Métodos iterativos de refinamento

Funcionam como os algoritmos de ali-

nhamento progressivo, mas os grupos de sequências são realinhados constantemente ao longo das análises, garantindo que o alinhamento inicial não defina o resultado final. O principal programa a utilizar este algoritmo como base para os cálculos de alinhamento é o MUSCLE.

Algoritmos genéticos

Estes algoritmos buscam simular o processo evolutivo no conjunto de sequências a serem alinhadas, aplicando conceito de seleção e recombinação. É ainda um método lento e, devido à aleatoriedade do processo, não garante o mesmo resultado para diferentes alinhamentos do mesmo conjunto de dados. O programa SAGA é um dos poucos a implementar algoritmos genéticos.

Modelos ocultos de Markov

Modelo baseado em probabilidades estatísticas, destacando os eventos de substituição e inserção ou deleção de caracteres.

3.6. Alinhamento múltiplo local

Na busca por regiões localizadas de similaridade entre diferentes sequências, são aplicados principalmente os seguintes algoritmos: análise de perfis, análise de blocos e análise de motivos.

Análise de perfis

A partir de um alinhamento primário de todas as sequências envolvidas na análise e utilizando uma matriz de custo padrão, o algoritmo seleciona as regiões altamente conservadas e produz uma nova matriz de pontuação (matriz de custo), chamada de perfil. A construção deste perfil pode ser realizada através de dois métodos diferentes (método das médias e método evolutivo) e inclui pontuações para *matches*, *mismatches* e lacunas. Assim que produzido, este perfil pode ser utilizado para alinhar sequências entre si utilizando as pontuações calculadas para avaliar a probabilidade em cada posição ou para buscar sequências com o mesmo padrão em um banco de dados.

A desvantagem do método de perfis está na especificidade da nova matriz de custo obtida. Se o alinhamento inicial contiver poucas sequências, pode não representar adequadamente a variabilidade de caracteres em uma determinada posição e prejudicar o algoritmo na busca por similaridade com outras sequências. Este método é principalmente utilizado para alinhamentos de aminoácidos.

Análise de blocos

Assim como a análise de perfis este método requer, inicialmente, a seleção da região de maior similaridade de um alinhamento múltiplo. Estas regiões podem ser chamadas de blocos e diferem dos perfis por não acomodarem *indels*, que serão automaticamente eliminados das análises. Este método é também capaz de realizar a busca de pequenas regiões de similaridade entre sequências, de maneira semelhante ao método de palavras.

Análise de motivos

Este método é especialmente utilizado na busca por motivos proteicos em sequências de aminoácidos. O método foi desenvolvido através do alinhamento de milhares de sequências de aminoácidos extraídas de grandes bancos de dados de proteínas. A partir deste alinhamento, analisou-se cada uma das colunas para buscar um padrão de substituição entre os aminoácidos. Estes padrões de mudança refletem uma maior probabilidade de substituição. Para proceder ao alinhamento, os algoritmos que aplicam a análise de motivos iniciam o processo por uma análise de blocos. As regiões de alta similaridade são então analisadas para buscar os padrões de substituição descritos inicialmente. O conjunto de padrões resultante da análise das colunas é chamado de motivo. A probabilidade de existência de cada motivo em uma sequência de proteína é estimada através do banco de dados do SwissProt.

3.7. BLAST

O BLAST, ou Ferramenta de Busca por Alinhamento Local Básico (*Basic Local Alignment Search Tool*) é um algoritmo capaz de realizar buscas baseadas em alinhamento que, apesar de não serem exatas, são confiáveis e muito rápidas, sendo estas suas vantagens em relação a outros métodos. Ele é um dos programas mais usados em Bioinformática devido à velocidade em que consegue responder a um problema fundamental em biologia celular e molecular: comparar uma sequência desconhecida com aquelas depositadas em bancos de dados.

O algoritmo do BLAST aumenta a velocidade do alinhamento de sequências ao buscar primeiro por palavras comuns (ou *k-tuples*) na sequência de busca e em cada sequência do banco de dados. Em vez de buscar todas as palavras de mesmo tamanho, o BLAST limita a busca àquelas palavras que são mais significantes. O tamanho de palavra é fixado em 3 caracteres para sequências de aminoácidos e em 11 para sequências de nucleotídeos (3 se as sequências forem traduzidas nos 6 quadros de leitura possíveis). Esses são os tamanhos mínimos para obter uma pontuação por palavras que seja alta o suficiente para ser significativa sem perder fragmentos menores, mas importantes, de sequência.

Funcionamento do algoritmo BLAST

Para funcionar, o BLAST necessita de uma sequência de busca (*query*) e de sequências alvo. Comumente, as sequências alvos são o conjunto de sequências depositadas em um banco de dados, local ou na *web*. Um dos conceitos principais empregados pelo BLAST é de que alinhamentos estatisticamente significantes contêm pares de segmentos de alta pontuação (HSP, *high-scoring segment pairs*), e são esses HSPs que o algoritmo busca entre a sequência sendo analisada e aquelas depositadas no banco de dados.

As principais etapas do funcionamento do algoritmo BLAST, para uma sequência proteica genérica incluem:

i. Remoção de repetições ou regiões de baixa complexidade na sequência de busca.

Uma região de baixa complexidade é definida como uma região composta por poucos tipos de elementos. Essas regiões normalmente apresentam pontuações altas que podem confundir o programa em sua busca por sequências com similaridade significativa. Por esse motivo, tais regiões são identificadas antes da próxima etapa e ignoradas.

ii. Estabelecer uma lista de palavras com k-letras.

Sendo este um caso envolvendo sequências proteicas, k = 3, ou seja, cada palavra tem tamanho 3. Como mostrado na Figura 10-3, são listadas palavras com comprimento de 3 caracteres, sequencialmente, até que a última letra da sequência de busca seja incluída.

Figura 10-3: Exemplo de lista de palavras geradas pelo BLAST.

iii. Listar as possíveis palavras correspondentes.

Diferente de outros algoritmos (como o FASTA), o BLAST considera apenas as palavras de maior pontuação. As pontuações são estabelecidas por comparação das palavras listadas na etapa *ii* com todas as outras palavras de 3 letras. Uma matriz de substituição (BLOSUM62) é usada para pontuar as comparações entre pares de resíduos. Existem 20^3 possíveis pontuações de correspondência considerando uma palavra de 3 letras. Como exemplo, a comparação das palavras PQG e PEG tem pontuação de 15, enquanto a comparação de PQG com PQA pontua como 12. A seguir, um limiar T para pontuação de palavras vizinhas é usado para reduzir o número de possíveis palavras correspondentes. As palavras cujas pontuações forem maiores que o limiar T serão mantidas na lista de possíveis correspondências, enquanto aquelas cujas pontuações

forem menores serão descartadas. Considerando o exemplo anterior, se T = 13, PEG será mantida, enquanto PQA será abandonada.

iv. Organizar as palavras de alta pontuação.

As palavras remanescentes, com alta pontuação, são organizadas em uma árvore de busca. Isso permite que o programa compare as palavras com as sequências do banco de dados de maneira rápida.

v. Repetir os passos *iii* e *iv* para cada palavra de k-letras originadas da sequência de busca.

vi. Varrer as sequências do banco de dados em busca de correspondências com as palavras remanescentes.

O BLAST realiza uma varredura das sequências depositadas no banco de dados, buscando pelas palavras de alta pontuação (como PEG, no exemplo anterior). Se uma correspondência exata for encontrada, ela será empregada para nuclear um possível alinhamento sem lacunas (*gaps*) entre a sequência de busca e a depositada no banco de dados.

vii. Estender as correspondências exatas entre pares de segmentos de alta pontuação.

A versão original do BLAST estende o alinhamento para a esquerda e para a direita de onde ocorre uma correspondência exata. A extensão é parada apenas quando a pontuação acumulada pelo HSP começa a diminuir (um exemplo pode ser visto na Figura 11-3).

Figura 11-3: Exemplo do esquema de pontuação empregado pelo BLAST.

Para acelerar o processo, a versão atual do BLAST (BLAST2 ou *Gapped* BLAST) emprega um limiar mais baixo para a vizinhança das palavras, mantendo a sensibilidade na detecção de similaridade de sequências. Assim, a lista de possíveis correspondências obtidas na etapa *iii* é maior. Como observado na Figura 12-3, as regiões de correspondência exata com distância menor que A na mesma diagonal serão unidas como uma nova região, mais extensa. Posteriormente, essas regiões são estendidas da mesma maneira como ocorre no BLAST original, com os HSPs sendo pontuados com base em uma matriz de substituição.

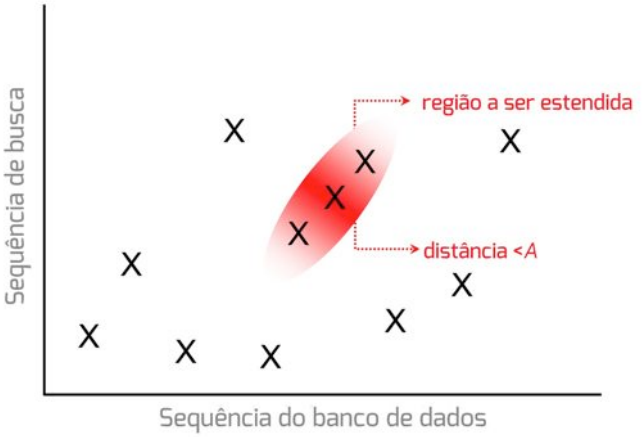


Figura 12-3: Esquema da extensão de zonas de correspondência entre sequências identificadas pelo BLAST.

viii. Listar todos os HSPs do banco de dados cuja pontuação seja alta o suficiente.

Nessa etapa são listados todos os pares de segmentos cuja pontuação seja maior que um determinado ponto de corte S. A distribuição de pontuações obtidas por alinhamento de sequências aleatórias é a base para determinação desse ponto de corte.

ix. Avaliar a significância da pontuação dos HSPs.

A avaliação estatística de cada par de segmentos de alta pontuação explora a Distribuição de Valores Extremos de Gumbel. O valor de confiança estatística e apresentado pelo BLAST, chamado de valor de expectativa, reflete o número de vezes que uma sequência não relacionada presente no banco de dados pode obter, ao acaso, um valor maior que S (ponto de corte). Ou seja, o e reflete o número de falsos positivos entre os resultados de similaridade encontrados. Para $p < 0{,}1$, o valor e se aproxima da distribuição de Poisson (ver item 4.8).

x. Transformar duas ou mais regiões de HSP em um alinhamento maior.

Em alguns casos, duas ou mais regiões de HSP podem ser combinadas em um trecho maior de alinhamento (uma evidência adicional da relação entre a

sequência de busca e a encontrada no banco de dados). Existem dois métodos para comparar a significância das novas regiões ligadas. Se, por exemplo, forem encontradas duas regiões de HSP combinadas com pares de pontuação (67 e 41) e (53 e 45), cada método se comportará de maneira diferente. O método de Poisson conferirá maior significância ao conjunto com valor mínimo maior (45 em vez de 41). O método de soma dos pontos, ao contrário, dará preferência ao primeiro conjunto, pois 108 (67+41) é maior que 98 (53+45). O BLAST original usa o primeiro método, enquanto o BLAST2 emprega o segundo.

xi. Exibir os alinhamentos locais entre a sequência de busca e cada uma das correspondências no banco de dados.

O BLAST original produz apenas alinhamentos sem lacunas (*gaps*), incluindo cada um dos HSPs encontrados inicialmente, mesmo que mais de uma região de correspondência seja encontrada numa mesma sequência do banco de dados. O BLAST2 produz um único alinhamento com lacunas, podendo incluir todas as regiões de HSP encontradas. É importante destacar que o cálculo da pontuação e do valor e leva em conta as penalidades por abertura de lacunas no alinhamento.

xii. Registrar as correspondências encontradas.

Quando o valor *e* dos alinhamentos encontrados entre a sequência de busca e as do banco de dados satisfazem o ponto de corte estabelecido pelo usuário, a correspondência é registrada. Os resultados da busca são apresentados de forma gráfica, seguidos por uma lista de correspondências organizada pela pontuação e pelo valor *e*, e finalizam com os alinhamentos. A Figura 13-3 traz um exemplo de resultado obtido pelo BLAST.

Diferentes tipos de BLAST

O BLAST constitui uma família de programas, que podem ser usados para diferentes fins, dependendo das necessidades do usuário. Esses programas variam quanto ao tipo de sequência de busca, o banco de dados a ser empregado, e o tipo de comparação a ser realizada. As diferentes aplicações disponíveis pelo BLAST incluem:

i. *blastn*: BLAST nucleotídeo-nucleotídeo. Usando uma sequência de DNA como entrada, dá como resultado as sequências de DNA mais similares presentes no banco de dados especificado pelo usuário.

ii. *blastp*: BLAST proteína-proteína. Usando uma sequência proteica como entrada, dá como resultado as sequências proteicas mais similares presentes no banco de dados especificado pelo usuário.

iii. *blastpgp*: BLAST iterativo com especificidade de posição (PSI-BLAST). Usado para encontrar proteínas distantemente relacionadas. Nesse caso, uma lista de proteínas proximamente relacionadas é criada. Essa lista serve de base para a criação de uma sequência média, que resume as características importantes do conjunto de sequências. A sequência média é usada para buscar sequências similares no banco de dados e um grupo maior de proteínas é encontrado. O grupo maior é usado na construção de uma nova sequência média e o processo é repetido. Ao incluir proteínas relacionadas na busca, o PSI-BLAST é muito mais sensível na percepção de relações evolutivas distantes que o BLAST proteína-proteína tradicional.

iv. *blastx*: tradução de nucleotídeos em 6 quadros-proteína. Compara os produtos de tradução conceitual nos 6 quadros de leitura de uma sequência de nucleotídeos contra o banco de dados de sequências proteicas.

v. *tblastx*: tradução de nucleotídeos em 6 quadros-tradução de nucleotídeos em 6 quadros. O mais lento dos programas BLAST, tem por objetivo encontrar relações distantes entre sequências de nucleotídeos. Ele traduz a sequência de nucleotídeo nos 6 possíveis quadros de leitura e compara os resultados contra a tradução nos 6 quadros de leitura das sequências de nucleotídeos depositadas no banco de dados.

vi. *tblastn*: proteína-tradução de nucleotídeos em 6 quadros. Compara uma sequência de proteína contra a tradução nos 6 quadros de leitura das sequências de nucleotídeos depositadas no banco

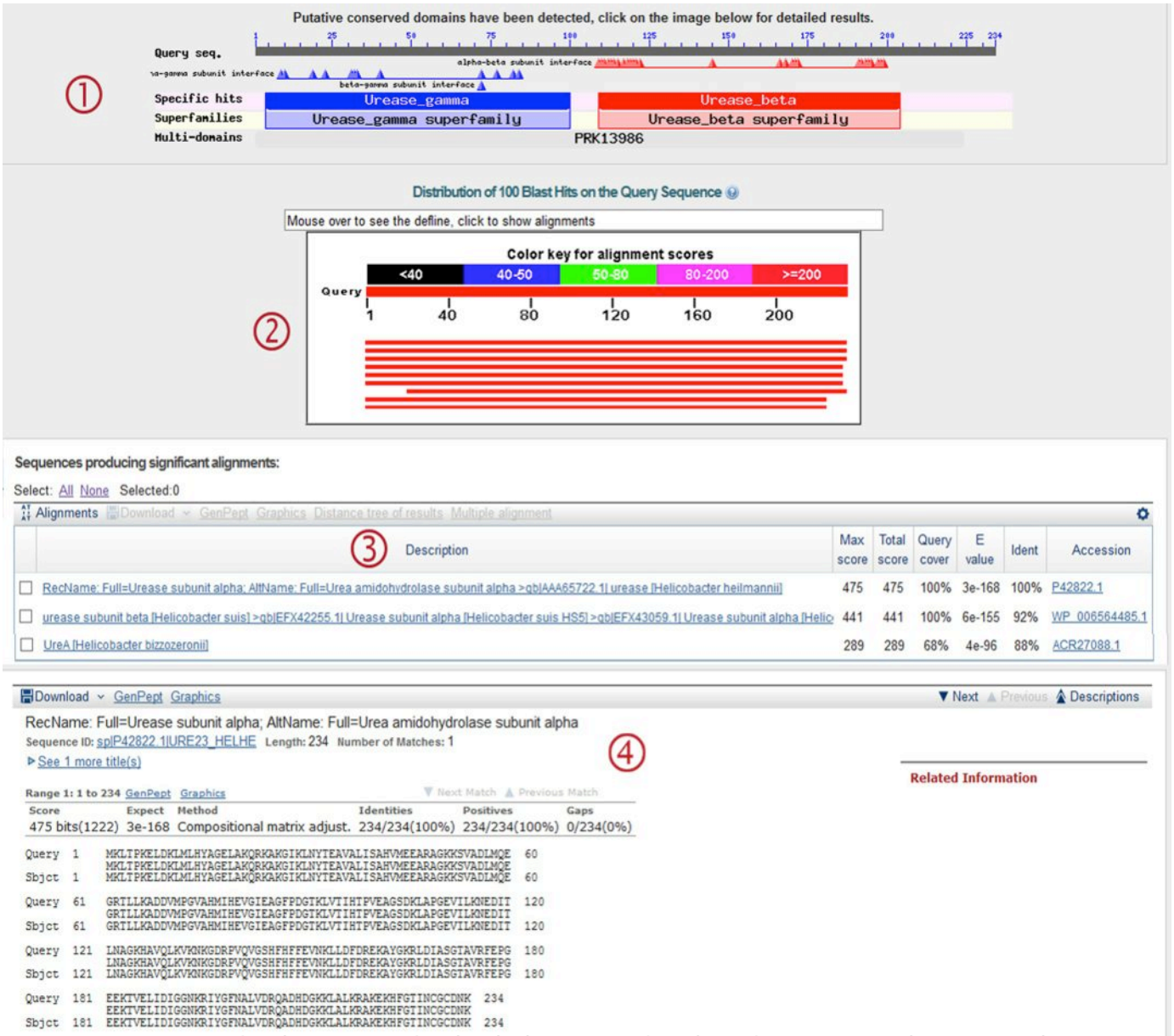


Figura 13-3: Exemplo de um resultado de busca realizada pelo BLAST. Diferentes informações são apresentadas: *1*) representação gráfica de domínios conservados identificados na sequência; *2*) representação gráfica de *matches*, indicando qualidade do alinhamento e cobertura das sequências identificadas; *3*) informações estatísticas dos resultados encontrados, incluindo identidade e valor *e*; *4*) alinhamento de cada sequência encontrada com a sequência de busca (*query*).

de dados.

vii. *megablast*: para empregar um grande número de sequências de busca. Quando se compara um grande número de sequências de busca (especialmente no BLAST por linha de comando), o megablast é muito mais rápido que o BLAST executado por várias vezes seguidas. Ele agrupa muitas sequências de busca, formando uma grande sequência, antes de realizar a busca no banco de dados. Os resultados são pós-analisados em busca de alinhamentos individuais.

3.8. Significância estatística

Em determinados casos, especialmente para buscar evidência de homologia entre sequências, o alinhamento é analisado sob o ponto de vista estatístico. Nessa óptica, podemos calcular quão bom pode ser um ali-

nhamento simplesmente levando em consideração as razões de chance de alinhamento entre nucleotídeos quaisquer. Para isso, sequências de nucleotídeos ou aminoácidos são geradas aleatoriamente, alinhadas em conjunto e avaliadas, segundo um determinado esquema de pontuação. Para alinhamentos globais, pouco se sabe a respeito destas distribuições randômicas. No entanto, felizmente, estas técnicas são bem entendidas para casos de alinhamentos locais e, atualmente, são amplamente utilizadas para a avaliação de similaridade, especialmente em bancos de dados que comportam grande quantidade de sequências.

Para analisar a probabilidade associada a determinado alinhamento é necessário, inicialmente, gerar um modelo aleatório das sequências em análise. Esses novos alinhamentos serão pontuados seguindo um determinado esquema de pontuação. Neste contexto, será calculada a probabilidade de se obter aleatoriamente uma pontuação pelo menos igual à pontuação do alinhamento original. O valor associado aos múltiplos testes realizados é chamado de valor e (*e-value*). Para banco de dados, este valor corresponde ao número de distintos alinhamentos, com uma pontuação igual ou melhor, que são esperados ocorrer na busca por sequências similares simplesmente por razões de chance (aleatórios). Estes cálculos estatísticos levam em consideração a pontuação do alinhamento e o tamanho do banco de dados. Quanto menor o valor e, menor o número de chances de uma determinada sequência ser alinhada aleatoriamente com outras e, portanto, mais significante é o resultado. Por exemplo, um valor e de 1e-3 ($1x10^{-3}$ ou 0,001) significa que há a chance de 0,001 de que a sequência alvo seja alinhada com uma sequência aleatória do banco de dados. Por exemplo, em um banco de dados que contém 10.000 sequências, neste caso, esperaríamos encontrar até 10 outras sequências que alinharão significativamente com a sequência alvo. É importante ressaltar que o fato de encontrarmos um valor e próximo de zero na comparação entre duas sequências não necessariamente denota a homologia destas sequências, dado que sequências não relacionadas podem conter similaridades devido à evolução convergente.

3.9. Alinhamento de 2 estruturas

O alinhamento de estruturas é um problema matematicamente complexo que só pode ser resolvido por algoritmos heurísticos. A Figura 14-3 apresenta um exemplo de alinhamento estrutural simples. Diferentes algoritmos oferecem resultados diferentes para o alinhamento, e algumas vezes essas diferenças são grandes. Por esse motivo é importante testar diferentes programas de alinhamento estrutural. Cada um deles tem pontos fortes e fracos, que podem ser explorados a partir da leitura dos artigos que os propuseram originalmente.

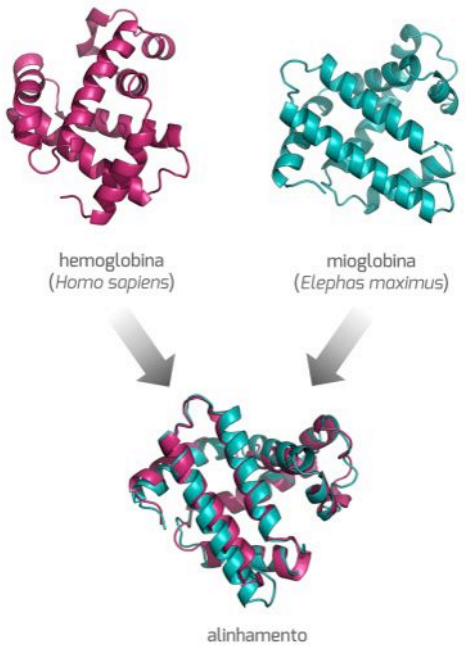


Figura 14-3: Exemplo de alinhamento de duas estruturas proteicas, oriundas de diferentes organismos: hemoglobina humana e mioglobina de elefante-asiático.

Existem três etapas essenciais para as diferentes estratégias de alinhamento estrutural: a representação, a otimização e a pontuação. A representação se refere às maneiras de representar as estruturas de uma forma que não seja dependente de coordenadas espaciais e que seja adequada ao ali-

nhamento. A otimização lida com a amostragem do espaço de possíveis soluções para o alinhamento entre as estruturas. A pontuação lida com a classificação dos resultados obtidos e com sua significância estatística. A seguir apresentamos as características específicas de alguns dos métodos mais utilizados para o alinhamento de duas estruturas.

DALI: emprega matrizes de distâncias para representar as estruturas, transformando as estruturas 3D em conjuntos 2D de distâncias entre Cα. Se imaginarmos a sobreposição das matrizes, as regiões de sobreposição na diagonal representam similaridades na estrutura 2ária (similaridades no esqueleto polipeptídico), e similaridades fora da diagonal representam similaridades na estrutura 3ária. As matrizes são então divididas em matrizes menores, de tamanho fixo, com base nas similaridades encontradas. Cada submatriz é unida a outras que sejam adjacentes para obter a matriz de sobreposição com maior abrangência. A significância estatística do alinhamento é calculada com base na distribuição encontrada em uma comparação de centenas de estruturas de baixa identidade. A pontuação é apresentada como número de desvios-padrão em relação a tal distribuição.

SSAP: cria vetores ligando resíduos a partir dos Cβ, representando a estrutura em duas dimensões, considerando posição e direção. Um algoritmo de programação dinâmica identifica similaridades entre as matrizes de vetores, gerando uma nova matriz que é posteriormente recalculada considerando as diferenças entre cada posição de similaridade encontrada na primeira etapa em relação às outras posições de similaridade, até que uma matriz ótima seja atingida. A pontuação do SSAP não é estatística, mas foi calibrada em relação ao banco de dados CATH. Assim, uma pontuação maior que 70 indica similaridade entre as estruturas comparadas.

VAST: cria vetores a partir de elementos de estrutura 2ária cujo tipo, direção e conexão estão relacionados com a topologia da proteína. Esses elementos (fragmentos) de estrutura 2ária são alinhados e comparados com alinhamentos gerados aleatoriamente. Alinhamentos com boa pontuação são agrupados e depois realinhados usando um procedimento de otimização por Monte Carlo. A significância estatística é dada pelo valor p (assim como ocorre no BLAST). O valor p é proporcional à probabilidade de se obter o alinhamento ao acaso.

SARF2: transforma as coordenadas em um conjunto de elementos de estrutura 2ária. Posteriormente, avalia pares desses elementos comparando o ângulo entre eles, a menor distância entre seus eixos e as distâncias mínimas e máximas entre cada elemento e a linha média. Um otimizador baseado em grafos é empregado para obter o maior número de conjuntos mutuamente compatíveis, e então o alinhamento final é calculado por adição de mais resíduos até que um valor mínimo de RMSD, definido pelo usuário, seja atingido. A pontuação final do alinhamento é calculada como função do RMSD e do número de Cα pareados entre as estruturas. A significância estatística é obtida por comparação à distribuição de pontuações obtidas pelo alinhamento da proteína leghemoglobina a centenas de estruturas não redundantes.

CE: representa as proteínas como conjuntos de distâncias entre Cα de oito resíduos consecutivos na estrutura. Primeiramente, são identificados todos os pares de octâmeros compatíveis entre as estruturas. Posteriormente, um algoritmo de extensão combinatória identifica e combina os pares mais similares entre as estruturas, adicionando mais pares a cada etapa do cálculo até a obtenção do melhor alinhamento. A significância estatística é dada por comparação às pontuações obtidas em um conjunto de alinhamentos entre estruturas com menos de 25% de identidade de sequência.

MAMMOTH: transforma as coordenadas da proteína em um conjunto de vetores unitários a partir dos Cα de heptâmeros consecutivos. A similaridade entre heptâmeros é calculada pela sobreposição de seus vetores, a matriz de similaridade ótima é identificada e então o melhor alinhamento local entre estruturas é identificado dentro de um valor de RMSD pré-definido. A significância estatística é dada pelo valor p, baseado na comparação com a pontuação de alinhamentos obtidos aleatoriamente.

SALIGN: representa as proteínas por um conjunto de propriedades ou características calculadas a partir da sequência e da estrutura ou definidas arbitrariamente pelo usuário. Tais propriedades incluem tipo de resíduo, distância entre resíduos, acessibilidade da cadeia lateral, estrutura 2ária, conformação local da estrutura e característica a ser definida pelo usuário. O programa calcula uma matriz de dissimilaridade entre propriedades equivalentes, e a pontuação da dissimilaridade é calculada pela soma das matrizes de cada característica. A melhor sobreposição de matrizes é

obtida por um algoritmo baseado em programação dinâmica. A significância estatística não é calculada pelo SALIGN e o usuário obtém apenas os valores da pontuação de dissimilaridade. O programa fornece, entretanto, um valor adicional de qualidade, apresentado como porcentagem de Cα cuja distância é menor que 3,5 Å entre os pares de estruturas alinhadas.

3.10. Alinhamento de >2 estruturas

A maior parte dos métodos disponíveis para o alinhamento múltiplo de estruturas inicia-se estabelecendo todos os alinhamentos entre pares de estruturas e, então, emprega-os para estabelecer um alinhamento consenso entre todas as estruturas. A Figura 15-3 apresenta um exemplo de alinhamento estrutural múltiplo. Os métodos para obter o alinhamento consenso variam entre os programas de alinhamento. A seguir apresentamos as características específicas de alguns dos métodos mais utilizados para o alinhamento de estruturas múltiplo.

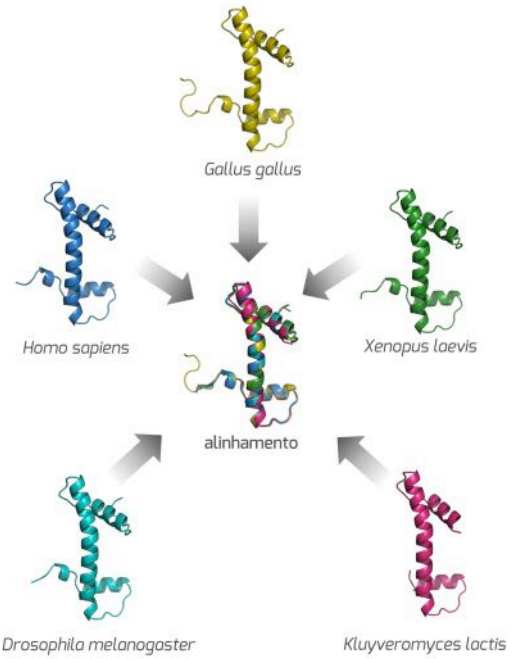


Figura 15-3: Exemplo de alinhamento de múltiplas estruturas proteicas, oriundas de diferentes organismos (histonas H3 de levedura, mosca-da-fruta, homem, frango, sapo-de-garras).

CE-MC: realiza o refinamento de um conjunto de alinhamentos de pares de estruturas empregando uma técnica de otimização de Monte Carlo. O algoritmo modifica o alinhamento múltiplo aleatoriamente, e as modificações são aceitas se houver melhoria na pontuação do alinhamento. O processo encerra quando o alinhamento múltiplo não puder mais ser melhorado por modificações aleatórias.

MAMMOTH-Mult: essa extensão do MAMMOTH gera inicialmente todos os alinhamentos de estruturas aos pares. Um procedimento de organização por médias é empregado para agrupar as estruturas com base em suas similaridades aos pares, gerando uma árvore. O alinhamento múltiplo é gerado por reorganização dessa árvore, onde ramos similares vão sendo agrupados aos pares, iterativamente.

SALIGN: pode realizar alinhamentos múltiplos de duas maneiras, baseado em uma árvore ou por alinhamento progressivo. O primeiro caso é muito similar ao MAMMOTH-Mult. No alinhamento progressivo, as estruturas são alinhadas na ordem em que são fornecidas para o programa. A vantagem desse método é o de seu custo computacional ser menor que o do método baseado em uma árvore.

3.11. Alinhamento flexível

O alinhamento de estruturas considerando sua flexibilidade está se tornando cada vez mais importante devido à melhor compreensão do enovelamento proteico. Cada vez mais, percebe-se que não existem enovelamentos estanques, mas sim um gradiente densamente populado por variantes conformacionais. Desta forma, torna-se mais difícil definir domínios proteicos, sendo mais adequado descrever as estruturas como conjuntos de estruturas supra-secundárias. Com base nessa proposta, a diferença entre proteínas relacionadas reside na orientação relativa desses subdomínios. A Figura 16-3 demonstra as diferenças que podem ser observadas ao alinhar um par de estruturas de maneira rígida ou flexível. A seguir apresentamos as características específicas de alguns dos métodos mais utilizados para este tipo de alinhamento de estruturas.

FATCAT: o algoritmo adiciona "torções" entre pares de fragmentos proteicos alinhados, que são tratados

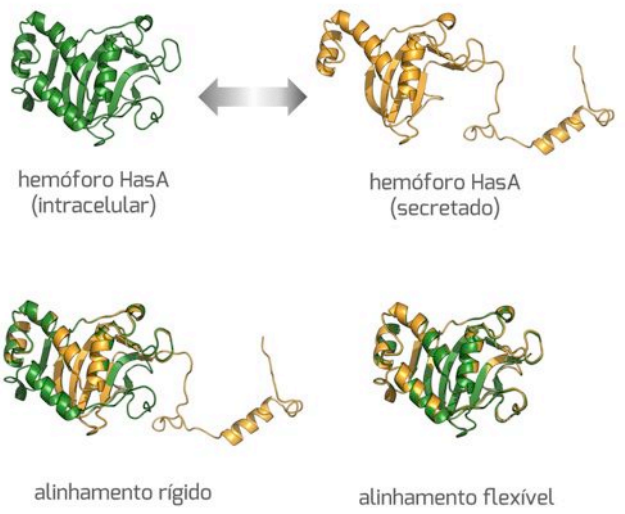


Figura 16-3: Comparação entre alinhamento estrutural rígido e flexível. A estrutura da proteína HasA (um captador bacteriano de grupamentos heme) foi obtida para suas formas intra- e extra-celular. Observe que o alinhamento rígido identifica similaridade parcial entre as estruturas, enquanto o alinhamento flexível detecta o rearranjo espacial de parte da proteína, evidenciando sua identidade.

como corpos rígidos. De maneira geral, o programa permite a inclusão dessas torções quando elas diminuem o valor final do RMSD, refletindo em um melhor alinhamento estrutural. O alinhamento final é obtido por programação dinâmica e se baseia na matriz de similaridade entre os fragmentos pareados, obtidos na primeira etapa do cálculo.

FLEXPROT: mantém uma das proteínas rígida, enquanto a outra pode sofrer alterações em busca de maior similaridade estrutural. As regiões potencialmente flexíveis da proteína são detectadas automaticamente e empregadas nas alterações conformacionais.

ALADYN: alinha pares de estruturas com base em sua dinâmica interna e similaridade entre seus movimentos de grande escala. O posicionamento ótimo entre as proteínas é encontrado ao maximizar as similaridades entre os padrões de flutuação estrutural, que são calculados pelo modelo de redes elásticas.

POSA: uma variante do FATCAT para o alinhamento múltiplo flexível de estruturas. Emprega uma metodologia combinada, introduzindo grafos de ordem parcial para visualizar e agrupar regiões similares entre as estruturas.

3.12. Conceitos-chave

Algoritmo: sequência lógica de instruções necessárias para executar uma tarefa.

Alinhamento: método de organização de sequências ou estruturas biológicas para evidenciar regiões similares e dissimilares. Estes métodos estão geralmente atrelados a inferências funcionais ou evolutivas.

Alinhamento Múltiplo: alinhamento que envolve mais de duas sequências ou estruturas

Alinhamento Simples: alinhamento que envolve apenas duas sequências ou estruturas.

BLAST: *Basic Local Alignment Search Tool* (Ferramenta de Busca por Alinhamento Local Básico), empregado para buscar sequências em bancos de dados com base em sua similaridade.

Homologia: é um termo essencialmente qualitativo que denota uma ancestralidade comum de determinada sequência.

HSP: pares de segmentos de alta pontuação (*high-scoring segment pairs*), zonas de similaridade entre sequências identificadas pelo BLAST.

Identidade: Porcentagem de caracteres similares entre duas sequências (excluindo-se as lacunas).

Indels: identifica inserções e deleções de caracteres ao longo do processo evolutivo.

Lacunas: regiões identificadas por hifens que representam a inserção/deleção de caracteres ao longo do processo evolutivo.

Matches: regiões que apresentam caracteres idênticos entre diferentes sequências.

Mismatches: regiões que apresentam caracteres não idênticos entre diferentes sequências.

Penalidades por lacuna (PL): conjunto de parâmetros necessários para atribuir a pontuação para uma lacuna em um sistema de alinhamento por pontuação.

RMSD: desvio médio quadrático.

Tradução: tradução (*in silico*) de uma sequência de mRNA em sua possível sequência proteica correspondente

3.13. Leitura recomendada

BOGUSKI, Mark S. A molecular biologist visits Jurassic Park. ***Biotechniques***, 12, 668-669, 1992.

CARUGO, Oliviero. Recent progress in measuring structural similarity between proteins. ***Curr. Protein. Pept. Sci.***, 8, 219-241, 2007.

MADDEN, Tom. The BLAST sequence analysis tool. In: McENTYRE, Jo; OSTELL, Jim (Org.). ***The NCBI Handbook***. Bethesda: National Center for Biotechnology Information, 2002.

MARTI-RENOM, Marc A.; et al. Structure comparison and alignment. In: GU, Jenny; BOURNE, Philip E. (Org.). ***Structural Bioinformatics***. 2.ed. Hoboken: John Wiley & Sons, 2009.

MAYR, Gabriele; DOMINGUES, Francisco S.; LACKNER, Peter. Comparative analysis of protein structure alignments. ***BMC Struct. Biol***., 7, 50, 2007.

MOUNT, David W. ***Bioinformatics: Sequence and Genome Analysis***. 2.ed. Cold Spring Harbor: Cold Spring Harbor Laboratory Press, 2004.

ROSSMANN, Michael G.; ARGOS, Patrick. The taxonomy of binding sites in proteins. ***Mol. Cell. Biochem.***, 21, 161-182, 1978.

Apêndice 06

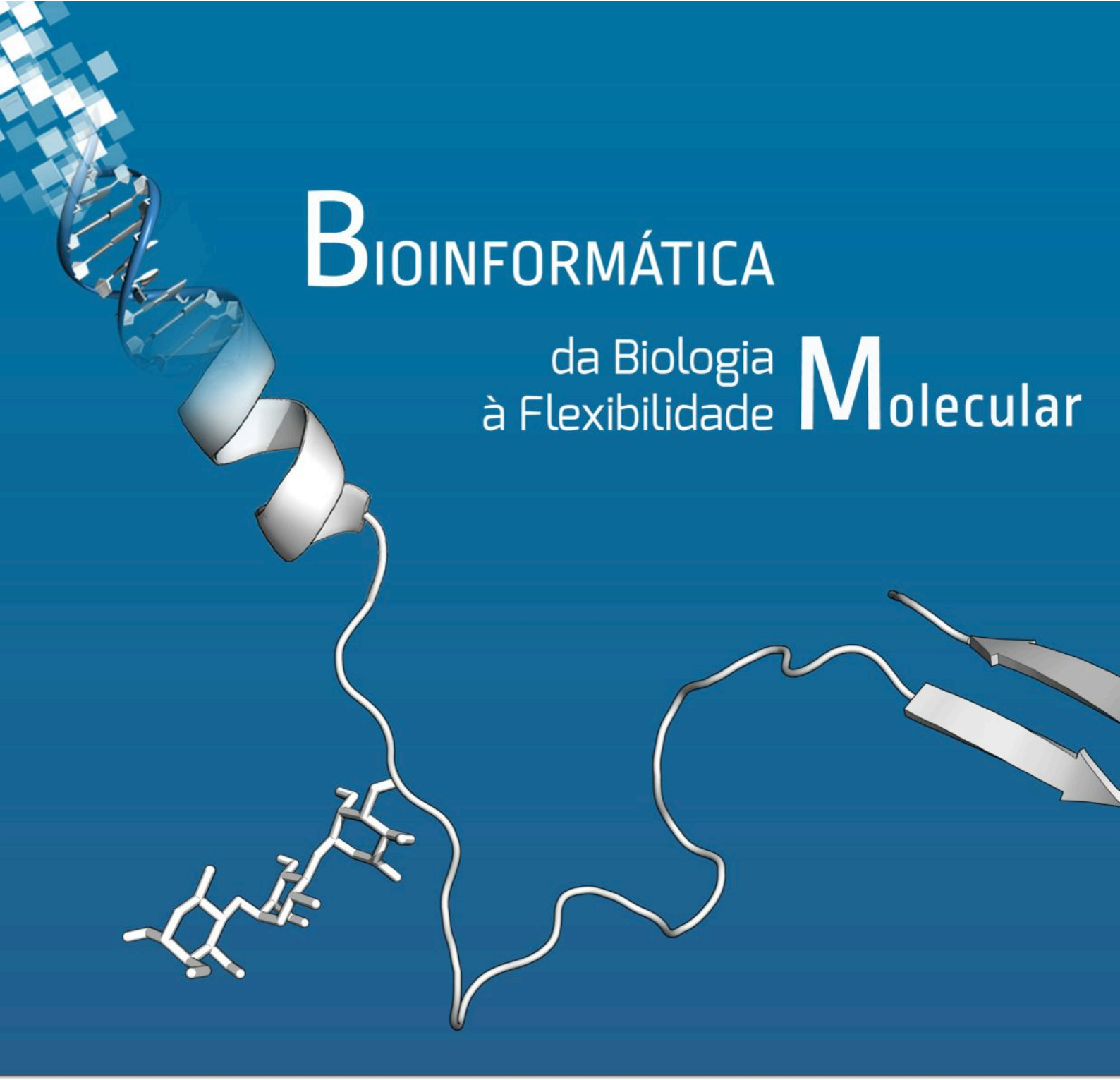
"Filogenia Molecular"

Braun RL, Junqueira DM, , Verli H

Bioinformática: da Biologia à Flexibilidade Molecular, 2014

Capítulo 05

E-book disponível em: http://www.ufrgs.br/bioinfo/ebook/



Bioinformática

da Biologia à Flexibilidade Molecular

Hugo Verli (org)

Apoio:

CNPq CAPES FAPERGS UFRGS Universidade Federal do Rio Grande do Sul

Conteúdos

Autores

Bruno César Feltes
Centro de Biotecnologia, UFRGS

Camila S. de Magalhães
Pólo de Xerém, UFRJ

Charley Christian Staats
Centro de Biotecnologia, UFRGS

Dennis Maletich Junqueira
Depto Genética, UFRGS

Diego Bonatto
Centro de Biotecnologia, UFRGS

Edwin A. Yates
Instituto de Biologia Integrativa, Universidade de Liverpool

Fabio Lima Custódio
Laboratório Nacional de Computação Científica

Fernanda Rabaioli da Silva
Centro de Biotecnologia, UFRGS

Fernando V. Maluf
Centro de Inovação em Biodiversidade e Fármacos, IFSC - USP

Glaucius Oliva
Centro de Inovação em Biodiversidade e Fármacos, IFSC - USP

Gregório K. Rocha
Laboratório Nacional de Computação Científica

Guilherme Loss de Morais
Laboratório Nacional de Computação Científica

Helena B. Nader
Departamento de Bioquímica, Unifesp

Hugo Verli
Centro de Biotecnologia, UFRGS

Isabella A. Guedes
Laboratório Nacional de Computação Científica

Ivarne L. S. Tersariol
Departamento de Bioquímica, Unifesp

João Renato C. Muniz
Grupo de Biotecnologia Molecular, IFSC - USP

Joice de Faria Poloni
Centro de Biotecnologia, UFRGS

Laurent E. Dardenne
Laboratório Nacional de Computação Científica

Luís Maurício T. R. Lima
Faculdade de Farmácia, UFRJ

Marcelo A. Lima
Departamento de Bioquímica, Unifesp

Marcius da Silva Almeida
Instituto de Bioquímica Médica, UFRJ

Priscila V. S. Z. Capriles
PPG Modelagem Computacional, UFJF

Raphael Trevizani
Laboratório Nacional de Computação Científica

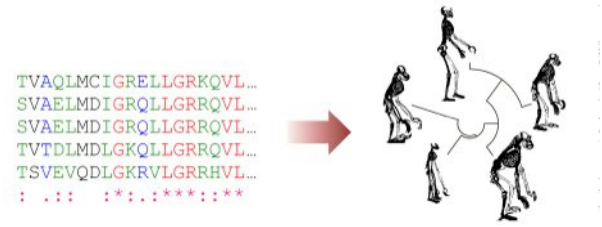
Rafael V. C. Guido
Centro de Inovação em Biodiversidade e Fármacos, IFSC - USP

Rodrigo Ligabue Braun
Centro de Biotecnologia, UFRGS

Rogério Margis
Centro de Biotecnologia, UFRGS

Yraima Cordeiro
Faculdade de Farmácia, UFRJ

5. Filogenia Molecular



Estabelecimento de relações evolutivas a partir de sequências de aminoácidos ou nucleotídeos.

Rodrigo Ligabue Braun
Dennis Maletich Junqueira
Hugo Verli

5.1. Introdução

Desde seus primórdios, a humanidade se mostrou inclinada a organizar e classificar o mundo à sua volta com o objetivo de facilitar o entendimento e a comunicação. Em relação ao mundo natural, diferentes sistemas foram empregados para compor métodos de organização e classificar os organismos, utilizando critérios naturais ou artificiais.

Um dos sistemas de maior influência no período pré-Darwiniano foi a Escala Natural de Platão. Neste sistema, do fogo ao ser humano, diferentes níveis eram organizados à maneira de uma escada. A ideia de ascensão estava associada à perfeição, representada em sua forma plena pelo homem. O sistema classificatório de Lineu, por sua vez, se baseava em características visíveis, arbitrariamente selecionadas para classificar os seres vivos (por exemplo, número de patas ou de pétalas), sendo o ser humano o organismo do topo da cadeia. Sistemas como este são considerados sistemas artificiais, pois estão sujeitos à tendência de seu autor em considerar um caractere em detrimento de outro(s), conforme sua vontade ou necessidade. Entretanto, como o próprio Lineu reconheceu, tais sistemas foram absolutamente necessários para a fase inicial (descritiva) da biologia, servindo de base para o sistema natural de classificação e para as hipóteses de similaridade que surgiriam a seguir.

Ao final do século XVIII e início do século XIX, surgem os sistemas naturais de classificação. Estes buscavam refletir sobre a ordem natural dos seres vivos através de poucas características intrínsecas, geralmente associadas à forma. No entanto, com o objetivo de tornar a classificação mais racional, tomaram lugar debates sobre a real necessidade de haver um sistema hierárquico de organização dos organismos. Opositores da ideia consideravam que a classificação era, muitas vezes, inadequada e desnecessária, e que não deveria ser um fim em si mesma, senão um método para o levantamento de novas perguntas à Biologia.

Em 1818, a introdução do conceito de homologia por E.G. Saint-Hillaire causa uma revolução nas ciências biológicas. Para ele e seus colegas, partes homólogas correspondiam às partes de animais diferentes com uma estrutura essencialmente semelhante, mesmo com forma ou função distintas. Por

exemplo, as asas de um morcego, as nadadeiras de uma baleia e os braços de um macaco, segundo esta lógica, são considerados órgãos homólogos e podem servir como critério para agrupar morcegos, baleias e macacos em um mesmo grupo. Assim, a homologia serviria como critério principal para uma classificação natural dos organismos.

A primeira árvore filogenética moderna (esboço de Darwin no manuscrito de A Origem das Espécies)

A partir da famosa publicação de Darwin, "A Origem das Espécies", em 1859, a classificação dos organismos passou a ser não apenas natural, mas também a apresentar uma condição essencial de ancestralidade comum. Segundo este pensamento, os organismos são derivados uns dos outros, desde o surgimento da vida na terra. Darwin representou este padrão através de um esquema de ramificação, onde os galhos representam o tempo entre o organismo ancestral e o novo organismo, e os nós representam os próprios organismos. Mais tarde, esta viria a ser a primeira árvore filogenética utilizada para representar processos evolutivos.

Com influência direta da teoria evolutiva de Darwin (e colaborações de Wallace e Lamarck), desenvolve-se a Taxonomia Evolutiva. Este sistema de classificação incorporou o vetor tempo (caráter temporal normalmente inferido por meio de fósseis) e, além disto, adicionou uma quantificação da divergência estrutural entre os grupos (a chamada distância patrística). Já em meados do século XX, inicia-se a Fenética (taxonomia numérica ou neodansoniana). Esta escola buscava incluir na classificação dos organismos o máximo possível de características, atribuindo-lhes o mesmo peso na tentativa de eliminar qualquer subjetividade ou arbitrariedade. Seu impacto, entretanto, foi limitado devido às dificuldades em traduzir os índices (valores) obtidos em informações relevantes do ponto de vista biológico (como a separação de espécies, por exemplo). Na mesma época, surge a Cladística (ou sistemática filogenética), liderada pelo entomólogo alemão Willi Hennig. Na proposta de Hennig (1950), organismos que compartilhassem características derivadas (apomórficas) poderiam ser considerados descendentes do organismo ancestral, na qual a característica em seu estado primitivo (ou plesiomórfico) passou para o estado derivado.

Desde a origem dos sistemas de classificação até a Cladística, os métodos baseavam-se essencialmente no fenótipo dos organismos, ou seja, em suas características físicas claramente discerníveis. Entretanto, com o advento dos métodos de sequenciamento, tanto protéico quanto genômico, cada vez mais os dados moleculares foram se tornando importantes nas análises evolutivas de ancestralidade. Neste sentido, a ciência passa de um ponto de vista macroscópico a um ponto de vista molecular de análise.

O método de sequenciamento de aminoácidos, iniciado por Sanger em 1954, abriu caminho para que proteínas de uma mesma classe, em diferentes organismos, pudessem ser comparadas quanto às suas origens evolutivas. Da mesma forma, ao decodificar a primeira longa sequência de DNA, em 1977, Sanger deu início à explosão do sequenciamento de ácidos nucleicos, permitindo a comparação de genes em larga escala. É importante destacar que as sequências moleculares podem tanto ser comparadas entre si, buscando conhecer a história evolutiva de um gene ou proteína (por exemplo, relações entre hemoglobinas de diferentes mamíferos), quanto podem ser associadas a outros dados na reconstrução da história evolutiva de organismos (por exemplo, associando as relações obtidas por comparação de DNA ribossomal de aves com datação de fósseis, buscando estabelecer relações de ancestralidade).

No entanto, ao lidar com sequências moleculares, diferentes questões podem surgir. Por exemplo, o conceito de gene é di-

nâmico e mudou muito desde sua primeira definição. Além disso, genes podem sofrer diferentes processos evolutivos que alteram sua estrutura e/ou função, como mutações e rearranjos, ou ainda duplicações e perdas de função. Esses fatores fazem com que a relação 1:1 entre gene e organismo seja perdida. Por exemplo, uma mesma leguminosa pode possuir duas cópias do gene para a proteína leghemoglobina (genes parálogos). Além disso, muitas sequências do genoma não chegam à etapa de tradução, podendo conter elementos regulatórios ou transponíveis. Tais variações aumentam a complexidade e dificultam a interpretação das relações de descendência.

5.2. Aplicações

Ao classificarmos os organismos, atribuímo-lhes uma história evolutiva. Essa história, entretanto, é frequentemente desconhecida. Sendo assim, é necessário inferir a sequência de mudanças que levaram ao surgimento de um novo organismo ou proteína. Contudo, existe apenas uma história verdadeira, que talvez jamais seja conhecida. Assim, ao empregarmos as técnicas filogenéticas, o objetivo é coletar e analisar dados capazes de fornecer a melhor estimativa para chegarmos à filogenia verdadeira. De certa forma, a obtenção de filogenias lembra a atuação de um historiador. Baseando-se em dados disponíveis no presente (tais como organismos vivos, fósseis e sequências moleculares), tenta-se obter uma imagem de como teria sido o passado.

Quando analisamos sequências de nucleotídeos ou aminoácidos para inferir uma filogenia, utilizamos informações derivadas das taxas evolutivas para determinar a sequência de eventos que levaram ao surgimento de novos organismos. A taxa de evolução molecular refere-se à velocidade na qual os organismos acumulam diferenças genéticas ao longo do tempo. Essa taxa é frequentemente definida pelo número de substituições por sítio (ou posição no alinhamento de sequências) por unidade de tempo e, portanto, são usadas para descrever a dinâmica das mudanças em uma linhagem ao longo de várias gerações.

As taxas evolutivas são empregadas quando se buscam estimativas temporais para datação de eventos evolutivos. Normalmente, se assume que as mudanças nas sequências se acumulam a uma taxa mais ou menos constante ao longo do tempo. Esse conceito é chamado de Hipótese do Relógio Molecular. Entretanto, é conhecido que as taxas evolutivas são dependentes de vários fatores, tais como o tempo de geração, o tamanho da população e do próprio metabolismo, o que normalmente viola o modelo estrito de relógio molecular. Com base nestas informações, diversos modelos foram propostos para lidar com desvios no comportamento temporal de diferentes linhagens moleculares e, hoje em dia, são referidos como relógios moleculares relaxados.

Atualmente, a inferência filogenética é um campo de pesquisa à parte das outras ciências. Tornou-se uma ferramenta complementar para diversas áreas e indispensável para outras. Apesar de ter sido idealizada para desvendar apenas as relações evolutivas entre organismos, atualmente a filogenética molecular é aplicada a problemas muito mais diversos que este. Com o advento do relógio molecular estrito, foi possível aplicar a estimativa de tempo às filogenias e datar surgimento de espécies, disseminação de organismos e, até mesmo, entender grandes eventos biológicos que ocorreram no passado. Com a abordagem relaxada do relógio molecular, iniciou-se a utilização de modelos de dinâmica populacional que comportam os eventos coletivos de grupos específicos. Ainda, com o avanço da capacidade de processamento computacional, vem sendo possível criar algoritmos capazes de reconstruir genomas ancestrais. Também a partir da filogenética molecular desenvolveu-se o campo da filogeografia. Segundo esta área do conhecimento, as filogenias podem ser utilizadas para verificar a distribuição geográfica de indivíduos. Neste contexto, outras técnicas, além das filogenias, são incorporadas às aná-

lises, incluindo a estruturação de genes, as análises de redes e as análises de haplótipos.

A filogenia molecular busca inferir a história evolutiva de organismos ou outras entidades biológicas (como proteínas e genes) a partir de sequências de ácidos nucleicos ou aminoácidos. Ao investigar as relações entre diferentes espécies, análises de genes ribossomais são comumente empregadas, pois independentemente da espécie ou do organismo, os indivíduos possuirão genes codificantes de RNA ribossômico. Em contrapartida, quando se busca compreender as relações entre diferentes enzimas de uma mesma família é necessário utilizar sequências de aminoácidos, e não de nucleotídeos. Em determinadas situações, o genoma completo pode ainda ser utilizado para inferir a filogenia. Este é o caso de diversos vírus, especialmente quando se busca compreender a origem de novas variantes ou a disseminação de uma cepa. O alvo de estudo (isto é, sequência de nucleotídeos ou aminoácidos, gene ou genoma) depende, exclusivamente, do objetivo da análise e é um dos principais fatores a ser definido primariamente pelo pesquisador.

Atualmente, as filogenias funcionam como importantes ferramentas para diferentes áreas do conhecimento, incluindo as áreas de evolução, genética, epidemiologia, microbiologia, virologia, parasitologia, botânica e zoologia, dentre outras. Adicionalmente, de maneira inédita, a inferência filogenética foi utilizada como evidência para a resolução de crime e principal prova durante um impasse internacional envolvendo diferentes países. Em resumo, dependendo do objetivo, os métodos de construção de filogenias (inferência filogenética) são a base para diversas áreas e importantes objetos para o avanço computacional na análise de dados biológicos.

5.3. Representação de árvores

A Filogenética (termo obtido por união dos termos gregos para tribo e origem) é a ciência que busca reconstruir a história evolutiva dos organismos, levando em conta as sequências de nucleotídeos ou aminoácidos. As hipóteses sobre a história evolutiva são o resultado dos estudos filogenéticos e se chamam Filogenia.

As filogenias ou árvores filogenéticas representam o contexto evolutivo dos organismos de forma gráfica. São formadas por nós (pontos) ligados por diversos ramos (linhas) (Figura 1-5). Os nós terminais, mais externos na filogenia, identificam os indivíduos, genes ou proteínas que foram amostrados e incluídos na análise filogenética. Geralmente representam o alvo de estudo do pesquisador e estão ligados aos nós mais internos na filogenia através de traços horizontais, chamados de ramos terminais (Figura 1-5).

Os nós internos, pelo contrário, representam indivíduos não amostrados. Eles identificam uma inferência evolutiva do ancestral comum mais recente dos ramos derivados daquele nó e se ligam a nós cada vez mais internos, através dos ramos internos. Por exemplo, na Figura 1-5, os grupos de nós terminais representados em verde possuem como ancestral comum o nó laranja, mais interno, enquanto os nós terminais azuis possuem como ancestral comum o nó lilás. Da mesma forma, o nó vermelho é a representação do indivíduo, gene ou proteína mais ancestral da filogenia que, através de processos evolutivos, deu origem aos nós laranja e lilás.

O tamanho dos ramos horizontais pode ter diferentes significados, dependendo do método para inferência da filogenia, conforme

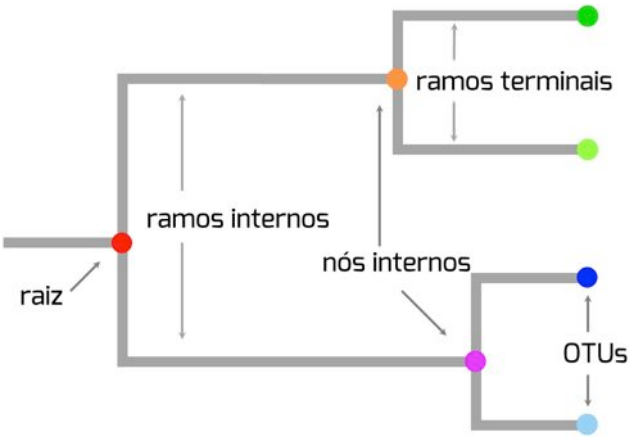


Figura 1-5: Nomenclatura associada a árvores filogenéticas.

veremos a seguir. No entanto, os ramos representados na vertical (Figura 1-5) não expressam qualquer significado, e seu tamanho não altera em nada a idéia filogenética. Como a análise pode ser feita em diferentes níveis, utilizando dados moleculares de genes, proteínas, indivíduos, espécies, gêneros, famílias, ou qualquer outro taxon, os nós terminais são amplamente denominados OTUs (*operational taxonomical units*), ou unidades taxonômicas operacionais (também chamados de folhas, Figura 2-5). A ordem e disposição exata das OTUs em uma filogenia é denominada topologia.

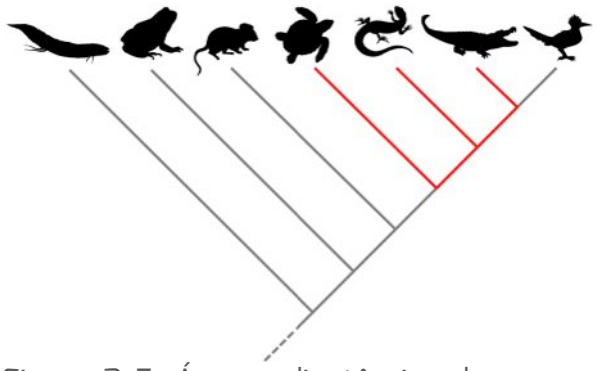


Figura 2-5: Árvore dicotômica dos grupos de vertebrados. As OTUs (nós terminais) estão representadas por ícones (peixes pulmonados, anfíbios, mamíferos, tartarugas, lagartos e serpentes, crocodilos e aves). Observe que o grupo dos répteis é parafilético (destacado em vermelho). O grupo seria considerado monofilético se incluísse as aves.

Além da forma gráfica, as árvores filogenéticas podem também ser descritas na forma textual. Em vez do diagrama com linhas e pontos, as relações evolutivas são representadas por notações com parênteses. A estrutura da árvore da Figura 2-5, por exemplo, pode ser descrita linearmente como (Peixes pulmonados, (Anfíbios, (Mamíferos, (Tartarugas, (Lagartos, (Crocodilos, Aves)))))) ou (Peixes pulmonados + (Anfíbios + (Mamíferos + (Tartarugas + (Lagartos + (Crocodilos + Aves)))))). Estas notações foram desenvolvidas para utilização computacional da informação filogenética. Algoritmos e programas que realizam análises moleculares necessitam da informação na forma textual e, quando necessário, fornecem a saída para o usuário na forma gráfica.

Partindo do princípio de derivação evolutiva, onde um organismo dá origem a outro (ou outros), podemos reconhecer dois principais processos na representação de filogenias: derivação dicotômica e derivação politômica. No primeiro caso, cada nó interno dá origem a apenas dois ramos. Para espécies, por exemplo, a ramificação de um ancestral comum em dois ramos evidencia o processo de especiação. No segundo caso, três ou mais ramos surgem de um mesmo nó interno.

Apesar de árvores dicotômicas serem mais comuns e normalmente esperadas, em alguns casos, como a dispersão explosiva do HIV e do HCV, árvores politômicas representam melhor o processo evolutivo. Casos como estes, onde um ancestral comum origina simultaneamente várias linhagens descendentes, são chamadas de politomias verdadeiras (*hard polytomies*). Por outro lado, as politomias falsas (*soft polytomies*) são casos onde a topologia não foi bem resolvida por não haver certeza do padrão de ancestralidade, tornando múltipla uma divisão que se esperaria ser formada por uma série de divisões dicotômicas.

Assim, ao agruparmos as OTUs segundo a sua ancestralidade, podemos reconhecer diferentes padrões: grupos monofiléticos, parafiléticos e polifiléticos (Figura 2-5). Os grupos monofiléticos incluem todos os membros descendentes de um único ancestral, assim como o próprio ancestral. Na Figura 2-5, por exemplo, as aves e os crocodilos são considerados um grupo monofilético, pois compartilham o mesmo ancestral comum. Da mesma forma, as aves, os crocodilos e os lagartos também podem ser considerados um grupo monofilético, pois se originaram de um mesmo ancestral. A análise das relações entre os grupos, neste caso, dependerá do objetivo do pesquisador. Adicionalmente, os grupos monofiléticos podem ser denominados clados por agruparem duas ou mais sequências que são descendentes de um mesmo ancestral (Figura 3-5a e b). A organização da topologia em que um clado está contido em outro é comumente chamada de clados aninhados ou clados embutidos (Figura 3-5c).

Os grupos parafiléticos, por sua vez, se

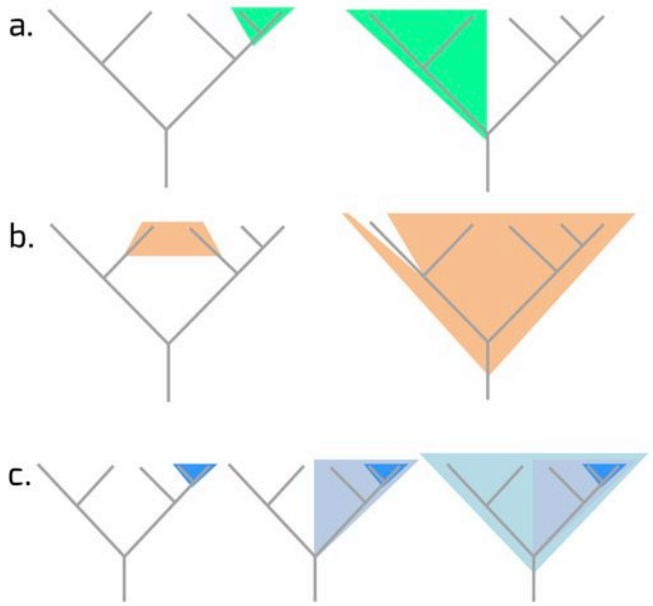


Figura 3-5: (a) Exemplos de clados destacados em verde. (b) Exemplos de organizações da topologia que não caracterizam a existência de um clado, destacados em laranja. (c) Diferentes níveis de clados que podem estar embutidos em um clado de maior ordem. Observe que os clados de diferentes ordens, quando embutidos, formam clados monofiléticos.

originam de um único ancestral, mas nem todos os organismos derivados deste ancestral fazem parte do grupo. Na Figura 2-5, os répteis são um grupo formado pelas tartarugas, lagartos e crocodilos, e seu ancestral comum está na base do ramo que dá origem às tartarugas. No entanto, este ancestral comum também deu origem às aves e, por isso, os répteis não podem ser considerados um grupo monofilético, mas um grupo parafilético.

Finalmente, os grupos polifiléticos provêm de dois ou mais ancestrais diferentes. Nestas relações se encontram OTUs que apresentam características comuns, mas que possuem diferentes ancestrais comuns. Por exemplo, a condição endotérmica (animais que mantém a sua temperatura corporal constante) é apenas apresentada por aves e mamíferos. Por este critério, poderíamos agrupar estes dois grandes grupos sem, no entanto, compartilharem o mesmo ancestral comum direto (Figura 2-5). A organização destes grupos permite descrever características resultantes de convergência evolutiva, pois uma mesma característica se desenvolveu independentemente em diferentes grupos.

Sabendo das relações evolutivas entre os táxons e da existência de ancestrais comuns, as árvores podem ser representadas de maneira a evidenciar o ancestral mais antigo (árvore com raiz ou enraizada), ou apenas destacar as relações evolutivas entre os táxons, sem destacar qual a OTU mais ancestral (árvore sem raiz ou não enraizada) (Figura 4-5).

A raiz da filogenia é a espécie ou sequência ancestral a todo o grupo que está sob análise. Quando presente, a raiz aplica uma direção temporal à árvore, permitindo observar o sentido das mudanças evolutivas da raiz (mais antigo) aos ramos terminais (mais modernos). Uma árvore não enraizada, pelo contrário, reflete apenas a topologia estabelecida entre as OTUs, sem indicar o ancestral do grupo. Árvores não enraizadas podem ser confusas, e sua interpretação requer mais cuidado devido à facilidade em cometer erros de análise (Figura 4-5).

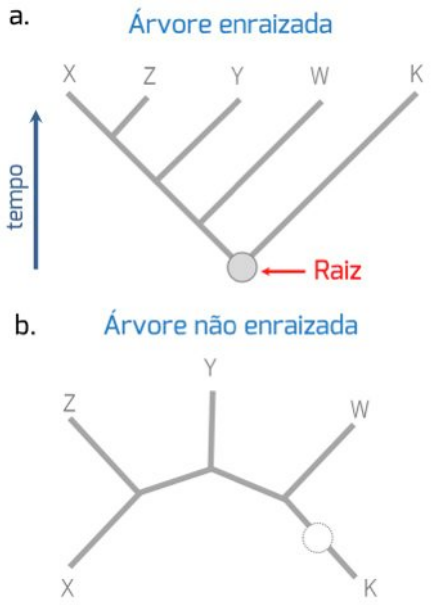


Figura 4-5: Comparação de árvores (a) enraizadas e (b) não enraizadas. No primeiro caso, é possível definir a direção das mudanças evolutivas, devido à presença do vetor tempo dado pela presença da raiz.

A identificação de uma raiz nas filogenias geralmente requer a inclusão de uma ou diversas OTUs que representem grupos externos. Os grupos externos devem ser ancestrais comuns das OTUs em estudo, já conhecidos, que indicarão caracteres presentes em organismos mais próximos aos ancestrais, provendo um direcionamento para a interpretação dos processos evolutivos. Para o caso do estudo de HIV, por exemplo, é comum que os vírus da imunodeficiência de símios (SIV) sejam utilizados como grupo externo nas filogenias, pois sabidamente estes vírus deram origem ao HIV.

A adição de grupos externos aumenta o número de topologias diferentes que uma filogenia pode assumir. O número de árvores possíveis varia com o número de OTUs e com a presença ou ausência de raiz. Para mais de duas OTUs, a quantidade de possíveis árvores com raiz é sempre maior que o número de árvores sem raiz. A possibilidade de inferência de diferentes topologias para os mesmos dados moleculares ressalta a extrema variabilidade de cenários possíveis na busca do verdadeiro evento evolutivo. É importante também ressaltar que, assim como a complexidade, o tempo computacional envolvido na construção das filogenias aumenta exponencialmente com o aumento de OTUs.

Em relação à topologia das árvores, a inversão de ramos derivados de um mesmo nó não altera a relação evolutiva apresentada pela árvore (Figura 5-5). Nesse sentido, a árvore filogenética pode ser comparada a um móbile: cada peça suspensa é livre para girar em seu eixo, ficando mais próxima ou mais distante espacialmente das outras peças, sem alterar a estrutura geral do objeto. Independentemente da posição destas OTUs, após o giro dos ramos, o mesmo ancestral comum será identificado e, por isso, não há qualquer alteração no significado da filogenia.

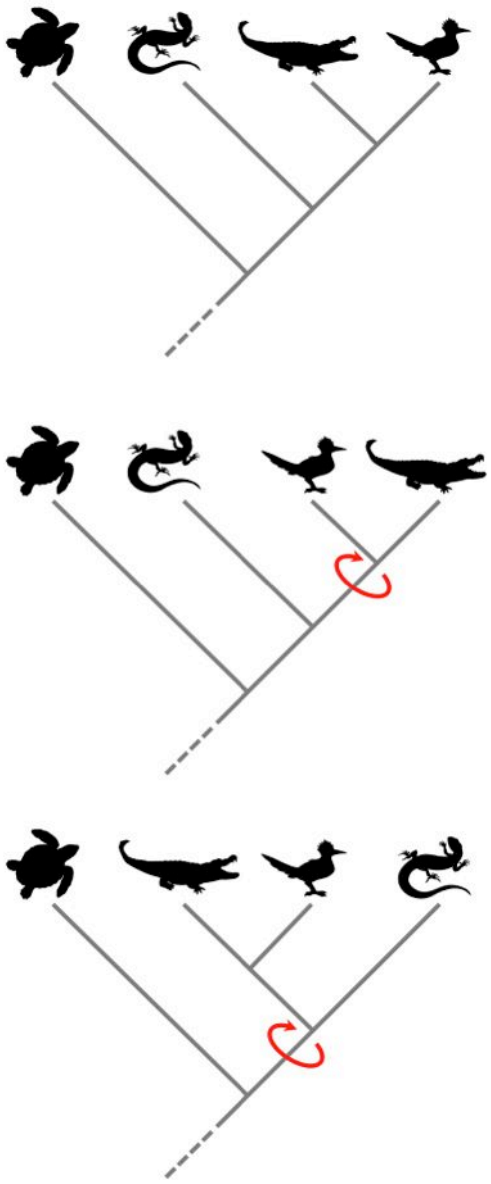


Figura 5-5: A porção terminal da árvore dos vertebrados (representada na Figura 2-5) foi rearranjada de diferentes maneiras (as setas indicam o ponto de rotação). Conforme a analogia de um móbile, todas elas representam a mesma relação evolutiva.

Quanto à nomenclatura de árvores filogenéticas, diferentes termos são empregados, tais como cladogramas, filogramas e dendrogramas (Figura 6-5). Um cladograma é uma árvore simples, que retrata as relações entre os nós terminais. Pelo contrário, uma árvore aditiva (árvore métrica ou filograma) apresenta informações adicionais, pois o comprimento dos ramos é proporcional a algum atributo, como quantidade de mudança. Por sua vez, uma árvore ultramétrica (ou dendrograma) constitui um tipo especial de filogenia devido aos seus ramos serem equidistantes da raiz. Os dendrogramas podem, desta forma, retratar o tempo evolutivo. É importante ressaltar que alguns autores denominam qualquer filogenia como cladograma, o que pode ser confuso.

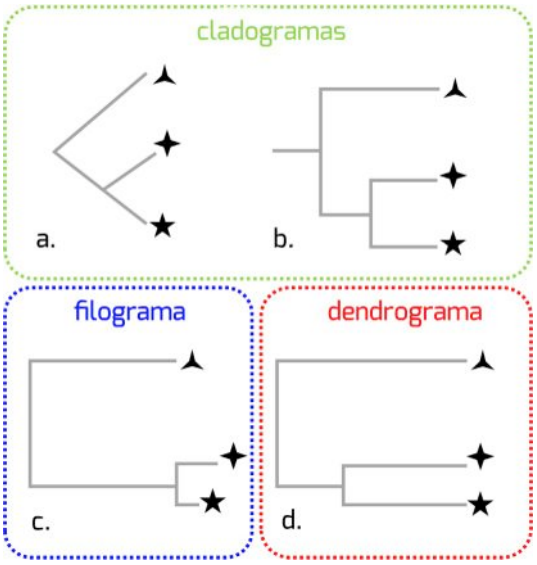


Figura 6-5: Nomenclatura de árvores filogenéticas. Observe que os cladogramas *a* e *b* são equivalentes, mas o filograma *c* e o dendrograma *d* não o são.

O tipo de dado molecular a ser empregado nas análises também deve ser levado em conta. Sequências de aminoácidos são mais conservadas que sequências de ácidos nucleotídeos em decorrência da degeneração do código genético. São, portanto, úteis em análises de produtos de genes ou espécies que visam entender fenômenos que aconteceram há amplos períodos de tempo evolutivo. Além disso, por formarem um conjunto de pelo menos 20 membros (contra quatro membros presentes em DNA ou RNA), sua variação pode ser mais significativa.

A despeito desta diferença no volume de informação, com a popularização do sequenciamento de ácidos nucleicos, especialmente DNA, sequências de nucleotídeos passaram a ser as mais empregadas em estudos de filogenia. Ácidos nucleicos são mais propensos a alterações, podendo sofrer transições (quando ocorre a troca de uma purina por outra purina, ou de uma pirimidina por outra pirimidina) e transversões (quando ocorre a troca de uma purina por uma pirimidina ou vice-versa), além de inserções ou deleções de pares de base que interferem no quadro de leitura. Essa variabilidade pode ser interessante no estudo de eventos mais recentes do ponto de vista evolutivo.

É preciso, assim, conhecer o caso de estudo e o tipo de pergunta que se busca responder com cada filogenia. Ao lidarmos com genes de diferentes espécies, por exemplo, é importante saber da existência e disposição de íntrons, da necessidade de lidar com o gene inteiro ou apenas parte dele ou da necessidade de incluir regiões regulatórias para a análise.

Um exemplo recente da aplicação de análises filogenéticas está no caso da identificação da origem da linhagem do vírus influenza H1N1, envolvido no surto de gripe de 2009. Para tanto, Smith e colaboradores empregaram genomas completos de influenza isolados de diferentes localidades e hospedeiros, e construíram árvores filogenéticas para cada uma das oito regiões do genoma buscando identificar a fonte de cada rearranjo presente no vírus envolvido no surto. Por meio das árvores obtidas, foi possível rastrear a contribuição genética dos vírus isolados de aves, suínos e humanos (Figura 7-5). Assim, o emprego da filogenia neste trabalho permitiu não apenas caracterizar o vírus do ponto de vista molecular, como também reconstruir a história evolutiva do agente etiológico de uma pandemia.

5.4. Distância genética

A formulação de modelos evolutivos é uma maneira de descrever matematicamente os processos que moldam as mudanças nas sequências de nucleotídeos ou aminoácidos dos organismos ao longo do tempo. Do ponto de vista molecular, estas mudanças podem ser resultado de diferentes forças evolutivas que reorganizam a sequência e a própria estrutura dos genes.

Um modelo geral para descrever de maneira eficaz estas alterações evolutivas deveria considerar os processos de substituição, inserção, deleção e duplicação, bem como ocorrência de transposição ou até mesmo de retrotransposição. Contudo, apesar de estes fenômenos serem claros agentes na modelagem dos genomas, matematicamente

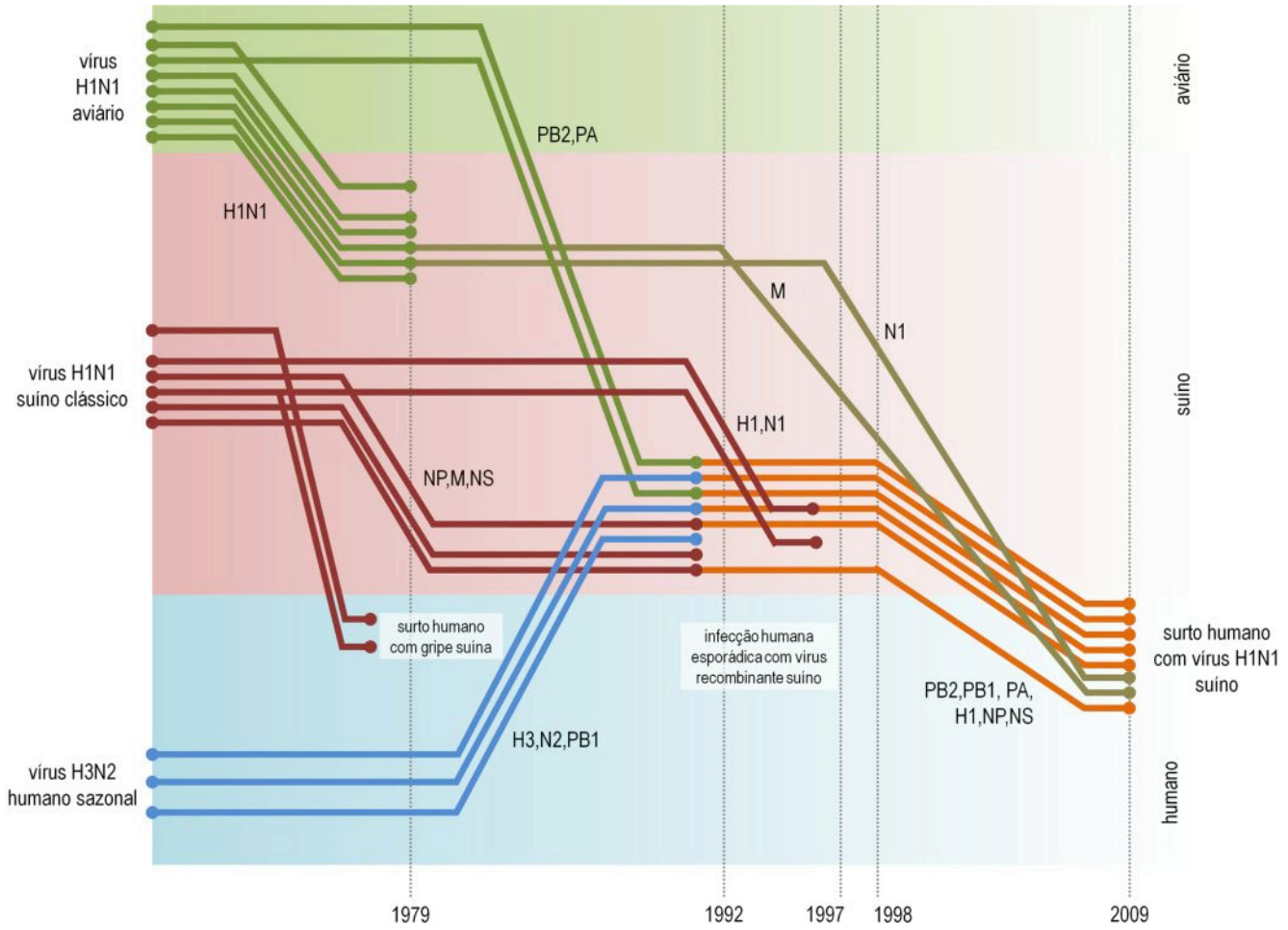


Figura 7-5: Representação esquemática das recombinações que originaram o vírus Influenza envolvido no surto de gripe suína em 2009. Diferentes linhas representam diferentes regiões do genoma do vírus. Observe a interação entre vírus de origens aviária, suína e humana em eventos que datam, pelo menos, desde 1990. Os eventos de recombinação e as análises temporais foram baseadas em análises filogenéticas (Adaptado de Smith e colaboradores, *Origins and evolutionary genomics of the 2009 swine-origin H1N1 influenza A epidemic. Nature, 459*, 1122-1125, **2009**).

ainda não é factível colocá-los como componentes de modelos que expliquem inteiramente o processo evolutivo.

Assim, devido à grande relevância dos mecanismos de substituição para a evolução dos genomas em diferentes organismos e da disponibilidade de modelos de probabilidade estatística que expliquem este processo, as trocas têm sido o principal alvo para o desenvolvimento de modelos matemáticos e compõem a base de diversos métodos de inferência filogenética.

Após a divergência de duas sequências a partir de seu ancestral comum, de forma dicotômica, fenômenos evolutivos garantirão as mudanças nas sequências de nucleotídeos de forma independente (Figura 8-5). Uma medida tradicional para expressar o número de substituições de nucleotídeos que se acumularam nas sequências desde a divergência é chamada de distância genética. Esta informação é uma medida quantitativa da dissimilaridade genética entre diferentes OTUs, e permite estabelecer uma estimativa relativa da quantidade de mudanças que ocorreram desde a divergência.

A distância é também um importante conceito na construção de filogenias, pois está diretamente relacionada com a relação evolutiva entre duas OTUs: uma menor distância

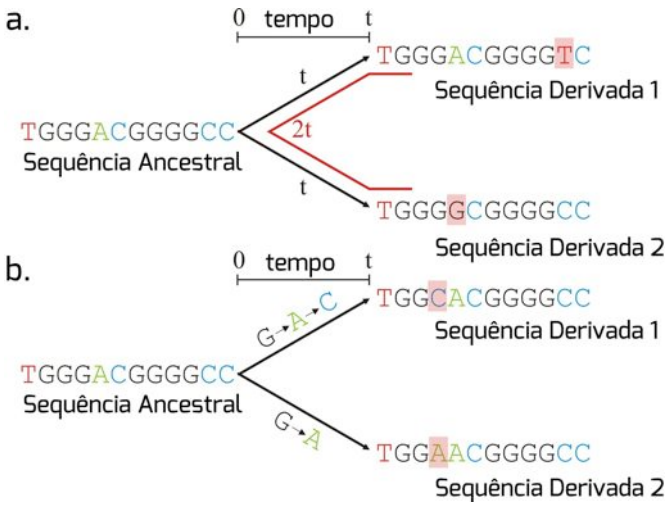


Figura 8-5: Após a divergência de dois organismos a partir de seu ancestral comum, seus genomas acumularão diferenças independentemente. (a) A medida da dissimilaridade genética entre duas sequências homólogas ao longo do tempo é chamada de distância genética, e a relação temporal entre duas sequências divergentes é dada por $2t$. (b) A ocorrência de múltiplas substituições ao longo do tempo na divergência de sequências homólogas pode mascarar as verdadeiras diferenças entre as sequências. Apesar de ocorrerem dois eventos de mutação na sequência derivada 1, apenas o último evento é observado, pois ocorreram no mesmo sítio. Os quadrados em vermelho evidenciam as diferenças em relação às sequências ancestrais.

genética indica uma relação evolutiva mais próxima, enquanto que um valor maior sugere uma derivação evolutiva proporcionalmente maior. Tipicamente, a informação da distância genética é incorporada à inferência filogenética na definição do tamanho dos ramos. No entanto, além desta informação é necessária uma escala de distância que especifique o número de mudanças que ocorreram ao longo do ramo.

O método mais simplista para avaliar a distância genética entre duas sequências é conhecido como distância p. Este método é baseado na contagem das diferenças dividida pelo número total de sítios do alinhamento. Se oito sítios são diferentes entre duas sequências homólogas com tamanho de 100pb, a distância p obtida será 0,08. Este resultado reflete a porcentagem de sítios diferentes em relação ao tamanho total da sequência, e geralmente é utilizado na especificação da escala de distância das filogenias (Figura 8-5).

A variação genética em um determinado sítio pode decorrer de diferentes processos e resultar em mais de uma substituição. As múltiplas substituições, ou *multiple hits*, ocorrem naturalmente e podem subestimar o verdadeiro número de mudanças no cálculo da distância p, já que "escondem" as diversas trocas de nucleotídeos ou aminoácidos. Na Figura 8-5b, por exemplo, apesar de ocorrerem duas substituições no mesmo sítio ao longo de um dos ramos, aparentemente a sequência derivada parece ter sofrido somente um evento evolutivo. Sendo assim, a relação entre as diferenças nas sequências e o tempo decorrido da divergência nem sempre é linear, especialmente devido à ocorrência das múltiplas substituições em um mesmo sítio.

Devido à ineficácia da distância p em efetivamente estimar a distância genética entre duas sequências, diferentes modelos probabilísticos foram desenvolvidos para descrever as mudanças entre os nucleotídeos e corrigir a distância observada. Tais modelos implicam no uso de diversas suposições simples a respeito das probabilidades de substituição de um nucleotídeo por outro, mas garantem uma aproximação da realidade quando sustentadas por uma taxa de mutação fidedigna.

Estas técnicas de correção são comumente conhecidas por modelos de substituição (ou matrizes de substituição), e garantem a conversão da distância observada em medidas de distâncias evolutivas próximas da realidade, permitindo reconstruir a história evolutiva dos organismos.

Diversos modelos de substituição foram propostos para explicar as trocas de nucleotídeos em sequências de DNA, reduzindo a complexidade do processo evolutivo a um padrão de mudança simples que consegue ser explicado através de poucos parâmetros. Todos estes modelos, no entanto, de alguma forma são interrelacionados, diferindo principalmente no número de

parâmetros utilizados para explicar estas substituições. Devido à influência do modelo de substituição na inferência de filogenias, a escolha de um método particular deve ser justificada. A estratégia mais simples é utilizar os modelos que comportam o maior número de variáveis, embora a complexidade não esteja diretamente relacionada à melhor qualidade de análise das sequências. Com o aumento de parâmetros, o sistema se torna mais complexo, aumentando a probabilidade de erro e exigindo um maior processamento computacional. Assim, é necessário verificar os alinhamentos caso-a-caso para atribuir o melhor modelo de substituição na inferência filogenética.

A substituição de nucleotídeos ou aminoácidos em uma sequência é usualmente modelada sob a forma de um processo quase aleatório. Devido ao caráter dinâmico desta aleatoriedade, é necessário enquadrar as substituições, seguindo certos pressupostos. Assim, as substituições são descritas por um processo de Markov homogêneo, onde a probabilidade de substituição de um nucleotídeo X pelo Y não depende do estado prévio do nucleotídeo X.

As probabilidades de mudança de um nucleotídeo para outro (ou de um aminoácido para outro) são especificadas através de uma matriz 4x4 das taxas de substituição (ou 20x20 no caso dos aminoácidos) que especificam com qual taxa cada um dos nucleotídeos ou aminoácidos poderá mudar para outro. É necessário assumir também que os eventos de substituição sejam independentes ao longo dos sítios das sequências, e ainda, possuam um caráter reversível. Além disso, devem especificar a frequência estacionária dos nucleotídeos, ou frequência de equilíbrio, onde será atribuída a provável proporção de cada um dos caracteres na sequência.

Para sequências de nucleotídeos, o modelo de substituição mais simples foi proposto por Jukes e Cantor em 1969 (JC69). Segundo este modelo, as mudanças entre os nucleotídeos podem ocorrer com a mesma probabilidade, assumindo uma frequência estacionária igual para todos (cada nucleotídeo tem 25% de chance de ocorrer na sequência).

Com o advento da publicação das primeiras sequências de genoma mitocondrial, na década de 1980, se observou que as transições eram muito mais comuns que as transversões. Devido à uniformidade do método proposto por Jukes e Cantor, foi necessário criar um modelo que acomodasse essas diferenças.

Assim, o modelo proposto por Kimura (K80 ou K2P) cria as variáveis α e β para representar, respectivamente, as taxas de transição e de transversão. Apesar da inclusão de dois parâmetros, as frequências de equilíbrio se mantêm constantes em ¼ para cada nucleotídeo. Em 1981, Kimura adiciona um terceiro parâmetro (γ) ao modelo já proposto, passando a ser identificado como K3P. A atualização do modelo permitiu dividir as taxas de transversão em duas variáveis.

Alguns genomas apresentam uma grande quantidade de guaninas e citosinas em relação a timinas e adeninas. Se algumas bases são mais frequentes que outras, será esperado que algumas substituições ocorram com mais frequência que outras. O modelo criado por Felseinstein (F81) acomoda essas observações e permite que as proporções individuais de cada nucleotídeo (frequência estacionária) sejam diferentes de ¼. É importante ressaltar que este modelo considerará a mesma proporção de bases em todas as sequências envolvidas no alinhamento. Se diferentes sequências possuem diferente composição de bases, a pressuposição principal do modelo será violada.

O modelo HKY85, proposto por Hasegawa, Kishino e Yano, essencialmente mistura os modelos K2P e F81. Além de supor que a frequência das bases é variável, este modelo permite que transições e transversões ocorram com taxas diferentes.

Posteriormente, o modelo GTR (*generalised time-reversible*), o mais complexo dos modelos aqui apresentados, foi desenvolvido a partir do HKY85 com o intuito de acomodar diferentes taxas de substituição e diferentes frequências de bases. Este modelo requer seis parâmetros para taxa de substituição e quatro parâmetros para a frequência das bases, misturando todos os modelos aqui descritos.

Atualmente, além destes mais de 200 modelos de substituição podem ser aplicados a alinhamentos de nucleotídeos. Alguns programas, como Modeltest e Jmodeltest, são capazes de selecionar o modelo de substituição que melhor se ajusta a um dado alinhamento.

Uma importante extensão desses modelos de substituição incorpora a possibilidade de variação nas taxas evolutivas entre os sítios, permitindo ao modelo mais realismo. Assim, para cada sítio no DNA será atribuída uma probabilidade de evolução a uma taxa contida em um intervalo discreto de probabilidades. O método que garante a heterogeneidade de taxas evolutivas é modelado através de uma distribuição gama (Γ), que considera um número específico de taxas de

evolução para os sítios do DNA.

A aplicabilidade deste modelo nas inferências filogenéticas é facilitada pela simplicidade do método, já que apenas um único parâmetro (α) controla a forma da distribuição gama. Quando $\alpha < 1$, existe um grande número de taxas de evolução entre os sítios das sequências em análise, ou seja, quanto maior α, menor a heterogeneidade. Algumas vezes, uma proporção de sítios invariáveis (I), no qual uma determinada proporção de sítios é assumida como incapaz de sofrer substituição, pode também ser usada para modelar a heterogeneidade entre os sítios.

Ao contrário dos modelos de substituição de nucleotídeos, os modelos que explicam as trocas de aminoácidos são tradicionalmente empíricos. A partir da análise de alinhamentos de proteínas com identidade mínima de 85% Dayhoff, em 1970, desenvolveu uma série de matrizes de probabilidade que explicavam as mudanças de aminoácidos ao longo do tempo.

As matrizes PAM, como ficaram conhecidas, correspondem a modelos de evolução nos quais os aminoácidos são substituídos aleatoriamente e independentemente, de acordo com uma probabilidade predefinida que depende do próprio aminoácido.

Em 1992, um novo modelo de substituição de aminoácidos é criado por Henikoff e Henikoff. A análise de sequências de proteínas distantes evolutivamente, possibilitada pelo modelo de Henikoff-Henikoff, estabeleceu as bases para a criação das matrizes BLOSUM. As matrizes desta série foram identificadas por números (por exemplo, BLOSUM62) que se referem à porcentagem mínima de identidade dos blocos dos aminoácidos utilizados para construir o alinhamento. Matrizes similares, como GONNET e JTT, surgiram na mesma época.

Em 1996, foi proposto um modelo de substituição específico para proteínas codificadas pelo DNA mitocondrial, onde foi observado desvio de transições entre aminoácidos em relação às proteínas codificadas pelo material genético nuclear. Essa matriz, criada por Adachi e Hasegawa, foi chamada de mtREV.

Finalmente, em 2001, Whelan e Goldman propõem a matriz WAG, baseada em combinação e ampliação de vários modelos de substituição anteriores. Tal matriz é considerada superior às suas antecessoras para descrever filogenias de proteínas globulares.

5.5. Inferência filogenética

A reconstrução filogenética, ou seja, a reconstrução da história evolutiva de organismos, é um complexo processo que envolve uma série de etapas. O alinhamento, além de ser o primeiro passo, é um importante ponto para a inferência de filogenias (ver capítulo 3). Um alinhamento preciso, além de garantir maior confiabilidade nas análises posteriores, é requerido por todos os métodos de inferência filogenética para construção da árvore.

Depois que o alinhamento foi proposto, diversos métodos podem ser usados para estimar a filogenia das sequências estudadas. Podemos dividir estes métodos em dois principais grupos: métodos quantitativos e métodos qualitativos (Tabela 1-5). Estes grupos diferem na forma como os dados são tratados, refletindo diretamente como os dados do alinhamento serão inicialmente processados.

Os métodos quantitativos se baseiam na quantidade de diferenças entre as sequências do alinhamento para calcular uma árvore final. Já os métodos qualitativos constroem diversas filogenias que são classificadas seguindo uma determinada qualidade (critério). A filogenia que obtiver o maior valor associado à tal qualidade será a filogenia resultante.

Os métodos quantitativos compreendem os métodos de distância. Estes métodos convertem o alinhamento em matrizes de distância par-a-par para todas as sequências incluídas. Dentro destes algoritmos destacam-se dois métodos principais: UPGMA e aproximação dos vizinhos. Devido à grande eficiência computacional, estes métodos geralmente são utilizados para construção de uma filogenia inicial, que posteriormente é submetida a algum método do grupo qualitativo. Como principal ponto negativo, estes métodos apresentam apenas uma filogenia como resultado final (ver adiante).

Idealmente, todas as possíveis árvores para um dado alinhamento deveriam ser analisadas para garantir a escolha da melhor filogenia. Para isso, é necessário atribuir certos parâmetros que avaliem, dentre todas as ár-

Tabela 1-5: Comparação entre os tipos de métodos para inferência de filogenias.

Tipo	Método	Princípio	Programa
Métodos Quantitativos	UPGMA	Agrupa sequencialmente as OTUs com menor distância evolutiva entre si	Geneious MEGA
	Aproximação dos vizinhos	Busca a árvore com a menor soma total de ramos	MEGA Geneious HyPhy
Métodos Qualitativos	Máxima Parcimônia	Busca a filogenia com menor número de eventos evolutivos	PAUP MEGA Mesquite
	Máxima Verossimilhança	Busca a árvore com o valor de maior verossimilhança entre todas as filogenias construídas	PAUP PAML phyML MEGA
	Estatística Bayesiana	Amostra um número representativo de filogenias a partir do espaço amostral total de árvores e busca a mais provável	Mr. Bayes BEAST BAMBE

vores, aquela que explica as relações evolutivas de forma mais precisa.

Assim, os métodos qualitativos envolvem algoritmos que atribuem um critério de otimização para escolher a melhor filogenia. Nestes métodos, diversas filogenias são construídas e, seguindo um critério definido pelo algoritmo utilizado, uma filogenia será identificada como a que melhor explica a relação evolutiva entre os OTUs. O critério é utilizado para atribuir um valor a cada filogenia e ordená-las segundo este valor.

Estes métodos têm a vantagem de requerer uma função explícita para escolha das filogenias, sendo portanto independente da escolha do operador. No entanto, devido ao caráter de sua análise, são métodos mais refinados e intrinsecamente mais demorados computacionalmente. Três critérios de otimização são tradicionalmente empregados na inferência de filogenias: (a) Máxima Parcimônia, (b) Máxima Verossimilhança e (c) Inferência Bayesiana.

Por se tratarem de métodos que buscam uma única filogenia entre diversas árvores, os métodos qualitativos exigem algoritmos que vasculhem o maior número possível de filogenias em busca da melhor árvore. Dois grupos de algoritmos são destacados: os algoritmos exatos e os algoritmos heurísticos. Atualmente, devido ao tempo e à exigência computacional, os métodos heurísticos são preferidos aos exatos. No entanto, qualquer um deles pode ser aplicado aos métodos qualitativos de inferência filogenética. Como desvantagem dos métodos qualitativos, repetidos processos de procura em um mesmo conjunto de sequências podem levar a resultados diferentes, dependendo da árvore que é construída inicialmente pelo algoritmo.

Os métodos exatos buscam todas as filogenias possíveis para um grupo de sequências. O funcionamento destes métodos geralmente envolve a seleção aleatória inicial de três OTUs para a construção de uma árvore filogenética não enraizada. Por tentativa, um a um, novas OTUs, também tomadas aleatoriamente do alinhamento, são inseridas em diferentes posições na árvore. Esse procedimento é repetido até todos os táxons serem inseridos, garantindo que todas as filogenias possíveis para o alinhamento dado sejam geradas.

A partir da aplicação de um critério de otimização (dado pelo método qualitativo) para classificar as filogenias e ordená-las segundo este valor, é possível organizar um espaço virtual que contém todas as filogenias possíveis para o alinhamento empregado. É importante lembrar que, tomando poucas sequências, milhões de árvores podem ser geradas. Este conjunto total de filogenias é comumente chamado de espaço amostral. Como exemplo, podemos organizar o espaço amostral de filogenias originadas a partir de um alinhamento de dez sequências em um gráfico bidimensi-

onal baseado no valor atribuído pelo critério de otimização a cada árvore (Figura 9-5). Nestas condições, será possível observar que algumas árvores possuem valores maiores que outras, formando picos que agrupam as melhores filogenias. Da mesma forma, entre diferentes picos existem vales representados por árvores com valores menores e, portanto, menos consistentes.

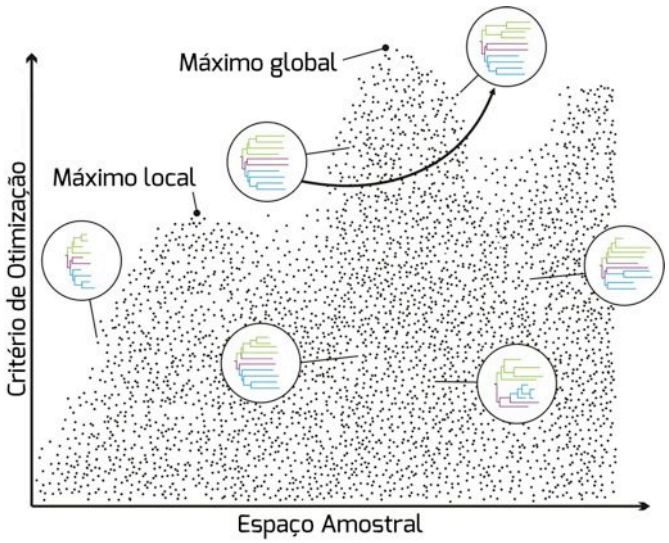


Figura 9-5: Descrição de parte do espaço amostral das possíveis filogenias para um determinado sistema, ordenadas segundo um valor atribuído pelo critério de otimização. Cada ponto no gráfico representa uma topologia diferente inferida a partir de um conjunto de dez sequências homólogas. O espaço amostral, neste caso, é definido por 2.027.025 filogenias e apresenta, segundo o critério de otimização, dois máximos locais e um máximo global, que contém as melhores filogenias. Em destaque, algumas filogenias exemplificando as possibilidades de arranjo dos ramos. A seta indica a mudança de topologia da filogenia e o consequente aumento de seu valor dado pelo critério de otimização.

Os métodos de busca exaustiva construirão um espaço amostral de árvores através de métodos específicos de modificação das filogenias. Por acumularem um grande número de resultados, estes métodos exigem um tempo computacional muito elevado, por vezes tornando-se proibitivos.

Os algoritmos de busca heurística procuram pela melhor filogenia em um subconjunto de todas as filogenias possíveis. Apesar de serem muito mais rápidos computacionalmente, estes métodos não garantem que a filogenia correta seja encontrada, pois apenas algumas árvores do espaço amostral total serão consideradas. Ainda assim, estes métodos tem mostrado grande eficiência.

Atualmente, os principais métodos qualitativos de inferência filogenética incorporam algoritmos de busca heurística para amostrar as filogenias do espaço amostral virtual. Usualmente, estes algoritmos de busca são executados em dois passos. Primeiramente, diferentes árvores são construídas e, após encontrar a melhor árvore guiada por um critério de otimização, aplica-se um algoritmo para modificar aleatoriamente o arranjo dos ramos. Este método permite testar se outros arranjos são ou não mais consistentes.

Devido ao grande número de métodos para inferência filogenética, a decisão quanto ao uso de cada um é de grande importância para a interpretação do resultado final: a filogenia. Ao escolher um método, é fundamental verificar o poder (tamanho e quantidade de sequências necessária para resolver a filogenia), a eficiência (habilidade de estimar a filogenia correta com um número limitado de dados), a consistência (habilidade de estimar a filogenia correta com um número de dados ilimitado) e a robustez (habilidade de estimar a filogenia correta quando certos pressupostos da análise são violados).

Até o momento, não existe um método que apresente todas estas características simultaneamente e garanta a reconstrução filogenética correta. É importante, sobretudo, conhecer a biologia do organismo (ou dos organismos) em questão para que a escolha do método tenha, além de tudo, uma justificativa biológica.

5.6. Abordagens quantitativas

UPGMA

O método baseado em distâncias UPGMA (*unweighted pair-group method using arithmetic averages*, ou método de agrupamento par a par usando médias aritméticas não ponderadas) foi proposto por Sneath e Sokal, em 1973, e é o método mais simples para reconstrução filogenética. O UPGMA

parte do pressuposto de que todas as linhagens evoluem a uma taxa constante (hipótese do relógio molecular).

No UPGMA, uma medida de distância evolutiva é computada para todos os pares de sequências utilizando um modelo evolutivo. Após, estas distâncias são organizadas na forma de uma matriz, conforme ilustrado abaixo:

Sequências	*1*	*2*	*3*	*4*
2	$d_{1,2}$			
3	$d_{1,3}$	$d_{2,3}$		
4	$d_{1,4}$	$d_{2,4}$	$d_{3,4}$	
5	$d_{1,5}$	$d_{2,5}$	$d_{3,5}$	$d_{4,5}$

O agrupamento das sequências é iniciado pelo par com menor distância. Supondo que $d_{1,2}$ seja a menor distância no exemplo acima, as sequências *1* e *2* são agrupadas com um ponto de ramificação na metade dessa distância ($d_{1,2/2}$). As sequências *1* e *2* são então combinadas em uma entidade composta, agora denominada *y*, e a distância entre esta entidade *y* e as outras sequências é computada (observe abaixo).

Sequências	$y_{(1,2)}$	*3*	*4*
3	$d_{y,3}$		
4	$d_{y,4}$	$d_{3,4}$	
5	$d_{y,5}$	$d_{3,5}$	$d_{4,5}$

Supondo que $d_{y,3}$ seja a menor distância, *y* e *3* são combinados em uma nova entidade composta, digamos, *z*. Seu ponto de ramificação é calculado levando em conta a distância de cada membro de *y* (*1* e *2*) em relação a *3* e dividindo por *2*, ou seja, $(d_{1,3}+d_{2,3})/2$. O mesmo procedimento se repete, calculando a menor distância entre *z* e outra sequência (suponhamos que seja a sequência *4*). Calculam-se a distância de cada membro de *z* até *4*, divide-se o somatório das distâncias por dois e cria-se uma nova sequência composta. O mesmo procedimento é repetido até que existam apenas duas sequências a serem agrupadas (comumente, uma sequência simples e uma entidade composta).

Ao empregar sequências de DNA ou proteína proximamente relacionadas, o UPGMA pode construir duas ou mais "árvores empatadas" (*tie trees*). Essas árvores surgem quando dois ou mais valores de distância na matriz se mostram idênticos. É possível representar todas as árvores empatadas, mas essa abordagem é pouco útil, uma vez que tais árvores são muito semelhantes e surgem por erros de estimativa das distâncias. Para tais casos, sugere-se apresentar uma única árvore, geralmente a árvore consenso do *bootstrap* (ver seção 5.8).

Por se basear na hipótese do relógio molecular, o UPGMA pode levar à obtenção de topologias falsas quando tal hipótese não for satisfeita pelos dados. Sabe-se que o método é muito sensível a variações nas taxas evolutivas entre linhagens, fato este que levou a proposição de métodos onde as variações são ajustadas para a obtenção de sequências que satisfaçam o relógio molecular. Apesar disso, devido ao surgimento de métodos mais robustos e mais eficientes em lidar com dados não uniformes, o UPGMA encontra-se praticamente abandonado como alternativa para reconstrução filogenética.

Aproximação dos Vizinhos

O método de aproximação dos vizinhos (*neighbor joining* ou NJ) foi proposto por Saitou e Nei em 1987. Este método se baseia em um aceleramento dos algoritmos de evolução mínima que existiam até então. Em sua versão original, estes algoritmos buscavam a árvore com menor soma total de ramos, de maneira que todas as árvores possíveis precisavam ser construídas para que se verificasse qual delas apresentava a menor soma. O algoritmo de NJ facilitou esse processo, tendo o princípio de evolução mínima implícito no processo e produzindo apenas uma árvore final.

Para construir a filogenia, o NJ começa por uma árvore totalmente não resolvida (topologia em estrela) (Figura 10-5). Tendo como base uma matriz de distâncias (semelhante à matriz inicial construída pelo método de UPGMA) entre todos os pares de sequências, construída a partir da aplicação de um modelo de substituição (conforme descrito na seção 5.4), o par que apresentar a menor distância é identificado, unido por um nó (que representará o ancestral comum deste par de sequências) e incorporado na árvore (na Figura 10-5, f e g são unidos pelo nó u). As distâncias de cada sequência do par são recalculadas em relação ao novo nó u, assim como as distâncias de todas as outras sequências são recalculadas em relação ao novo nó u. O algoritmo reinicia, substituindo o par de vizinhos unidos pelo novo nó e usando as distâncias calculadas no passo anterior.

Quando duas somatórias de ramos são iguais, a decisão sobre quais ramos unir depende do programa empregado. Alguns optam pela primeira sequência apresentada no arquivo de dados, enquanto outros escolhem aleatoriamente qual dos pares deve ser unido primeiro. Árvores empatadas (*tie trees*) são raras com o uso de NJ, e recomenda-se o emprego da árvore consenso do *bootstrap* (ver seção 5.8) para evitá-las. Uma variação do algoritmo NJ, o BIONJ tem se mostrado ligeiramente melhor que o NJ em casos pontuais; no entanto, conserva o mesmo princípio do algoritmo.

5.7. Abordagens qualitativas

Parcimônia

O princípio de parcimônia foi proposto por Guilherme de Occam (ou *William of Ockham*) no século XVII. Occam defendia que a natureza é por si só econômica e opta por caminhos mais simples. O pensamento se espalhou por diversas áreas do conhecimento e, atualmente, seu princípio é conhecido como Navalha de Occam.

Historicamente, a parcimônia teve um papel muito importante no estabelecimento da disciplina de filogenética molecular. Desde 1970, foi o critério de otimização mais utilizado para inferência de filogenias.

Contudo, atualmente a máxima parcimônia foi substituída por outros métodos, como máxima verossimilhança e inferência Bayesiana devido, principalmente, às simplificações nos processos evolutivos assumidas pelo método e, sobretudo, nas limitações de seu uso. Apesar disso, a máxima parcimônia ainda está integrada ao campo da inferência filogenética por ser um método rápido e, em alguns casos, muito efetivo.

A aplicação do princípio de máxima parcimônia nas reconstruções filogenéticas é conceitualmente simples: dentro de um conjunto de filogenias, aquela filogenia que apresentar o menor número de eventos evolutivos (substituições) deve ser a mais provável para explicar os dados do alinhamento.

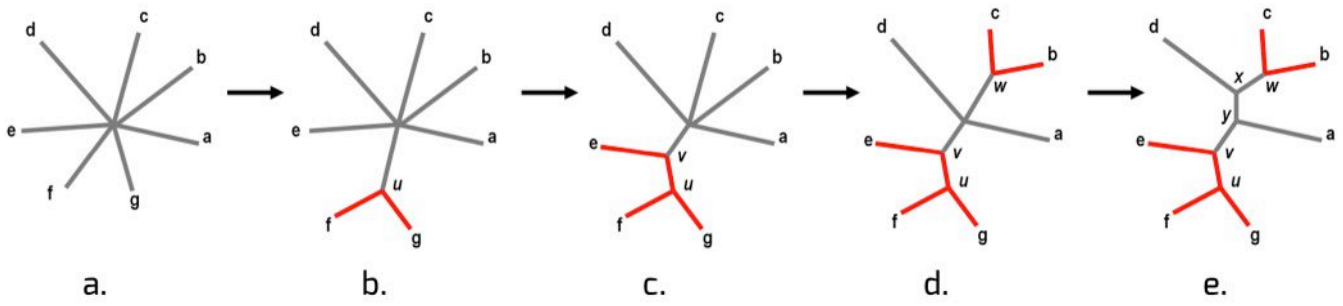


Figura 10-5: Começando com uma árvore em estrela (a), a matriz de distâncias é calculada para identificar o par de nós a ser unido (nesse caso, f e g). Estes são unidos ao novo nó u (b). A porção em vermelho é fixada e não será mais alterada. As distâncias do nó u até os nós a-e são calculadas e usadas para unir o próximo vizinho. No caso, u e e são unidos ao recém criado nó v (c). Mais duas etapas de cálculo levam à árvore em (d) e então à árvore em (e), que está totalmente resolvida, encerrando o algoritmo.

Metodologicamente, o critério de parcimônia deve determinar a quantidade total de mudanças na filogenia, descrevendo o tamanho dos ramos. Adicionalmente, a parcimônia guia a busca, entre todas as árvores possíveis, daquela filogenia que minimiza os passos evolutivos de forma máxima sendo, portanto, a filogenia de máxima parcimônia.

Assim que uma determinada filogenia é proposta, o método calculará as probabilidades de mudanças dos nucleotídeos desde os ramos terminais até os ramos mais ancestrais da árvore. Por se tratar de um método qualitativo, a parcimônia considera cada sítio do alinhamento individualmente e calcula as probabilidades de ocorrência dos quatro nucleotídeos nos táxons ancestrais.

Devido ao caráter probabilístico do método, é necessário que certas pressuposições sejam estabelecidas para especificar o custo de substituição dos nucleotídeos. A forma mais simples do método (Parcimônia de Wagner) assume que as substituições de nucleotídeos tem custo 1, enquanto que a não alteração não é penalizada (Figura 11-5a). No entanto, esquemas um pouco mais complexos que levam em consideração as questões biológicas envolvidas no processo evolutivo foram propostas. Um esquema comum de matriz com custo desigual, proposto para especificar as transições e as transversões, leva em consideração a diferença na probabilidade de mudança entre purinas e pirimidinas (Figura 11-5b). Comumente, a matriz é especificada sem que constem os respectivos nucleotídeos, no entanto, por convenção são atribuídos nas linhas e colunas em ordem alfabética (A, C, G e T).

Para o método de parcimônia, apenas sítios variáveis são considerados informativos. Estes sítios devem apresentar dois caracteres diferentes presentes em, no mínimo, dois indivíduos (Figura 12-5b). Aqueles sítios que não apresentam variação ou apresentam autapomorfias (caracter diferente presente em apenas um indivíduo) serão descartados automaticamente das análises.

Devido ao tamanho dos alinhamentos e ao número de OTUs incluídas para a inferência de filogenias, foi necessário que algoritmos fossem desenvolvidos para acelerar os cálculos na busca pela árvore de máxima parcimônia. Algoritmos de programação dinâmica são capazes de lidar com a atribuição de custos e realizar os devidos cálculos para escolha da filogenia com o menor custo. Diversos algoritmos foram desenvolvidos, embora a parcimônia de Sankoff, desenvolvida em 1975, tenha se tornado uma das mais populares.

a.

Matriz de custo igual =

	A	C	G	T
A	0	1	1	1
C	1	0	1	1
G	1	1	0	1
T	1	1	1	0

b.

Matriz de custo desigual =

	A	C	G	T
A	0	4	1	4
C	4	0	4	1
G	1	4	0	4
T	4	1	4	0

Figura 11-5: Matrizes de custo aplicadas ao método de máxima parcimônia para penalizar as substituições de um nucleotídeo por outro. (a) Matriz de custos iguais para todas as mudanças entre nucleotídeos. (b) Matriz de custo desigual, considerando a maior probabilidade de ocorrência de transições em relação às transversões ao longo do processo evolutivo.

Após a atribuição de uma matriz de custo e a proposição de uma filogenia, o algoritmo utilizará cada um dos sítios informativos do alinhamento independentemente para cálculo dos custos (Figura 11-5).

Considere a matriz desigual da Figura 11-5b e a filogenia inicialmente proposta na Figura 12-5a. O esquema demonstra que para cada sítio informativo será construída uma filogenia com a mesma topologia da árvore proposta em 12-5a (ver adiante).

Tomando, por exemplo, o sítio 28, identificamos a presença de três ancestrais não amostrados que, no entanto, para o cálculo dos custos, terão que ter seus caracteres inferidos. Segundo o algoritmo de Sankoff, os cálculos devem iniciar tomando os clados mais derivados (isto é, mais recentes). Em 12-

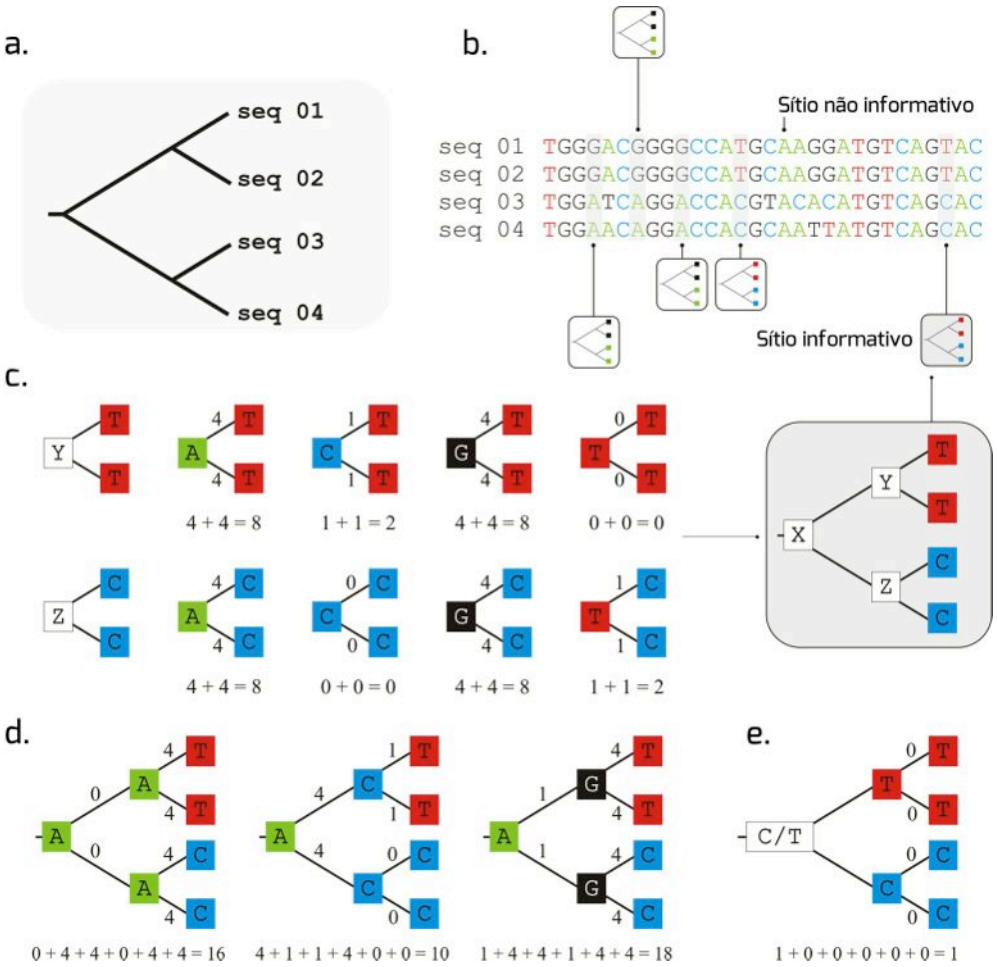


Figura 12-5: Determinação dos custos de substituição pelo método de parcimônia para um sítio do alinhamento de nucleotídeos. (a) Topologia da filogenia proposta para quatro táxons (ver adiante). (b) Alinhamento de nucleotídeos de quatro sequências homólogas. Destacados em cinza estão os sítios informativos para o método de parcimônia. Os demais sítios são considerados não informativos e serão descartados durante os cálculos. (c) Cálculo dos custos para os dois clados presentes na filogenia proposta em "a". O método supõe que a posição "Y" possa ser ocupada por qualquer um dos quatro nucleotídeos. (d) Exemplo do procedimento adotado pelo método, supondo que a posição "X" na filogenia foi ocupada pelo nucleotídeo A. É necessário considerar todas as possibilidades de caracteres nos sítios ancestrais e calcular os respectivos custos. (e) Arranjo de menor custo para a posição 28 do alinhamento de nucleotídeos.

5c, a posição "Y" da filogenia necessariamente foi ocupada por um dos quatro nucleotídeos. Em cada uma das proposições (A, C, G ou T), o custo associado à substituição é consultado na matriz. No primeiro caso, a hipótese para ocupação da posição "Y" é A. O custo da substituição em cada um dos ramos deve ser verificado e somado. Por exemplo, a substituição de A por T possui custo 4. Como a mesma substituição ocorreu em dois ramos diferentes, somamos o custo total, que totaliza 8. O mesmo procedimento será repetido considerando os outros três nucleotídeos na posição "Y".

Após o cálculo dos custos para as posições "Y" e "Z", é necessário verificar os custos de substituição de "X" para "Y" e "X" para "Z". A Figura 12-5d apresenta a primeira hipótese para ocupação da posição "X": o nucleotídeo A. Aqui, o algoritmo somará os custos de substituição de todos os ramos, novamente considerando cada um dos quatro

nucleotídeos na posição "X", mas também considerando a variação nas posições "Y" e "Z". A Figura 12-5e identifica a filogenia com o menor custo para o sítio 28. Note que o caractere mais ancestral pode ser tanto o nucleotídeo T quanto C. Os mesmos cálculos serão realizados para todos os sítios do alinhamento, tomando a topologia dada em 12-5a e, ao final, os menores custos para cada sítio serão somados para encontrar o tamanho dos ramos da árvore. A árvore que possuir os ramos mais parcimoniosos será tomada como a árvore de máxima parcimônia.

Computacionalmente, o cálculo dos tamanhos de ramos mais parcimoniosos não é um problema. O desafio da maioria dos métodos de reconstrução filogenética está na inferência da topologia. Assim como no método de máxima verossimilhança, discutido a seguir, o método de máxima parcimônia contará com algoritmos heurísticos para arranjo das topologias. A filogenia é então proposta pelo algoritmo, e o critério de parcimônia avalia a árvore. A partir de perturbações realizadas nesta topologia, uma nova topologia é proposta e novamente o critério qualifica a filogenia.

Apesar de velozes, os métodos de parcimônia falham ao estimar a relação evolutiva entre um grande número de táxons, especialmente se diferentes linhagens possuem taxas evolutivas variáveis ou taxas evolutivas muito rápidas. Nestes casos, é comum que o método agrupe incorretamente os táxons com maiores taxas de evolução, levando à inferência da filogenia errada (atração de ramos longos).

Ainda, por não ter um modelo de substituição especificado, o método de parcimônia é incapaz de considerar mutações reversas ou múltiplas substituições. Métodos que geram diferentes hipóteses a partir do alinhamento, considerando as observações biológicas na seleção do modo de substituição dos nucleotídeos e, assim, lidam com eventos aleatórios de probabilidade, substituíram o uso da máxima parcimônia e, atualmente, são os principais métodos utilizados para a inferência de filogenias.

Máxima Verossimilhança

Idealmente, os métodos de inferência filogenética devem resgatar o máximo de informações contidas em um dado conjunto de sequências homólogas, buscando desvendar a verdadeira história evolutiva dos organismos.

Quando um grande número de mudanças evolutivas em diferentes linhagens é demasiadamente desigual, o método de máxima parcimônia tende a inferir filogenias inconsistentes, proporcionalmente convergindo à árvore errada quanto maior o número de sequências no alinhamento. Assim, abre-se espaço para uma técnica de inferência filogenética mais robusta, que alie as informações do alinhamento a um modelo estatístico capaz de lidar com a probabilidade de mudança de um nucleotídeo para outro de maneira mais completa.

Dentro do campo da filogenética computacional, o método de máxima verossimilhança primeiramente ocupou este espaço e, desde então, tem sido amplamente utilizado devido à qualidade da abordagem estatística empregada.

A implementação de uma concepção estatística para a máxima verossimilhança, originalmente desenvolvida para estimar parâmetros desconhecidos em modelos probabilísticos, se deu entre 1912 e 1922 através dos trabalhos de A. R. Fisher.

Apesar de utilizado para dados moleculares na década de 1970, o método de máxima verossimilhança só se tornou popular na área da filogenética a partir de 1981, com o desenvolvimento de um algoritmo para estimar filogenias baseadas no alinhamento de nucleotídeos. Atualmente, diversos programas implementam este método para realizar a inferência filogenética, incluindo PAUP, MEGA, PHYLIP, fastDNAml, IQPNNI e METAPIGA, dentre outros (Tabela 1-5).

O objetivo principal do método da máxima verossimilhança é inferir a história evolutiva mais consistente com relação aos dados fornecidos pelo conjunto de sequências. Neste

modelo, a hipótese (topologia da árvore, modelo de substituição e comprimento dos ramos) é avaliada pela capacidade de predizer os dados observados (alinhamento de sequências homólogas). Sendo assim, a verossimilhança de uma árvore é proporcional à probabilidade de explicar os dados do alinhamento. Aquela árvore que com maior probabilidade, entre as outras árvores possíveis, produz o conjunto de sequências do alinhamento, é a árvore que reflete a história evolutiva mais próxima da realidade, mais verossímil e, por isso, de máxima verossimilhança.

É importante ressaltar que diferentes filogenias podem explicar um determinado conjunto de sequências, algumas com maior probabilidade e, outras, com menor probabilidade. No entanto, a soma das verossimilhanças de todas as árvores possíveis para um determinado conjunto de sequências nunca resultará em 1, pois não estamos lidando com as probabilidades de que estas filogenias estejam corretas, mas avaliando a probabilidade de explicarem o alinhamento que foi fornecido.

Se, por exemplo, aplicássemos o método de máxima verossimilhança para inferir a árvore filogenética de um grupo de sequências homólogas que incluem porções recombinantes, encontraríamos uma árvore filogenética com um determinado valor de verossimilhança. A utilização do método, por si só, garantiria como resultado a inferência de uma filogenia. No entanto, sabemos que esta árvore, apesar de ser a mais plausível para explicar o alinhamento dado, não tem qualquer relação com a realidade evolutiva do organismo, já que eventos de recombinação aconteceram no decorrer do tempo e impedem a explicação sob a forma dicotômica de uma filogenia.

A aplicação do método de máxima verossimilhança exige a construção de uma filogenia inicial, geralmente obtida por métodos quantitativos. Como exemplo, considere a árvore filogenética proposta inicialmente e o respectivo alinhamento de nucleotídeos da Figura 13-5. Para calcularmos a verossimilhança desta filogenia será necessário utilizar um modelo evolutivo, que será importante para atribuir valores e parâmetros às substituições e ajudará no cálculo da probabilidade de que uma sequência X mude para uma sequência Y ao longo de um segmento da árvore.

Dado um determinado modelo evolutivo (JC69, K2P, F81, HKY ou GTR, por exemplo), e assumindo que cada sítio do alinhamento evolui de maneira independente dos demais, podemos calcular o valor de verossimilhança para cada um destes sítios e, posteriormente, multiplicar os valores de cada sítio para encontrar a verossimilhança da árvore dada (Figura 13-5 e a Figura 14-5). Sítios que apresentam deleções serão eliminados da análise.

Como os nós internos destas árvores, geradas a partir de cada sítio do alinhamento, são a representação de OTUs não amostrados (isto é, ancestrais) e, por conseguinte, não se conhecem suas sequências de nucleotídeos, será necessário considerar a ocorrência de todos os nucleotídeos (A, T, C e G) nestas posições da árvore (Figura 13-5c).

Por certo, alguns cenários são mais prováveis que outros; no entanto, todos devem ser considerados durante os cálculos de verossimilhança, pois apresentam alguma probabilidade de terem gerado as sequências dadas no alinhamento. Adicionalmente, além de calcular a probabilidade de todas as mudanças possíveis para cada um dos sítios do alinhamento (Figura 13-5c), a expressão matemática da verossimilhança ainda incluirá o tamanho dos ramos, dentre outros elementos do modelo de substituição, como um fator determinante para o cálculo (Figura 13-5d).

A probabilidade de ocorrência de cada um dos quatro nucleotídeos no nó mais interno da árvore será igual à respectiva frequência estacionária dada pelo modelo de substituição, já que este parâmetro especifica a proporção esperada de cada um dos quatro nucleotídeos. No modelo de Jukes e Cantor, por exemplo, assume-se que os quatro nucleotídeos ocorrem em proporções iguais de 25%.

Conforme o exemplo da Figura 13-5d, a equação utilizada para calcular a verossimilhança da filogenia

proposta no sítio 28, inicialmente, leva em consideração a frequência estacionária do nucleotídeo G, já que este é o nucleotídeo que está sendo considerado como presente no nó mais ancestral da árvore. A probabilidade de este G ser substituído por um A (P_{GA}), ou permanecer G (P_{GG}) será dada pelo modelo de substituição escolhido. Da mesma forma, serão os casos P_{GT}, P_{AC} (repetido duas vezes cada pelo fato de existirem dois ramos terminais com o mesmo nucleotídeo).

O tamanho dos ramos entre dois nós será multiplicado pelas probabilidades de substituição dos nucleotídeos, levando em conta variações em parâmetros do modelo de substituição. Apesar da dificuldade de cálculo computacional, os algoritmos aplicados à inferência filogenética (baseados no princípio de Pulley) automaticamente estimarão o tamanho de cada ramo de modo que este maximize o valor da verossimilhança da árvore filogenética em construção. Nestes casos, o algoritmo atribui diversos valores de distância para um ramo e, a cada valor, verifica a verossimilhança da árvore, buscando aqueles valores que resultam na filogenia com a maior verossimilhança.

A probabilidade de observar os dados em um sítio particular é a soma das probabilidades de todos os possíveis nucleotídeos que poderiam ser observados nos nós internos da árvore (Figura 13-5c). O número de

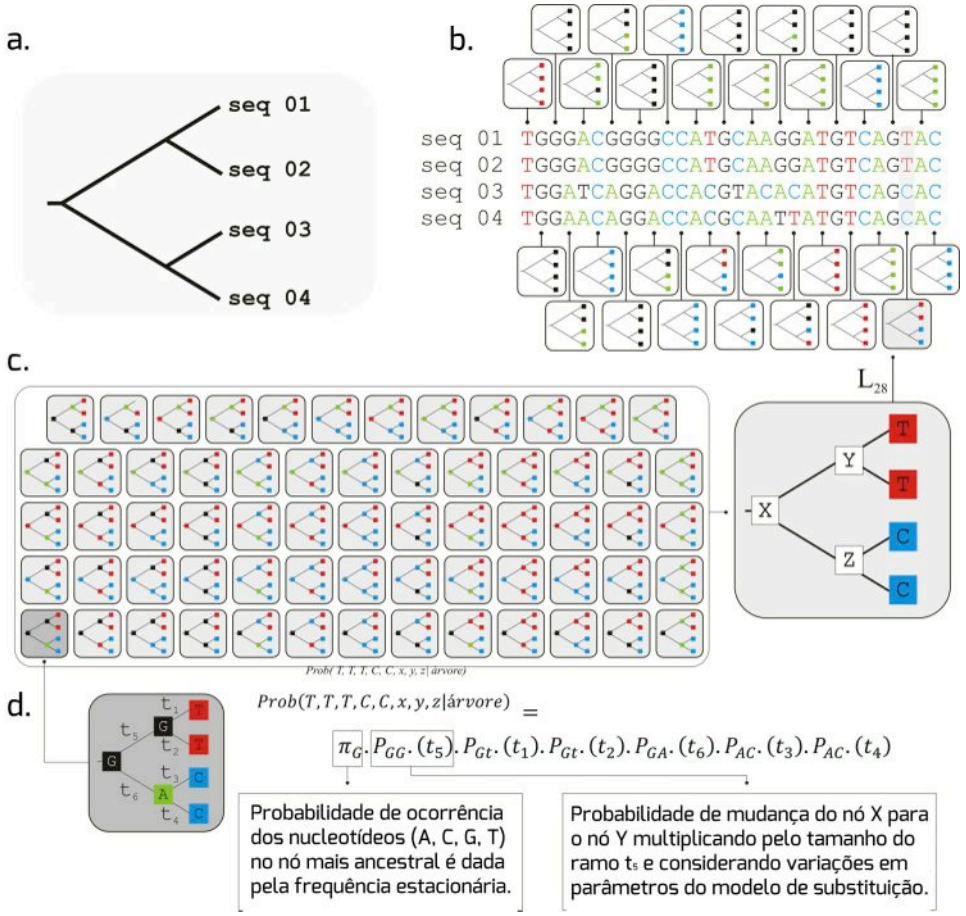


Figura 13-5: Esquema do cálculo da verossimilhança para uma filogenia e seu respectivo alinhamento de nucleotídeos. (a) Árvore filogenética proposta inicialmente para o alinhamento em "b". (b) Para cada posição do alinhamento é destacada a organização dos quatro sítios do alinhamento na árvore proposta em "a". Como exemplo, apenas o sítio do alinhamento destacado em cinza será considerado para o cálculo da verossimilhança. Os quadrados pretos, azuis, verdes e vermelhos nos ramos terminais das filogenias representam, respectivamente, os nucleotídeos guanina, citosina, adenina e timina. (c) Probabilidade de cada uma das 64 possíveis combinações de nucleotídeos nos nós internos da árvore, já que estes representam os sítios de táxons ancestrais não amostrados (P_{XY}, P_{YT}, P_{XZ}, P_{ZC}). (d) O esquema para o cálculo da máxima verossimilhança leva em conta a multiplicação do tamanho dos ramos (t_1, t_2, t_3, t_4, t_5 e t_6) pelas respectivas probabilidades de transição (P_{GG}, P_{GT}, P_{GA} e P_{AC}), além da frequência estacionária dos quatro nucleotídeos no nó mais ancestral (π_X).

nós internos rapidamente se torna muito grande com o aumento do número de OTUs. Felizmente, através de um algoritmo criado por Felseinstein (algoritmo de "poda"), que se aproveita da própria topologia da filogenia, esses cálculos podem ser realizados de uma maneira computacionalmente eficiente.

Neste processo, propõe-se que os cálculos da verossimilhança de uma determinada árvore sejam feitos a partir de sub-árvores dos ramos terminais em direção aos nós internos, semelhante ao algoritmo usado para o cálculo da parcimônia. No entanto, quando aplicado este método à inferência por máxima verossimilhança é necessário garantir que os modelos de substituição, não presentes no método de máxima parcimônia, sejam reversíveis, ou seja, que a probabilidade de mudança de A para T (P_{AT}) seja a mesma que T para A (P_{TA}). A introdução deste método permitiu que as análises de verossimilhança pudessem ser aplicadas a grandes conjuntos de sequências, de forma mais rápida e efetiva.

Ao final, multiplicamos os valores de verossimilhança de todos os sítios e encontramos o valor de verossimilhança da árvore (Figura 14-5):

A expressão matemática acima indica que a verossimilhança (L) é igual à multiplicação (Π) das probabilidades de cada sítio i (D^i, calculado conforme Figura 13-5), dada a árvore filogenética (topologia, modelo evolutivo e tamanho dos ramos). Aquela árvore que tiver o maior valor de verossimilhança entre todas as árvores possíveis para um determinado alinhamento de sequências será a árvore que melhor explica o alinhamento e, por isso, a árvore de máxima verossimilhança. Por fim, é importante ressaltar que, apesar de estarmos avaliando nucleotídeos neste exemplo, o mesmo raciocínio poderia ser aplicado para a inferência filogenética para um alinhamento de aminoácidos.

Até o momento vimos, em linhas gerais, como realizar o cálculo de verossimilhança para uma dada filogenia (Figura 13-5). No entanto, outra função importante dos métodos computacionais de inferência filogenética é apontar a topologia e encontrar a árvore de máxima verossimilhança entre todas as árvores possíveis para o conjunto de dados. Infelizmente, não existem algoritmos que garantam a localização da árvore real devido ao grande espaço amostral de árvores possíveis (Figura 9-5).

Após uma árvore ser construída, é necessário calcular sua verossimilhança e comparar este valor com todas as árvores já construídas. Como é impossível testar a verossimilhança para todas as filogenias possíveis, os algoritmos de máxima verossimilhança incluirão buscas heurísticas para solucionar este problema (estes métodos construirão diferentes filogenias a partir do mesmo conjunto de dados do alinhamento).

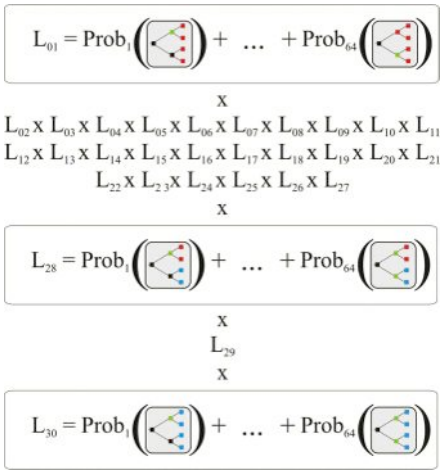


Figura 14-5: Cálculo da máxima verossimilhança de uma dada filogenia, considerando seu respectivo alinhamento de nucleotídeos contendo quatro táxons e 30 bases (Figura 13-5b). Para cada sítio (L_{01}, L_{02}, ..., L_{30}) será calculado um valor de probabilidade que envolve a consideração de todos os quatro nucleotídeos em cada um dos ramos ancestrais da filogenia. Posteriormente, os valores de verossimilhança de cada sítio serão multiplicados para encontrar a verossimilhança total da filogenia.

Na problemática das filogenias, diferentes programas têm proposto as mais diversas alternativas para avaliar o maior número de árvores do espaço amostral total e encontrar aquela com o maior valor de verossimilhança. No entanto, como regra geral, a maioria dos programas de máxima verossimilhança segue alguns passos comuns:

i) Uma filogenia preliminar com determinada topologia é construída (geralmente são utilizadas árvores construídas pelo método de aproxima-

ção de vizinhos);
ii) Os parâmetros para esta árvore são modificados buscando maximizar a verossimilhança (em alguns casos, a filogenia vai sendo construída pela adição de novos táxons aleatoriamente). Para a modificação da filogenia, os algoritmos podem implementar técnicas de rearranjos de ramos, conforme descrito em 5.4;
iii) O valor de máxima verossimilhança para esta árvore é armazenado;
iv) Outras topologias são construídas e seus parâmetros também são avaliados;
v) Finalmente, a filogenia que possuir o valor de máxima verossimilhança será a melhor estimativa evolutiva para o dado conjunto de sequências.

Embora estes processos simplifiquem os verdadeiros fenômenos biológicos que governam a evolução de uma sequência, apresentando assim dificuldades em identificar a árvore com o maior valor de verossimilhança, eles são normalmente robustos o bastante para estimar as relações evolutivas entre táxons.

Como estes métodos implicam em encontrar a árvore com o valor máximo de verossimilhança entre todas as árvores amostradas, o resultado final sempre fornecerá apenas uma filogenia, ao contrário dos métodos Bayesianos que serão vistos a seguir. Cabe ressaltar que, devido ao uso de diferentes algoritmos, na prática, um mesmo conjunto de sequências submetido a diferentes programas para inferência filogenética por máxima verossimilhança dificilmente resultará na mesma árvore. Por isso, é necessário ser cauteloso ao interpretar árvores geradas pelo método de máxima verossimilhança.

Análises Bayesianas

A estatística Bayesiana nasceu com a publicação de um ensaio matemático do reverendo Thomas Bayes, em 1793. Nesta publicação, o reverendo apresenta o desenvolvimento de um método formal para incorporar evidências prévias no cálculo da probabilidade de acontecimento de determinados eventos.

Inicialmente, este método foi aplicado apenas no campo da matemática e, só a partir de 1973, passa a ser incorporado no pensamento biológico e na inferência filogenética. Com o advento de diversos programas de acesso livre para realizar a inferência de filogenias por estatística Bayesiana, o método se difundiu e, atualmente, tornou-se um campo de estudo específico dentro da filogenética computacional.

A inferência Bayesiana engloba o método de máxima verossimilhança (Tabela 2-5) mas, adicionalmente, inclui o uso de informações dadas *a priori*. Estas informações refletem características a respeito da filogenia, do alinhamento ou dos táxons, que o pesquisador sabe de antemão.

Entre os principais parâmetros que podem ser conhecidos antes da reconstrução filogenética pode-se destacar a taxa evolutiva, tipo de relógio molecular, parâmetros do modelo de substituição, datas de coleta das amostras, datas para calibração da filogenia (achados fósseis, datação por carbono-14, aproximações arqueológicas, etc.), distribuição geográfica, organização monofilética de um grupo de indivíduos ou, até mesmo, parâmetros de dinâmica populacional.

Os valores atribuídos *a priori* são incorporados à estatística Bayesiana na forma de probabilidades e comporão o termo chamado de probabilidade anterior (*prior probability*). Se sabemos de antemão que um determinado grupo de organismos é ancestral em relação a outro, podemos atribuir uma maior probabilidade àquelas filogenias que relacionam estes organismos da maneira como sabemos *a priori*.

Qualquer informação útil, que é fornecida pelo pesquisador antes da própria reconstrução da filogenia, poderá ser convertida em uma probabilidade anterior para ser inserida nas análises de inferência Bayesiana. No entanto, as informações cedidas *a priori* devem

Tabela 2-5: Comparação entre os métodos de máxima verossimilhança e inferência Bayesiana.

Método	Vantagens	Desvantagens
Máxima Verossimilhança	Captura totalmente a informação dos sítios do alinhamento para construção das filogenias	Comparativamente ao método Bayesiano, o algoritmo para reconstrução por máxima verossimilhança é mais lento
Estatística Bayesiana	Tem grande ligação com a máxima verossimilhança, sendo, no entanto, geralmente mais rápida. Modelos populacionais podem ser incluídos para inferência das filogenias	Os parâmetros para as probabilidades anteriores devem ser especificados e pode ser difícil especificar quando as análises são satisfatórias

ser distribuições de números prováveis (mínimo e máximo), e não números exatos. Quando estes valores não são conhecidos ou quando, por exemplo, não se quer atribuir maior probabilidade a uma determinada topologia, o parâmetro terá uma distribuição uniforme de probabilidades.

Na maioria dos aplicativos que lidam com inferência Bayesiana existem distribuições uniformes associadas às probabilidades anteriores que assumem que todos os valores possíveis são dados pela mesma probabilidade.

Além das probabilidades anteriores, a inferência Bayesiana é baseada nas probabilidades posteriores de um parâmetro como, por exemplo, a topologia. Através da probabilidade posterior é possível verificar a probabilidade de cada uma das hipóteses (árvores filogenéticas). Sendo assim, ao final das análises, é possível estabelecer uma estimativa da probabilidade dos eventos retratados por uma determinada filogenia, ou seja, a probabilidade de cada filogenia. As probabilidades posteriores são calculadas utilizando a fórmula de Bayes:

$$\mathrm{L}(H \mid D) = \frac{\mathrm{L}(H)\,\mathrm{L}(D \mid H)}{\mathrm{L}(D)}$$

O termo $\mathrm{L}(H \mid D)$ é chamado de distribuição de probabilidades posteriores, e é dado pela probabilidade da hipótese (topologia da árvore, modelo de substituição e comprimento dos ramos) a partir dos dados disponíveis (alinhamento de sequências). O termo $\mathrm{L}(D \mid H)$ descreve o cálculo de máxima verossimilhança, enquanto o multiplicador $\mathrm{L}(H)$ é a probabilidade anterior. Para o termo que envolve a função de máxima verossimilhança, é ainda necessário considerar também todos os tópicos já discutidos na seção anterior. O denominador $\mathrm{L}(D)$ é uma integração sobre todas as possibilidades de topologias, tamanhos de ramo e valores para os parâmetros do modelo evolutivo, o que garante que a soma da probabilidade posterior para todos eles seja 1. O denominador atuará como um normalizador para o numerador. Reescrevendo, temos:

$$\mathrm{L}(filogenia|alinhamento) = \frac{\mathrm{L}(filogenia)\,\mathrm{L}(alinhamento|filogenia)}{\sum_{\mathrm{H}} \mathrm{L}(filogenia)\,\mathrm{L}(alnhamento|filogenia)}$$

onde o termo filogenia descreve a topologia da árvore, o modelo de substituição e o comprimento dos ramos. Assim, através da multiplicação das probabilidades anteriores pela verossimilhança, divididos pelo fator de normalização, o método busca a hipótese (topologia da árvore, o modelo de substituição e o comprimento dos ramos) em que a probabilidade posterior é máxima.

O objetivo da inferência Bayesiana é calcular a probabilidade posterior para cada filogenia proposta. No entanto, para cada árvore diversos parâmetros devem ser especificados pelo usuário, incluindo topologia, tamanho dos ramos, parâmetros do modelo de substituição, parâmetros populacionais, relógio molecular, taxa evolutiva, etc. Dada uma filogenia, todos os parâmetros terão sua probabilidade posterior calculada. Se dadas 1000 filogenias, teremos 1000 valores de probabilidade posterior para cada parâmetro.

Devido à impossibilidade de construção de todas as filogenias possíveis para a maioria dos alinhamentos, a análise Bayesiana se aproveita de técnicas de amostragem para estimar os valores esperados de cada parâmetro.

Neste sentido, os métodos de inferência

Bayesiana utilizam as Cadeias de Markov Monte Carlo (MCMC, *Monte Carlo Markov Chain*) para aproximar as distribuições probabilísticas em uma grande variedade de contextos. Esta abordagem permite realizar amostragens a partir do conjunto total de filogenias, relacionando cada filogenia a um valor probabilístico. Sem a aplicação de um método que obtenha amostras do espaço de possíveis filogenias, como o modelo de MCMC, a estimativa de todos os parâmetros se tornaria analiticamente impossível nos atuais computadores.

Um dos métodos de MCMC mais usados na inferência filogenética é uma modificação do algoritmo Metropolis, chamado de Metropolis-Hastings. A ideia central deste método é causar pequenas mudanças em uma filogenia (topologia, tamanho dos ramos, parâmetros do modelo de substituição, etc.) e, após a modificação, aceitar ou rejeitar a nova hipótese de acordo com o cálculo de razão das probabilidades. Este método garante que diversas árvores sejam amostradas do espaço total de filogenias, amostrando filogenias com probabilidade posterior mais alta (Figura 15-5):

i) Inicialmente, o algoritmo MCMC gera uma filogenia aleatória X, arbitrariamente escolhendo o tamanho dos ramos para dar início à cadeia;

ii) O valor de probabilidade associado a esta filogenia é calculado (probabilidade posterior calculada através da fórmula de Bayes);

iii) Perturbações aleatórias são realizadas nesta filogenia inicial X (mudanças na topologia, no tamanho dos ramos, nos parâmetros do modelo de substituição, etc.) e geram uma filogenia Y;

iv) A probabilidade posterior é calculada para a filogenia Y;

v) A filogenia Y é tomada ou rejeitada para o próximo passo baseado na razão R (probabilidade posterior de Y dividida pela probabilidade posterior de X). Se R é maior que 1, a filogenia Y é tomada como base para o próximo passo. Se R é menor que 1, um número entre 0 e 1 é tomado aleatoriamente. Se R é maior que o número aleatório gerado, a filogenia será tomada, no entanto se for menor, a filogenia Y é rejeitada;

vi) Se a nova proposta Y for rejeitada, retorna-se ao estado X e novas modificações serão realizadas nesta filogenia;

vii) Supondo que a proposta Y tenha sido aceita, ela sofrerá uma nova perturbação a fim de gerar uma nova filogenia;

viii) Todas as árvores amostradas são armazenadas para posterior comparação. Os pontos visitados formam uma

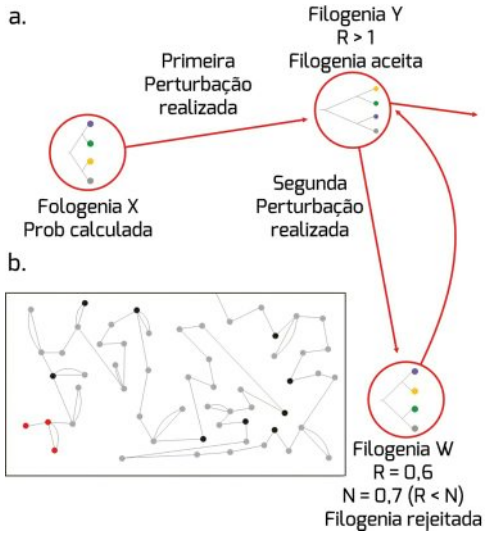


Figura 15-5: Esquema de amostragens MCMC aplicada à inferência filogenética pelo método Bayesiano utilizando o algoritmo de Metropolis-Hastings. (a) Após a proposição de uma filogenia inicial X, perturbações aleatórias são realizadas para gerar a filogenia Y. Devido à razão R>1, a nova filogenia é aceita. Nova perturbação é realizada para gerar a filogenia W e, devido a razão de probabilidades R resultar em um número menor que 1, um número aleatório N é sorteado. Sendo R<N, a nova proposição é rejeitada e a cadeia retorna à filogenia Y. (b) Andamento da cadeia na amostragem de filogenias. Cada círculo destaca uma nova filogenia que é proposta após a perturbação. As linhas conectando os círculos evidenciam a direção do andamento da cadeia. Apesar de a cadeia percorrer muitos passos, apenas alguns serão registrados para análise final (círculos pretos). Os círculos em vermelho são aqueles evidenciados em (a).

espécie de cadeia ao longo do espaço amostral total de filogenias.

O principal objetivo da cadeia é amostrar filogenias com probabilidades crescentes. No entanto, é importante que o algoritmo utilizado para tal permita que algumas árvores com menor probabilidade sejam amostradas para evitar que a cadeia fique "presa" em picos de máximo local (Figura 9-5).

Sendo assim, o cálculo da razão R considerando um valor aleatório entre 0 e 1 garantirá que, em determinados momentos, uma filogenia com menor probabilidade seja aceita. Por este método, é possível amostrar filogenias da região de um vale passando, por exemplo, de um pico de ótimo local para o pico de ótimo global (Figura 9-5).

A proposta de novas árvores na cadeia de Markov é uma etapa crucial para uma boa amostragem de filogenias. Na abordagem Bayesiana, uma boa amostragem inclui um grande número de filogenias, suficientemente diferentes entre si. Se filogenias muito diferentes são propostas, serão rejeitadas com muita frequência, pois é provável que tenham menor probabilidade posterior. Pelo contrário, se filogenias muito similares forem geradas, o espaço amostral não será varrido adequadamente e a cadeia deverá "correr" por muitos passos (amostrar um maior número de filogenias), aumentando o tamanho da cadeia e o tempo computacional.

Estimar o quanto a cadeia deve percorrer para amostrar um número suficiente de filogenias para as sequências dadas (espaço de árvores) é um fator fundamental para obter bons resultados em uma análise Bayesiana. Na maioria dos programas que utilizam estatística Bayesiana para inferir filogenias, o usuário deve especificar o tamanho da cadeia. Esse número é de grande subjetividade, e depende diretamente da distribuição das probabilidades anteriores, do número de táxons incluídos na filogenia e da relação evolutiva entre eles.

A Figura 16-5 exemplifica o andamento da amostragem da MCMC em um espaço de filogenias. Supondo que os quadrados em a, b e c representam um espaço amostral de filogenias, semelhante ao apresentado na Figura 15-5b, e que os pontos pretos sejam as filogenias que vão sendo amostradas com o desenvolvimento da MCMC vemos que, ao final do processo, depois de empregados 100 mil passos (Figura 16-5c), um grande número de filogenias foi amostrado.

Ainda, na região delimitada por um círculo, assumimos que estão as filogenias com maior probabilidade de explicar a história evolutiva de um grupo de organismos, ou seja, as filogenias reais. Note que quanto maior o número de passos percorridos pela cadeia, maior a amostragem do espaço de filogenias e maior o número de amostras dentro da região com filogenias de alta probabilidade.

Ao final, após o término da cadeia, a distribuição das probabilidades posteriores de todos os parâmetros deve ser verificada. No

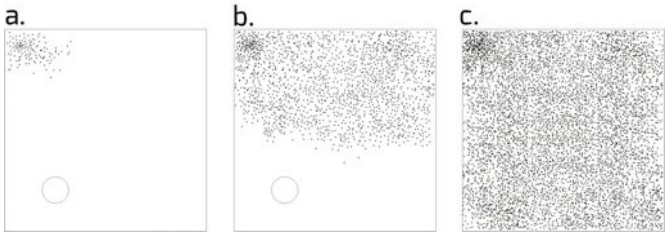


Figura 16-5: Espaço de possíveis árvores analisadas pela MCMC. Considerando que os quadrados descrevem o espaço amostral de todas as filogenias possíveis para um dado conjunto de sequencias, os pontos pretos representam as filogenias que foram amostradas ao longo da cadeia. Os círculos presentes no canto esquerdo inferior representam a região de máximo global (isto é, maior probabilidade) neste espaço amostral. O andamento da cadeia neste exemplo é o mesmo apresentado na Figura 15-5b (a) cento e trinta passos percorridos pela cadeia; (b) trinta mil passos percorridos pela cadeia; (c) cem mil passos percorridos pela cadeia. Nota-se que quanto maior o número de passos percorridos, maior a amostragem de filogenias no espaço. Da mesma forma, aumenta a probabilidade de a cadeia amostrar aquelas filogenias de máximo global.

entanto, as amostras tomadas no início da cadeia são tipicamente descartadas, pois estão sob forte influência do local de início da cadeia. As filogenias do início da cadeia estão muito longe de pontos máximos no espaço amostral e, por isso, é provável que todas as novas filogenias sugeridas subsequentemente sejam tomadas para o próximo passo (qualquer árvore proposta será mais provável que as árvores iniciais semelhantes àquela gerada aleatoriamente).

Esta fase inicial é conhecida como período de *burn in* (Figura 17-5). Conforme a cadeia avança, espera-se que a probabilidade das árvores amostradas aumente e, quando um número suficiente de filogenias for amostrado, chegue a uma distribuição estacionária. Em termos Bayesianos, espera-se que a cadeia atinja a convergência.

Um dos primeiros indicativos de que a cadeia convergiu para a distribuição correta está na estabilidade dos valores de probabilidade dos parâmetros da cadeia (cada parâmetro da filogenia poderá ter uma distribuição independente). Portanto, a representação gráfica dos valores das probabilidades e dos respectivos passos da cadeia (*trace plot*) é uma importante ferramenta para monitorar o desempenho da MCMC (Figura 17-5).

Devido ao aumento brusco de probabilidade das filogenias que são visitadas pelo andamento da cadeia, os gráficos necessariamente incluirão os valores medidos em escala logarítmica (ln L, Figura 17-5). Em estatística Bayesiana, é comum que seja atribuído um intervalo de credibilidade de 95% para os parâmetros amostrados. Estes valores são obtidos através da eliminação de 2,5% dos valores mais baixos e de 2,5% dos valores mais altos para um determinado parâmetro. Um intervalo de credibilidade contém o valor correto com 95% de probabilidade; no entanto, não se trata de um intervalo de confiança.

Adicionalmente, outros métodos são úteis para diagnosticar a convergência da cadeia, tais como o exame do tamanho amostral efetivo (ESS) e a comparação de amostras resultantes de diferentes cadeias (várias cadeias de MCMC são aplicadas para o mesmo conjunto

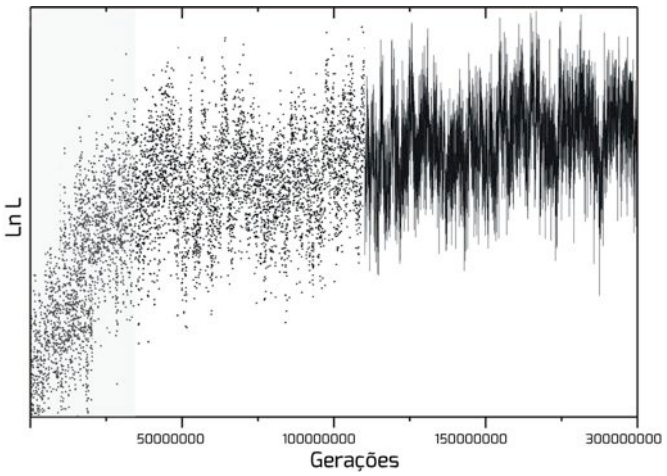


Figura 17-5: Representação gráfica das probabilidades das filogenias na cadeia ao longo de 300 milhões de amostragens. O esquema demonstra duas visualizações possíveis: à esquerda, são mostrados apenas os pontos referentes às amostras tomadas ao longo da cadeia e, à direita, as amostragens sucessivas são ligadas umas as outras para facilitar a visualização do comportamento da cadeia. Em cinza, a fase inicial de *burn in* da Cadeia de Markov Monte Carlo.

de dados). Apesar de ser computacionalmente intensiva, a última alternativa parece ser a mais confiável para verificar a convergência. Contudo, o exame de ESS é, ainda hoje, o método mais utilizado. O tamanho amostral efetivo é uma estimativa para verificar o número de amostras independentes existentes na cadeia, ou seja, quantas amostras não similares foram tomadas. Atualmente, um ESS maior que 200 é um indicativo de que a cadeia convergiu adequadamente.

A técnica de *Metropolis Coupling*, conhecida como MCMCMC ou $(MC)^3$, através da introdução da corrida simultânea de duas cadeias, pode ajudar na amostragem de máximos globais e beneficiar na convergência da cadeia. Nesta técnica uma cadeia, chamada de quente (*hot chain*), permite aproximar os valores de máxima e mínima probabilidade das amostras para que a cadeia possa, de forma mais rápida, "saltar" entre picos de probabilidade, especialmente de máximos locais para máximos globais. O aquecimento da cadeia é dado pelo parâmetro β e visa diminuir a altura dos picos locais no espaço amostral. Uma segunda cadeia simultânea, chamada de fria (*cold chain*), utiliza as informações destes saltos da cadeia quente para melhorar a sua

amostragem e garantir a convergência.

Os métodos Bayesianos de inferência filogenética ainda têm a vantagem de aplicar modelos que envolvem diferentes tipos de relógios moleculares.

As distâncias genéticas, depois de "tratadas" pelos modelos de substituição, não tem qualquer significado sozinhas quando se deseja estimar, por exemplo, a idade do ancestral comum mais recente de duas OTUs. Esta e outras questões podem ser avaliadas quando aplicamos uma medida de tempo nas inferências, a fim de calibrar as taxas evolutivas. Sequenciamentos de amostras isoladas em diferentes épocas podem fornecer a calibração adequada para inferências temporais, pois se assume uma taxa evolutiva constante ao longo de um tempo t para todos os ramos de uma filogenia (relógio molecular estrito).

As taxas evolutivas dependem de diversos fatores e podem variar, nem sempre seguindo a constância proposta por este modelo. Após a introdução de um tipo específico de relógio molecular relaxado, as taxas de evolução podem variar ao longo da árvore para diferentes grupos e não são correlacionadas, ou seja, grupos evolutivamente próximos não necessariamente terão taxas de evolução semelhantes (relógio molecular relaxado não correlacionado).

Complexos modelos de dinâmica populacional podem ser analisados sob uma perspectiva Bayesiana. Quando o conjunto de sequências submetido às analises são isolados de uma população homogênea, os parâmetros de história demográfica podem ser usados para modelar as mudanças populacionais ao longo do tempo. Desta forma, através da estatística Bayesiana é possível, além da inferência filogenética, refinar as análises e datar filogenias e ramos específicos (Figura 18-5), inferir caracteres ancestrais e analisar a dinâmica populacional sob uma ótica evolutiva.

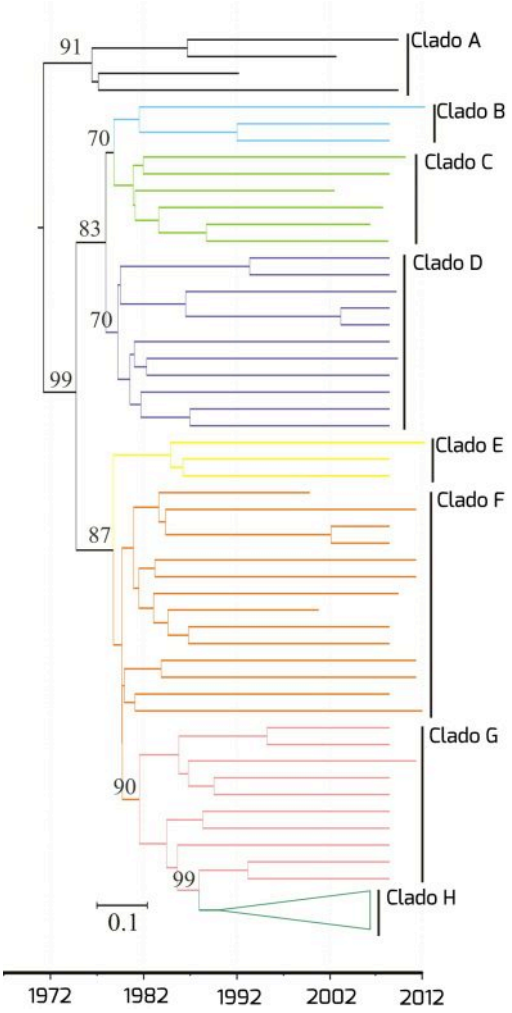


Figura 18-5: Árvore filogenética consenso gerada por inferência Bayesiana para 70 sequências de nucleotídeos. As cores nos ramos representam diferentes clados (B-H). O grupo externo está identificado como clado A. O Clado H foi agrupado para facilitar a representação. Nos nós estão especificados os valores de probabilidade posterior acima de 70. Abaixo, é apresentada a escala temporal inferida a partir da utilização de um relógio molecular relaxado.

5.8. Confiabilidade

O papel principal das técnicas de inferência filogenética é desvendar as relações evolutivas reais através de dados moleculares, buscando garantir que esta reconstrução seja fidedigna. Além da inferência das relações evolutivas entre os táxons, é igualmente importante que a filogenia possua precisão. Esta característica está relacionada ao número de filogenias que podem ser excluídas, a partir do conjunto total de filogenias, por não serem "verdadeiras". Quanto maior o número de filogenias excluídas neste processo, mais preciso é o método.

Em geral, na maioria dos casos de reconstrução filogenética, a falta de precisão das filogenias está relacionada ao conjunto de dados que está sendo fornecido no alinha-

mento. O gene considerado, o tamanho das sequências, o número de indivíduos e o grupo externo são atribuições fundamentais para uma reconstrução filogenética precisa e dependem, especialmente, do objetivo do estudo e da própria disponibilidade de informação.

Em muitos casos, o pesquisador é ainda dependente do número de amostras e do sucesso de coleta em campo, sobretudo, quando seu objeto de estudo se trata de uma espécie rara ou de indivíduos de difícil amostragem. No entanto, apesar de toda a informação relacionada ao conjunto de dados, a dificuldade de amostragem de indivíduos parece ser, sem dúvida, o principal problema relacionado a precisão das filogenias, pois a falta de dados de variabilidade genética compromete a inferência de história evolutiva coerente.

Como é possível saber se a amostragem foi suficiente e a filogenia é confiável? Usualmente, a resposta para esta questão consiste na reamostragem de dados. Se novas amostras forem tomadas e a mesma filogenia for reproduzida, a filogenia proposta tem seu valor reforçado. No entanto, na maioria dos casos, a reamostragem de dados da forma usual (coletas de novos espécimes, reamostragens em campo, achado fóssil diferente, etc) não é factível. Assim, algoritmos que produzem diferentes amostragens utilizando o mesmo conjunto de dados foram desenvolvidos para possibilitar a verificação da confiabilidade nos clados das filogenias. Destaca-se entre estes algoritmos o método de *bootstrap*.

Bootstrap é um método de reamostragem utilizado para realizar comparações da variabilidade das hipóteses filogenéticas, oferecendo medidas de confiabilidade aos clados propostos. A reamostragem é realizada a partir do mesmo conjunto de dados, e novas amostras fictícias com o mesmo tamanho serão geradas.

Segundo este método, cada sítio do alinhamento será tratado de forma independente. Conforme a Figura 19-5, inicialmente o algoritmo reconstruirá a filogenia a partir do alinhamento dado e, posteriormente, diversas replicatas serão reconstruídas. As colunas, representando os sítios do alinhamento, serão aleatoriamente tomadas (amostradas) pelo algoritmo e, em seguida, serão agrupadas uma ao lado da outra de maneira a formar um novo alinhamento (com o mesmo número de sítios do alinhamento original, Figura 19-5).

Por este método, é possível que um mesmo sítio seja amostrado mais de uma vez e, portanto, alguns sítios não serão selecionados para o novo alinhamento. Um número fornecido pelo usuário especificará o número de pseudoreplicatas (novos alinhamentos) que serão construídas. Assim que uma pseudoreplicata for criada, o algoritmo constrói a filogenia correspondente.

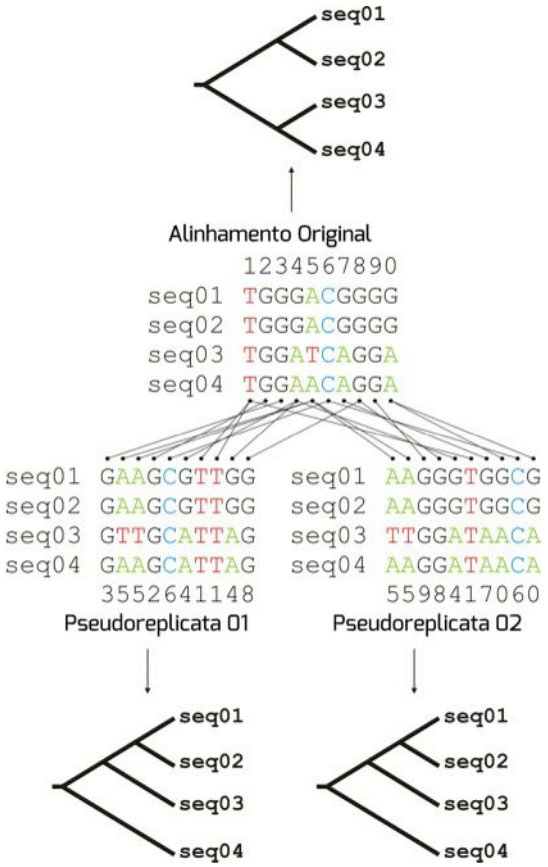


Figura 19-5: Método de *boostrap* para filogenias. A partir do alinhamento original, as colunas que representam os sítios serão aleatoriamente amostradas para construir pseudoreplicatas (um mesmo sítio pode ser sorteado diversas vezes). Estas, por sua vez, serão utilizadas para a inferência de filogenias, da mesma forma que o alinhamento original.

É importante ressaltar que a inferência destas filogenias será realizada pelo método de construção especificado pelo usuário, seja aproximação de vizinhos, máxima parcimônia ou máxima verossimilhança (para árvores bayesianas, veja adiante). Ao final, o algoritmo analisará os clados e automaticamente verificará a presença de determinados agrupamentos em todas as filogenias construídas. Se, por exemplo, encontramos as sequências 1 e 2 formando um clado em 70% das filogenias construídas, atribuiremos a confiabilidade de 70 ao clado formado por estas duas sequências. Comumente, o valor de confiabilidade dos clados é colocado próximo ao ancestral comum do clado (Figura 18-5).

A partir dos resultados de confiabilidade dos clados é possível também construir filogenias baseando-se na árvore consenso gerada pela regra da maioria (*majority-rule consensus tree*). Neste método, o algoritmo tabulará todos os clados formados em todas as replicatas geradas. Aqueles clados que mais aparecerem servirão para montar a filogenia consenso.

Ao contrário dos métodos de aproximação de vizinhos, máxima parcimônia e máxima verossimilhança, a confiabilidade de filogenias construídas através de estatística Bayesiana é inerente ao processo. Como diversas filogenias são amostradas ao longo do desempenho da Cadeia de Markov, não é necessário nenhum método para simular reamostragens do mesmo conjunto de dados. As amostras serão resumidas a partir da distribuição posterior de filogenias como frequência de clados individuais e serão identificadas por um número próximo ao ancestral comum daqueles clados (Figura 18-5). Portanto, o valor de probabilidade posterior de um clado representa uma inferência a respeito da probabilidade daquele clado.

A comparação dos valores de *bootstrap* e de probabilidade posterior dos clados para filogenias construídas a partir do mesmo alinhamento utilizando máxima verossimilhança e o método Bayesiano, respectivamente, leva a conclusão de que o método Bayesiano superestima a confiança aos clados. A confiança atribuída pela probabilidade posterior é geralmente maior que aquela atribuída pelo método de *bootstrap*. Por isso, enquanto uma confiança acima de 70 é considerada sustentada para o *bootstrap*, apenas valores acima de 90 podem ser considerados relevantes para os métodos Bayesianos.

5.9. Interpretação de filogenias

Árvores filogenéticas são diagramas que denotam a história evolutiva de diferentes OTUs a partir de seu ancestral comum. Mais do que isso, as filogenias moleculares são ferramentas que ajudam no entendimento dos diversos processos evolutivos que moldam o genoma dos organismos. Desta forma, a interpretação das implicações evolutivas associadas a um, ou a um conjunto de táxons, está diretamente relacionada à disposição dos ramos internos e externos de uma árvore. Independentemente do método de inferência, ou da forma como a árvore é apresentada, a interpretação dos resultados será baseada nos mesmos pressupostos, ainda que métodos diferentes possam originar filogenias diferentes.

Inicialmente, é necessário observar a presença de uma raiz. Como já discutido, o método de enraizamento pelo grupo externo é o mais comum e utiliza organismos sabidamente relacionados ao grupo em evidência, servindo para orientar o algoritmo em relação às características mais ancestrais do grupo. O grupo externo ajudará a evidenciar o tempo evolutivo. Na Figura 20-5, por exemplo, o grupo externo é dado pelo orangotango, pois este compartilha o mesmo ancestral comum que o restante do grupo. No caso de filogenias sem raiz, é necessário ter cautela nas interpretações, pois este tipo de diagrama apenas revela a relação entre os táxons.

Depois de encontrada a raiz da filogenia, é preciso avaliar os ramos. Dependendo do método, os ramos podem ter significados diferentes. Na Figura 18-5, os ramos evidenciam o tempo real, apresentando OTUs amostradas no passado. Pelo contrário, na Figura 20-5, os ramos evidenciam apenas um

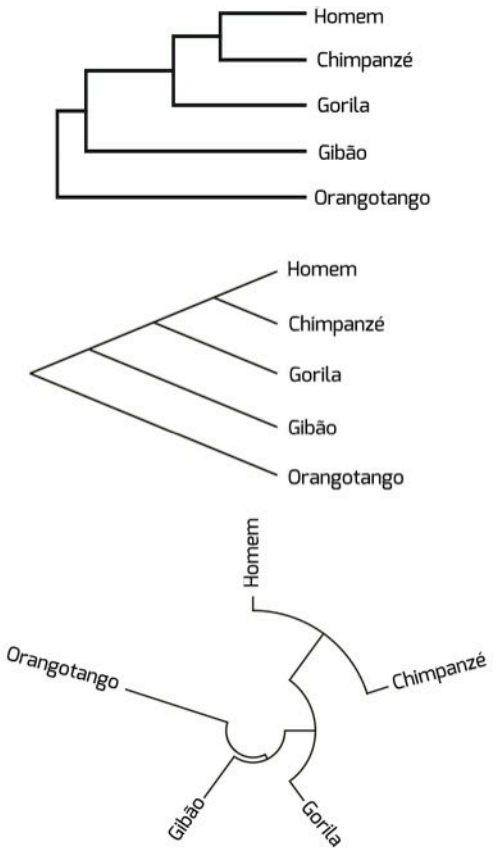


Figura 20-5: Diferentes representações da filogenia dos primatas.

tempo evolutivo representado pelo número de modificações genômicas, desde o organismo ancestral até os ramos terminais. Além disso, deve-se perceber a escala na qual os ramos foram representados, pois estes indicam o número de substituições que provavelmente ocorreram ao longo do processo evolutivo e podem ajudar na interpretação das taxas evolutivas.

Conclusões evolutivas baseadas em árvores filogenéticas devem ser sustentadas em árvore confiáveis e, por isso, a medida de confiabilidade dos ramos deve ser denotada. Inicialmente, é necessário verificar o método utilizado para reconstrução da filogenia e, quando necessário, verificar o algoritmo utilizado para gerar a confiabilidade dos clados. Ramos com maiores valores de confiabilidade gerarão conclusões mais confiáveis, enquanto que clados com baixos valores deverão ser interpretados com maior cuidado. No entanto, não é necessário negar totalmente conclusões baseadas em filogenias com baixa confiabilidade nos ramos. O tipo de método, a forma de amostragem e o número de OTUs podem ser fatores de interferência e, assim, podem prejudicar a valorização dos ramos.

O padrão de organização dos ramos de uma filogenia denota o padrão de ancestralidade. As filogenias não são escadas, onde alguns organismos são "mais evoluídos" que outros, mas uma representação da história da derivação de OTUs. Na Figura 18-5, por exemplo, é possível observar que os clados B, C, D, E, F e G possuem um ancestral comum que compartilha um outro ancestral com o clado A. Já o clado H, representado por um triângulo para evidenciar um grande número de táxons naquele ponto da filogenia, teve um ancestral comum dentro do clado G. Este padrão sugere que o clado H se originou a partir do clado G. Da mesma forma, podemos observar a disposição do clado G em relação ao F e concluir que o primeiro se originou a partir do segundo.

No caso da Figura 20-5, observamos que humanos e chimpanzés tiveram um mesmo ancestral comum. Com base nestes dados, é incorreto pensarmos que humanos são derivados de chimpanzés, ou que humanos são mais evoluídos que chimpanzés. Estes organismos estão apenas formando um mesmo clado dentro da filogenia dos primatas.

Por último, é fundamental saber o objetivo do estudo filogenético a ser realizado. Árvores filogenéticas devem ser construídas para responder uma determinada questão, que pode envolver apenas um, ou diversos organismos.

Quando possível, é importante reconstruir a filogenia utilizando diferentes métodos de inferência e compará-las entre si. A conclusão desta forma será melhor sustentada. Além disso, atualmente, a história retratada em uma filogenia não é por si só satisfatória. Outras ferramentas podem ser utilizadas para complementar e sustentar a interpretação de uma filogenia, incluindo análises de recombinação, pressão seletiva e estruturação populacional, verificação de coespeciação, construção de redes filogeográficas, compa-

ração com dados de fósseis, eventos geológicos, dados históricos e, até mesmo, análises de dados comportamentais.

Um exemplo da combinação de análises filogenéticas com dados históricos veio na confirmação da origem e disseminação humana a partir da África. Através da utilização de dados histórico-antropológicos (como vestígios materiais de hominídeos ancestrais), fósseis de hominídeos e análises de DNA mitocondrial de representantes de diferentes etnias, os pesquisadores puderam traçar as rotas de disseminação humana a partir da África.

Outro exemplo está na solução de um enigma que perturbou zoólogos por um longo período: a posição taxônomica do panda-gigante entre os mamíferos carnívoros. Apesar de esta espécie ser fisicamente muito similar a um urso, outras características, como dentição e anatomia das patas, levaram à proposição de uma hipótese antes não imaginada.

Tal hipótese propunha que o panda-gigante (*Ailuropoda melanoleuca*) seria proximamente relacionado ao o panda-vermelho (*Ailurus fulgens*), um mamífero de pequeno porte, semelhante ao guaxinim. Com o emprego de diferentes dados, incluindo fósseis, anatomia de mamíferos atuais, distribuição geográfica, sequências de DNA de diferentes porções do genoma, sequências de aminoácidos de diferentes proteínas e mapeamento cromossômico, foi possível estabelecer uma história evolutiva plausível, capaz de descrever a origem evolutiva do panda-gigante (Figura 21-5).

Por meio dessa análise combinada de dados, se propôs que o panda-gigante, um urso, derivou do ancestral comum dos ursos há cerca de 24 milhões de anos, muito antes das derivações que originaram todos os outros ursos existentes hoje. Além disso, observou-se que os ursos e os procionídeos (grupo que inclui o guaxinim e o panda-vermelho) possuem um ancestral comum que deu origem às duas linhagens há aproximadamente 30 milhões de anos.

A filogenia molecular é uma ferramenta útil quando empregada isoladamente, mas que pode se beneficiar de diferentes tipos de dados para propor uma história evolutiva. Em última análise, a decisão sobre que tipos de

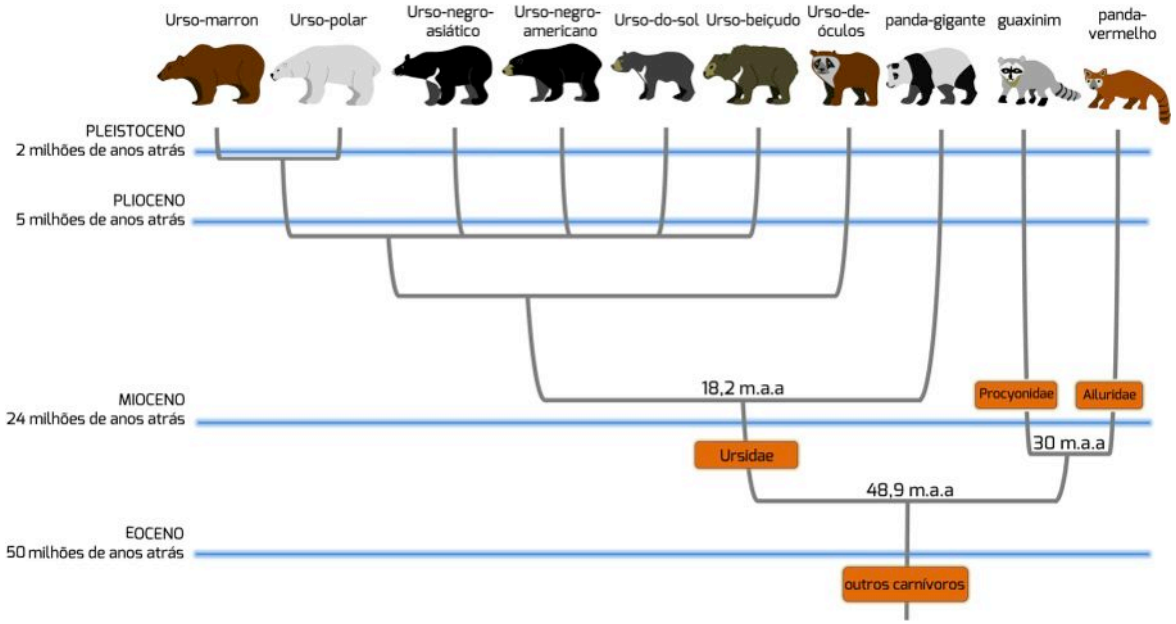


Figura 21-5: Posição filogenética do panda-gigante, baseada na combinação de diferentes tipos de dados. Baseado em BININDA-EMONDS, Olaf R.P. *Phylogenetic position of the giant panda*. Em: LINDBURG, D.G. & Baragona, K. *Giant pandas: Biology and conservation*. Berkeley: University of California Press, 2004; e em EIZIRIK, Eduardo e colaboradores: *Pattern and timing of diversification of the mammalian order Carnivora inferred from multiple nuclear gene sequences*. *Mol Phylogenet Evol*, 56, 49, 2010.

dados (além dos moleculares) serão empregados na análise filogenética dependerá da pergunta a ser respondida com essa técnica. Não existem regras pré-estabelecidas, e as estratégias analíticas precisam ser propostas caso a caso.

5.10. Conceitos-chave

Ancestral: organismo ou sequência que originou novo(s) organismo(s) ou sequência(s). Em alguns casos pode ser considerado o mesmo que primitivo.

Apomórfico: refere-se a um caractere novo adquirido ao longo do processo evolutivo, uma inovação. Uma apomorfia pode servir de diagnóstico para separação de clados.

Aproximação dos vizinhos: *neighbor joining* (NJ), método de inferência filogenética quantitativo baseado em distância genética.

Autapomorfias: apomorfias específicas e restritas a um clado.

Bootstrap: método de reamostragem que permite verificar a confiabilidade dos ramos de uma filogenia.

Cadeias de Markov Monte Carlo: método utilizado pela estatística Bayesiana para amostrar as probabilidades de distribuição de diferentes parâmetros das filogenias.

Clado: grupo formado por um ancestral e todos seus descendentes, um ramo único em uma árvore filogenética.

Derivado: que se originou de um ancestral e é mais recente no tempo evolutivo (nota: deve-se evitar o termo "mais evoluído" e, em seu lugar, empregar "derivado").

Distância Genética: medida quantitativa da divergência genética entre organismos.

Espaço Amostral de Filogenias: espaço teórico que inclui todas as filogenias possíveis (com raiz ou sem raiz) para um determinado alinhamento.

Frequência de equilíbrio: ponto em que não existe mais alteração nas frequências dos alelos.

Grupos irmãos: clados que dividem um ancestral comum.

Homologia: similaridade originada por ancestralidade comum.

Inferência filogenética Bayesiana: método qualitativo de inferência filogenética baseado na estatística Bayesiana. Através da Cadeia de Markov Monte Carlo este método buscará as árvores mais prováveis dentro das filogenias amostradas.

Máxima Parcimônia: método qualitativo de inferência filogenética que busca a árvore que minimiza o número total de substituição de nucleotídeos.

Máxima Verossimilhança: método qualitativo de inferência filogenética que busca a árvore com a máxima verossimilhança.

Monofilia: associação entre o ancestral comum e todos os seus descendentes, formando um clado monofilético.

Múltiplas Substituições: eventos múltiplos de substituição de nucleotídeo localizado em um mesmo sítio do DNA.

Modelos de Substituição: modelos matemáticos utilizados para descrever o processo evolutivo ao longo do tempo, podendo ser aplicados ao alinhamento de nucleotídeos ou aminoácidos.

Ortólogo: genes homólogos em diferentes organismos e que mantém a mesma função.

OTU: unidade taxonômica operacional, folha ou nó terminal em uma árvore filogenética.

Parafilia: associação entre o ancestral comum e apenas parte de seus descendentes, formando um clado parafilético.

Parálogo: genes homólogos de um mesmo organismo que divergiram após duplicação.

Plesiomórfico: dotado de características do ancestral que são conservadas nos descendentes.

Polifilia: associação entre diferentes OTUs sem a necessidade de um único ancestral comum, frequentemente originada por convergência evolutiva.

Primitivo: diz-se de características ou organismos ancestrais, anteriores no tempo evolutivo a organismos ou características mais recentes.

Probabilidades Anteriores: distribuição dos valores de um parâmetro filogenético que é sabido de antemão pelo pesquisador.

Probabilidades Posteriores: conjunto da distribuição dos valores de parâmetros filogenéticos resultantes do método de inferência Bayesiana.

Sistemática: estudo da diversificação das formas vivas e suas relações ao longo do tempo.

Taxonomia: estudo que busca agrupar os organismos com base em suas características e nomear os grupos obtidos, classificando-os em alguma escala.

Taxon: grupo (de qualquer nível hierárquico) proposto pela taxonomia.

Topologia: descreve a ordem e a disposição exata das OTUs em uma filogenia.

UPGMA: *unweighted pair-group method using arithmetic average*, método de inferência filogenética quantitativo baseado em distância.

5.11. Leitura recomendada

FELSENSTEIN, Joseph. ***Inferring Phylogenies***. Sunderland: Sinauer, 2004.

GREGORY, T. Ryan: ***Understanding Evolutionary Trees***. Evo. Edu. Outreach, 2008, 1,121-137.

LEMEY, Philippe; SALEMI, Marco; Vandamme, Anne-Mieke (Org.). ***The Phylogenetic Handbook***. 2.ed. Cambridge: Cambridge University Press, 2009.

MATIOLI, Sergio Russo; FERNANDES, Flora M.C. (Org.). ***Biologia Molecular e Evolução***. 2.ed. Ribeirão Preto: Holos, 2012.

NEI, Masatoshi; KUMAR, Sudhir. ***Molecular Evolution and Phylogenetics***. Nova Iorque: Oxford University Press, 2000.

PABÓN-MORA, Natalia; GONZÁLEZ, Favio. A classificação biológica: de espécies a genes. In: ABRANTES, Paulo C. (Org.), ***Filosofia da Biologia***. Porto Alegre: Artmed, 2011.

SCHNEIDER, Horacio. ***Métodos de Análise Filogenética: Um Guia Prático***. 3.ed. Ribeirão Preto: Holos, 2007.