CAMILA MALTA ROMANO

CARACTERIZAÇÃO E DINÂMICA EVOLUTIVA DE RETROVÍRUS ENDÓGENOS DA FAMÍLIA K (ERV-K) EM GENOMAS PRIMATAS

Tese apresentada ao Departamento de Microbiologia do Instituto de Ciências Biomédicas da Universidade de São Paulo, para obtenção do Título de Doutor em Ciências.

São Paulo
2009

CAMILA MALTA ROMANO

CARACTERIZAÇÃO E DINÂMICA EVOLUTIVA DE RETROVÍRUS ENDÓGENOS DA FAMÍLIA K (ERV-K) EM GENOMAS PRIMATAS

Tese apresentada ao Departamento de Microbiologia do Instituto de Ciências Biomédicas da Universidade de São Paulo, para obtenção do Título de Doutor em Ciências. Área de concentração: Microbiologia

Orientador: Prof. Dr. Paolo Marinho de Andrade Zanotto

São Paulo
2009

DADOS DE CATALOGAÇÃO NA PUBLICAÇÃO (CIP)
Serviço de Biblioteca e Informação Biomédica do
Instituto de Ciências Biomédicas da Universidade de São Paulo

Romano, Camila Malta.
Caracterização e dinâmica evolutiva de retrovírus endógenos da família K (ERV-K) em genoma de primatas / Camila Malta Romano. -- São Paulo, 2009.

Orientador: Paolo Marinho Andrade zanotto.

Tese (Doutorado) – Universidade de São Paulo. Instituto de Ciências Biomédicas. Departamento de Microbiologia. Área de concentração: Microbiologia. Linha de pesquisa: Virologia.

Versão do título para o inglês: Characterization and evolutionary dynamics of endogenous retroviruses K (ERV-K) in primate genomes .

Descritores: 1. Retrovírus endógenos 2. Evolução 3. Primatas 4. Filogenia 5. Evolução Molecular 6. Coevolução I. Zanotto, Paolo Marinho Andrade II. Universidade de São Paulo. Instituto de Ciências Biomédicas. Programa de Pós-Graduação em Microbiologia III. Título.

ICB/SBIB0176/2009

UNIVERSIDADE DE SÃO PAULO
INSTITUTO DE CIÊNCIAS BIOMÉDICAS

Candidato(a): Camila Malta Romano.

Título da Tese: Caracterização e dinâmica evolutiva de retrovírus endógenos da família K (ERV-K) em genoma de primatas .

Orientador(a): Paolo Marinho Andrade zanotto.

A Comissão Julgadora dos trabalhos de Defesa da Tese de Doutorado, em sessão pública realizada a ...../...../....., considerou

( ) Aprovado(a) **( ) Reprovado(a)**

Examinador(a): Assinatura: .....
Nome: .....
Instituição: .....

Examinador(a): Assinatura: .....
Nome: .....
Instituição: .....

Examinador(a): Assinatura: .....
Nome: .....
Instituição: .....

Examinador(a): Assinatura: .....
Nome: .....
Instituição: .....

Presidente: Assinatura: .....
Nome: .....
Instituição: .....

À minha família

Agradecimentos

Ao Prof. Paolo Zanotto, meu orientador, e exemplo. Obrigada pela chance de aproveitar esses últimos anos da melhor e mais intensa forma possível; A oportunidade de passar por experiências fantásticas. Por ver a sua paixão por Ciência, pelo conhecimento, e verdade. Por saber que no fundo, é isso o que importa. Por tudo o que me ensinou nesse tempo, que vou levar pra sempre.

A minha família, que sempre me apoiou, me 'empurrou' pra frente, acreditando e dando suporte para tudo. Pai, mãe e Thais. Por estarem sempre aqui.

A minha segunda família, os amigos do LEMB que fizeram toda a diferença: Amiga Carla, Frankola, Ali, Ju Vela, Atila, Moto, Ana Vit. A todos que passaram por lá, e também fizeram diferença: Tatis, Taís, Corelio, Adalbs, Vanessa e Vanessa. Ao caçula da familia, CJ. E aos agregados, os amigos do depto de Micro, especialmente Angélica, Clau e Silvana sempre dispostas a darem a mão.

A "turma da Cerveja", que sempre compartilharam muito mais do que cerveja, mas amor, companheirismo, idéias, amizade, carinho. Por todas as horas, por serem Amigos.

A Ana Vit e ao Frank, meus *reviewers* particulares, pacientes, atenciosos e queridos.

Ao Prof. Eddie Holmes, que me deu quatro meses de experiência que valeram por 40, pelo cuidado, presteza, e exemplo de pesquisador.

Aos meus ex orientadores, de IC e de Mestrado, que fizeram parte dessa história.

“Na vida não há prêmios nem castigos. Somente consequências.”

Robert Green Ingersoll

RESUMO

Romano CM. Caracterização e dinâmica evolutiva de retrovírus endógenos da família K (ERV-K) em genoma de primatas [Tese]. São Paulo: Instituto de Ciências Biomédicas da Universidade de São Paulo; 2009.

O genoma de primatas é repleto de sequências de retrovírus endógenos (ERVs). Estes são elementos derivados de vírus exógenos, que em algum momento infectaram células germinativas e proliferaram no genoma do hospedeiro. A família K é uma das maiores famílias de ERV, integrada no genoma de primatas após a separação das linhagens que originaram os primatas do Velho Mundo e os do Novo Mundo. Os ERVs podem alterar o padrão de expressão de genes vizinhos devido a atividade dos seus promotores. Além disso, promovem rearranjos e duplicações gênicas, sendo fundamentais para a evolução dos genomas. Uma vez que a família K possui um grande número de provirus completos e está presente apenas em genomas de primatas, esse trabalho teve como objetivo realizar uma investigação detalhada da distribuição e dinâmica evolutiva desses elementos nos diferentes hospedeiros. Utilizando ferramentas de bioinformática para buscar provirus completos de ERV-K nos genomas de primatas, foram identificados e caracterizados 58 ERV-K em humanos, 38 em chimpanzés, 35 em orangotangos e 19 em macaco rhesus. Estimativas do tempo de integração revelaram diversos elementos recentemente integrados nos genomas de humanos, e possivelmente, no genoma de orangotango. Análises filogenéticas evidenciaram dois grupos principais, Grupo O/N, que compreende os provirus com data de integração mais antiga e também os mais recentes, e Grupo I, que contém provirus com tempo de integração intermediário. Os resultados também mostraram que a dinâmica de espalhamento de ERV-K é diferente entre os hospedeiros. Aparentemente, a fixação e eliminação dos ERV-K é resultado de fatores demográficos e populacionais, como gargalos de garrafa e expansões sofridas por cada linhagem ao longo da sua evolução. Por fim, as análises de quais provírus são potencialmente ativos em pacientes HIV positivos e com cancer demonstrou que, em cada patologia, distintos provirus são transativados, sugerindo que a superexpressão de determinados ERVs possa ter alguma consequencia biológica para o hospedeiro. Além disso, os resultados sugeriram que a atividade não depende exclusivamente do tempo de integração, mas sim da integridade de regiões específicas contidas na LTR.

Palavras-chave: Retrovírus endógenos. Evolução Molecular. Primatas. Filogenia

ABSTRACT

Romano CM. Characterization and evolutionary dynamics of endogenous retroviruses K (ERV-K) in primate genomes [Thesis]. São Paulo: Instituto de Ciências Biomédicas da Universidade de São Paulo; 2009.

Primate genomes carry thousands of copies of retroelements at different levels of integrity, such as retrotransposons and ERVs, which are remains of ancient viral infection in the germ line cells and subsequent vertical transmission. ERV-K family integrated in the primate genome after the separation of Old and New World monkeys at 30 to 45 million years before present. Due to the presence of promoter regions in their LTRs, ERVs can affect the expression of nearby genes. Moreover, they play a fundamental role on genome evolution and foster variability. Since the K family is present only in primate genomes and has several complete elements, this work investigated their distribution and evolutionary dynamics in distinct primate hosts. We found 55 complete ERV-K genomes in the human genome, 38 in chimpanzee, 35 in orangutan and 19 in Rhesus monkey. Integration time estimates revealed several recently integrated proviruses in both human and orangutan genome. Two main groups were recovered by phylogenetic inference, named Group O/N, comprising the newest and the oldest integrated proviruses and, Group I, enclosing those with intermediate integration time. Interestingly, although the primary integration took place in the ancestral lineage of all primates investigated, their evolutionary dynamic was different among them. I propose that ERV-K fixation and purging depends on the flutuations of the host demography experienced throughout their evolution. This work also investigated the putative source of proviral transcripts previously detected in HIV carries and cancer patients. The differential expression found under these conditions suggested a biological role of the ERV-K overexpression. Finally, the results also showed that the ERV-K overexpression does not depend exclusively on integration time, but on the integrity of specific promoters in their LTR.

Key words: Endogenous retroviruses. Molecular Evolution. Primates. Phylogeny.

LISTA DE ABREVIATURAS E SIGLAS

APOBEC – apolipoprotein B mRNA editing enzyme

BLAT – Blast-Like Alignment tool

Blast – Basic Local Alignment Search Tool

ERV – *Endogenous retrovirus* (retrovirus endógenos)

LTR – *Long terminal repeat* (sequência repetitiva terminadora)

m.a.a. – milhões de anos atrás

ORF – *Open reading frame* (fase aberta de leitura)

Pb – pares de base

TE- Elemento transponível

TMRCA – Tempo do ancestral comum mais recente

SUMÁRIO

1 INTRODUÇÃO

Grande parte do genoma dos organismos eucariotos, com exceção de alguns protozoários Apicomplexa, é composta de fragmentos de DNA capazes de se translocar de uma região a outra do genoma por um processo conhecido como transposição (Wicker *et al*., 2007). Descoberto por McClintock como fragmentos móveis presentes no genoma do milho, esses elementos são chamados transposons (TE- *transposable elements*) (McClintock, 1950). O sucesso evolutivo dos transposons se deve principalmente a sua habilidade de se autorreplicar (retrotransposons) e de se manter no genoma hospedeiro. Além disso, por serem capazes de se inserir em qualquer região do genoma, incluindo regiões reguladoras e codificadoras de proteínas, a mobilidade de TEs frequentemente resulta em efeitos deletérios para o hospedeiro (Finnegan, 1992). Por esse motivo, TEs eram anteriormente classificadas como "DNA-parasitas" ou *selfish DNA* (Doolittle e Sapienza 1980). Atualmente, entretanto, é bem reconhecido o seu papel na evolução dos genomas hospedeiros, participando da regulação gênica e contribuindo para o aumento de variabilidade (Rouzic e Capy 2005).

Existem inúmeras famílias de TEs entre os diferentes organismos, algumas vezes representando uma substancial parte dos seus genomas (3-20% em fungos e 3- 45% em metazoários) (Hua-Van *et al*., 2005). Nesse sentido, cerca de 45% do genoma humano é formado por elementos repetitivos e transposons (Lander *et al*., 2001). Em 1989, Finnegan propôs um sistema de classificação de TEs baseado no mecanismo de transposição. Os elementos pertencentes a Classe I compreendem os retrotransposons, que utilizam transcriptase reversa, geralmente codificada pelo próprio elemento, para fazerem cópias de si mesmo a partir de um RNA transcrito e reintegrarem no genoma hospedeiro. A Classe II compreende os TEs sem intermediário de RNA, que se movem no genoma por mecanismo de excisão e inserção (Finnegan, 1989). Os retrotransposons ainda são classificados em quatro ordens, com diferentes superfamílias (Wicker *et al*., 2007). Entre os diversos tipos de retrotransposons, destacam-se os retrovírus endógenos (ERVs), que possuem a mais o gene de envelope (*env*), além de capsídeo (*gag*), protease e polimerase (*pro-pol*) e regiões flanqueadoras não-codificadoras denominadas LTRs (*Long Terminal Repeat*).

Análises comparativas da estrutura e dos domínios presentes nos retrotransposons sugerem que retrotransposons com LTR (principalmente *gypsy*) possam ser os ancestrais dos retrovírus (Bucheton, 1995). Inclusive, os próprios retrotransposons com LTR teriam se originado a partir de um ancestral comum de estrutura bem mais simples, sem LTR (Xiong e

Eickbush, 1990). O aumento da complexidade dos retrotransposons teria acontecido pela aquisição de genes endógenos de seus hospedeiros, como por exemplo, a aquisição da transcriptase reversa, endonucleases, *gag* e *env* (Capy *et al.,* 1998).

Embora o padrão de distribuição de TEs não reflete obrigatoriamente sua filogenia, na maioria dos casos o grau de divergência entre as sequências é proporcional à distancia evolutiva entre os hospedeiros (Voytas *et al.,* 1992).

1.1 Retrovírus Endógenos

A família *Retroviridae* compreende vírus que utilizam a RNA-dependente-DNA-polimerase, ou seja uma transcriptase reversa (TR) (Baltimore, 1970; Temin, 1970), como particularidade de sua estratégia replicativa no organismo hospedeiro. Os retrovírus normalmente infectam células somáticas e possuem uma fase denominada "provírus", representada pela integração do genoma viral no genoma do hospedeiro (Coffin *et al.,* 1997). De acordo com a classificação do Comitê Internacional de Taxonomia de Vírus (ICTV), existem sete gêneros dentro da familia *Retroviridae*: Alpharetrovírus, Betaretrovírus, Deltaretrovírus, Gammaretrovírus, Epsilonretrovírus, Spumavírus e Lentivírus (Coffin *et al.,* 1997). Os vírus pertencentes aos diferentes gêneros se distinguem pela organização do seu genoma, pela presença/ausência de oncogenes e genes acessórios, forma do capsídeo e o organismo hospedeiro.

ERVs, do inglês *endogenous retroviruses*, representam a fase proviral dos respectivos retrovírus exógenos que, em algum momento, se integraram em células germinativas do seu hospedeiro. Verticalmente transmitidos aos descendentes, os provirus colonizaram definitivamente seus genomas (Li *et al.,* 1993; Coffin *et al.,* 1997). O genoma humano, por exemplo, possui mais de 98.000 sequências (completas ou parciais) de retrovírus endógenos de diversas famílias (Paces *et al*., 2002; Lander *et al*., 2001).

Os ERVs foram descritos apenas nos genomas dos vertebrados, sendo a maior variabilidade encontrada em genomas de mamíferos. Entretanto, elementos similares a ERVs em tunicados já foram descritos (Britten *et al.*, 1995). Os ERVs são agrupados em três grandes classes, de acordo com a similaridade aos retrovírus exógenos. Os ERVs da classe I são relacionados aos gamaretrovírus, como o Murine leukemia vírus (MLV) e inclui, entre outras, as famílias ERV-W e ERV-H. ERVs da classe II estão relacionados aos betaretrovírus, como Mouse mammary tumor vírus (MMTV) e inclui os ERV-K. A classe III é relacionada aos spuma retrovírus e inclui ERV-L e ERV-S (Mayer e Meese, 2002; Tristem,

2000). O nome das famílias de ERVs é dado de acordo com o tipo de RNAt utilizado pelo PBS-*primer binding site* (região iniciadora da transcrição reversa), como por exemplo no caso da família K: Lisina. Um típico retrovírus endógeno codifica genes de capsídeo (*gag*), de protease e polimerase (*pro-pol*) e de envelope (*env*) flanqueados por duas LTRs (*Long Terminal Repeat*). Até bem pouco tempo atrás, haviam sido identificados retrovírus endógenos de todos os gêneros, com exceção de Lentivírus e Deltaretrovírus (Tristem e Gifford, 2003). Dois recentes trabalhos, entretanto, demonstraram a presença de lentivírus endógenos em genoma de coelhos (RELIK) e primatas (pSIVgml) (Katzourakis *et al*., 2007; Gifford *et al*., 2008).

Após a endogenização, os ERVs retêm a capacidade de replicação dentro do genoma hospedeiro por algum tempo. Essa replicação pode ocorrer por retrotransposição (replicação ativa mediada pelo promotor LTR), ocorrendo na mesma célula, ou ainda por reinfecção, que depende do envelope para invadir outras células (Coffin *et al.,* 1997). Mesmo inativo pelo acúmulo de substituições, os ERVs podem ainda aumentar em número de cópias através de duplicações de regiões cromossômicas (Li *et al.,* 1993). A velocidade de proliferação de um ERV tende a ser muito maior no período após a infecção inicial, declinando com o tempo. Isso acontece principalmente em razão dos mecanismos supressores da patogenicidade exercidos pelo organismo hospedeiro, que acaba levando ao acúmulo de substituições deletérias ao longo do genoma proviral (Coffin *et al.,* 1997). Atualmente, embora a maioria dos ERVs sejam inativos, muitos deles possuem importante papel na biologia do hospedeiro, participando do controle de transcrição e regulação gênica (Medstrand *et al*., 2005), bem como outras diversas funções descritas mais adiante.

1.1.2 A família ERV-K

Em 1980, Kurth e colaboradores reportaram a existência de partículas virais que brotavam de células de linhagem de teratocarcinoma, e as chamaram de HTDV (human teratocarinoma-derived particles) (Kurth *et al*., 1980). Anos depois, Callahan e colegas encontraram similaridade entre regiões do genoma de primatas com genes de *gag* e *pol* dos vírus MMTV, descrevendo assim, pela primeira vez provirus similares aos MMTV (Callahan *et al*., 1985), e por isso foram chamados de HML (human MMTV-like). Um ano depois, Ono e colegas observaram que esses provirus utilizavam RNA transportador específico de lisina, e assim chamaram esses elementos de HERV-K (Ono *et al*., 1986). Foi somente em 1993, que

os provirus da família K foram associados às partículas encontradas anteriormente por Kurth em teratocarcinoma (Boller *et al.*, 1993).

Os provirus K possuem um genoma de 9.2 kb, e estão presentes apenas em primatas do Velho Mundo (Steinhuber *et al*., 1995). Isso implica que a integração desses elementos ocorreu em algum momento depois da divergência dos Platyrrhines (primatas do Novo mundo) dos Catarrhines (primatas do Velho mundo), há cerca de 35 milhões de anos atrás (Jones *et al*., 1994), porém antes da separação dos cercopithecoides dos hominoides. Embora não tenha sido comprovada a existência de ERV-K em primatas do Novo mundo, Kim *et al.* (1999) utilizando oligos conservados para a região promotora identificou alguns elementos em poucas espécies. No entanto, esse resultado jamais foi reproduzido em nenhum outro trabalho.

Um dos primeiros métodos de classificação de ERV-K foi feito com base em uma deleção de um fragmento de 292 nucleotídeos entre os genes de polimerase e envelope (Ono *et al.,* 1986). Essa deleção divide os provirus em tipo 1, os que possuem a deleção, e em tipo 2, os que não a possuem. Provirus do tipo 1 codificam uma oncoproteína chamada Np9, que é superexpressa em certos tipos de câncer e células de linhagem tumorais (Ambruester *et al*., 2002). Os provirus do tipo 2 codificam uma proteína similar, a Rev de HIV-1, denominada c-ORF ou Rec (Lower *et al*., 1995).

Representando as distintas integrações dos respectivos vírus exógenos no genoma hospedeiro, a família K ainda se divide em 10 subgrupos, denominados HML-1 a HML-10 (Anderson *et al*., 1999; Medstrand e Blomberg, 1993). Os provirus que foram classificados em tipo 1 e tipo 2 são pertencentes a subfamíla HML-2. A maioria dos genomas de provírus completos (contendo *gag, pol* e *env,* além das LTRs) encontrados no genoma humano são pertencentes a família K, subfamília HML-2, e dentro dela, o elemento mais conservado é o HERV-K10, com apenas um codon terminador localizado entre os genes *gag* e *env* (Ono *et al*., 1986). Existem, entretanto, algumas controvérsias em relação a nomenclatura dos ERVs, principalmente em relação a família K. Por exempo, desde sua descoberta até hoje, os trabalhos publicados que incluem ERV-K denominam diferentes nomenclaturas, tais como HERV-K10, HTDV/HERV-K, HERV-K(HML-2), HML-2HOM, HERV-K e HERVK. Além disso, uma vez que esses vírus pertencem a família *Retroviridae*, é errado denominar "K", "R"ou "H" como sendo famílias de ERV. Assim, 'K" e "H" seriam no máximo grupos. Por essas razões, um trabalho publicado este ano sugeriu que a classificação dos ERVs fosse reformulada e unificada (Bloomberg *et al*., 2009).

Os membros do grupo HML-2 são os mais conservados entre si, apresentando diversos provirus ativos até hoje. De acordo com isso, há diversos *loci* exclusivamente humanos, indicando integração após a divergência das linhagens que deram origem aos gêneros *Homo* e *Pan* (Barbulescu *et al*., 1999; Tristem *et al*., 2003). Alguns desses elementos inclusive são expressos em diversos órgãos, como placenta, pulmões, testículos, tecidos tumorais e principalente em células germinativas (Parseval *et al.,* 2001).

1.2 Impactos da atividade dos retroelementos no genoma hospedeiro

1.2.1 Patologias relacionadas a expressão anormal de ERVs

ERVs fazem parte do transcriptoma humano, e embora apresentem variações quanto ao número de transcritos e a família da qual fazem parte, em geral eles são expressos em baixos níveis na grande maioria dos tecidos já testados (Seifarth *et al*., 2003). No entanto, devido a capacidade de replicação e re-integração em regiões não específicas do genoma hospedeiro, os ERVs podem ao se integrar, causar disruptura em algum gene, ou ainda, alterar o padrão de expressão de genes vizinhos devido a atividade das regiões promotoras contidas nas LTRs (Boeke e Stoye, 1997). Justamente por essa razão, os efeitos deletérios dos ERVs nos genomas hospedeiros teem sido amplamente discutidos. Ademais, devido ao carater polimórfico de certos provirus, a presença de alguns deles é considerada como um provável fator de risco para determinadas doenças (Margerat *et al*., 2004; Magistrelli *et al*., 2004; Moyes *et al*., 2005).

Transcritos de HERV-K foram detectados em diferentes tipos de células tumorais (Herbst *et al*., 1996; Wang-Johanning *et al*., 2001; Serafino *et al*., 2009; Golan *et al.*, 2008) em desordens neurológicas, como esquizofrenia e transtorno bipolar (Frank *et al*. 2005), diabetes tipo 1 (Margerat *et al*., 2004), doenças autoimunes (Magistrelli *et al*., 2004; Christensen 2005) e em infecções por vírus exógenos (Stevens *et al.*, 1999; Kwun *et al*., 2002). Porém, uma real associação entre expressão de ERV com o desenvolvimento ou progressão dessas doenças nunca foi estabelecida, permanecendo a questão se a superexpressão de ERVs é causativa, ou simplesmente consequência da alteração da malha gênica celular, que ocorre diante do stress causado por essas patologias. Diante disso, um recente trabalho discute o possível papel biológico da atividade de vírus endógenos em infecções por HIV, onde sua expressão diferencial teria um papel evolutivamente selecionado por trazer consequências benéficas ao hospedeiro (Garrison *et al*., 2007). Neste trabalho, foi

observado que em uma infecção por HIV-1, a superexpressão de ERVs estimula a produção e recrutamento de células T CD8+ ERV-específicas, que conseguem atacar as células infectadas que estejam expressando esses epítopos.

1.2.2 Participação de ERVs na malha gênica do hospedeiro

Pode ser verdade que, se por um lado a atividade de provirus realmente traz consequências negativas a curto prazo, a longo prazo a manutenção da atividade de certos elementos seja vantajosa. Retrotransposons e retrovirus endógenos codificam enzimas como transposases, integrases, transcriptases reversas e proteínas de envelope. Uma vez que um genoma é invadido por esses elementos, não é impossível supor que todo este repertório de novas funções poderia ser interessante para o hospedeiro. O termo "domesticação molecular" tem sido amplamente utilizado para descrever processos onde o genoma do hospedeiro acaba incorporando proteínas e promotores de retroelementos na sua malha gência (Jordan *et al*., 2003; Volff, 2006).

O mais famoso exemplo de domesticação molecular em humanos diz respeito a expressão do gene de envelope de um provirus da família W (HERV-W) localizado no cromossomo 7. Por possuir propriedades fusogênicas, essa proteína é relacionada a formação do sinsiciotrofoblasto durante o desenvolvimento da placenta (Mi *et al*., 2000). Interessantemente, esse é o único locus da família W que ainda possui uma ORF (*open reading frame*) intacta do gene de envelope (Blond *et al*., 1999). Outra vantagem da expressão de genes de origem retroviral na placenta poderia estar relacionada à mediação da resistência a ataques por vírus exógenos (Voisset *et al*., 2000). Proteínas de envelope de ERVs também podem ser detectadas na superfície de ovócitos, o que supostamente influenciaria na imunomodulação do sistema imune, evitando o ataque do óvulo fertilizado (Prudhomme *et al*., 2005). A expressão de ERVs, bem como a regulação de genes do hospedeiro por promotores provirais, é muito mais frequente em células germinativas, testículos e placenta. Portanto, é válido supor que a superexpressão de ERVs, especificamente nesses tecidos, pode ser não mais do que um artefato do mecanismo de endogenização, visto que os retrovirus que infectam e são mais ativos em células germinativas têm maior probabilidade de, mais tarde, se tornarem retrovirus endógenos (Cohen *et al*., 2009).

Diversos estudos têm demonstrado que a presença de retroelementos alteram não só a estrutura dos genes e seus produtos, mas também sua regulação (Leib-Moch *et al*., 1996;

Reiss *et al*., 2007; Brosius, 2000). A regulação da expressão de genes do hospedeiro por promotores provirais foi descrita em diversas situações, e é entendida como um importante processo para a diversificação e evolução de genes de mamíferos (Jordan *et al*., 2001; van de Lagemaat *et al*., 2003; Bejerano *et al*., 2006; Buzdin *et al*., 2006). Jordan e colaboradores (2001) revelaram que 25% dos promotores do genoma humano que foram experimentalmente caracterizados continham sequências derivadas de TEs, incluindo elementos regulatórios. No mesmo ano, outro trabalho mostrou que 15% dos transcritos do gene Apolipopoteína C-I, sabidamente superexpresso em células do fígado, derivavam de promotor retroviral (Medstrand *et al*., 2001).

Um extensivo trabalho recentemente publicado (Faulkner *et al*., 2009) mostrou que a participação de retroelementos na regulação do transcriptoma humano é maior do que se imaginava, atuando principalmente na expressão diferencial de genes entre tecidos distintos, *splicing* alternativos, *enhancers*, promotores alternativos e na geração de RNA de interferência (RNAi). Ademais, a influência de TEs na origem, biogênese e modo de ação dos RNAi tem sido cada vez mais reconhecida (Piryapongsa *et al*., 2007). Nesse contexto, inúmeros precursores de RNAi contém, ou mesmo derivam de TEs (Smalheiser *et al*., 2005).

Outro importante papel dos retrotransposons na regulação da expressão gênica seria como mediador de variações fenotípicas. Whitelaw e Martin (2001) mostraram que a cor da pelagem em camundongos pode ser influenciada por fatores epigenéticos, dependendo da atividade de certos retrotransposons. A partir disso, eles propuseram um modelo onde o silenciamento imperfeito de certos retrotransposons durante a embriogênese produziria um padrão de expressão tipo mosaico nas células somáticas.

1.3 O papel dos ERVs na Evolução Genômica

1.3.1 Geração de variabilidade

O papel de ERVs na geração de novos genes, bem como no aumento da variabilidade entre os organismos, talvez seja um dos mais relevantes efeitos positivos da presença desses elementos no hospedeiro. Retroelementos com integração polimórifica, assim como solo–LTRs (LTRs solitárias geradas por recombinação) contribuem com a variação alélica na população (Lower *et al*., 1996; Barbulescu *et al*., 1999). Dois deles, o HERV-K113 e o HERV-K115 foram descritos em 1999 por Barbulescu como sendo presentes apenas em uma pequena porcentagem da população humana (Barbulescu *et al*.,1999). A prevalência de

ambos é maior em afro-descendentes, e menor em povos da Oceania, chegando aos extremos de 43% em certos países da África contra zero indivíduos em Papua Nova Guiné (Moyes *et al*., 2005).

Uma das regiões do genoma humano que apresentam maior variabilidade e densidade de retrovirus endógenos é o MHC (complexo principal de histocompatibilidade) humano. Localizado no cromossomo 6, o MHC estende-se por quase 4x 10^6 pares de bases e é a região mais genicamente densa do genoma. O MHC, principalmente nas regiões que compreendem os genes das classes I e II de humanos, abriga diversos fragmentos de ERVs de diferentes famílias em uma proporção de 10:1 em relação ao restante do genoma (Kulski *et al*., 1999). Além disso, diversos elementos são polimórficos na população, como o HERV-K(C4), que, localizado no intron 9 do gene do complemento C4 de primatas, está presente em apenas 70% da população humana (Dangel, 1999). Por ser produzido um RNA antissenso desse provirus capaz de interagir com sequências virais *in vitro*, esse polimorfismo foi associado a proteção por infecções virais exógenas (Mack *et al*., 2004).

De maneira geral, ao longo do genoma, a maioria das inserções de ERVs ocorre em regiões intergênicas, e quando próximas a genes ou em regiões intrônicas, ocorrem ainda no sentido antissenso (van de Lagemaat *et al*., 2006). Porém, o excesso de ERVs no sentido senso ao longo do complexo MHC sugere a existência de algum tipo de seleção, favorecendo a integração nesse sentido nessa região.

Finalmente, por serem linhagem ou, espécie-específicos e ainda polimórficos em termos de presença/ausência em indivíduos de uma mesma população, os elementos transponíveis já foram utilizados como marcadores filogenéticos e populacionais (Minghetti e Dugaiczyk, 1993; Ray *et al*., 2006).

1.3.2 Origem de novos genes e evolução dos primatas

Após a separação das linhagens dos Platyrrhines dos Catarrhynes, houve uma verdadeira invasão do genoma de primatas do Velho Mundo por retrovirus endógenos e retrotransposons (Goodchild *et al*., 1993). Inclusive, a fração do genoma humano hoje ocupada por ERVs (6 a 8%) é quatro vezes maior do que as regiões dedicadas as sequências codificadoras de proteínas (Lander *et al*., 2001; Belshaw *et al*., 2004).

A origem de novos genes, que ocorre através de rearranjos e duplicação gênica é fundamental para a evolução dos genomas, possibilitando desde variabilidade fenotípica até especiação (Rouzic e Capy, 2005). Já é consenso que inversões cromossômicas, rearranjos,

inserções e deleções são os principais responsáveis pelas diferenças observadas entre as diferentes linhagens de primatas (Minghetti e Dugaiczyh, 1993; Sverdlov, 2000; Frazer *et al*., 2003). Visto que retroelementos atuam como *hotspots* para esse eventos (Schwartz *et al.,* 1998; Barbulescu *et al.,* 1999; Sverdlov, 2000), sua participação na evolução dos genomas de primatas parece evidente. Um trabalho onde foram encontrados sete novos retrogenes (genes originados a partir de duplicações de retroelementos) específicos de primatas, discute a possibilidade de que esse número deve chegar a 76 (Marques *et al*., 2005).

Com o sequenciamento do genoma humano, do chimpanzé, e mais recentemente de macaco rhesus e orangotango, as análises genéticas entre esses organismos tornaram-se ainda mais acessíveis. Diversos trabalhos direcionados as análises de regiões ortólogas têm sido feitos com elementos integrados, além de prováveis funções biológicas e efeitos estruturais causados por integrações e retroposições (Polavarapu *et al*., 2006; Han *et al*., 2007; Lee *et al*., 2008). Análises comparativas mostraram que retroelementos são responsáveis por 44% dos *loci* com inversões observadas entre *Homo* e *Pan* (Lee *et al*., 2008). Dentre essas, três inversões incluindo exons estão associadas as variações fenotípicas entre as linhagens. Frazer e colaboradores encontraram diversas deleções e inserções em um fragmento de 9 Mb no cromossomo 21, e discutem o quanto isso contribui para as diferenças entre chimpanzés e humanos (Frazer *et al*., 2003). A capacidade de ERVs gerarem duplicações gênicas foi também claramente demonstrada no presente trabalho (Romano *et al*., 2006). Um ERV-K presente no cromossomo X de chimpanzé e humanos pode ter sido o responsável por um evento de recombinação e geração de uma nova LTR que, duplicada, aparece inserida no sentido inverso apenas no cromossomo de humanos.

A invasão massiva do genoma de primatas por ERVs também pode ter contribuído para a evolução do sistema de defesa contra virus exógenos. Recentemente, dois sistemas de defesa antiretroviral que limitam infecções por HIV-1 *in vitro* foram descritos em primatas: APOBEC3 (Sheehy *et al*., 2002) e TRIMα (Stremlau *et al*., 2004). TRIMα é um fator de restrição pós-entrada, dependente do capsídeo do vírus, e foi inicialmente descrito em macaco rhesus (*Macaca mulatta*). Já a enzima APOBEC3 é uma deaminase e causa hipermutação em genomas retrovirais (Teng *et al*., 1993). APOBEC é o componente catalítico do *C*omplexo de *E*dição da *APO*lipoproteína *B*, que especificamente deamina C → U no RNA mensageiro da apolipoproteína B (Teng *et al*., 1993). Enquanto a maioria dos mamíferos codifica apenas um gene dessa família, o genoma humano e de outros primatas codifica pelo menos cinco tipos diferentes de APOBEC3, localizadas em um único locus no cromossomo 22. Além disso,

tanto o TRIMα quanto APOBEC3G, estão sob forte seleção positiva ao longo da evolução dos primatas (Sawyer *et al*., 2004; Liu *et al*., 2005). Juntos, esses dados sugerem que essa família gênica apareceu relativamente a pouco tempo na escala evolutiva, e continua evoluindo por duplicação gênica (Cullen, 2006). Finalmente, já foi demonstrado que membros dessa família inibem a retrotransposição de LINE-1 e Alu (Bogerd *et al*., 2006) e de retrovírus endógenos (Esnault *et al*., 2005), sugerindo a possibilidade de que a evolução dessas famílias gênicas tenham sido, ao menos em parte, dirigida pela constante atividade de diferentes retroelementos.

Finalmente, considerando todos os aspectos mencionados, comparações entre retrovirus endógenos de organismos filogeneticamente relacionados contribuem para uma melhor compreensão acerca da dinâmica evolutiva desses elementos. Além disso, estudos dessa natureza trazem informações sobre como os organismos hospedeiros entendem a invasão de seus genomas por elementos móveis. O reconhecimento da participação desses elementos transponíveis nos processos evolutivos dos primatas do Velho Mundo vem contrtibuir ao melhor entendimento das possiveis funções biológicas por eles exercidas, assim como sua contribuição na evolução estrutural e funcional desses genomas.

2 OBJETIVOS

A proposta de buscar sequências de ERV-K em outros primatas além de humanos visou uma investigação o mais completa possível no sentido descritivo e comparativo da distribuição e dinâmica evolutiva desses elementos nos diferentes sistemas. Dessa forma o objetivo central desse trabalho foi estudar os processos evolutivos da interação ERV-hospedeiro, a partir da comparação de genomas provirais completos em humanos e primatas do Velho mundo.

Os objetivos específicos foram:

(i) Através de ferramentas de bioinformática, identificar e caracterizar sequências completas de retrovírus endógenos da família K (ERV-K), em diferentes gêneros de primatas Catarrhini (famílias Pongidae, Hominidae e Cercopithecidae).

(ii) Avaliar impactos estruturais causados nos genomas hospedeiros devido a integração dos provírus.

(iii) Estudar a dinâmica evolutiva de ERV-K frente aos diferentes hospedeiros, levando-se em conta parâmetros populacionais e a biologia de cada um.

(iv) Investigar mais detalhadamente a atividade dos elementos no genoma do hospedeiro humano. Descrever quais elementos são potencialmente ativos e investigar as possíveis razões que dirigem a perda ou manutenção da atividade.

3 METODOLOGIA

3.1 Busca e mapeamento *in silico* de genomas completos de ERV-K em primatas

Em um trabalho prévio de iniciação científica realizado neste laboratório, o estudante Rodrigo Ramalho deu início a procura de ERV da família K nos genomas humano e de chimpanzé, utilizando as ferramentas Blastn e Fastacmd (Altschul *et al*., 1990). Desse trabalho resultou um total de 40 provírus humanos e um de chimpanzé, alguns dos quais já previamente descritos. A partir disso, este trabalho prosseguiu com a procura de elementos ainda não descritos nestes mesmos genomas como também no de outros primatas que foram subsequentemente depositados no banco de dados público. O material disponível para a busca de ERVs em banco de dados compreendeu os genomas humano (*Homo sapiens*, montagem 35 e 36.3), do chimpanzé (*Pan troglodytes*, montagens 1 e 2), do macaco rhesus (*Macaca mulatta*, montagem 1.1) e do orangotango (*Pongo pigmaeus abelii*, montagem 1).

A escolha da abordagem experimental adotada para a busca de ERV-K foi feita exclusivamente com o propósito de encontrar provirus completos, de um grupo específico, nos genomas de hospedeiros primatas. Entretanto, um levantamento mais genérico de retroelementos, completos ou não, seria inviável apenas com essas metodologias. Para isso, programas como RepeatMasker (http://www.repeatmasker.org) ou LTR_STRUC (McCarthy e McDonald, 2003) seriam de maior utilidade. Estes programas foram desenhados com o propósito de encontrar elementos repetitivos em genomas, e com eles, alguns trabalhos onde novas famílias de ERV foram descritas em genomas de primatas foram gerados (Polavarapu *et al.,* 2006a; Polavarapu *et al*., 2006b). Por utilizar um algoritmo que busca elementos que contenham ambas as LTRs e os dois sítios de duplicação, o LTR_STRUC encontra genomas de retroelementos baseado em estrutura, e não em similaridade. Isso implica que retrovírus endógenos de todas as famílias são igualmente encontrados, sendo necessária uma investigação posterior para classificá-los. Além disso, elementos que não contém os sítios de duplicação conservados são ignorados pelo programa, o que também não seria vantagem para este trabalho. Por outro lado, ferramentas como Blastn (http://www.ncbi.nlm.nih.gov/blast/Blast.cgi) por exemplo, são baseadas apenas em busca por similaridade, o que confere a vantagem recuperar somente sequências relacionadas aquela utilizada como sonda durante o processo de busca. Além disso, essas ferramentas oferecem a possibilidade de escolha de diferentes bancos de dados público para as buscas, permitindo

que um genoma, ainda em fase muito precoce de montagem, seja também investigado. Assim, a especificidade dessas ferramentas,aliada a gama de possibilidades de busca em diferentes genomas e banco de dados, bem como sua fácil utilização, determinaram sua escolha como metodologia padrão de busca de ERV-K neste trabalho.

Duas ferramentas de busca *on line* foram utilizadas: Blastn (http://www.ncbi.nlm.nih.gov/blast/Blast.cgi) e BLAT search (Blast-Like Alignment Tool) disponível no domínio http://www.genome.ucsc.edu/cgi-bin/hgBlat. Além dessas duas, uma ferramenta chamada "*ERV-Finder*" desenvolvida no nosso laboratório foi utilizada na procura de ERV-K. Esse *script* funcionava da seguinte maneira: inicialmente era feita uma busca no genoma hospedeiro por sequências que obedecessem ao padrão *gag-pol-env*, similares a sonda utilizada de um provírus completo (HERV-K), com um valor de *score* de similaridade pré-estabelecido. Após a identificação de uma região que obedecesse esse padrão, o programa estendia a busca para 10 mil pares de bases a jusante e a montante do fragmento encontrado, até encontrar que encontrasse regiões repetitivas flanqueando-a (LTRs). Na última etapa, o programa extraía do genoma a sequência encontrada e a adicionava em um banco de dados, onde as coordenadas da mesma eram anotadas e serviam como fator de eliminação para subsequente busca. Assim, o programa não deveria 'encontrar' duas vezes o mesmo elemento. A busca automatizada viabilizada por essa ferramenta foi bastante útil nas primeiras etapas do trabalho, entretanto, sua utilização foi interrompida em virtude de problemas que se tornaram limitantes. O primeiro deles consistiu no fato de que a ferramenta apenas encontrava elementos que estivessem muito bem conservados, contendo as três regiões *gag-pol-env*. Dessa forma, provírus que tivessem deleções ou inserções em um dos genes não eram identificados. O segundo impedimento para sua utilização veio da existência de diferentes montagens dos genomas, ou ainda, *contigs* com diferentes números de acesso e coordenadas distintas para a mesma região. Uma vez que as coordenadas de localização dos provírus são dependentes da fonte (ID, *contig* ou cromossomo), essas acabavam sendo entendidas pelo programa como um 'provírus distinto'. Ao final, a ferramenta "*ERV Finder*" foi utilizada apenas para a busca inicial de ERV-K de humanos.

Empregando principalmente as ferramentas *on line* e, como sonda, as sequências de HERV-K completas já disponibilizadas no banco de dados público do GenBank, foram então feitas buscas por similaridade nos genomas de primatas por novas sequências completas de ERV-K. Com a finalidade de encontrar toda a sorte de elementos com diferentes níveis de integridade, a sonda utilizada era substituída a cada nova rodada pelo elemento que obteve o menor *score* dentre todos na busca anterior. Essa estratégia permitiu que fosse encontrado

um grande número de elementos completos com a maior variabilidade possível. Após encontradas e extraídas do banco de dados em formato fasta, as sequências eram verificadas em relação a presença das duas LTRs que flanqueiam o provírus. Utilizando a ferramenta *DN*Astrider (disponível no link http://cellbiol.com/soft.htm), foram feitas matrizes de similaridade com o *DN*A dos provírus, onde os mesmos eram comparados com a sua própria sequência. Este exercício resultava em matrizes onde era possível observar a presença/ausência das duas LTRs (Figura 1). Em decorrência das LTRs serem idênticas, estarem localizadas nas extremidades do provírus e possuírem cerca de 1000 pares de bases, a presença delas em uma matriz aparece evidenciada nas extremidades dos gráficos, permitindo fácil visualização e localização do início e fim das mesmas.

Após a verificação da integridade dos provírus, os mesmos foram mapeados nos respectivos cromossomos. Além disso, devido ao avançado estágio de anotação do genoma humano, o mapeamento dos ERV-K de humanos (HERV-K) foi extremamente específico, podendo ser determinada a exata região cromossômica em que o elemento se encontrava.

Todos os genomas dos primatas utilizados neste trabalho estão disponíveis em sua última fase de montagem no domínio: http://www.ncbi.nlm.nih.gov/genomes/leuks.cgi?p3=12:Mammalsetaxgroup=11:l12:Mammals.

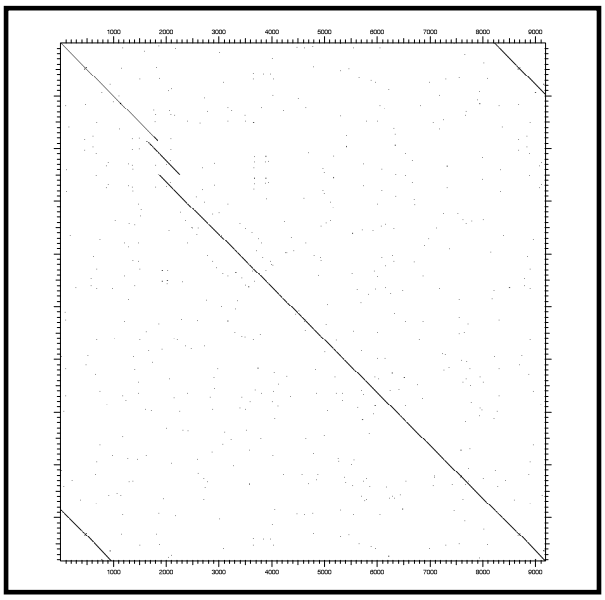


Figura 1. Matriz de similaridade contruída com o HERV-K10 no programa DNA Strider. Notar a presença das duas LTRs com 1000 pares de bases nas extremidades 5' e 3' (canto inferior esquerdo e superior direito).

3.2 Alinhamento dos genomas

Para o alinhamento de genomas completos foram utilizados os programas BlastAlign (Belshaw e Katzourakis, 2005) e Muscle (Edgar, 2004). Estes dois programas de alinhamento múltiplo minimizam o tempo e a memória computacional requerida por utilizarem um algoritmo rápido de alinhamento progressivo. O programa de alinhamento múltiplo ClustalX (Thompson *et al*., 1994) foi utilizado apenas para realinhar pequenos trechos das sequências, e para alinhar as LTRs e regiões flanqueadoras em análises posteriores. A partir do alinhamento inicial de genomas completos, que compreendeu os ERV-K de humanos e de chimpanzés, foi construído um alinhamento parcial designado *partgen*, contendo apenas os genes de capsídeo (*gag*) e polimerase (*pol*) dos provírus. Os ajustes finos dos alinhamentos, bem como as edições e separação das regiões codantes das não codantes foram feitas manualmente com o programa Se-Al (Rambaut, 1996).

Com a publicação do genoma do macaco rhesus e, posteriormente do orangotango, novas sequências de ERV-K foram encontradas e adicionadas ao *partgen*. Para facilitar a união dos alinhamentos dos novos provírus ao alinhamento do *partgen* original, a sequência do HERV-K10 foi inserida no arquivo em formato fasta (poli fasta) dos novos provírus e alinhada junto aos demais. Assim, uma vez que o HERV-K10 estava presente nos dois conjuntos de alinhamento (*partgen* original e novo alinhamento), ele servia de base para inserir manualmente o novo arquivo ao conjunto dos dados original. Para tal foi utilizado o programa Se-Al (Rambaut, 1996). Para as análises filogenéticas, os alinhamentos foram exportados em formato nexus e utilizados como entrada para os programas de reconstrução.

3.3 Reconstruções Filogenéticas

A Sistemática Filogenética teve origem com a publicação, em 1966, da edição inglesa do livro de Hennig "Phylogenetic Systematics", escrito em alemão em 1950 (Henning 1950; Henning 1966). A Sistemática Filogenética é parte da ciência de Biologia comparativa e o objetivo primário dela é descrever a diversidade entre taxa e reconstruir sua hierarquia ou relações filogenéticas (Henning, 1966). Foi nessa época que Henning instituiu o termo “Cladística” para representar as relações filogenéticas entre os seres vivos. A Cladística é uma escola de análise das relações evolutivas entre grupos de organismos, baseado em sinapomorfias compartilhadas entre eles (Henning, 1950), podendo ser esquematicamente representada pelo o que se chama de cladograma. Este é um esquema dicotômico,

representando uma hipótese sobre as relações filogenéticas de um grupo de taxa. Por se basear em caracteres, os cladogramas explicitam as relações evolutivas entre os organismos considerando os estados ancestrais e derivados para a reconstrução filogenética. Na Cladística de Henning, a reconstrução filogenética que requer menos "passos evolutivos" é mais parcimoniosa em relação a outra que requer maior numero de "passos". A análise cladística ordena as sinapomorfias de modo a obter uma classificação hierárquica dos taxa, recorrendo, sempre, à solução mais simples. Esse método, por sua vez, é conhecido como máxima parcimônia. Basicamente existem três grandes conjuntos de métodos de inferências filogenéticas: Métodos de parcimônia, métodos de distância e métodos de verossimilhança (Felsenstein 1988). Os demais métodos podem ser entendidos como variações ou aperfeiçoamento destes.

O primeiro estudo a formalizar o uso da parcimônia em cladística foi de Kluge e Farris (1969). Mas, foi somente na segunda metade da década de 80, que a cladística recebeu um grande impulso como conseqüência da popularização dos programas de análise por parcimônia, como o Henning86 (Farris, 1988), PAUP (Swofford, 2000) e o PHYLIP (Felsenstein, 1989).

3.3.1 Método de Parcimônia

Parcimônia deriva do latim *parsimonia*, e é definida como o principio lógico do “mais simples”, ou “princípio da pluralidade desnecessária”. Este conceito foi introduzido por William of Ockham, e posteriormente popularizado como navalha de Ockham (Wiley, 1981). O conceito em si determina que a explicação para qualquer fenômeno deve assumir apenas as premissas estritamente necessárias à sua elucidação, e eliminar todas as que não causam qualquer diferença aparente nas predições da hipótese. Esse princípio tem aplicações em diversas áreas, como física, estatística, economia e sistemática. Em sistemática, a parcimônia, ou máxima parcimônia (MP), é a base do critério de optimalidade da cladística (Hennig, 1966), que visa a reconstrução de árvores a partir da premissa do conceito da minimização do número de mudança de estados ao longo da árvore. Ou seja, o princípio da parcimônia em cladística assume que a melhor hipótese filogenética é aquela que requer o menor numero de eventos (mudança de um estado para outro em um um caracter - substituição de nucleotídeos) para explicar a distribuição de estados observados nos nós apicais (terminais) da filogenia (ou seja o dado amostrado). Foi inicialmente introduzida para a reconstrução de topologias por Edwards e Cavalli-Sforza (1963), que o definiram como o

método da evolução mínima. Mais tarde Hennig (1966) concebeu a Sistemática Filogenética, também conhecida como Cladística.

Embora o princípio seja a minimização do numero de passos, existem diferentes modalidades que diferem quanto às restrições impostas a determinados tipos de mudanças de caracteres, como variações na pesagem e ordenação dos caracteres e construção das matrizes de mudança. Para os casos em que a probabilidade de mudança de estados do caracter é simétrica, ou seja, permite a reversibilidade de estados, os métodos de Fitch e de Wagner podem ser empregados. O algoritmo de Fitch (1971) não impõe qualquer restrição para transformações entre estados de caracteres, sendo o critério mais adequado para o tratamento de caracteres multi-estado não ordenados. Em outras palavras, cada estado pode ser derivado de qualquer outro, e em qualquer sequência, sendo considerado apenas como um passo a mais. Já o critério de Wagner (Wagner 1961) impõe restrições mínimas às mudanças de caracteres. No caso da parcimônia de Wagner, o custo das transformações ao longo das sequências é cumulativo, diferente da de Fitch, na qual o custo de qualquer mudança entre estados é igual a 1. A possibilidade de reversibilidade de caracteres assumida por esses critérios resulta em matrizes de passos completamente simétricas.

Em outros casos, quando as probabilidades de mudanças entre estados são assumidas como assimétricas, ou seja, atribui-se custos diferentes para determinados tipos de mudanças, outros critérios de parcimônia podem ser empregados. Descrito por Farris (1977), a parcimônia de *Dollo* proíbe a ocorrência de convergências, assumindo que a presença do caracter é unicamente derivada. Esse critério é restrito a certos tipos de dados comparativos, onde a probabilidade de ocorrência de perdas secundárias é maior que a de origens independentes. Ademais, o conceito da parcimônia generalizada foi desenvolvido por Swofford e Olsen (1990), e entende os diferentes modelos descritos como aplicações particulares de um método maior (generalizado). Este se caracteriza por assumir um determinado custo para cada mudança de estado, de acordo com as restrições impostas pelo modelo adotado em particular. Um parâmetro que deve ser levado em conta ao se avaliar um método de reconstrução é a consistência, ou seja a obtenção do mesmo resultado quando aumentamos o conjunto de dados, que em alguns casos é limitada pela evidência disponível.

Entre as vantagens oferecidas pelo método da parcimônia está a relativa rapidez com que a análise pode ser realizada e sua robustez quando as sequências são provenientes de taxa relativamente próximos, ou seja, quando os ramos da árvore são curtos (Lewis, 2001). Entretanto, existindo uma considerável variação nos comprimentos dos ramos, causados por diferentes taxas de substituição entre as linhagens, o método de parcimônia se torna

inconsistente, causando um artefato onde se observa a aproximação dos taxas com ramos mais longos. Esse fenômeno é conhecido como atração dos ramos longos (Felsenstein, 1988), em que homoplasias em ramos longos, durante a otimização de número de passos visando máxima parcimônia, tornam-se sinapomorfias causando agrupamentos espúrios de taxa. Em outras palavras, o erro sistemático induzido por ramos longos é causado por homoplasias quando estas tornam-se mais comuns do que as sinapomorfias. Ou seja, ramos longos, causados por taxas de evolução muito desiguais ou pelo longo tempo de divergência, que acabam sendo aproximados em uma análise porque as chances de substituições paralelas passam a ser maiores do que substituições únicas em ramos curtos. Ramos longos são especialmente prováveis em radiações rápidas e antigas, sendo menos significante ao se analisar organismos experimentando cladogênese mais recente, ou onde há preservação do sinal filogenético relativo ao processo de incorporação serial de sinapomorfias, independente de escala temporal e ou mutacional.

3.3.2 Métodos de Distância

Métodos de reconstrução por distância fazem uso de uma matriz de distância para estimar a relação entre as linhagens, podendo ser empregado tanto para sequências de DNA, como de aminoácidos (Felsenstein, 1981). Resumidamente, métodos de distância constroem uma árvore a partir de uma matriz de distância genética estimada para cada par de sequência. Essa distância é um valor único (tendência central expressa como a distâncias sítio a sítio normalizada pelo numero de sítios), estimado com base na proporção das posições em que as sequências variam (distância–p ou *p- distance*) (Felsenstein, 2004). Para a reconstrução de uma árvore, algoritmos de agrupamento ("*grouping*") ou agregação ("*clustering*") baseados na matriz de distância são utilizados.

3.3.2.1 Algoritmos de agrupamento ou agregação

O primeiro método de busca de árvores a partir de matrizes de distância eram baseados em mínimos quadrados. O *least-square* (LS), ou método dos mínimos quadrados, é uma técnica de otimização matemática que consiste em computar a mínima soma dos quadrados das diferenças entre as distâncias pareadas observadas e as estimadas para uma dada topologia, e com base nisso, escolher a topologia com o menor valor de soma. O primeiro método baseado em *least-square* foi o de Fitch-Margoliash, introduzido por Fitch e

Margoliash (1967). Com o objetivo de reduzir a inacurácia nas medidas de distância entre sequências menos relacionadas, pesos são atribuídos as sequências mais próximas, e os valores de distância são normalizados para prevenir artefatos. As distâncias calculadas por esse método são lineares, e o critério de linearidade requer que os valores esperados para os comprimento de ramos entre dois taxa deva ser igual ao valor esperado para a soma da distância de cada um desses taxa. Correção adequada das distâncias observadas é feita através de modelos de substituição de complexidade crescente, desde o modelo de Jukes-Cantor (1969) até modelos reversíveis no tempo, que consideram a as probabilidades de transição instantânea entre nucleotídeos, a composição das seqüências (expressa em frequência de nucleotídeos), a proporção de sítios invariantes e a heterogeneidade de taxas de mutação ao longo do sítios (ver seção **3.3.3.2** abaixo para um tratamento detalhado dos modelos de substituição comumente usados em reconstrução filogenética).

O método conhecido como UPGMA (*unweighted pair group method with arithmetic mean*) é outro algoritmo de clusterização por distância. Este método foi idealizado para agrupar sequências fenotipicamente similares, mas hoje é empregado para reconstruções a partir de dados moleculares, utilizando matrizes de distância genética. Este algoritmo consiste em agrupar hierarquicamente os pares de sequências com menor distância entre si, considerando uma taxa constante de evolução para as linhagens (Michener & Sokal 1957). Nesse algoritmo, as distâncias entre os taxa são calculadas como a média simples da distância entre cada um, facilitando o processo do agrupamento. A topologia gerada é ultramétrica e enraizada no táxon mais divergente, não permitindo o enraizamento opcional com outro taxa. A não ser em casos particulares onde o relógio biológico é constante entre linhagens, assumir taxa constante de evolução entre todas as linhagens, pode causar erros sistemáticos, frequentemente resultando em agrupamentos errôneos.

Uma variação do método de UPGMA é o WPGMA, (*weighted pair group method with arithmetic mean*). Basicamente, o algoritmo trabalha da mesma forma que o UPGMA, mas com a diferença de que, no lugar das medias simples calculadas entre as distancias dos taxas, as medias são "pesadas", de modo que cada taxa contribua igualmente para o resultado.

Outro algoritmo muito utilizado é o *Neighbor-Joining* (NJ), ou agrupamento de vizinhos (Saitou e Nei 1987). Diferentemente do UPGMA, o NJ não requer taxa constante de evolução entre as linhagens, fornecendo assim uma topologia com comprimento de ramos. O princípio desse método é encontrar pares de taxa que minimizem o comprimento total de ramos da árvore a cada estágio da clusterização (Saitou e Nei 1987). Em uma árvore inicial não enraizada (em formato de estrela), é clusterizado o par de taxa que possui menor

distância entre si (ou seja, os "vizinhos mais próximos") (Figura 2). Com eles, é criado um novo nó na árvore juntando os dois nós mais próximos (os dois nós estão ligados por seu nó ancestral comum). O próximo passo então, é calcular a distância de cada um dos nós do par para seu nó ancestral.

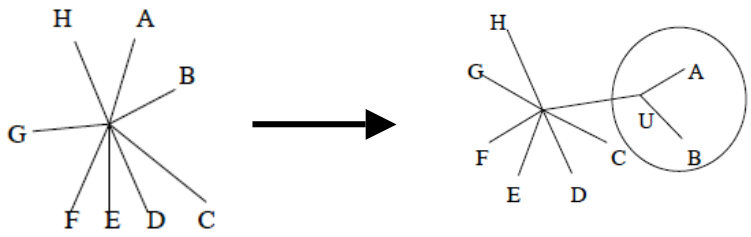


Figura 2. Esquema da árvore-estrela com oito taxa que é construída e utilizada para iniciar o agrupamento por NJ., e a topologia resultante da primeira junção dos vizinhos

A distância entre os demais nós também é calculada em relação ao seu nó ancestral. O algoritmo então reinicia considerando agora o par de vizinhos como um único táxon e usando as distâncias calculadas na etapa anterior (os nós terminais são substituídos por seus nós ancestrais, e o nó ancestral é então tratado como um nó terminal). Estas junções são realizadas sequencialmente, até que a todos os taxa sejam adicionados. Ao final, o número de junções será N−2 (N = número de linhagens). Por ser baseado no critério de mínima evolução, a topologia que possui o menor valor resultante da soma dos comprimentos de ramos, é a preferida para o próximo passo (adição de um novo nó). Dessa forma, a topologia final representará o menor valor resultado da somatória de todos os comprimentos de ramos.

Embora métodos de distância sejam computacionalmente eficientes, e permitam correções com modelos de substituição de nucleotídeos, grande parte da informação contida nas sequências é perdida, pois toda a informação de variação de divergência observável ao longo dos sítios é reduzida a um único valor que sumariza a distância entre seqüências (Steel *et al*., 1988). Além disso, por serem métodos aditivos, as reconstruções feitas por algoritmos UPGMA ou NJ produzem apenas uma reconstrução, não permitindo a avaliação de outras possibilidades de topologias dentro do universo de árvores possíveis.

3.3.3 Máxima verossimilhança (MV)

O conceito de verossimilhança (MV) foi proposto por Fisher em 1922 (Fisher, 1922). Somente mais tarde a idéia de utilizar métodos de máxima verossimilhança para inferências filogenéticas foi introduzida por Cavalli-Sforza e Edwards (1967), com dados de frequência gênica. Anos depois, Felsenstein desenvolveu um algoritmo prático para a reconstrução de uma árvore filogenética a partir de MV (Felsestein, 1981).

Abaixo irei exemplificar e definir mais precisamente a maximização da verossimilhança, mas neste momento vale avaliar a diferença entre probabilidade frequentista e verossimilhança. A probabilidade P dada a hipótese H geradora dos dados D observáveis segue uma distribuição probabilística conhecida, é expressa como $P(D|H)$. Por sua vez a verossimilhança L da hipótese H dado um modelo específico para os dados obtidos D é expressa como $L(H|D)$ (Edwards, 1992). $L(H|D)$ é proporcional a $P(D|H)$ mas a constante de proporcionalidade é arbitrária. Ademais, para $P(D|H)$, H é constante e D é variável, enquanto que para $L(H|D)$, para um conjunto de dados D constante, buscamos a H dentre várias que maximiza L. A arbitrariedade ou desconhecimento da constante de proporcionalidade, pela qual transformamos L em P, não impede o uso da mesma definição de L para variáveis discretas e contínuas, cuja principal utilidade é a comparação de verossimilhanças. Ainda mais, dado o mesmo conjunto de dados D, o mesmo modelo específico e para diversas H ($H1$, $H2$,..., Hn), temos a mesma constante de proporcionalidade, o que não é limitante pois não se propõe uma comparação absoluta entre hipóteses diferentes a partir de dados diferentes (Edwards, 1992).

O método da verossimilhança pode ser exemplificado por meio de um estudo de caso probabilístico simples (Huelsenbeck e Crandall, 1997), como por exemplo, em um experimento de lançamentos independentes de uma moeda e observação do resultado. Neste experimento com as moedas, dois resultados são possíveis, cara (C) ou coroa (K), e no caso de a moeda não ser viciada, cada face da moeda tem a probabilidade de ocorrência de ½. Porém, na maioria das situações práticas, os eventos simples do espaço amostral não são equiprováveis e não podemos calcular probabilidades usando a definição clássica. Neste caso, vamos calcular probabilidades como a frequência relativa de um evento.

Considerando-se as premissas de independência entre os experimentos e obtenção de probabilidades constantes, pode-se atribuir à probabilidade desconhecida de cara a quantidade p, e à probabilidade desconhecida de coroa a quantidade $1-p$, sendo que o valor de p varia no intervalo [0,1]. Supondo que a moeda tenha sido lançada onze vezes, o seguinte resultado

CCKKCKCCKKK (cinco caras e seis coroas) foi observado. Com base agora neste resultado, podemos então calcular a probabilidade de ocorrência desse dado. Ou seja, calcula-se a chance de se observar esse resultado para diferentes valores da quantidade *p* no intervalo [0,1].

$$L = \text{Prob}\,(D|p) = pp\,(1-p)\,(1-p)\,p\,(1-p)\,pp\,(1-p)\,(1-p)\,(1-p) = p^5\,(1-p)^6$$

Deve-se ressaltar que existe um resultado conhecido CCKKCKCCKKK e um modelo probabilístico que o descreve: $p^5\,(1\text{-}p)^6$. O valor de *p* no intervalo [0,1] para o qual se tem a maior chance de se observar o resultado 5 C e 6 K é ~0,454, ou seja, 5/11. Esse mesmo valor pode ser obtido analiticamente, plotando *p* contra *L*, onde temos as probabilidades do mesmo dado D para diferentes valores de *p* (Figura 3). Para isso, faz-se a derivada da primeira função igual a zero, ou seja, $df(p)/dp = 0$. Assim, $f(p) = p^5 \times (1-p)^6$. Para a condição de máxima verossimilhança, temos um *p* de 4545 , $0 \cong p$. Este valor da quantidade *p* é a que proporciona maior chance de se observar a seqüência CCKKCKCCKKK no experimento dos lançamentos.

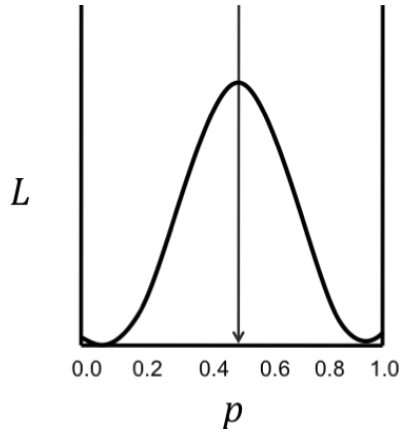


Figura 3. Likelihood (eixo y) para a probabilidade (eixo x) de caras em uma serie de 11 lançamentos independentes, resultando em cinco caras e seis coroas. $p = 5/11$

Como visto acima, o método da verossimilhança possui três elementos: dados D, hipóteses variáveis *H1*, *H2*, ... *Hn* e um modelo probabilístico específico que geraria *D*|*H*. Os dados são os resultados observados para um determinado experimento e o modelo probabilístico é sempre fornecido. Os dados *D* consistem de observações $x1,...,xn$ cuja distribuição tem função de probabilidade dada por $f(X, \Theta 1,...,\Theta k)$. A função *f* tem forma conhecida e depende de parâmetros desconhecidos $\Theta 1,...,\Theta k$ e o conjunto de valores

admissíveis do parâmetro é denominado Espaço Paramétrico. Para obter-se estimativas de um parâmetro, são utilizados os dados disponíveis e uma função que opera sobre esses dados. Essa função é um mapeamento do espaço de dados para o espaço de valores admissíveis do parâmetro. Dessa forma, um estimador é uma função no espaço das observações ($x1,...,xn$) com valores no espaço paramétrico ($\Theta1,...,\Theta k$). A verossimilhança (do inglês, *Likelihood*) é proporcional à probabilidade dos dados, D, dada uma hipótese H: $L = P(D|H)$. Voltando ao exemplo do lançamento da moeda, os dados são representados pelo resultado CCKKCKCCKKK, o modelo probabilístico é dado por $f(p) = p^5(1-p)^6$ e as hipóteses possuem valores dentro do intervalo [0,1] para o parâmetro p. Assim, espera-se encontrar o valor de p que maximize a verossimilhança de H assumindo um modelo probabilístico. Essa estimativa de verossimilhança máxima do parâmetro Θ (ou p, no caso do exemplo) à uma dada amostra observada é obtida encontrando o valor de Θ que maximiza a função de verossimilhança. A função pode ser definida como:

$$L(\Theta) = f(x1,\Theta) \times f(x2,\Theta) \times ... \times f(xn,\Theta)$$

sendo $x1$, $x2$,..., xn uma amostra aleatória de uma distribuição dada por $f(X, \Theta)$, e Θ um parâmetro desconhecido.

3.3.3.1. Busca de árvores através de ML

O método de máxima verossimilhança (MV), ou *maximum likelihood* (ML), para a reconstrução de árvores utiliza matrizes de probabilidade instantâneas de substituições entre nucleotídeos (ou seja, modelos específicos), as quais explicitam modelos evolutivos para a inferência dos estados ancestrais a partir dos estados observados (Harris *et al*., 1998). Dessa forma, para as reconstruções, é possível avaliar as probabilidades de cada substituição observada no conjunto de dados, sitio a sitio. Os parâmetros considerados, entretanto, não consistem apenas no formato da topologia propriamente dita, mas também nos comprimentos dos ramos para cada reconstrução obtida. Dessa forma, a verossimilhança é maximizada a partir de correções nos comprimentos de ramos (Nei e Kumar, 2000). Finalmente, a partir de uma distribuição de valores de probabilidade (lnL) atribuído para cada reconstrução, a topologia final passa a ser a que possui maior verossimilhança.

O método requer a especificação do dado e de um modelo de substituição que descreva o dado dada uma topologia. A probabilidade de observar o dado sob a luz de um

modelo pré-assumido vai mudar dependendo das perturbações topológicas e alterações dos comprimentos de ramos durante a busca pela "melhor" (mais verosímil) árvore. Conceitualmente, a função verossimilhança em filogenia segue os mesmos passos do exemplo do lançamento de uma moeda (Felsenstein, 2004; Huelsenbeck e Crandall, 1997), onde os dados são, novamente, interpretados como variáveis randômicas. Para o exemplo das sequências de DNA, a função pode ter uma distribuição multinomial, ao invés de uma função binomial. A distribuição multinomial é uma generalização da binomial e tem a forma:

onde n_i é o numero de observações do i-ésimo padrão de sitio, p_i é a probabilidade do sitio i ocorrer, s é o numero de padrões de sítios possíveis (no caso de nucleotídeos, $s = 4$). A topologia da árvore e o comprimento de ramos constituem a hipótese a ser verificada dado um modelo específico de substituição de nucleotídeos ou aminoácidos (matriz Q instantânea de substituição). A verossimilhança, portanto, não se refere à probabilidade de que a árvore seja correta, mas sim, a probabilidade de se observar o dado, dado uma hipótese (topologia) e um modelo (matriz Q).

Diferentemente dos métodos de distância onde a divergência entre um par de sequências é interpretado como um único valor, a MV estima a divergência para cada caracter (sítio) ao longo da árvore. Isso torna imprescindível a utilização de modelos de evolução que possam predizer as probabilidade de ocorrência de cada transformação de um estado para outro. Por exemplo, quando a divergência entre as sequências observadas é baixa, a distância (p) é um estimador fiel da real distância genética. Porém, quanto maior é a divergência entre as sequências, mais a distância observada pode ser subestimada, o que leva a necessidade de uma correção estatística (Figura 4). Essa correção é feita por modelos de substituição de nucleotídeos.

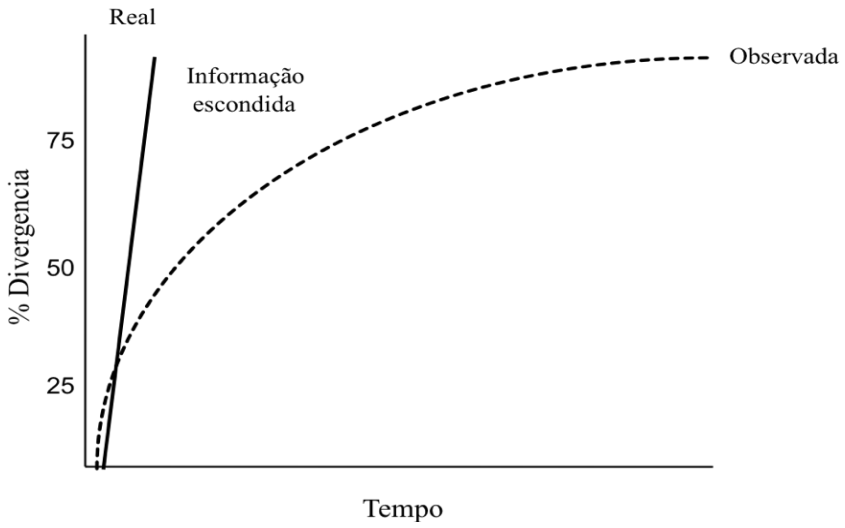


Figura 4. Representação gráfica de divergência real e observada ao longo do tempo para sequências de DNA. Notar que quanto maior a distância real entre elas ao longo do tempo, mais informações acerca da divergência é perdida.

Portanto, para a inferência filogenética por MV em particular e por outros métodos quantitativos em geral é crucial o uso de modelos de substituição específicos adequados.

3.3.3.2 Modelos de substituição

O ponto de partida para iniciar uma reconstrução por MV é a escolha de um modelo de substituições de nucleotídeos adequado para os dados. A partir da observação de sequências atuais, é impossível determinar todas as substituições realmente ocorridas durante a evolução das mesmas. Isso ocorre pelo fato de que em uma única posição pode ter havido mais que uma substituição. Os modelos de substituição servem para recuperar a informação perdida (em termos de substituições de nucleotídeos) ao longo do tempo, assim como corrigir para a ocorrência de múltiplas substituições em um mesmo sitio. Modelos de evolução são utilizados para reconstruções filogenéticas feitas por diferentes métodos. Por exemplo, em métodos de distância, a divergência entre as sequências pode ser corrigida através de modelos que levem em conta as probabilidades de diferentes substituições em cada sítio. Porém, é na MV que a utilização de modelos faz-se mais necessária. Como já discutido anteriormente, a MV reconstrói uma filogenia baseada nas mudanças dos caracteres (cada posição dos nucleotídeos na sequência é um sitio diferente, e é considerado um caracter independente). Assim, diferentemente dos métodos de distância onde a divergência é transformada em um único valor, a MV estima uma distância a partir de cada caracter.

O primeiro modelo de substituição, e também o mais simples, foi criado por Jukes e Cantor (1969), e é conhecido como JC69. Esse modelo assume que as freqüências de base são iguais na sequência de nucleotídeos ( ) e que a taxa de substituições de um nucleotídeo para outro é a mesma para qualquer tipo de mudança. Pode ser representado da seguinte forma, como na Figura 5:

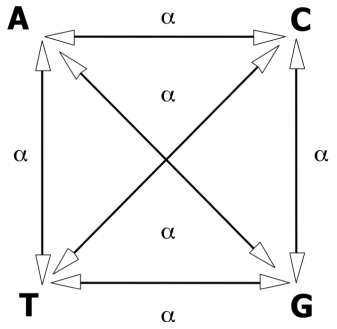


Figura 5. Esquema da matriz de substituição assumida pelo modelo JC69, onde todas as taxas de substituição α (transversão e transição) são iguais.

O modelo Kimura 2-parâmetros (Kimura, 1980) é praticamente uma extensão do JC69, caracterizando-se por distinguir entre dois tipos de substituição: purinas e pirimidinas. Dessa forma, o número de substituições por sitio (*d*) pode ser estimado a partir da seguinte fórmula:

$$d = -1/2 \ln(1-2P-Q)\sqrt{1-2Q}$$

onde *P* é a proporção de transições observadas e *Q* é a proporção de transversões observadas, entre duas sequências.

Hasegawa, Kishino and Yano (1985) sugeriram um modelo que posteriormente Tamura e Nei (1993) extenderam, o modelo HKY85. Esse modelo leva em conta distintas taxas de tranversões e transições, como o Kimura 2 parâmetros, e considera frequências independentes de nucleotídeos na sequência. Outra categoria de modelos de substituição são os *Time reversible,* ou seja, assumem que a freqüência relativa de cada caracter não muda ao longo do tempo. Isso é importante pois, ao analisar um dado biológico, a intenção é reconstruir o estado ancestral, o qual não se tem informação. O modelo GTR (*generalized time reversible*) foi descrito em 1986 por Tavaré (Tavaré, 1986). Esse modelo assume uma

matriz de substituição assimétrica, ou seja, (i) A muda para T a mesma taxa que T muda para A; (ii) cada sítio pode evoluir a diferentes taxas e (iii) os nucleotídeos ocorrem a freqüências distintas. Outros parâmetros podem ser adicionados ao GTR, como proporção de sítios invariantes (GTR+I) e assumindo variação da distribuição *gama* entre os sítios (GTR+I+Γ).

Atualmente, para a escolha do melhor modelo filogenético de substituição de *DN*A em um conjunto de sequencias, é utilizado o programa Modeltest v.3.7 (Posada e Crandall, 1998). Este software compara de forma hierárquica diferentes modelos de substituição de nucleotídeos, onde 56 modelos evolucionários são testados, e matrizes de distância filogenética são construídas a partir destes modelos. Um arquivo denominado *modelscore* é então gerado, e um valor de verossimilhança (*likelihood* value, lnL) é atribuído para cada um dos 56 modelos testados. Dessa forma, como a análise é feita de forma hierárquica, sempre um modelo complexo é comparado a um modelo mais simples, com o objetivo de se verificar se a inclusão de um parâmetro a mais (em relação ao mais simples) explica melhor o dado. Dessa forma, os modelos são comparados, valores de lnL são atribuídos e a escolha é feita com base na premissa de que a adição de mais parâmetros não resulta num incremento no valor de likelihood em relação ao modelo anterior. Essa comparação entre valores de lnL atribuídos a diferentes modelos é feita por um teste denominado *Likelihood ratio test* ou LRT. No entanto cabe considerar que este procedimento é baseado em comparar modelos dada uma filogenia possivelmente sub-ótima.

A razão de verossimilhança LRT entre duas hipótese é um teste estatístico que compara diferentes hipóteses frente a um dado, dado um modelo específico. Isso porque o dobro da diferença entre as verossimilhanças de 2 modelos aninhados, expressa por 2(lnL1-lnL2) segue uma distribuição χ^2, com um número de graus de liberdade dado pela diferença entre o número de parâmetros dos dois modelos. Por exemplo, na comparação entre HKY85 e GTR, o GTR difere do primeiro pela adição de quatro parâmetros adicionais. Assim, se os valores de lnL para ambos são respectivamente: -lnL = 1787.08 (HKY85) e -lnL = 1784.82 (GTR), logo, LR = 2 (1787.08 - 1784.82) = 4.53. Tendo em mente que a adição de quatro parâmetros implica em quatro graus de liberdade, temos então que o valor critico para assumir que há diferença significante entre eles é 9.49 (P = 0.05). Portanto, um LR de 4.53 não exclui o modelo mais simples. Como sugerido acima, embora ainda muito utilizado, há críticas em relação ao Modeltest. Estas devem-se principalmente ao fato de que toda a análise é feita com base em uma única arvore do tipo NJ, previamente construída, implicando que todos os parâmetros sejam estimados em cima de uma árvore simples. A alternativa seria

a utilização de um programa que estime e otimize os parâmetros durante as interações. Recentemente, foi criado o programa de reconstruções filogenéticas por MV chamado Garli (Genetic Algorithm for Rapid Likelihood Inference- Garli) (Zwickl, 2006). Este programa utiliza um algoritmo que otimiza os parâmetros iniciais a cada geração, levando a um incremento ou queda no valor de lnL a cada nova reconstrução. O critério de otimização busca topologias e os valores de modelos de substituição que maximizam a verossimilhança, e o programa automaticamente para a busca após um determinado número de gerações (fornecido pelo usuário), durante o qual não ocorrer incremento no valor de lnL ao se perturbar topologias, concomitante com a otimização de modelo e ajuste comprimento de ramos.

3.3.4 Algoritmos de busca

As técnicas de busca empregam algoritmos que procuram árvores em um universo de árvores possíveis para um determinado conjunto de taxa. Essas buscas podem ser classificadas em dois tipos: (i) exatas, que garantem a otimização e são geralmente demoradas e (ii) aproximadas, que não garantem a otimização, mas são muito mais rápidas, permitindo sua utilização para dados que possuem múltimplos taxa. Dentre as exatas, destacam-se os métodos exaustivos e *branch-and-bound*.

3.3.4.1 Busca exaustiva

São investigadas todas as topologias não-enraizadas possíveis para um determinado conjunto de taxa. Acontece que, mesmo com poucos taxa, o número de árvores possíveis é astronômico em uma busca exaustiva (Tabela 1), e pode ser calculado através da fórmula $N = (2n - 5)! / 2^{n-3}(n-3)!$, sendo N o numero de árvores não enraizadas.

Tabela 1. Numero de árvore enraizadas e não-enraizadas Possíveis para cada numero de taxa em um conjunto de dados.

# de taxa	# árvores enraizadas	# árvores não- enraizadas
2	1	1
3	3	1
4	15	3
5	105	15
6	945	105
7	10.395	945
8	135.135	10.395
9	2.027.025	135.135
10	34.489.707	2.027.025
15	2E14	8E12
20	8E21	2E20
50	2.8E76	3E74

Um valor ótimo (dependendo do critério de otimização) é associado a cada uma dessas topologias e aquela(as) que melhor atenderem o critério, são escolhidas. No caso da parcimônia, o tamanho total da árvore é calculado para cada uma das possíveis topologias, e a menor é escolhida. Essa técnica garante atender o critério de otimização, mas utiliza muita memória computacional, demanda tempo, e é geralmente inviável para mais de 12 taxa.

3.3.4.2 Branch-and-bound

Similar à busca exaustiva, mas muito mais rápida, o algoritmo de *branch and bound* foi aplicado pela primeira vez em buscas de árvores por parcimônia (Hendy e Penny 1982). A medida em que se percorre um caminho para a busca da melhor árvore, ajusta-se um limite (*bound*), descartando caminhos que não apresentem nenhuma chance de chegar ou superar o ótimo. Com isso, economiza-se muito tempo de computação, mas ainda assim, demanda bastante tempo. Isso é feito construindo-se uma árvore (não necessariamente ótima, mas sub-ótima) por algum algoritmo rápido (NJ por exemplo). Atribui-se um valor de *score* pra essa árvore, e o algoritmo começa a busca tendo esta como ponto de partida. Ao visitar o espaço

amostral para a busca da melhor árvore, as árvores que não apresentarem um *score* melhor do que a inicial não serão visitadas.

3.3.4.2 Busca Heurística

Os algoritmos de busca heurística são os mais empregados em inferência filogenética. Estes partem de uma estrutura inicial e fazem alterações nessa estrutura de modo a atingir o ótimo, consistindo de duas etapas. A primeira é a construção de uma árvore inicial e a segunda consiste no rearranjo dos ramos (*branch-swapping*), buscando a otimização. A árvore inicial pode ser obtida por diferentes métodos, (i) adição sequencial de taxa (*stepwise addition*), (ii) por agrupamento de vizinhos (NJ), ou (iii) pode ser arbitrária (informada pelo usuário). O método mais comum é a adição sequencial aleatória, com múltiplas replicações, visando aumentar as chances de se atingir o ótimo global. Os rearranjos dos galhos, por sua vez são feitos por diferentes algoritmos.

O algoritmo heurístico NNI (*nearest neighbor interchange*, ou troca de vizinhos mais próximos) é um algoritmo de perturbação topológica que consiste em trocar de lugar os taxa que aparecem separados por apenas um ramo interno e reconstruir a topologia, atribuindo um valor de *likelihood* (lnL). Ao final, a topologia escolhida será a que maximizar o valor de lnL. A Figura 6 demonstra como o algoritmo trabalha. Considere o ramo interno da primeira árvore uma árvore binária, não enraizada, que tem quatro "sub-árvores" a ela conectada. O algoritmo simplesmente desconecta essas sub-arvores, e rearranja-as em outras posições. Isso é feito para cada ramo interno da árvore, até que não há melhora no *score* da árvore, de acordo com o critério de optimalidade escolhido.

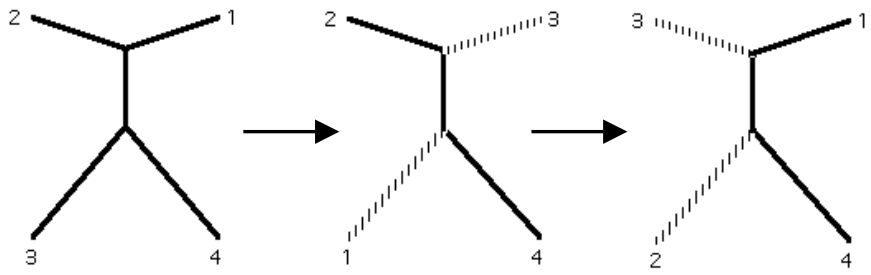


Figura 6. Representação do processo da "troca de vizinhos" do algoritmo *nearest neighbor interchange*, onde um ramo interno é desfeito e as quatro sub-árvores conectadas a ele são isoladas e re-conectadas em outras posições.

Um segundo, e mais elaborado algoritmo de perturbação topológica é chamado *Subtree Prunning and Regrafting* (SPR), descrito por Swofford e Olsen (1990). Este consiste em remover um ramo inteiro de uma arvore (seja interno ou externo), que esteja conectado a outros ramos (sub-arvore). Essa sub-árvore é então re-inserida em todos os possíveis lugares da árvore original. A Figura 7 mostra uma árvore inicial com onze taxa, dos quais um ramo contendo cinco taxa são removidos e inseridos em uma das nove possíveis posições da arvore remanescente. Ao final, para onze taxas, existirão 288 possíveis arvores sob o algoritmo SPR.

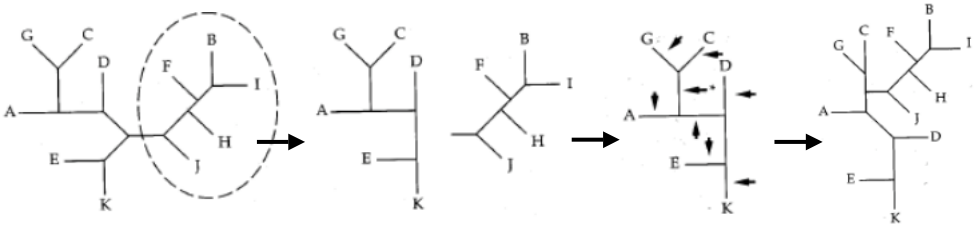


Figura 7. Rearranjo do tipo SPR (*Subtree pruning and regrafting*). As setas pequenas indicam as nove possibilidades de onde os ramos extraídos (dentro do circulo na primeira arvore) podem ser re-inseridos.

Outro algoritmo utilizado para perturbações topológicas é o TBR (*tree bisection and reconnection*), e este é considerado uma extensão do SPR. Nesse algoritmo, um ramo interno é quebrado e os dois fragmentos resultantes são considerados duas árvores distintas. Dessa forma, todos as possíveis conexões são feitas a partir dessas duas arvores. Para os mesmos onze taxa da figura anterior, teremos agora 296 possíveis arvores construídas a partir do algoritmo TBR (Figura 8).

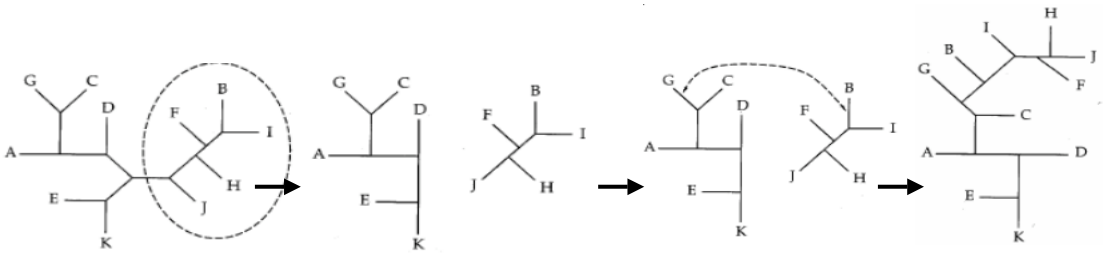


Figura 8. Rearranjo do tipo TBR (*tree bisection and reconnection*). O ramo dentro do círculo é destacado, e ambas as árvores resultantes são reconectadas pela adição de ramos em todos os possíveis nós.

Como já descrito anteriormente, o universo de árvores possíveis é muito grande, e isso pode ser representado por uma paisagem (*landscape*) de árvores onde as várias topologias se encontram distribuídas. Como é praticamente impossível percorrer toda essa paisagem, as buscas heurísticas utilizam estratégias para visitar apenas algumas, de modo a se atingir o ótimo. Um dos grande problemas ocorre em função da topografia dessa paisagem, que não é plana. Ou seja, existem "ilhas" de árvores localmente ótimas (picos menores dos morros, Figura 9), e apenas um pico que corresponde as globalmente ótimas (pico mais alto). Assim, técnicas de busca heurística não garantem encontrar a melhor árvore do maior pico, mas sim, a melhor árvore de um dos picos dos quais a busca foi realizada. Isso acontece devido ao fato de que, por causa do próprio algoritmo, ao amostrar uma árvore em um determinado pico, não é possível a transição para outro pico sem que se passe pelo 'vale' (onde as piores árvores se encontram).

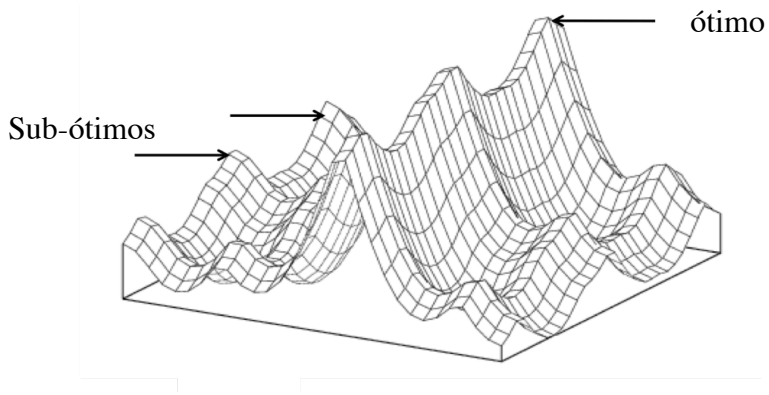


Figura 9. Superfície representando o universo de árvores possíveis, onde os picos representam as reconstruções com valores melhores, e os vales representam as reconstruções não ótimas.

3.3.5 Testando diferentes hipóteses

Alguns estudos de simulação sugerem que a técnica de verossimilhança é mais robusta e acurada que outros métodos de inferência (Kuhner e Felsenstein 1994; Huelsenbeck, 1995). Mesmo assim, a técnica de verossimilhança permite testar se a hipótese (a melhor árvore) é realmente a mais verossímil. Como já descrito no item **3.3.3.2**, o LRT (likelihood ratio test) é um teste estatístico que permite comparar entre duas hipóteses frente a um dado (modelos de substituição, árvores diferentes, relógio molecular, etc). O valor de *L* (likelihood) é maximizado tanto sob a hipótese nula como a alternativa, dessa forma, o teste fornece uma medida de suporte para o dado de uma hipótese, versus a outra. O teste é feito atribuindo-se os valores de lnL para as duas árvores, e calculando a razão entre a árvore real e

a hipótese nula (Z). Assim, se Z > 1, então o dado é mais provável sob a hipótese nula. Da mesma forma que a hipótese alternativa é mais provável que a nula se Z<1. Normalmente porém, a hipótese nula é uma árvore com probabilidade bem próxima a hipótese a ser testada, ou seja, duas árvores com valor de *L* bem parecidas. Nesse caso, o valor de Z será sempre <1 e o -2logZ segue uma distribuição X^2 com *q* graus de liberdade, onde *q* é a diferença no número de parâmetros entre as duas hipóteses. Alternativamente, a probabilidade de observar um dado Z se a hipótese nula for a correta (significância) pode ser calculado por simulações de Monte Carlo. Esse teste exige alguns parâmetros, como por exemplo, que a hipótese nula esteja dentro do espaço amostral que contem a hipótese alternativa.

Embora o LRT seja um teste bastante utilizado, existem alguns problemas na sua utilização para certos casos. Vamos tomar como exemplo o teste de LRT entre duas árvores completamente resolvidas (sem politomias) e não enraizadas. Ambas as árvores terão o mesmo numero de graus de liberdade, implicando que a diferença entre eles será zero, portanto, sendo impossível calcular a diferença de *L* entre elas. Uma alternativa sugerida por Felsenstein, (1988) seria um teste mais conservativo, ou seja, comparar cada árvore resolvida com uma não resolvida. Por exemplo, supondo que exista uma árvore com o seguinte agrupamento ((Humano, Chimpanzé), Gorila). A comparação poderia ser entre esta e uma alternativa trifurcada (Humano, Chimpanzé, Gorila). Entretanto, isso não leva em consideração a hipótese ((Chimpanzé, Gorila), Humano). Assim, o próprio Felsenstein considera que isso não representa o melhor teste que pode ser feito para avaliar hipóteses dentre diferentes alternativas.

3.3.5.1 Técnicas de Reamostragem paramétricas e não paramétricas

Em inferências estatísticas, existem muitos problemas cujas soluções analíticas não podem ser determinadas. Reamostragem não usa a distribuição de probabilidades assumida, mas calcula uma distribuição empírica de estatísticas estimadas. Criando múltiplas amostras da amostra original, a reamostragem requer apenas poder computacional para estimar um valor de uma estatística para cada amostra. A diferença entre os vários métodos de reamostragem é se as amostras são extraídas com ou sem reposição. Os métodos de reamostragem mais conhecidos para avaliar o suporte de ramos em uma filogenia são o *Jackknife* e o *Bootstrap*, os quais diferem na maneira como eles obtém a amostra.

3.3.5.1.1 Bootstrap e Jackknife

As técnicas de reamostragem *Jackknife* e *Bootstrap* são técnicas estatísticas para, empiricamente, se estimar a variância de uma estimativa (Felsenstein 2004). Elas permitem precisamente usar uma amostra para estimar a quantidade de interesse através de uma estatística e avaliar as propriedades da distribuição dessa estatística. Sua aplicação em filogenia começou em 1982, com os trabalhos de Mueller e Ayala (1982). O *Jackknife* trabalha recalculando o valor da estatística de interesse em cada uma de N pseudo amostras de tamanho *n*-1, formadas a partir de todas as observações da amostra original, exceto a primeira. Assim, a segunda pseudoamostra exclui a segunda observação, e assim por diante até a n-ésima. Em outras palavras, o *jackknife* é uma técnica de reamostragem sem reposição, sempre de tamanho n-1. A crítica a esse método porém é devido ao fato de que a remoção sempre de apenas um sítio tem pouco impacto estatístico durante a técnica de reamostragem. E isso fica mais critico quanto maior for o alinhamento, pois uma vez que a perturbação ocorrerá sempre com n-1, o impacto na remoção de apenas um sitio por vez em uma sequência longa será quase insignificante.

O *bootstrap* foi introduzido por Efron (1979), e é um método mais versátil que o *Jakknife*. Este é um conjunto de técnicas que obtém informações acerca das características da distribuição de algum parâmetro de variável aleatória. Em filogenia, um alinhamento de sequências com número de sítios (nucleotídeos) *n* representa o conjunto de dados observados, onde os diferentes sítios de um alinhamento são considerados para a reamostragem. O *bootstrap* existe tanto na forma paramétrica como não paramétrica, dependendo do conhecimento do problema. O não paramétrico gera um número X de árvores, construídas a partir de um conjunto de dados onde alguns caracters são removidos, e um número igual de outros caracteres ocupam o espaço, ou seja, amostragem com reposição. A premissa do *bootstrap*, assim como do *jakknife*, é a independência dos caracteres. Ao determinar a quantidade de vezes que cada ramo aparece na árvore em uma determinada posição, é atribuído um valor de confiabilidade da repetibilidade do observado. Uma árvore final então pode ser construída representando a sumarização do que foi mais frequentemente observado nas repetições. Essa árvore é chamada de *majority rule consensus tree*, ou seja, a árvore que representa o consenso da maioria. A maior critica feita a técnica de *boostrap* é que assumir a independência total dos caracteres, sendo que há a remoção e substituição de vários sitios de uma só vez pode causar algum tipo de viés na amostra, visto

que biologicamente, a distribuição dos caracteres na sequência não é igual e completamente independente.

No caso do paramétrico, quando se tem informação suficiente sobre a forma da distribuição de dados, a amostra *bootstrap* é formada realizando-se amostragem a partir da distribuição paramétrica que originou os dados. Ou seja, os *datasets* que são construídos para as réplicas são obtidos por simulações a partir da melhor árvore, e não simplesmente por reamostragem de colunas a partir do dado original.

3.3.6 Análise Bayesiana

Um dos grandes problemas das inferências filogenéticas sempre esteve relacionado ao grande numero de árvores possíveis que podem descrever a relação entre os taxa que se está analisando. Isso é especialmente agravado quando se considera os diferentes processos de evolução entre os taxa e também entre cada sitio, levando frequentemente a obtenção da árvore errada. Dessa forma, métodos de reconstrução que explicitam modelos de substituição, corrigindo assim para o problema de múltiplas substituições em um mesmo sitio, ou variações nas taxas dentro de um mesmo *dataset* são mais indicados para corrigir inconsistência estatística. Um exemplo de método de inferência que leva isso em conta é o já discutido método da verossimilhança. Outro método que leva em consideração a variação de taxas dentro de um conjunto de sequências é o método Bayesiano. O método não procura a árvore ótima, mas sim, um conjunto de árvores que possuem probabilidade de representar os dados, dado um modelo de evolução e premissas admitidas *a priori*. Em outras palavras, conhecimentos *a priori* são utilizados para prever hipóteses *a posteriori*. As aplicações do método Bayesiano em filogenética não se destinam apenas a reconstruções de árvores. Trabalhos foram feitos com estudo de variações entre taxas de substituição entre ramos (Huelsenbeck *et al*., 2000), tempo de divergência entre linhagens (Thorne *et al.* 1998), relógio molecular (Suchard *et al.,* 2001; Romano *et al*., 2008), e assim por diante.

A técnica de reconstrução por inferência Bayesiana foi proposta em 1996 (Rannala & Yang, 1996), mas somente nos últimos anos é que esta metodologia tem sido mais amplamente empregada. Inferência filogenética a partir de métodos bayesianos se baseiam em uma medida chamada de probabilidade posterior. Isso pode ser expresso em uma ligeira modificação do teorema de Bayes onde este é utilizado para combinar a probabilidade a *priori* de uma filogenia (Pr[tree]) com a verossimilhança (Pr[Data|Tree]) para produzir uma probabilidade posterior de uma árvore (Pr[Tree|Data].

$$Pr[Tree \mid Data] = Pr[Data \mid Tree] x Pr[Tree] / Pr[Data]$$

A probabilidade posterior agora pode ser interpretada como a probabilidade de que a árvore seja a correta (considerando todas as condicionais). Além do conjunto de sequências, um modelo de substituição é dado (pode ser por exemplo escolhido pelo programa *Modeltest*) e premissas são assumidas a *priori*. Essas premissas dizem respeito ao conhecimento prévio do problema a ser tratado, como por exemplo distribuição de comprimento de ramos, informação temporal acerca da origem das amostras, tempo máximo ou mínimo que essas sequências estão divergindo umas das outras e assim por diante. Embora a probabilidade posterior pareça fácil de formular, envolve a sumarização de todas as árvores possíveis, considerando variações na posição e comprimento de cada ramo na árvore, assumindo *a priori* que todas as árvores são igualmente prováveis. Ou seja o denominador será uma somatória, e deve ser calculado assumindo um número astronômico de possibilidades. Isso inclusive, já foi discutido anteriormente (item 3.3.4.1), onde o numero de arvores possíveis cresce exponencialmente de acordo com o numero de taxa. Analiticamente, isso é impossível de calcular. A opção oferecida por esse método então é visitar por Monte Carlo, via cadeia de Markov (MCMC), em fase estacionária, uma grande quantidade (preferívelmente) de árvores do universo amostral.

3.3.6.1 Método Monte Carlo e cadeia de Markov

O método de Monte Carlo (MMC) é um método estatístico utilizado em simulações estocásticas, como forma de obter aproximações numéricas de funções complexas ou a exploração de universos probabilísticos desconhecidos. Envolve a geração de observações de alguma distribuição de probabilidades e o uso da amostra obtida para aproximar a função de interesse. A idéia do método é escrever a integral que se deseja calcular como um valor esperado. A idéia do MCMC é literalmente varrer o universo de topologias possíveis, mas mais próximas à melhor árvore, avaliando a probabilidade posterior de cada uma delas frente aos parâmetros conhecidos.

A definição da propriedade Markoviana, também chamada de memória markoviana, é que os estados anteriores são irrelevantes para a predição dos estados seguintes, desde que o estado atual seja conhecido. Uma cadeia de Markov é uma seqüencia X1, X2, X3, ... de variáveis aleatórias. O escopo destas variáveis, isto é, o conjunto de valores

que elas podem assumir, é chamado de espaço de estados, onde Xn denota o estado do processo no tempo n. Se a distribuição de probabilidade condicional de Xn+1 nos estados passados é uma função apenas de Xn, então:

$$Pr(X_{n+1}=x|X_0, X_1, X_2, ..., X_n) = Pr(X_{n+1}=x \mid X_n),$$

Onde x é algum estado do processo. Em filogenia, o algoritmo MCMC envolve duas etapas: (i) Uma árvore é proposta por perturbação estocástica de uma dada árvore, e (ii) esta árvore pode ser aceita ou rejeitada, de acordo com o valor de probabilidade posterior a ela atribuído. Se a árvore é aceita, então ela é o ponto de partida para a perturbação seguinte, de acordo com o algoritmo de Metropolis-Hasting (Metropolis 1953; Hastings, 1970).

O algoritmo de Metropolis-Hasting tem como objetivo determinar valores esperados de propriedades do sistema simulado, através de uma média sobre uma amostra. Em inferências bayesianas, este algoritmo é adaptado para permitir o cálculo da probabilidade a *priori* de uma arvore Prob(T) e os likelihoods Prob ($D|T$). A taxa de 'aceitação" é a taxa entre as probabilidades da arvore proposta e da arvore atual, multiplicada pela razão de likelihood entre essas arvores. Para visualizar como funciona esse algoritmo em filogenia, imagine todas as arvores possíveis a partir de um conjunto de dados, conectadas umas as outras como se fosse uma rede. A escolha de uma árvore é feita (T_i). Amostrando a árvore que está logo ao lado desta, no universo de árvores conectadas, esta receberá o nome de T_j. Agora, as razões das probabilidades de T_i e de T_J são computadas, sendo então $R= f(T_J)/(T_i)$. Se $R \geq 1$, então a nova árvore é aceita, se $R \leq 1$, então um valor randômico é escolhido (entre 0 e 1). Se este valor for agora $\leq$ que R, aceita-se a nova arvore como a melhor, e assim por diante. Este é um processo Markoviano pelo fato de que o processo que ocorre no próximo passo não depende do anterior, ou seja não ha memória do passo anterior para considerar nenhum nova proposta, dependendo apenas do estado atual. A figura 11 ilustra o descrito acima, onde o número de vezes que há a transição do estado i para o estado j é igual ao número de vezes em que há a transição do estado j para o estado i.

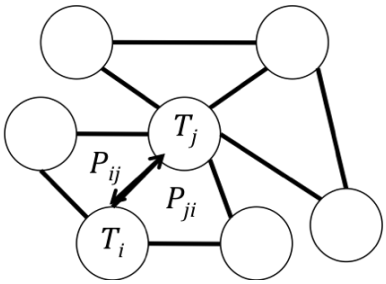


Figura 11. Representação das árvores conectadas no universo de árvores possíveis. As arvores T*i* e T*j* são visitadas em cada cadeia, com probabilidades de amostrar em qualquer direção (P).

O processo todo de busca de árvores por esses algoritmos funciona da seguinte maneira. Para calcular a distribuição de probabilidades a priori, consideremos uma árvore inicial. Em cima desta árvore, são feitas propostas para alteração de alguns parâmetros, usando o algoritmo de Metropolis para aceitar ou rejeitar essas mudanças. O processo começa com um '*burn-in*", que consiste em amostragens em espaços distintos, e deverá levar o processo até um estado de equilíbrio, o que implica que agora, proposições de mudanças em cima de uma árvore próxima ao ótimo serão visitadas. Depois desse período de "pré-aquecimento", começam as cadeias propriamente ditas, onde serão armazenadas árvores a cada *S* passos. O valor de *S* é escolhido para que se armazene uma quantidade suficiente de árvores (por exemplo, em 10 milhões de cadeias, armazenar arvores a cada 1000 passos). Essas arvores armazenadas podem ser entendidas como as que possuem valores de probabilidade posterior muito próximos ao ótimo (assumindo ausência de co-relação após o período de *burn-in*). As probabilidades posteriores de cada ramo será dada pelo intervalo de vezes em que esse ramo foi consistente nas arvores armazenadas.

Como em todos os métodos de inferência, a técnica Bayesiana também apresenta controvérsias para seu uso. A primeira controvérsia diz respeito ao uso de informações *a priori*. Se por um lado a informação prévia pode auxiliar na formulação da hipótese mais provável, por outro lado , os resultados podem ser seriamente afetados se algum *prior* não verdadeiro for assumido. Assim, sugere-se que não se assuma determinados *priors* se não houver absoluta certeza de que eles correspondem a realidade dos dados. Entretanto, o uso de *flat priors*, ou seja, não informativos, implica que a distribuição posterior será proporcional ao likelihood. Outra critica feita ao método bayesiano, e que parece ser bem mais séria do que assumir prior errados, é o problema do tipo da exploração no universo de árvores (Larget e Simon, 1999). Como a Figura 9 do item 3.3.4.2, o universo de árvores remete a uma ilha,

onde existem vales e picos. Os picos podem ser entendidos como regiões onde os valores (likelihood ou posterior) são maximizados, e os vales são os valores não ótimos. Ocorre que, após o período de "*burn-in*", uma região estável, de valores próximos ao ótimo é atingida (plato) e as arvores agora visitadas por MCMC serão apenas as que estão em volta deste mesmo pico. Uma vez que propostas que resultem na diminuição do valor de probabilidade posterior não serão aceitos, a mudança de um pico a outro sem passar por um vale é impossível. Dessa forma, se a amostragem estiver sendo feita em um pico que não representa o ótimo global (mas sim, um pico sub-ótimo), a melhor árvore jamais será visitada. Para contornar este problema porém, foi criado o chamado Metropolis Coupled MCMC (MCMCMC). A idéia neste caso é, no lugar de rodar apenas uma cadeia de Markov, várias cadeias são rodadas simultaneamente, sendo algumas delas chamadas "heated-MCMC chains". As "cadeias quentes" permitem que de tempos em tempos os diferentes picos da paisagem de árvores sejam cruzados sem a necessidade de passar por vales.

3.3.7 A escolha do método

A escolha da metodologia de reconstrução filogenética adotada neste trabalho foi feita com base nas vantagens e desvantagens oferecidas por cada um dos métodos. Como discutido anteriormente, métodos de distância possuem a desvantagem de reduzir toda a variabilidade da sequência a valores únicos de distância genética. Retrovírus endógenos de uma mesma família, mas presentes em hospedeiros evoluindo diferentemente, possuem certa variabilidade ao longo dos sítios que certamente trazem informação acerca dos processos evolutivos ocorridos ao longo do tempo. Assim, métodos de distância, embora rápidos e simples, não foram preferidos neste trabalho.

Em 1997, Tuffley & Steel mostraram que a MV e a máxima parcimônia (MP) podem resultar em reconstruções equivalentes sob modelos de substituição de nucleotídeos extremamente simples e simétricos (Tuffley e Steel, 1997). Entretanto, alterações nas premissas desses modelos, como nas probabilidades de substituições por exemplo, são suficientes para que os dois métodos não produzam mais árvores equivalentes. De fato, um dos principais problemas nas reconstruções por MP, é que, por basear-se simplesmente na mudança de um caracter para outro, variações nas taxas de substituição, probabilidades de substituição de bases (transversões e transições) e variação na composição de nucleotídeos podem gerar árvores enviesadas (Collins *et al.,* 1994). Nesse sentido, métodos que utilizam modelos de substituição, como a MV, que levem em conta a possibilidade e principalmente, a

probabilidade de ocorrência desses eventos, são preferidos. Especialmente no caso das análises deste trabalho, devemos levar em conta que as sequências de ERVs estão evoluindo no genoma do organismo hospedeiro há milhões de anos. Por se comportarem como pseudogenes, na maioria dos casos, os ERV-K certamente acumularam muito mais substituições do que pode ser de fato observado nas sequências. Por outro lado, uma vez que alguns elementos são extremamente conservadas, deve-se considerar que há seleção atuando em determinados sítios. Com base nisso, é altamente provável que certos caracteres mudaram mais de uma vez, e alguns inclusive reverteram para a condição original, e outros caracteres, por restrições funcionais, apresentam probabilidades de substituição diferenciadas. Fica claro que isso deve influenciar fortemente ao se estimar a distância genética entre um taxa e outro. Essa condição apenas pode ser levada em consideração com o uso de modelos mais complexos de substituição de nucleotídeos. Assim, o método de máxima verossimilhança (MV), ou *maximum likelihood* (ML), foi adotado em detrimento aos demais. A vantagem de ser um método estatístico, e que permite o uso de modelos explícitos de substituição de nucleotídeos foram determinantes na sua escolha e de teste explícito de hipóteses. As desvantagens apresentadas pelo método de MV, como tempo computacional e escolha de modelos de substituição com parâmetros adequados foram tranquilamente contornadas.

Inicialmente para a escolha do melhor modelo filogenético de substituição de *DN*A, foi utilizado o programa Modeltest v.3.7 (Posada e Crandall, 1998), e as reconstruções foram feitas no programa PAUP* versão 4.0b10 (Swofford, 2002). Com o modelo evolucionário adequado para o conjunto de sequências, foram geradas topologias iniciais pelo método de agrupamento de vizinhos (Neighbor Joining (NJ). Em seguida, as topologias sofreram rearranjos (perturbações) através do algoritmo heurístico NNI (*nearest neighbor interchange*, ou troca de vizinhos mais próximos). Esse algoritmo de perturbação topológica permite a escolha da melhor árvore, e consiste em trocar de lugar os taxa que aparecem separados por apenas um ramo interno e reconstruir a topologia, atribuindo um valor de *likelihood* (lnL). Este algoritmo foi escolhido em detrimento ao TBR e ao SPR principalmente devido ao tempo computacional dedicado, que neste caso, é muito mais viável.

No final do ano de 2006 foi disponibilizado um novo programa de reconstruções filogenéticas por máxima verossimilhança denominado Garli (Zwickl, 2006). Como já descrito anteriormente, este programa otimiza os parâmetros iniciais a cada geração, levando a um incremento ou queda no valor de lnL a cada nova reconstrução. Apenas as topologias que obtêm valores melhores aos anteriores são salvas, e o programa automaticamente para de

gerar novas topologias quando um determinado número de gerações (fornecido pelo usuário) ocorre sem incremento no valor de lnL. Parâmetros como frequência de bases, taxa de transição e transversão e valor da distribuição *gama* (função composta, onde se estima a heterogeneidade de taxas de substituição entre os sítios) são estimados a partir dos dados, ou alternativamente, podem ser fornecidos pelo usuário. Os rearranjos topológicos são otimizados através dos algoritmos NNI (descrito acima) e SPR (Subtree Pruning Regrafting). Por atribuir valores de probabilidade para essas reconstruções, uma topologia ótima é encontrada. Os valores de lnL encontrados com o programa PAUP são geralmente extremamente parecidos aos encontrados pelo programa Garli, para uma mesma topologia. Esse fato, aliado ao tempo computacional extremamente reduzido, a facilidade para a utilização deste programa, e principalmente, a possibilidade de estimativa dos parâmetros não baseada em árvores de NJ (como no ModelTest) levou ao seu emprego em detrimento do PAUP para o restante das análises.

3.4 Determinação do período de integração dos ERV-K

Após a integração, um provírus começa a acumular substituições ao longo do seu genoma a uma taxa similar a do seu hospedeiro, sendo consequentemente, inativado após certo período de tempo. Da mesma forma, as LTRs 5' e 3', que são idênticas no momento da integração, acumulam substituições independentes e em sítios distintos, divergindo uma da outra ao longo do tempo. Assim, é possível deduzir o tempo médio de integração de um provírus pela simples relação entre a distância genética entre as duas LTRs do mesmo provírus e a taxa de substituições de nucleotídeos a qual ele evolui multiplicado por dois. Ou seja, T = k/2r, onde T é o tempo de integração em anos, k é a distância genética entre as LTRs e r é a taxa de substituição de nucleotídeos por sítio por ano (s/s/a) (Li e Graur, 1993). Dessa forma, desde que um provírus mantivesse ao menos parte de ambas as LTRs para a comparação, era possível estimar a idade média do mesmo (Dangel *et al.,* 1995). Foi estimado que os ERV-K acumulam substituições no seu genoma a uma taxa que varia de 2.3 a $5x10^{-9}$ s/s/a, dependendo da região do genoma analisada e do organismo hospedeiro (Johnson e Coffin, 1999). Assim, para uma análise uniforme de todos os ERV-K deste trabalho, foi escolhida a taxa aproximada de $3.3x10^{-9}$ s/s/a.

3.5 Identificação dos eventos de espalhamento dos ERV-K nos genomas de primatas e determinação de provírus ortólogos

As análises de ortologia e do padrão de integração (por replicação ativa ou por duplicação de fragmentos genômicos) dos provírus foram feitas de duas maneiras. A primeira consistiu em construir uma árvore de LTRs 5' e 3' de todos os provírus. Um arquivo polifasta contendo as sequências das LTR 5' e 3' de todos os elementos foi submetido a um alinhamento múltiplo no ClustalX e o arquivo de saída, convertido para o formato nexus, foi utilizado como entrada para os programas de reconstrução filogenética. Entretanto, alguns eventos, principalmente os ocorridos mais distantes do presente, não são determinados com grande precisão através dessa técnica. Assim, uma segunda metodologia foi necessária.

Um segundo arquivo foi construído compreendendo 1000 pares de bases das regiões flanqueadoras das extremidades 3' e 5' de cada elemento. Com o objetivo de comparar os flancos dos diferentes provírus entre si, esse arquivo foi submetido a análises de similaridade contra ele mesmo, com uso da ferramenta Blast local (Blast2seq). Essa ferramenta permitiu analisar massivamente todas as sequências de uma só vez, resultando em um arquivo de saída contendo todos os "pares", com os respectivos *e-values*, de sequências que encontraram similaridade com outra do mesmo arquivo. Foram considerados ortólogos todos os elementos cujas regiões flanqueadoras 5' e 3' dos provírus eram similares (> 85%) às mesmas regiões de um provírus de outro hospedeiro. Este mesmo resultado, quando encontrado entre provírus do mesmo organismo, indicava duplicação do provírus carreado por duplicação de regiões gênomicas. Finalmente, quando a árvore de LTR agrupava os promotores de provírus distintos, mas a análise dos flancos não confirmava duplicação, o resultado era interpretado como transposição replicativa.

3.6 Análise de Seleção

Com o objetivo de determinar se os ERV-K estavam sob pressão seletiva, diferentes métodos de análise de seleção foram utilizados. Os métodos de detecção de seleção mais sensíveis utilizam a razão entre o número de substituições não-sinônimas por sítio não-sinônimo (*dN*) e o número de substituições sinônimas por sítio sinônimo (*dS*). Assim, o regime de seleção o qual uma sequência está exposta é determinado pela razão *dN/dS* (ω).

Por exemplo, em caso de restrições funcionais ou estruturais, espera-se encontrar um menor número de variações nos sítios não sinônimos em relação aos sinônimos (*dN*<*dS*), ou seja, na razão *dN*/*dS*, ω <1, indicando portanto seleção negativa ou purificadora. Por outro lado, na presença de seleção positiva, ou diversificadora, espera-se encontrar um maior número de variações nos sítios não sinônimos (*dN*>*dS*), sendo ω >1. Em caso de regime neutro, onde sequências teoricamente não sofrem nenhum tipo de pressão seletiva, temos ω =1

Para poderem ser utilizadas nas análises de seleção, as sequências dos genes provirais previamente alinhadas foram editadas manualmente e colocadas em fase de leitura com o programa Se-Al (Rambaut, 1996), onde codons de terminação e gaps foram removidos. A primeira investigação de seleção no *partgen* e no envelope de humanos e chimpanzés foi feita através do método de distância par-a-par, ou, método Nei-Gojobori (Nei e Kumar, 2000). Utilizando o programa PAUP, matrizes de distância pareada foram construídas para as 1as e 2as posições dos codons (assumindo-se que estas são posições potencialmente não sinônimas), para as 3as posições (consideradas sinônimas) e para todas as posições. As distâncias foram plotadas em um gráfico, onde o eixo X representava as 1as e 2as posições dos codons, e o eixo Y as 3as posições. Assim, se houvesse algum tipo de seleção agindo sobre as sequências, deveríamos encontrar um desvio da condição neutra, onde todos os sítios evoluem igualmente.

Estimativas de *dNdS* por verossimilhança, baseadas em modelos explícitos de substituição entre codons (Goldman e Yang, 1994) também foram obtidas. Entretanto, visto que regiões de uma mesma proteína desempenham funções distintas, com diferentes restrições funcionais e/ou estruturais, não é realista assumir que todos os sítios evoluem da mesma maneira. Assim, um método implementado no programa CODEML, do pacote PAML v.3.14 (Yang, 1997) permite que cada sítio (codon) de uma proteina possa assumir valores de *dN* e *dS* independentes. Por utilizar correção filogenética para descrever o processo de evolução dos códons, o método de verossimilhança proporciona maior confiabilidade no suporte da hipótese formulada (Freckleton, *et al*., 2002). O programa CODEML utiliza vários modelos evolutivos que diferem entre si na distribuição de *dN*/*dS* (ω) entre os códons. Os modelos empregados foram M0 (invariante), M1 (neutro), M2 (seleção positiva), M7 (dez categorias de códons com valores de distribuição beta) e o M8 que incorpora uma décima primeira categoria em relação ao modelo M7.

Além do programa CODEML, as sequências foram submetidas a análise de seleção através do servidor de domínio público Datamonkey (Sergei *et al*., 2005) disponível no sítio

http://www.datamonkey.org/. O método estatístico utilizado aqui é o mesmo implementado no programa HyPhy (Pond *et al.,* 2005), que também utiliza verossimilhança para a detecção de seleção em um conjunto de sequências. A partir do servidor Datamonkey, três métodos de análise de seleção são possíveis, SLAC (*single likelihood ancestor counting*), FEL (*fixed effect likelihood*) e REL (*randon effect likelihood*). O SLAC é um método rápido que leva em conta a probabilidade de um codon estar sob seleção, dado sua sequência ancestral. O FEL estima diretamente as substituições sinônimas e não sinônimas por sítio, e o REL é uma variação do FEL, onde permite variações nas taxas de substituições sinônimas e não sinônimas de acordo com uma distribuição pré-definida. Com o objetivo de poder comparar os sítios selecionados dos genes provirais de cada hospedeiro, apenas as regiões comuns para todos os ERV-K foram utilizadas. Dessa forma, o *dataset* do gene *gag* continha 310 codons, o *dataset* do gene *pol* 824 codons e o envelope (apenas a região p36) compreendia 234 codons.

3.7 Análise Demográfica de ERV-K nos genomas de primatas

Com o objetivo de um melhor entendimento a respeito do comportamento dos ERV-K nos diferentes hospedeiros (mecanismos de ganho, perda (*purgin*) e manutenção), foram utilizadas técnicas baseadas no princípio da coalescência para recuperar eventos cladogênicos ocorridos no passado. Para isso, removendo das análises os provírus sabidamente originados por duplicação, foram construídos *datasets* a partir do *partgen* com 2.530 nucleotídeos dos provírus de cada hospedeiro, separadamente. Isso foi feito para minimizar o número de sítios não informativos, e também para que todos os taxa compartilhassem os mesmos sítios. Com isso, a partir de quatro conjuntos de dados, (19 RhERV-K, 21 CERV-K, 31 HERV-K e 32 PongERV-K) foram feitas as análises. Além das análises realizadas para os diferentes hospedeiros, conjunto de dados de HERV-K construídos separadamente para os Grupos I e O/N também foram sujeitos a análises demográficas.

Como já descrito anteriormente, a técnica Bayesiana permite a incorporação de informações previamente conhecidas (*priors*) para a estimativa dos parâmetros populacionais. O programa BEAST (Drummond e Rambaut 2007), utiliza do processo de Monte Carlo em combinação com um modelo de substituição de nucleotídeos previamente escolhido para inferir processos demográficos, onde parâmetros como taxas de evolução, probabilidades de transição entre nucleotídeos, topologia, tamanho da população efetiva ao longo do tempo e

seus priores Bayesianos (*priors*) são encapsulados em cadeias de Markov. Durante as sucessivas iterações (da ordem de dezenas de milhões) pelo método de Monte Carlo, estes parâmetros são modificados e novas cadeias são sucessivamente propostas e avaliadas por seu incremento de probabilidade posterior. Após a estabilização dos valores dos parâmetros, estes tendem a convergir ao mesmo tempo em que o espaço topológico é visitado e amostrado em intervalos suficientemente distantes para evitar auto-correlação entre as cadeias, desta forma fornecendo valores independentes para a estimativa apropriada da filodinâmica. A técnica utiliza basicamente as informações genealógicas para inferir processos demográficos ocorridos no passado. Assim, partindo de um conjunto de sequências devidamente datadas ou, de uma taxa de substituições de nucleotídeos previamente estimada, a história demográfica da população é exibida em forma de um *plot* que descreve as flutuações nos tamanhos populacionais ao longo do tempo (BSL). A vantagem da utilização deste método é que, adicionalmente aos valores médios e medianos obtidos para cada parâmetro, valores de incerteza são estimados. Valores máximo e mínimo com 95% de probabilidade posterior (HPD- *high posterior density*) são fornecidos, além de um gráfico que denota o padrão de distribuição da amostragem para cada um dos parâmetros estimados. Com isso é possível verificar a densidade obtida no espaço amostral disponível e explorado durante o Monte Carlo com cadeias de Markov. A convergência dos parâmetros foi verificada através dos valores de ESS (*effective sampling size*), onde geralmente valores superiores a 100 são considerados aceitáveis. Detalhes adicionais da utilização de *priors* para as corridas estão detalhados no ítem "Methods" do Anexo C.

3.8 Análise de HERV-K ativos sob diferentes condições

A atividade de ERV-K em humanos tem recebido cada vez mais atenção não só da pesquisa básica, mas principalmente da pesquisa nas áreas médica e clínica devido a relação de sua expressão com o aparecimento de doenças. Com a intenção de amplificar transcritos de HERV-K, trabalhos anteriores utilizaram material biológico de pacientes HIV positivos e pacientes portadores de certos tipo de câncer (Contreras-Galindo *et al*., 2006a; Contreras-Galindo *et al*., 2006b; Contreras-Galindo *et al*., 2008). Esses transcritos foram sequenciados e submetidos ao banco de dados público.

Com o objetivo de investigar quais os provírus que ainda são capazes de serem transcritos, e, reativados sob específicas condições, foi feita uma busca *in silico* de transcritos

de genes de HERV-K no genoma humano. *Datasets* contendo as sequências de HERV-K transcritas, mais os genes provirais, foram construídos separadamente para *gag*, *pol* e *env*, e, com eles, genealogias foram reconstruídas. Esses arquivos foram alinhados com o programa Muscle v.3.52 (Edgar, 2004), e utilizados como entrada para o programa Garli de reconstruções genealógicas. Para a determinação da relação entre os transcritos com os respectivos HERV-K, um critério de clusterização de no máximo 2% de divergência entre os taxa irmãos foi adotado. Esse valor foi escolhido por incorporar as taxas de erro introduzidos por PCR e sequenciamento em ambas as sequências e os polimorfismos de nucleotídeos entre as populações humanas (Deb *et al*., 1998; Zhao *et al*., 2000).

A integridade dos promotores dos provírus ativos, ou seja, aqueles identificados como provável fonte dos transcritos, e dos não ativos foi avaliada. Os elementos presentes nas LTRs foram escolhidos com base em trabalhos que descrevem essas regiões como sendo importantes para a transcrição de HERV-K, ou ainda, por apresentarem semelhança com regiões descritas em promotores de genes humanos. Dessa forma, foram analisadas as seguintes regiões: (*i*) o sítio de ligação (*binding site* BS) para o fator de transcrição celular OTF-2, (*ii*) TATA box, (*iii*) o sítio iniciador da transcrição Inr, (*iv*) o promotor a montante da região iniciadora (UP) localizado na região regulatória e (*v*) o sítio *enhancer* para o fator de transcrição celular YY1, localizado na região U3 da LTR (Kovalskaya *et al*., 1991; Seto *et al*., 1991; Knossl *et al*., 1999). A presença do sítio e sua conservação foram avaliadas, e as regiões foram classificadas apenas como ‘conservada’ ou ‘não conservada’. Considerando cada região como um caracter independente, os eventos de ganho e perda para cada um desses caracteres (“ganho” para região íntegra e “perda” para a perda da integridade do mesmo) foram mapeados com o programa MacClade (Maddison e Maddison, 2003) ao longo da árvore de ML contruída a partir do *partgen*.

Dando continuidade a essas análises, foi construído um gráfico onde foram plotadas as distâncias genéticas de cada taxa até a raiz contra a integridade das suas LTRs. Os níveis de integridade foram categorizados de 1 a 5, onde 1 correspondia as LTRs com apenas uma região conservada, e 5 correspondia a integridade das cinco regiões. As distâncias genéticas de cada taxa em relação a raíz da árvore foram calculadas pelo programa TreeStat (http://evolve.zoo.ox.ac.uk/software/treestat). Com isso objetivou-se verificar se a inativação das LTRs estaria relacionada com a inativação do provírus como um todo.

4 RESULTADOS

4.1 Busca in silico e mapeamento de genomas completos de ERV-K nos genomas de primatas

A abordagem experimental adotada para a busca de ERV-K se mostrou eficiente para o propósito de encontrar provírus completos nos genomas de hospedeiros primatas. Entretanto, um levantamento mais genérico de retroelementos, completos ou não, seria inviável apenas com essas metodologias. Para isso, programas como RepeatMasker (http://www.repeatmasker.org) ou LTR_STRUC (McCarthy e McDonald, 2003) seriam de maior utilidade.

A busca por genomas de ERV-K em primatas foi restrita aos genomas dos organismos já sequenciados, montados ou em fase de montagem disponíveis nos bancos de dados públicos. Isso compreendeu, além do genoma humano (*Homo sapiens*), os genomas do chimpanzé (*Pan troglodytes*), macaco rhesus (*Macaca mulatta*) e o último disponibilizado durante a execução deste trabalho, o genoma do pongídeo orangotango (*Pongo pigmeus*). O genoma do sagui *Callithrix jacchus*, um primata do Novo mundo, foi sequenciado e disponibilizado no GenBank nesse último ano. Embora não tenha sido descrito ERV-K em primatas do Novo Mundo, uma busca foi conduzida também no genoma de sagui. Como esperado, este trabalho confirmou a inexistência de ERVs da família K nesses animais.

Os genomas de ERV-K encontrados foram nomeados de acordo com as letras que precedem o gênero, ou o nome popular de cada organismo hospedeiro. Assim, os elementos de chimpanzé foram chamados CERV-K, os de macaco rhesus RhERV-K e os do pongídeo orangotango, PongERV-K. Como descrito no item 3.1 da Metodologia, a procura por genomas provirais foi feita através de ferramentas de busca de sequências por similaridade, como Blast e BLAT, sempre usando como "sonda" um genoma de ERV-K completo. Isso resultou em um total de 58 HERV-K, 38 CERV-K, 19 RhERV-K e 35 PongERV-K (Tabela 1). Exceto por alguns poucos elementos, apenas os provírus possuindo as duas LTRs flanqueando suas extremidades foram selecionados. Embora um grande número de provírus tenha sido excluído das análises com a adoção deste critério, isso foi feito para possibilitar a datação dos provírus, que só é possível com a presença de ambas as LTRs (ver item 3.3 da Metodologia).

As sequências genômicas de primatas utilizadas neste trabalho já se encontravam em adiantado estágio de montagem e anotação. Assim, foi possível mapear os ERV-K nos

respectivos cromossomos de origem e, para os HERV-K, as regiões cromossômicas puderam também ser determinadas (Anexo A).

Tabela 1. ERV-K de primatas.

Provírus	Cromossomo	Tempo de Integração (milhões)	Duplicação	Ortólogo
HERV-K113	19	0.10		
HERV-K4	1	0.15		
HERV-K10	5	0.25		
HERV-K102	1	0.3	HK50a	
HERV-K50a	1	0.3		
HERV-K68	3	0.3		
3-HML	7	0.3		
2-HML	7	0.3		
HERV-K101	22	0.37		
HERV-K50b	3	0.37		
HERV-K60	21	0.375		
HERV-K36	11	0.5		
HERV-K41	12	0.5		
HERV-K37	11	0.62		
HERV-K109	6	1.0	HK115	
HERV-K115	8	1.0		
HERV-K104	5	2.3	HK50d	
HERV-K50d	5	2.3		
HERV-KII	3	2.4		
HERV-KI	3	2.75		
HERV-K42	12	3.1		
HERV-K23	6	4		CK9
HERV-K33	10	4.1		Ckold12309

(Continua)

Tabela 1. ERV-K de primatas (continuação).

Provírus	Cromossomo	Tempo de Integração (milhões)	Duplicação	Ortólogo
HERV-K18	1	5.0		CK60
HERV-K63	X	6.3		CK1
HERV-K11	3	6.75		PK7
HERV-K18b	5	6.8		CK73
HERV-K31	9	7.4		CK72
HERV-K71	14	7.4		CK71
HERV-K50f	19	7.5		
HERV-K12309	19	7.5	HK50f	CK59; PK11
HERV-K77	4	7.5	sétuplo	
HERV-K51	19	7.6		
HERV-K29	8	8.0		
HERV-K17	17	9	sétuplo	
HERV-K59	20	9.5		PK33
HERV-K30	9	9.7		CK30; PK13
HERV-K130352	8	9.8	sétuplo	
HERV-K50c	4	9.8	sétuplo	
HERV-K35	11	9.8		CK35; PK27
HERV-K7	11	10	sétuplo	
HERV-K27	8	10.3	sétuplo	
HERV-K17b	4	10.7	sétuplo	
HERV-K5	4	13.8		
HERV-K8	5	15.3		PK6

(Continua)

Tabela 1. ERV-K de primatas (continuação).

Provírus	Cromossomo	Tempo de Integração (milhões)	Duplicação	Ortólogo
HERV-K6	1	15.7	HK76	CK70; CK22; CK24; PK17; PK18; PK19; PK26; PK32
HERV-K76	1	15.7		
HERV-K52	19	16		PK9
HERV-K43	8	17.2	HK70	
HERV-K70	8	17.2		
HERV-Kold345	22	18.2		
HERV-K20	6	18.5	HK69	CK20; PK31
HERV-K69	6	18.5		
HERV-K12	1	22		
HERV-K50e	12	28		CK52
HERV-K57	19	30		CK57; PK28
HERV-K34	19	31.07		CK34; PK34
HERV-Kold35587	6	0		CKold35587; PK21
CERV-K100	Y	0.3		
CERV-K25	12	1.06	CK102	
CERV-K102	10	1.4		
CERV-K101	Y	1.5		
CERV-K9	5	4.5		HK23
CERV-Kold35587	5	4.5		HKold35587
CERV-K1	X	5.45		HK63
CERV-K10	1	5.7		
CERV-K73	4	7.3		HK18b

(Continua)

Tabela 1. ERV-K de primatas (continuação).

Provírus	Cromossomo	Tempo de Integração (milhões)	Duplicação	Ortólogo
CERV-Kold12309	8	7.11		
CERV-K23	9	7.20	CK72	HK31
CERV-K36	4	8,11		HK5; PK30
CERV-K72	11	8.7		
CERV-K37	6	9.09		PK31; HK20; HK69;
CERV-K30	11	9.1		HK30; PK13
CERV-K32	4	9.38	sétuplo	
CERV-K38	4	9.56	sétuplo	
CERV-K33	4	10.09	sétuplo	
CERV-K2	1	10.2		
CERVK21	8	10.3	sétuplo	
CERV-K35	11	10.6		HK35; PK27
CERV-K29	8	10.92		
CERV-K59	19	11.14		HK50f; HKold12309
CERV-K60	1	11.17		HK18
CERV-K31	7	11.4	sétuplo	
CERV-K24	1_ran	12.83	CK22; CK70	HK6; HK76
CERV-K22	1	16.59		
CERV-K71	15	16.9		HK71
CERV-K70	1_rand	17.1		
CERV-K20	6	18.5		
CERV-K26	Y	21.78	CK27; CK28; CKold3	

(Continua)

Tabela 1. ERV-K de primatas (continuação).

Provírus	Cromossomo	Tempo de Integração (milhões)	Duplicação	Ortólogo
CERV-K27	Y	22.00		
CERV-K28	Y	22.26		
CERV-Kold3	Δ	22.6		
CERV-K52	Δ	28.6		HK50e
CERV-Kold2	Δ	28.9		PK12
CERV-K57	19	30		HK57; PK28
CERV-K34	19	Δ		HK34; PK34
RhK5	12	0		
RhK2	1	0.210		
RhK4	11	2.69		
RhK14	8	3.29		
RhK9	2	4.17		
RhK6	16	4.52		
RhK16	4	5.59		
RhK1	1	6.36		
RhK7	17	8.49		
RhK15	X	9.98		
RhK18	6	10.0		
RhK17	3	10.79		
RhK12	5	11.1		
RhK11	5	11.12		
RhK10	4	15.6		
RhK8	19	20.2		

(Continua)

Tabela 1. ERV-K de primatas (continuação).

Provírus	Cromossomo	Tempo de Integração (milhões)	Duplicação	Ortólogo
RhK3	1	23.08		
RhK19	9	42.36		
PongERV-K3	X	1.9		
PongERV-K1	12	2.24		
PongERV-K2	2b	2.8		
PongERV-K5	3	5.48		
PongERV-K10	Δ	6.35		
PongERV-K21	6	6.8		Hkold35587; CKold35587
PongERV-K13	9	8.0		HK30 CK30
PongERV-K11	19	9.84		HK50f
PongERV-K22	4	10.1		
PongERV-K4	12	10.6		
PongERV-K20	14	10.7	PK35	
PongERV-K15	Δ	11.47		
PongERV-K35	1_rand	12.36		
PongERV-K7	3	12.7		HK11
PongERV-K14	2a	13.85		
PongERV-K31	6	14.5		HK20; HK69; CK37
PongERV-K6	5	18.48		HK8
PongERV-K32	1_rand	18.9		CK22
PongERV-K17	1_ran	19.1	PK18; PK19; PK26; PK32	HK6; HK76
PongERV-K18	1_rand	19.9		HK6; HK76

(Continua)

Tabela 1. ERV-K de primatas (continuação).

Provírus	Cromossomo	Tempo de Integração (milhões)	Duplicação	Ortólogo
PongERV-K16	9	21.9		
PongERV-K19	1_rand	23		HK6; HK76
PongERV-K33	20	25.89		HK59
PongERV-K12	Δ	30.7		CKold2
PongERV-K27	11	33.29		HK35; CK35
PongERV-K9	19	34.17		HK52
PongERV-K28	19	37.5		HK57; CK57
PongERV-K34	19	45.25		HK34; CK34
PongERV-K30	4	57		HK5; CK36
PongERV-K23	1	Δ		
PongERV-K24	19_rand	Δ		
PongERV-K25	8_rand	Δ		CK1
PongERV-K26	1_rand	Δ		
PongERV-K29	11	Δ		
PongERV-K8	6	Δ		

Δ - ERV-K com uma das LTRs ausente, ou parcialmente presente.

(Conclusão)

4.2 Alinhamento dos genomas e reconstruções filogenéticas de ERV-K de primatas

A partir dos alinhamentos dos genomas dos provírus, foram construídos diferentes conjuntos de dados (*datasets*) para as reconstruções filogenéticas. O primeiro *dataset* compreendeu apenas as regiões codantes do genoma (capsídeo, protease e polimerase, com 4130 nucleotídeos), e foi designado *partgen*. A região correspondente ao envelope não foi incluída devido a grande variabilidade presente nesse gene. Entretanto, um alinhamento somente de sequências de envelope também foi construído e utilizado em análises posteriores. Após a escolha do melhor modelo evolucionário (HKY+Γ), os arquivos foram submetidos as recontruções filogenéticas nos programas PAUP (Swofford, 2002) e Garli (Zwickl, 2006).

Os dados descritos no Anexo B resumem os resultados obtidos na fase inicial do trabalho. Nessa etapa, apenas provírus de humanos e chimpanzés haviam sido descritos e analisados. A filogenia representada nas Figuras A.1 e B.1 do Anexo B foi reconstruída a partir do *partgen* de 76 provírus, utilizando como grupo externo retrovírus exógenos de símios (SRV1, SRV2, SMRV-HLB e MPMV).

Com a disponibilização do genoma do macaco rhesus (*Macaca mulatta*) no final do ano de 2006, dezenove provírus completos (RhERV-K) foram encontrados e adicionados ao *dataset* original. Durante essa etapa, mais doze genomas provirais foram identificados em chimpanzé e incluídos nas análises. A árvore filogenética representada na Figura C.1 do Anexo C, foi contruída a partir do alinhamento múltiplo (*partgen*) de 106 sequéncias (55 HERV-K, 32 CERV-K e 19 RhERV-K) e foi congruente ao obtido previamente (Figura B.1, Anexo B).

O último genoma de primata do Velho mundo sequenciado e disponibilizado publicamente (durante a execução desse trabalho) foi o genoma do pongídio *Pongo pigmeus*. Assim, uma terceira reconstrução filogenética foi feita e incluiu além dos 106 genomas provirais já descritos, trinta e cinco provírus de orangotango (PongERV-K), mais três CERV-K e dois HERV-K. A reconstrução filogenética mais completa contém 147 *taxa* e está representada na Figura 1, onde os ramos estão distinguidos por cores que variam para cada primata. As filogenias foram unânimes em mostrar que os ERV-K podem ser divididos em dois grandes grupos, denominados neste trabalho de Grupo O/N (cluster superior) e Grupo I (cluster inferior) (Figura 1) (Romano *et al.,* 2006). O Grupo O/N recebeu esse nome por incluir os elementos com tempo de integração mais antigos (Grupo *Old*) e também os mais recentes (Sub-grupo *New*). O cluster denominado Grupo I abriga os elementos com tempo de

integração intermediário (*Intermediate*) em relação aos do Grupo O/N. Um cluster de repetidas duplicações de provírus que teve início em chimpanzé e também ocorreu em humanos aparece em destaque nas filogenias (sétuplos). Nesse cluster há sete provírus humanos e cinco de chimpanzés. Dentro do subgrupo N existe um cluster específico de ERV-K de orangotango, sugerindo que houve atividade recente dentro dessa linhagem. O cluster dentro do Grupo O com 14 RhERV-K também é sugestivo de atividade recente. Entretanto, visto que não foram encontrados ortólogos em outros primatas, é possível que o agrupamento destes elementos seja apenas consequência da ausência de relação de RhERV-K com ERV-K de outros primatas.

Além da divisão dos ERV-K em dois grandes clusters, as recontruções mostraram uma divisão parcial de elementos dos tipos 1 e 2. Relembrando, elementos do tipo 1 apresentam uma deleção de 292 nucleotídeos entre os genes de polimerase e envelope. O cluster inferior da árvore, Grupo I contém apenas elementos do tipo 2, enquanto o cluster superior contém elementos de ambos os tipos nas quatro linhagens.

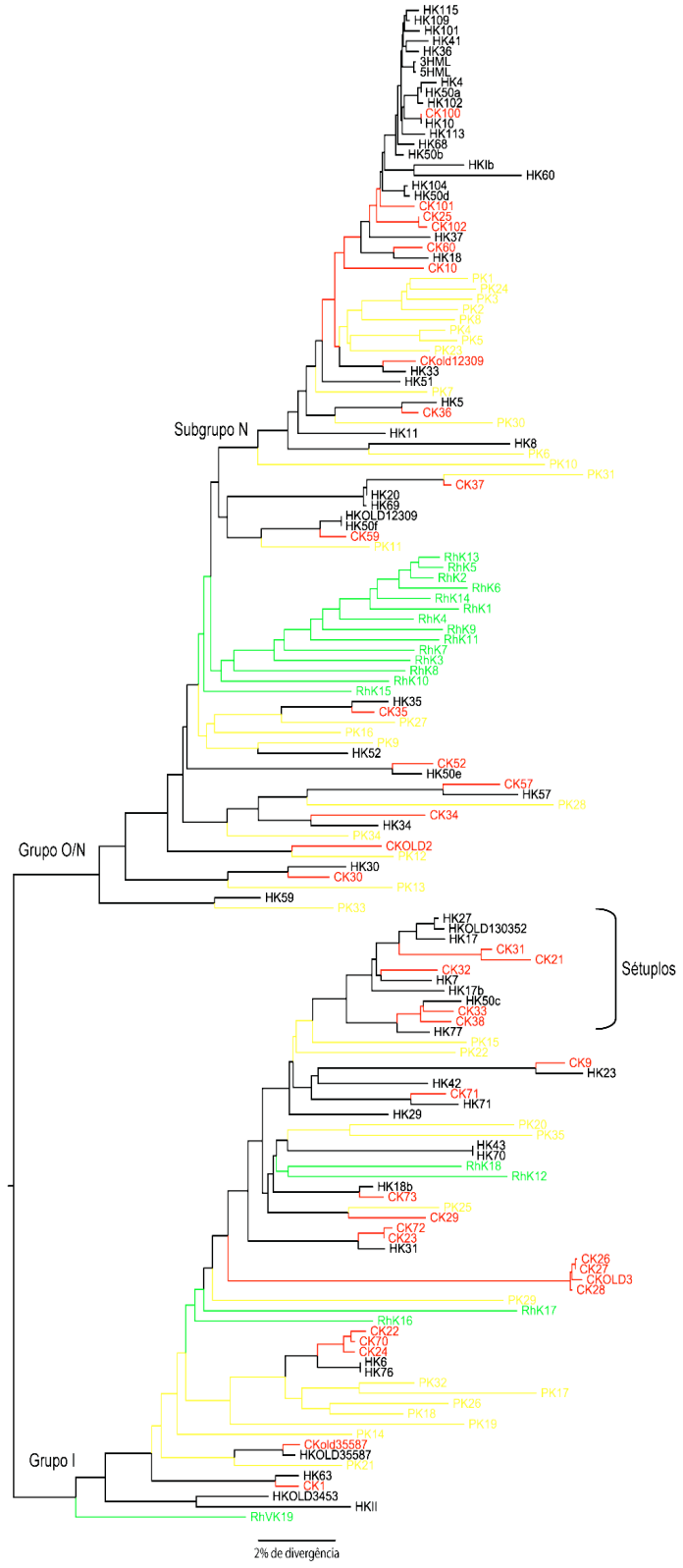


Figura 1. Reconstrução filogenética por máxima verossimilhança a partir de 147 sequências de ERV-K. Os ramos da árvore foram coloridos de acordo com o gênero do organismo hospedeiro: *Homo sapiens* em preto, *Pan troglodytes* em vermelho, *Macaca mulatta* em verde e *Pongo pigmaeus* em amarelo. Os três principais clusters estão indicados pelos respectivos nomes no nó mais basal que os origina.

4.3 Determinação do período de integração dos ERV-K

Devido ao mecanismo de replicação dos retrovírus, no momento da sua integração no genoma do hospedeiro as duas LTRs são idênticas (Coffin *et al*., 1997). No entanto, a menos que a preservação da sua atividade seja favorecida por alguma razão, os provírus tendem a acumular substituições ao longo do seu genoma logo após sua integração, o que acaba ocasionando sua inativação. Essas substituições, que também ocorrem independentemente nas duas LTRs, possivelmente se acumulam a uma taxa aproximada a de substituições de nucleotídeos de pseudogenes. Dessa forma, é possível estimar o tempo de integração de um provírus no genoma hospedeiro através da divergência entre as duas LTRs (em termos de substituições de nucleotídos). Como descrito no item 3.3 da Metodologia, a data de integração dos provírus foi deduzida a partir da fórmula $T = k/2\mu$, onde T é o tempo de integração (em milhões de anos), k é a distância genética estimada entre as LTR 5' e 3' de um mesmo provírus e μ é a taxa de substitituição de nucleotídos por sítio por ano. Na tabela 1 estão descritos os tempos estimados de integração para todos os ERV-K encontrados e que puderam ser datados. Na tabela 1 do Anexo B no entanto, é possível notar que algumas datas de integração divergem das descritas na tabela 1 do texto. Isso é explicado devido ao fato de que anteriormente, ao serem analisados elementos que eram sabidamente ortólogos, optamos por uma calibração mais fina da taxa de substituição para a região específica do genoma a qual o provírus estava inserido. Assim, assumindo 6 milhões de anos como sendo o tempo médio de divergência das linhagens ancestrais de *Homo* e *Pan*, estimamos taxa de substituições de nucleotídeos para todas as regiões flanqueadoras dos provírus que apresentavam ortólogos. Após a inclusão de provírus de outros primatas, decidiu-se utilizar uma taxa de substituição comum para todos os elementos, com o objetivo de minimizar os erros nas estimativas. Entretanto, analisando os tempos de integração de provírus que possuem ortólogos, observou-se que alguns deles apresentaram resultados utópicos, revelando uma data de integração mais recente do que a própria data estimada de divergência entre as linhagens em questão.

4.4 Identificação de elementos ortólogos a partir de reconstruções filogenéticas

O padrão de clusterização de certos provírus na filogenia pode ser um bom indicativo de ortologia. Apenas observando as filogenias, diversos ortólogos entre humanos, chimpanzés e orangotango puderam ser identificados. Como exemplo, dentre os vários eventos distribuídos ao longo da reconstrução, foi observado um cluster de três provírus no Grupo N, com o PongERV-K30, sendo o ancestral do CERV-K36 e HERV-K5, todos inseridos no cromossomo 4. Proporcionalmente, mais ortólogos foram encontrados nos grupos O e I, possivelmente devido aos provírus do Sub-Grupo N serem mais recentes, muitos dos quais exclusivamente humanos. Não foram encontrados ortólogos em rhesus a partir dos genomas completos analisados neste trabalho. Entretanto, uma análise mais cuidadosa dos sítios de integração dos elementos em todos os primatas estudados revelou que alguns provírus integraram-se no genoma da linhagem ancestral a rhesus, mas estes foram posteriormente purgados do seu genoma, restando uma solo-LTR ou o provírus parcial.

Embora diversos ortólogos possam ser sugeridos baseando-se apenas na filogenia, conclusões definitivas nesse sentido não podem ser obtidas apenas com base na reconstrução. Além disso, os filogramas não permitem discernir entre ortologia, eventos de duplicação e replicação ativa dos provírus. Por essa razão, o mapeamento físico dos ERV-K nos respectivos cromossomos, a análise e comparação das regiões flanqueadoras dos provírus e a reconstrução filogenética a partir das LTRs dos elementos foram cruciais para o entendimento da dinâmica dos ERV-K. Na tabela 1 estão descritos todos elementos que possuem ortólogos.

4.5 Confirmação de provírus ortólogos e determinação dos eventos de espalhamento de ERV-K

Antes mesmo de recontruir as filogenias, a busca de ERVs no banco de dados do BLAT search já sugeria quando um provírus possuía ortólogos, e em quais gêneros. A Figura 2 mostra um exemplo de um provírus de orangotango, que aparece presente no genoma humano, e deletado no genoma de chimpanzé, restando apenas as LTRs.

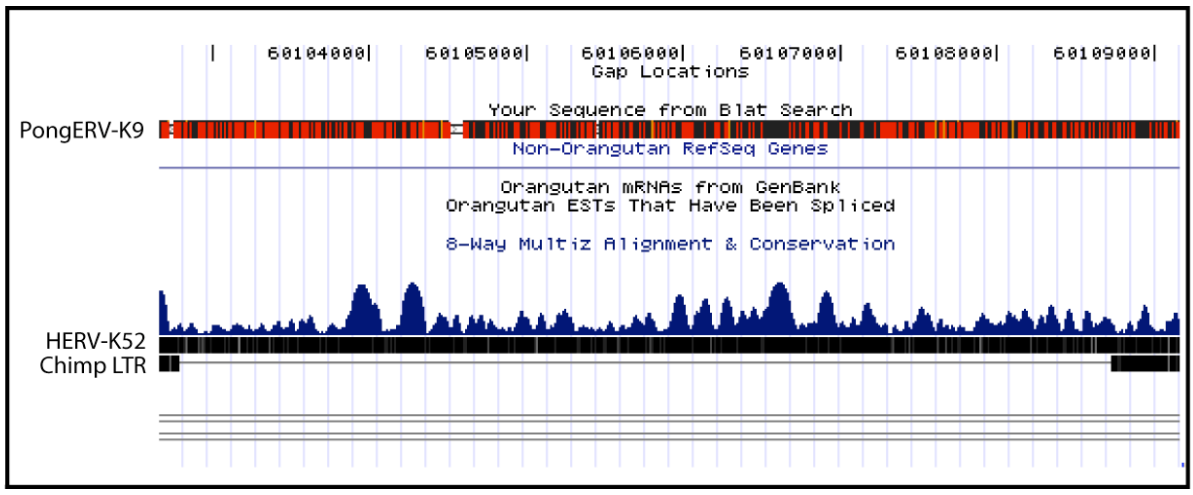


Figura 2. Representação gráfica da interface do programa BLAT mostrando regiões ortólogas com provírus em diferentes linhagens. A barra vermelha corresponde ao Pongerv-K e as pretas ao mesmo ERV em humano e chimpanzé. O nível de conservação da região analisada aparece em azul.

Além da indicação de ortologia, essa descrição gráfica permitia, indiretamente, inferir o tempo de integração aproximado. A primeira metodologia de verificação, utilizada principalmente para determinar os eventos ocorridos em humanos e chimpanzés, foi a análise de uma árvore construída a partir das sequências de LTRs 3' e 5' dos provírus. Existe um padrão básico esperado para cada tipo de evento, o que permite diferenciá-los em uma genealogia construída a partir das LTRs: (1) elementos originados por duplicação cromossômica geralmente apresentam maior similaridade entre as LTRs dos provírus irmãos, do que entre as suas próprias LTRs; (2) elementos que se originam por replicação ativa apresentam, além de similaridade entre suas próprias LTRs, pequena distância entre as mesmas com a LTR 5' do provírus parental e, (3) provírus ortólogos geralmente apresentam maior similaridade entre as LTR dos ortólogos do que entre as próprias LTRs 5' e 3'. A reconstrução filogenética feita a partir apenas das LTRs de ERV-K de humanos e chimpanzés está representada na Figura 3, onde é possível identificar eventos clássicos de duplicação e ortologia. Em alguns casos, além de ortologia, pode-se observar a duplicação passiva do ERV-K somente no genoma humano, ou seja, depois da divergência das linhagens.

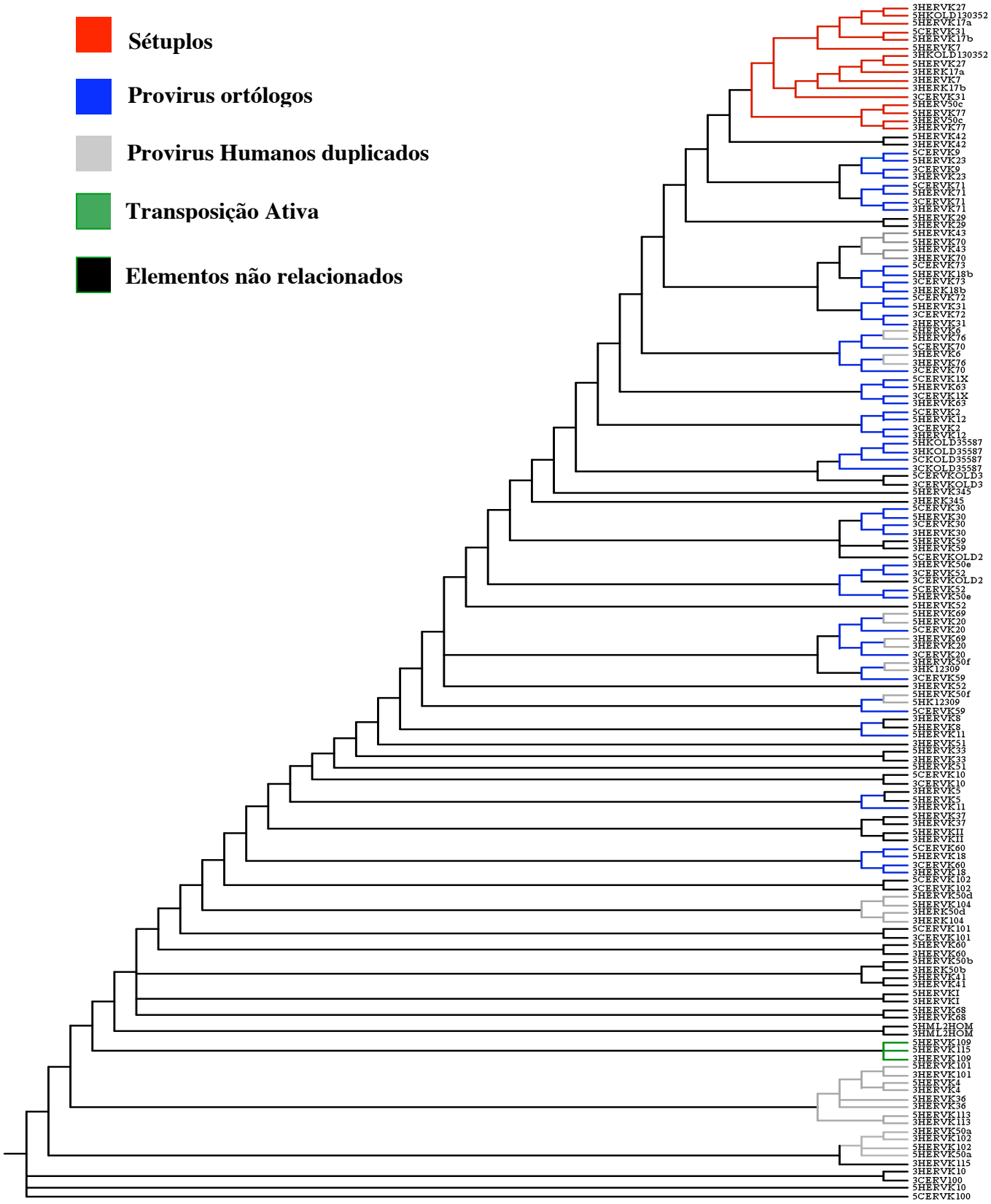


Figura 3. Cladograma reconstruído a partir das LTRs 5' e 3' de provírus de humanos e chimpanzés. Os eventos indicados pelo padrão de clusterização das LTRs estão indicados por diferentes cores (legenda no topo da figura).

Na tabela 1 estão descritos os principais eventos confirmados de duplicação e ortologia entre todos os ERV-K, bem como suas localizações cromossômicas. Entretanto, é importante mencionar que a localização dos CERV-K mapeados durante a primeira etapa do trabalho e descritos na tabela 1 do Anexo B seguiu a regra de nomenclatura antiga estabelecida para os cromossomos de chimpanzé. Por exemplo, HERV-K71 presente no cromossomo 14 é ortólogo do CERV-K71, mapeado no cromossomo 15 e o HERV-K18b no cromossomo 5 possui como ortólogo o CERV-K73 no cromossomo 4.

Os ERV-K de rhesus e orangotango também foram submetidos aos mesmos procedimentos para a verificação e confirmação de ortologia e duplicações. Para isso, uma árvore contendo ambas as LTRs de todos os ERV-K dos quatro primatas foi reconstruída. Embora a árvore tenha sido capaz de reconstruir diversos eventos envolvendo provírus dos diferentes primatas, é importante ressaltar que, por diversas razões discutidas posteriormente, nem todos os provírus sabidamente relacionados foram evidenciados a partir dessa árvore. Essa metodologia, embora eficiente para a detecção e distinção de eventos de origem dos provírus não teve suficiente sensibilidade para estimar com precisão eventos ocorridos mais distantes do presente, visto que os provírus com integração mais antiga podem apresentar distorções nos padrões descritos acima. Assim, para confirmar as relações entre os provírus sugeridas pelas reconstruções, bem como encontrar outras não evidenciadas nas mesmas, utilizamos uma segunda metodogia, a análise das regiões flanqueadoras dos provírus.

A partir da análise e comparação de 1000 pares de bases que flanqueiam os provírus em ambas as extremidades, foi possível confirmar e determinar a origem de diversos elementos. O valor mínimo de 85% de similaridade foi atribuído como limite mínimo por abranger qualquer evento que pudesse ser genuinamente verdadeiro, visto que mesmo os organismos hospedeiros mais distantes, comparados neste trabalho, não devem apresentar mais do que 15% de divergência entre si entre regiões ortólogas, mesmo que altamente variáveis (Magness et al., 2005). Os resultados mostraram diversos eventos de duplicação e ortologia que não haviam sido sugeridos pelas reconstruções filogenéticas. Em alguns casos, foram encontrados mais de dois provírus originados a partir de um mesmo parental. O caso mais interessante envolvendo duplicação e ortologia revelou sete HERV-K orginados por duplicação, sendo que ao menos os primeiros eventos de duplicação ocorreram em chimpanzé, o qual apresentou cinco ortólogos entre os sete provírus humanos. Os provírus originados desse mesmo evento foram encontrados nos cromossomos humanos 4, 8, 11 e 17 e nos cromossomos 4 e 8 de chimpanzés, sugerindo que a integração nos cromossomos 11 e 17 de humanos ocorreu após a divergência das linhagens.

A utilização dessa metodologia foi, de fato, crucial para a determinação de certos ortólogos, os quais não foram evidenciados pela filogenia. Além disso, todos os elementos dos quais a ortologia fora confimada não apresentaram mais do que 10% de divergência entre as regiões flanqueadoras. Os resultados descritos na tabela 1 correspondem aos eventos encontrados através das três metodologias aqui descritas. Nessa tabela, os provírus originados por duplicação ativa ou passiva também foram descritos e identificados com clareza através da análise dos flancos e da árvore de LTR.

O conjunto dos resultados sugeriu que um grupo de pelo menos cinco ERV-K de orangotangos (cluster de oito elementos representados em vermelho no subgrupo N) se originaram por atividade replicativa. O padrão de clusterização na árvore de *partgen*, a semelhança entre suas LTRs, a data de integração estimada para os mesmos (entre 1 e seis milhões de anos atrás) e a ausência de similaridade entre as regiões flanqueadoras dos mesmos sugerem fortemente que esses provírus originaram-se por atividade, e não por simples duplicação da região genômica.

4.6 Análise de Seleção

As primeiras estimativas de evidência de seleção feitas neste trabalho estão evidenciadas nas Figuras 3A e 3B do Anexo B. Os *plots* mostraram um excesso de substituições ocorrendo na 3ª posição dos codons em comparação as 1as e 2as posições, no envelope e também no *partgen*, indicando portanto seleção purificadora agindo nos genes provirais.

Os valores de *dN/dS* calculados separadamente para os ERV-K dos distintos primatas a partir dos genes de *gag*, *pol* e *env* não apresentaram variações significativas entre os diferentes hospedeiros, variando de 0.5 no gene *gag* até o valor máximo de 0.74 no gene de envelope. No gene *gag* os valores foram $\omega = 0.5$ para todos os ERV-K, com excessão dos de orangotango, onde $\omega = 0.6$. Em relação ao gene da polimerase, enquanto obtivemos $\omega = 0.6$ para os ERV-K de pongídeos e chimpanzés, em humanos e rhesus o valor de ω foi de 0.5. Por fim, a proteína p36 do gene de envelope apresentou valores de ω em torno de 0.7 para pongídeos e rhesus, 0.68 para humanos e de 0.6 em chimpanzés. Entretanto, quando calculados codon a codon, o número de sítios positiva e negativamente selecionados variou bastante entre as linhagens e entre os genes. A tabela 2 mostra o número de codons sob seleção positiva e negativa para todos os genes dos ERV-K dos diferentes primatas. Apenas

os codons sob seleção negativa, que são compartillhados com outros hospedeiros, estão evidenciados nessa tabela.

Tabela 2. Codons sob seleção positiva e negativa em ERV-K de primatas.

Provirus	*Gag*		*Pol*		*Env* (p36)	
	+	-	+	-	+	-
HERV-K	62, 350	27, 36, 50 64, 134, 203, 254, 256, 363	251	63, 65, 66, 76, 87, 194, 212, 265, 449, 769	80, 97, 168, 204, 212, 228	16, 40, 73, 84, 171, 193, 199
CERV-K	350	27, 36, 50, 64, 256	251, 535	63, 65, 66, 76, 87, 194, 212, 265, 449, 769	204	16, 40, 73, 84, 170, 193, 199
PongERV-K	75	36, 64, 254, 256	456	76, 194	52	114
RhERV-K	0	134, 203, 363	148	76, 449	0	114, 193, 199

Todos os codons sob seleção negativa em mais de um primata na região do *gag* mapeiam nos domínios da Gag_p10 e p24. Um dos codons em comum entre *Homo* e *Pan* positivamente selecionado em *gag* está localizado em uma região bastante conservada entre retrovirus exógenos também (AIR1–Arginine methyltransferase-interacting protein), e é responsável por modificações pós-traducionais, tráfego e secreção de proteínas. Os codons que apareceram sob seleção negativa na polimerase estão nas regiões da transcriptase reversa (RT) e RNAseH.

No gene de envelope, seis codons positivamente selecionados, sendo um deles presente também em chimpanzé e outro presente em pongo foram encontrados em ERV-K de humanos. De todos os genes analisados, a proteína p36 foi a que apresentou o maior número de codons positivamente selecionados (seis codons, contra um em *pol* e dois em *gag*). Uma vez que, em qualquer infecção viral, o envelope é sabidamente alvo do sistema imune do hospedeiro, o excesso de seleção positiva nesse gene sugere mutações adaptativas (mutações de escape), o que pode estar indiretamente associado a atividade recente.

4.7 Análise Demográfica de ERV-K nos genomas de primatas

Após a descrição física, filogenética e de alguns parâmetros evolutivos dos provírus, buscou-se o entendimento de como é a dinâmica de espalhamento desses elementos nos diferentes genomas hospedeiros. Para isso, utilizando técnicas baseadas em coalescência, implementadas no programa BEAST (Drummond e Rambaut 2007), investigamos as histórias demográficas dos ERV-K em cada um dos primata. Os *plots* criados nesse programa (BSLs) revelaram flutuações nas populações de ERV-K em todos os primatas analisados (Figura C.3 do Anexo C e Anexo D). No entanto, algumas diferenças na dinâmica de espalhamento dos mesmos puderam ser observadas. Por exemplo, enquanto um crescimento populacional dos RhERV-K parecia ter ocorrido nos últimos 25 milhões de anos (m.a), um sinal de aumento na população dos CERV e HERV-K foi observado somente há 17 m.a. atrás, seguido por uma redução no número de cópias em ambos, porém mais acentuada em humanos (Romano *et al*., 2007). Já em *Pongo*, foi possível notar um aumento constante na diversidade, sem posteriores flutuações (Anexo D). Ainda, foi possível observar uma grande semelhança na reconstrução demográfica obtida a partir dos RhERV e dos HERV-K, e em menor escala, dos PongERV-K. Além disso, os tamanhos de população efetiva dos elementos (*Ne.g*) também foi semelhante em rhesus e humanos, que por sua vez, são cerca de 3 vezes menores que os CERV-K. Entretanto, diferentemente de RhERV e CERV-K, os HERV-K apresentaram um visível aumento populacional ocorrido nos últimos 2 milhões de anos, que claramente condiz com a explosão dos elementos mais recentes pertencentes ao subgrupo N.

4.8 Análise de atividade de HERV-K

Uma vez que os resultados apontaram para diferenças bem marcantes no que diz respeito aos padrões de clusterização e dinâmica de HERV-K, foi feita nessa etapa uma análise da demografia dos HERV-K para os dois grupos separadamente. Os resultados mostrados na Figura 1 do Anexo E (BSL acima das árvores) mostram que, de acordo com o observado na filogenia, houve um acentuado crescimento populacional do Grupo I no passado, porém, sem sinais de atividade mais perto do presente. Os elementos do Grupo O/N, por outro lado, após um período de estabilidade voltou a expandir (elementos do subgrupo N) bem próximo ao presente. Portanto, buscando identificar quais provírus são

ativamente transcritos no genoma humano, foi feita uma associação *in silico* de transcritos de genes de HERV-K depositados no GenBank com os provírus mapeados e anotados neste trabalho.

Alguns trabalhos realizados anteriormente tiveram sucesso ao amplificar transcritos de HERV-K em material de pacientes HIV positivos e pacientes portadores de certos tipo de câncer. Esses transcritos foram sequenciados e submetidos ao banco de dados público. Assim, *datasets* incluindo essas sequências e os genes provirais, foram construídos separadamente para *gag*, *pol* e *env*, e, com eles, genealogias foram reconstruídas. As análises feitas por máxima verossimilhança mostraram que, em termos de quantidade, não houve diferença significativa em relação a associação de transcritos obtidos a partir das diferentes fontes (HIV e câncer) com os HERV-K analisados. Entretanto, uma grande diferença foi encontrada em termos de expressão de provírus dos diferentes Grupo (O/N e I). Enquanto apenas um dentre os 19 provírus do Grupo I foi relacionado a transcritos (5%), 80% dos provírus pertencentes ao grupo O/N pareciam estar ativos. Esses dados resultaram em um $\chi^2 = 10.0$ ($p<0.001$, d.f. = 1), mostrando que a diferença de atividade entre os dois grupos foi altamente significante.

Interessantemente, expressão diferencial entre provírus do tipo 1 (que codifica para a proteína Np9) e tipo 2 (que codifica para a proteína Rec) também pareceu ocorrer entre as duas patologias analisadas. Embora transcritos de ambos os tipos de HERV-K tenham sido detectados nas duas condições, houve uma significante associação entre HERV-K do tipo 1 com transcritos obtidos de pacientes com câncer (80%).

Em vista da expressão diferencial dos elementos em termos de grupo e tipo, foi feita uma análise das regiões promotoras de transcrição contidos nas LTRs. A tabela 1 do Anexo E descreve os provírus que tiveram transcritos detectados em pacientes HIV positivo e com câncer; o tipo do provírus (1 ou 2); a data aproximada de integração no genoma humano; e a análise de integridade dos cinco elementos estudados nas LTRs. A Figura 1 do Anexo E mostra claramente que existe uma associação entre expressão e integridade de regiões promotoras. Enquanto o Grupo O/N possui um alto nível de conservação das regiões promotoras, os provírus do Grupo I não. Além disso, os caracteres 1 e 5 (representando as regiões de enhancer e Inr respectivamente) parecem ser fundamentais para a expressão do genoma proviral, visto que não estão preservadas em nenhum elemento do Grupo I. Procurando ainda investigar se a inativação das LTRs correlaciona com as substituições acumuladas ao longo do genoma proviral, um gráfico onde foram plotadas as distâncias

genéticas dos elementos contra o nível de integridade das suas LTRs foi construído. O gráfico (Anexo E, Figura E.2) indicou uma correlação direta entre as duas variáveis. Uma vez que a data de integração dos elementos do Grupo I não são necessariamente mais distantes do presente do que a dos elementos do Grupo O/N, não se justifica um acúmulo de substituições maior nesses elementos. O que poderia explicar isso é que, após a inativação e perda de expressão, os provírus do Grupo I começaram a acumular mais substituições ao longo do seu genoma.

Resumidamente, os dados apoiam a teoria de que um maior acúmulo de substituições é consequência da inatividade dos mesmos, que só ocorre após a inativação dos elementos promotores presentes nas LTRs.

5 DISCUSSÃO

5.1 Datação de ERV-K através das LTRs

Inicialmente, apenas com os genomas de *Homo* e *Pan* disponíveis nos bancos de dados público, as estimativas de tempo de integração pareciam coerentes e não foi evidenciado nenhuma incongruência. Entretanto, com a disponibilização dos genomas de outros primatas, as datações por essa metodologia começaram a mostrar alguns problemas. Embora existam trabalhos que mostrem que as taxas de substituições variam entre os diferentes primatas, diminuindo ao longo da evolução dos Hominideos (Seino *et al*., 1992; Steiper *et al*., 2004), neste trabalho adotamos uma mesma taxa de substituição para estimar o tempo de integração dos ERV-K de todos os primatas, na tentativa de diminuir o erro incorporado causado por variações nas taxas dentro do próprio organismo. Outra importante razão para utilizarmos a mesma taxa provém do fato de que os provírus humanos e de chimpanzés, que compartilham ortólogos em espécies mais distantes, como Rhesus ou orangotango, estão evidentemente evoluindo a taxas mais lentas apenas há alguns milhares de anos. Sendo assim, foram construídos dois alinhamentos múltiplos a partir de sequências do gene *BSM1L* (*ribosome assembly protein*) de todos os primatas, um deles compreendendo os *exons* e outro apenas os *introns*. A partir das distâncias (k) estimadas para cada par de sequência e das datas de divergência entre os organismos, foram estimadas taxas de substituição para ambas as regiões empregando-se a fórmula $\mu = k/2T$, onde μ é a taxa de substituição de nucleotídeos por sítio por ano. A média calculada dessa taxa foi de 3.5E-09, o que está de acordo com a taxa utilizada aqui.

Assim, a partir de uma taxa de substituições média, deveríamos encontrar valores bem similares de integração entre elementos ortólogos, e que correspondessem a um tempo igual ou anterior a data de divergência das linhagens mais antigas que possuem o provírus. Assumindo que a divergência de *Pongo* da linhagem que deu origem a *Homo* e *Pan* ocorreu há aproximadamente 13 m.a (12 a 15 m.a) (Glazko e Nei, 2003), os provírus presentes nas três linhagens deveriam apresentar no mínimo 12 m.a. Porém, quatro deles apresentaram data de integração inferior a dez milhões de anos, um deles chegando a seis milhões. É importante lembrar que, utilizando uma taxa mais rápida apenas para os provírus de pongídeos, esses valores seriam ainda mais improváveis.

O PongERV-K21 (ortólogo dos CERV e HERV-Kold35587) é o caso mais interessante e merece ser discutido com mais detalhes. Os três provírus apresentaram o

mesmo padrão, onde a data de integração foi claramente subestimada por possuírem as LTRs quase idênticas. Recombinação entre provírus distintos não explica isso, pois as LTRs estão conservadas em todos os organismos desde pongídeos, indicando que, caso tenha ocorrido, foi na linhagem de *Pongo* (ou ancestral a ele) e, portanto, há tempo suficiente para que se observasse certa divergência entre as LTRs. Sabe-se que, por conter elementos reguladores e influenciar diretamente na expressão de genes vizinhos (Dunn *et al*., 2005; Landry *et al*., 2002; Medstrand *et al*., 2001), as LTRs podem acumular substituições em uma frequência diferente do restante do genoma proviral, o que em última análise as levaria a evoluir de maneira não neutra.

Uma cuidadosa investigação da região cromossômica onde esse provírus está inserido mostrou que existe um gene ribossomal, o RPL7L1, localizado há menos de 8 Kb a jusante da LTR 3', conservado em diversos mamíferos. Uma possibilidade plausível seria que a inserção de um provírus próximo a esse gene, carregando um promotor forte, poderia ter oferecido alguma vantagem evolutiva para o organismo hospedeiro, aumentando a expressão desse gene. Dando suporte a esse argumento, uma posterior análise incluiu um alinhamento de dez mil nucleotídeos compreendendo o gene RPL7L1, e mostrou que as taxas de substituição de nucleotídeos nessa região são muito próximas da média utilizada neste trabalho. Ou seja, independente da taxa de substituição da região cromossômica deste provírus, as LTRs estão evolutivamente conservadas.

Outro caso curioso foi o do provírus PongERV-K13. Este compartilha ortólogos em todos os primatas analisados, embora em rhesus ele tenha sido quase totalmente eliminado. Em humanos e chimpanzés, ele recebe o nome de ERV-K30, e o tempo de integração para todos foi estimado por volta de oito a dez milhões de anos atrás, portanto, muito mais recente do que o esperado. Uma análise da região flanqueadora mostrou a presença de um gene, Lipocalina 15 (*LCN15*), que participa do metabolismo de lipídeos. Novamente, poderíamos propor que a expressão desse gene poderia estar sendo aumentada pelos promotores do provírus. Infelizmente, não foram encontrados no GenBank transcritos desses genes iniciando a partir das LTRs provirais. Porém, sabe-se que a transcrição gênica é tecido-específica, principalmente no caso de genes expressos por promotores provirais (Tanaka *et al*., 2003; Seifarth *et al*., 2005). Assim, a falta de RNAs mensageiros originados por promotores provirais nos bancos públicos pode ser consequência de problemas de amostragem. Experimentos utilizando a metodologia de RLM-RACE para RNAs mensageiros extraídos de diferentes tecidos poderiam ajudar a elucidar essa questão.

5.2 Reconstruções Filogenéticas

Filogenias são fundamentais para o entendimento não somente da história evolutiva contida naas sequencias, mas também para a compreensão dos mecanismos evolutivos que moldaram essas sequencias.

As filogenias construídas a partir do *partgen* mostraram que os ERV-K (HML2) podem ser divididos em dois grandes clusters, chamados Grupo N/O e Grupo I (Romano *et al.,* 2006). Entretanto, Belshaw e colaboradores (2004) já haviam observado a presença de dois principais clusters em filogenias de ERV-K: (*i*) cluster exclusivo de provírus humanos e (*ii*) cluster contendo elementos de outros primatas além de humanos. Porém, como utilizamos um maior número de amostras obtidas a partir de quatro primatas, observamos que ambos os clusters continham ERV-K de primatas do velho mundo.

Um dos primeiros métodos de classificação de ERV-K foi feito baseado em uma deleção de um fragmento de 292 nucleotídeos entre os genes de polimerase e envelope (Ono *et al.,* 1986). Essa deleção divide os provírus em tipo 1 (os que possuem a deleção) e tipo 2 (os que não possuem). Dentre todos os elementos que foram mapeados nesse trabalho, treze provírus de humanos, três de chimpanzé e apenas um de orangotango são do tipo 1. Os elementos encontrados em macaco rhesus, por sua vez, são todos pertencentes ao tipo 2. Assim, visto que foi encontrado apenas um elemento do tipo 1 em *Pongo* e, assumindo que foi amostrado um número suficiente de elementos, podemos dizer que o evento da deleção ocorreu em algum momento após divergência dos Hominoidea dos Cercopithecoidea, ou seja, entre 12 e 21 milhões de anos atrás (Glasko e Nei, 2003). Esses dados confirmam os resultados obtidos por Mayer e colaboradores (1998), que não obtiveram sucesso ao tentar amplificar provírus do tipo 1 em amostras de babuínos e de *Macaca*.

O aparecimento da deleção nas filogenias ocorre a partir do PongERV-K6, o qual é ortólogo do HERV-K8. Abaixo disso, há somente clusters de elementos do tipo 2. Assim, assumindo que esse evento de deleção ocorreu apenas uma vez, e a partir disso os provírus de ambos os tipos continuaram a crescer em número, esperaríamos que a deleção fosse uma sinapomorfia, ou seja, uma característica ancestral que definisse grupos monofiléticos, distinguíveis pela presença ou ausência da deleção. Entretanto, provírus dos tipos 1 e 2 ocorrem nos mesmos clusters diversas vezes, com pequena distância entre si, principalmente nos ramos terminais da árvore. Observado também por Costas (2001) em ERV-K de humanos, esse é um resultado inesperado para a evolução de um elemento com duas linhagens principais que divergiram há mais de 12 milhões de anos. Isso apenas pode ser

explicado por eventos de recombinação não alélica ou conversão gênica entre os provírus. A recombinação poderia ainda ocorrer durante a transcrição reversa, entre duas moléculas de RNA co-encapsidadas, como ocorre em retrovírus exógenos.

Elementos transponíveis atuando como *hotspots* para recombinação, duplicação ou conversão gênica não é novidade (Schwartz *et al.,* 1998; Barbulescu *et al.,* 1999; Sverdlov, 2000). Além de contribuírem para a criação de novos genes ou promotores, (Lower *et al.,* 1996), recombinação entre provírus pode ser extremamente útil para recuperar a atividade de um elemento, perdida pelo acúmulo de mutações. Essa hipótese parece plausível diante do cenário de atividade recente dos HERV-K demonstrado pela filogenia, onde elementos de ambos os tipos estão ativos.

Além do padrão de clusterização observado em provírus humanos (Grupo N), indicando recente atividade, um cluster de elementos de orangotango dentro deste grupo sugere que estes também foram originados por replicação ativa, e não por simples duplicação. Isso foi comprovado nos resultados das análises das regiões flanqueadoras, mostrando que as regiões de integração são distintas entre estes elementos. Interessantemente, os RhERV-K basais ao Grupo N exibem um padrão similar, que também poderia indicar espalhamento por atividade, visto que nas análises de flancos não foram encontrados elementos originados por duplicação. Infelizmente, porém, a baixa qualidade do genoma de *M. mulata* no presente momento não permite nenhuma conclusão definitiva.

Foi proposto recentemente um modelo de evolução para HERV-K, no qual as chances de fixação de um elemento na população dependem diretamente do seu impacto no *fitness* do hospedeiro (Belshaw *et al.,* 2005). Esse modelo parte do pressuposto que os elementos ativos são potencialmente mais deletérios para o hospedeiro, uma vez que apresentam capacidade de se integrarem aleatoriamente no genoma e possuem promotores ativos (LTR) (Dunn *et al*., 2005; Landry *et al*., 2002; Medstrand *et al*., 2001). Portanto, de acordo com Belshaw e colaboradores (2005), somente os elementos que adquirissem mutações deletérias, por recombinação, deleção ou *snps* (*single nucleotide polymorphism*), seriam fixados na população. Esse modelo, embora explique o grande número de elementos inativos no genoma humano, não explica a fixação de muitos provírus ativos do Grupo N (Barbulescu *et al*., 1999; Turner *et al*., 2001).

Por outro lado, embora muito deles estejam presentes em caráter homozigoto na população humana, alguns ainda são polimórfico em termos de presença/ausência na população, sendo que, se forem de fato deletérios, mesmo que fracamente, podem ainda ser perdidos. Além disso, se em média a fixação de um novo alelo leva 4N gerações (N =

tamanho da população) (Beiguelman, 1994), em uma população pequena de 1000 indivíduos, com um tempo de geração de aproximadamente 20 anos, um ERV levaria cerca de 80 mil anos até ser completamente fixo ou perdido. Em decorrência da variação aleatória das frequências gênicas ao longo das gerações, tanto pode haver a eliminação como a fixação de um alelo, independente do valor adaptativo que este confere a população (deriva genética). Dessa forma, a atual prevalência dos provírus que se integraram antes da saída do homem da África, deve ser função da prevalência inicial do provírus na população que migrou para outro continente, bem como seu tamanho efetivo. A deriva deve ter tido um papel fundamental para a fixação de alelos nas populações primitivas, cujo tamanho efetivo não deveria exceder 100 indivíduos (Beiguelman, 1994). Suportando isso, o padrão de polimorfismo encontrado para dois dos mais recentes provírus (HERV-K113 e HERV-K115) entre diferentes populações é condizente com o padrão de migração humana (Macfarlane e Simmonds, 2004), onde se observa maior frequência em Africanos do que em não Africanos.

5.3 Seleção como indicação de atividade recente

Existem diferentes mecanismos pelos quais os ERVs podem proliferar no genoma hospedeiro. A proliferação pode ocorrer por retrotransposição, onde é necessária apenas a atividade da transcriptase reversa (RT) da Integrase. Nesse caso, a replicação pode ocorrer em *cis,* onde o próprio provírus supre todas as proteínas necessárias para a replicação, ou em *trans,* onde proteínas de outros provírus (endógenos ou exógenos), concomitantemente expressos na célula, complementam a replicação dos vírus defectivos. Alternativamente, ERVs podem formar partículas infecciosas e infectar outras células (o que requer a atividade do gene *env*). A distinção entre os processos acima descritos, bem como a inatividade dos provírus, pode ser sugerida a partir da detecção de seleção nos diferentes genes (Belshaw *et al.*, 2004). Enquanto reinfecção requer funcionalidade dos três genes, além da integridade do seu promotor, retrotransposição requer somente funcionalidade de *gag* e *pol* (não necessariamente no mesmo provírus se houver complementação em *trans*). Assim, se existe ou existiu atividade recente, esperaríamos encontrar sinais de seleção purificadora nos genes de ERV-K.

Já nas análises iniciais, os baixos valores de *dN/dS* para os diferentes genes (variando de 0.6 a 0.7) sugeriam que os mesmos se encontravam sob seleção negativa. Entretanto, apenas isso não permitia discernir quais e quantos códons estavam sob restrição funcional, e ainda, se os códons sob restrição eram os mesmos entre os primatas. As análises de seleção

por códons revelaram que alguns deles estão sob seleção em todos os primatas investigados. Embora os ERV-K de orangotango tenham apresentado um número de códons sob seleção equiparável ao de humanos em *gag*, o número de códons negativamente selecionados nos genes de polimerase e envelope de ERV-K de humanos foi consideravelmente maior do que nos demais hospedeiros, sugerindo que, além de replicação intracelular, como aparentemente ocorreu em todos os primatas, a amplificação por reinfecção pode ter sido um dos principais mecanismos de proliferação dos ERV-K em humanos (Belshaw *et al.*, 2004). Assumindo isso, deveríamos encontrar também seleção positiva agindo em alguns códons no envelope, como consequência da atividade do sistema imune do hospedeiro. De fato, foram encontrados seis códons sob seleção positiva no gene p36 do envelope dos HERV-K, contra apenas um em CERV-K, um em PongERV-K e nenhum em RhERV-K.

Inegavelmente, seleção atuando em genomas de retrovírus endógenos é um forte indicador de atividade. Embora existam diversos estudos de atividade de ERV-K em humanos, infelizmente ainda são escassos os trabalhos que investigam o mesmo em primatas não humanos (Greenwood *et al.,* 2005; Stengel *et al*., 2006). Porém, o número de códons sob seleção encontrado em PongERV-K concordam com a filogenia, que mostra um grupo relativamente recente originado por atividade. Os resultados ficam mais evidentes ainda para os ERV-K de humanos. Além disso, assumindo que o aumento em número de cópias pode também ter ocorrido por complementação em *trans*, a atividade de diversos elementos com ao menos um dos genes íntegros não surpreenderia. Finalmente, uma vez que a atividade de determinados provírus pode trazer consequências benéficas ao hospedeiro (Blond *et al*., 2000; Mi *et al*., 2000; Medstrand *et al*., 2001), a seleção por parte do hospedeiro atuando para a manutenção de alguns elementos não é inesperada.

5.4 História Demográfica de ERV-K em diferentes primatas

A teoria da coalescência, na qual alguns dos métodos utilizados nesse trabalho são baseados, é um modelo retrospectivo que pode ser usado para estimar a história demográfica de uma população até seu ancestral comum mais recente (Kingman, 2000). Previamente descrito por Kingman (1982), e independentemente por Tajima (1983) e Hudson (1983), o coalescente assume ausência de seleção, ausência de recombinação e de fluxo gênico entre as populações. Retrovírus endógenos, como qualquer população natural, não satisfazem todas essas premissas. Embora atualmente a maioria dos ERVs seja inativa, evoluindo portanto de maneira neutra, alguns deles permanecem ativos e estão sob seleção (Katzourakis *et al*., 2005;

Romano *et al*., 2006). Além disso, como discutido anteriormente, conversão gênica e recombinação também ocorrem com certa frequência. Dessa forma, os resultados observados nas descrições demográficas dos ERV-K poderiam ser reflexo da seleção natural, e não produto da real dinâmica dos provírus. No entanto, se a violação dessas condições exercesse de fato uma grande influência nas análises, esperaríamos encontrar assinaturas demográficas bem similares para todas as populações de ERV-K, uma vez que eles compartilham ortólogos entre os hospedeiros, e seus genes apresentaram valores gerais de *dN/dS* bastante similares entre as linhagens. Contudo, mesmo com alguma influência causada pela violação das premissas embutidas na teoria da coalescência, obtivemos diferentes dinâmicas de ERV-K entre os primatas. Ao contrário do que se imaginava, hospedeiros que são evolutivamente mais próximos não compartilharam as assinaturas mais parecidas de dinâmica de retroelementos. Se o impacto de um elemento móvel no genoma de organismos filogeneticamente relacionados é parecido, os dados então parecem sugerir que outros fatores, independentes da biologia do hospedeiro, influenciaram na dinâmica de espalhamento e fixação dos provírus.

Como discutido anteriormente, o fato de que os provírus humanos ainda não fixados aparecem nas populações em prevalências concordantes com o padrão de migração humana (Macfarlane e Simmonds, 2004), sugere que fatores demográficos exercem forte influência na fixação de ERV-K. Além disso, o tamanho efetivo de população influencia não só a fixação ou perda de alelos por deriva, mas também influencia na fixação de elementos transponíveis (TEs) (Rouzic *et al*., 2007) e nas taxas de transposição e seleção de TEs (Langley *et al*., 1983; Whitlock, 2003). Em vista disso, a dinâmica de ERV-K observada neste trabalho pode ser resultado das flutuações populacionais dos hospedeiros ao longo do tempo (como gargalos de garrafa e expansões), sofridas diferentemente por cada hospedeiro ao longo da sua evolução.

Por exemplo, o aumento de diversidade iniciado há 2.5 milhões de anos até o presente nos BSLs de ERV-K de humanos coincide com o aparecimento de *H. erectus* na África, que é provavelmente o ancestral do *H. sapiens* (Jones *et al.,* 1994). Crucialmente, o aumento de diversidade de HERV-K mais recente, evidenciado mais claramente nas filogenias, pode ser relacionado ao aparecimento do homem moderno há aproximadamente 200 mil anos, sua migração *out-of-Africa* e finalmente, pela sua posterior expansão populacional (Jones *et al.*, 1994). Interessante notar que a diminuição de diversidade, que seria esperada devido ao gargalo sofrido com a saída do homem da África não é detectada. Uma explicação para isso pode residir na atividade dos provírus. Os ERVs, além de aumentarem em número de cópias no genoma por duplicação de regiões cromossômicas, se multiplicam também por atividade.

Uma vez que há diversos elementos ativos no genoma, podemos imaginar um cenário onde, após as ondas migratórias, as populações isoladas sofreram, independentemente, um aumento na frequência de ERVs por replicação ativa, acumulando diversidade. O posterior fluxo gênico entre essas populações permitiu a reintrodução de alelos que possam ter sido perdidos por uma população (ou sequer existiram), mas fixados por outra. Interessantemente, suportando o argumento da relação entre demografia e assinatura genética, Kaessmann e colaboradores (2001) detectaram uma grande expansão da população humana nesse mesmo período através da análise de genes do cromossomo X de diferentes Hominoidea.

A dinâmica observada para os ERV-K de rhesus também pode ser explicada por fatores populacionais. A diminuição da diversidade observada nos últimos dez milhões de anos coincide com a emergência do gênero *Macaca,* que por sua vez é um dos grupos com maior número de espécies entre os *Cercopithecidae* (Brandon-Jones *et al.,* 2004). Depois disso, há cerca de 2 m.a., *M. mulatta* divergiu da linhagem que deu origem a *M. fascicularis*, ocupando a partir daí uma extensa área, do oeste da India até a costa da China. Uma vez que especiação, por produzir efeitos similares a gargalos de garrafa, tem como conseqüência imediata uma diminuição da diversidade, os eventos de especiação sofridos por esse grupo tiveram efeitos similares a gargalos de garrafa na arquitetura dos genomas, refletindo em uma dinâmica de ERV-K similar a de humanos, exceto pelo recente crescimento.

No caso de ERV-K de chimpanzés, a estabilidade observada nos últimos dez milhões de anos é condizente com a ausência de evidências de gargalos ocorridos no gênero *Pan*. O grande tamanho efetivo observado (*Ne.g*), bem maior dos que de ERV-K de humanos e rhesus, pode ser explicado pelo fato de que os chimpanzés possuem uma diversidade genética três vezes maior do que a observada pra humanos (Kaessmann *et al*., 2001). Além disso, o chimpanzé da África Central (*P. t. troglodytes*), do qual foi feito o sequenciamento do genoma utilizado nesse estudo, possui maior diversidade ainda do que *P. t. verus*, do Oeste da África (Fisher *et al*., 2004). O maior tamanho efetivo da linhagem ancestral de *Pan* (3,2 vezes maior do que *Homo*)*,* bem como a maior densidade por quilômetro quadrado, também suportam as diferenças observadas entre as duas dinâmicas (McKeown, 1988; Kaessmann *et al.*, 2001).

A dinâmica de ERV-K de orangotangos não mostrou sinais de perda de diversidade em nenhum momento. Ao contrário, esta só aumentou desde o tempo da integração no genoma ancestral. O gênero *Pongo* possui apenas duas espécies, *P. pygmaeus* (da ilha de Bornéu) e *P. abelii* (de Sumatra), que divergiram entre 2.7 e 5 milhões de anos atrás (Steiper, 2006). Essa última espécie, da qual foi feito o sequenciamento do genoma completo, possui

uma maior população efetiva e pelo menos o dobro da diversidade dos orangotangos de Bornéu (Steiper, 2006). Ademais, os orangotangos possuem a maior diversidade intrapopulacional dentre os *Hominoidea* (Kaessmann *et al.,* 2001), apenas se equiparando aos chimpanzés. Após a erupção de um vulcão que devastou Sumatra há aproximadamente 74 mil anos, a ilha foi recolonizada por animais vindos de Bornéu, sudeste da Ásia e Java. Uma vez que os efeitos da deriva possivelmente resultaram na fixação de ERV-K distintos entre essas populações, a união delas deve refletir uma grande diversidade de ERV-K, muitos dos quais talvez ainda não fixados. Assim, a expansão populacional ocorrida com os animais de Bornéu, associada a fusão de três populações distintas, pode explicar a crescente diversidade observada na dinâmica dos seus ERV-K. Além disso, como já foi discutido anteriormente, as datações e análises de padrão de integração sugerem que um grupo específico de PongERV-K aparentemente é produto de atividade replicativa dos elementos, o que evidentemente contribui para o aumento de diversidade perto do presente.

5.5 Atividade de ERV-K

No presente trabalho foi demonstrado que, com exceção de dois elementos, todos os provírus com data de integração menor que sete milhões de anos, pertencentes ao Grupo N, podem ter originado boa parte dos transcritos de HERV-K analisados, originados de pacientes com câncer e pacientes HIV-1 positivos. Não surpreende, entretanto, o fato de que os provírus mais recentes são preferencialmente expressos, uma vez que são mais conservados. Por outro lado, provírus mais antigos, pertencentes ao Grupo O e basais ao Grupo N (HERV-K20, K69, K51, etc.), também foram relacionados a transcritos, sugerindo que o nível geral de conservação de um provírus, que é geralmente proporcional ao tempo de sua integração, não é o único fator determinante de atividade. Dando suporte a essa idéia, provírus do Grupo I, também integrados há menos de sete milhões de anos, não clusterizaram com nenhum transcrito analisado. Com um chi-quadrado altamente significante, os dados parecem sugerir distinções funcionais entre os Grupos N/O e I.

Os resultados também evidenciaram uma diferença significativa em relação aos elementos que estão expressos em pacientes com câncer, onde 80% dos provírus ativos nestes pacientes são do tipo 1. Provírus do tipo 1 codificam uma oncoproteína chamada Np9, que é expressa somente em certos tipos de câncer e células de linhagem tumorais (Ambruester *et al*., 2002) e parece ser capaz de interferir nas vias metabólicas de diferenciação celular (Ambruester *et al.,* 2004). Provírus do tipo 2, por outro lado, codificam uma proteína similar

a Rev do HIV-1, denominada Rec, também detectada em tecidos tumorais, embora em menores quantidades (Yang *et al.,* 1999). Corroborando os resultados encontrados neste trabalho, a detecção de altos níveis de Np9 em detrimento aos baixos níveis de Rec em diversos tecidos transformados (Ambruester *et al.,* 2002) sugerem um possível envolvimento da proteína Np9 na tumorogênese. Um aspecto interessante é que, enquanto que o inativo Grupo I é constituído apenas de elementos do tipo 2, o Grupo N/O, com elementos recentes e ativos, possui provírus de ambos os tipos. Já foi discutido que os provírus dos dois tipos clusterizam juntos no Grupo N/O possivelmente devido a eventos de recombinação entre elementos do tipo 1 e 2. Talvez, os eventos de recombinação que certamente ocorreram com alguns deles foram determinantes para a recuperação da atividade de alguns, conforme postulado por Lower *et al.* (1996).

Embora os resultados tenham mostrado que os promotores dos provírus do Grupo N/O sejam bem mais conservados dos que os dos elementos do Grupo I, não foram encontradas distinções entre os promotores dos diferentes tipos de provírus. Se alguma diferença fosse encontrada, isso justificaria a expressão diferencial em pacientes com câncer, visto que os genes expressos não são os mesmos entre diferentes tecidos, dispondo assim de diferentes cofatores celulares. Visto que já foi reportada a interação física e funcional entre Np9 e outras proteínas, resultando em alterações na proliferação celular (Denne *et al.,* 2007), podemos especular que proteínas expressas diferencialmente em células transformadas interajam fisicamente com o gene Np9, formando complexos que inibem algum supressor de expressão ou, estimule diretamente a atividade promotora da LTR. Embora ainda não seja possível elucidar as razões para essa expressão diferencial, é igualmente especulativo discutir se a superexpressão de HERV-K nesses tecidos é causativa ou consequência de alterações na malha gênica celular.

Embora já existam evidências de que proteínas de vírus exógenos, como herpes simplex 1 (HSV-1) tenham a capacidade de interagir fisicamente com promotores de ERV-K (Kwum et al., 2002), não foram encontradas evidências de mecanismos de transativação de HERV-K diretamente por proteínas expressas pelo HIV (Lower et al., 1996b). Por essa razão, especula-se a possibilidade de transativação dos HERVs por algum co-fator que seja diferencialmente expresso em células infectadas por HIV. Em um trabalho publicado em 2007, Garrison e colaboradores encontraram evidências para o que podemos chamar de "efeito positivo" da expressão de retrovírus endógenos em pacientes infectados por HIV-1 (Garrison *et al.*, 2007). Os autores verificaram que a expressão de HERVs é capaz de incitar a resposta imune celular, e ainda, encontraram células que reconhecem epítopos

compartilhados entre HERVs e HIV-1, aumentando dessa forma a resposta imune contra o vírus exógeno. Além disso, os autores demonstraram que as células T-HERV-específicas são capazes de destruir células que estejam expressando os peptídeos correspondentes. Em vista disso, podemos imaginar um cenário onde a co-expressão de HERVs funcionaria como um mecanismo alternativo de controle contra infecções exógenas. Ademais, se os co-fatores celulares responsáveis pela ativação dos HERVs expressos em células infectadas por HIV forem capazes de atuar em células vizinhas, não necessariamente infectadas, os HERVs dessas células também serão ativados, ocasionando um aumento no estímulo da resposta imune, bem como no recrutamento de células T CD8+. Consequentemente, a eliminação das células alvo próximas à célula infectada pode constituir um mecanismo alternativo de controle da infecção pelo hospedeiro.

A atividade de ERVs já foi exaustivamente relacionada a patologias como câncer, desordens neurológicas (Frank *et al*., 2005), diabetes tipo 1 (Margerat *et al*., 2004), doenças autoimunes (Magistrelli *et al.*, 2004; Christensen, 2005) e superexpressão em infecções por vírus exógenos (Stevens *et al.*, 1999). Por outro lado, existe o argumento de que a superexpressão, ou expressão diferencial de retrovírus endógenos talvez não traga apenas conseqüências deletérias para o hospedeiro. Em plantas (Wessler, 1996; Grandbastien, 1998) e em *Drosophila* (Viera e Biemont, 1996), a atividade de retrotransposons está ligada diretamente resposta imediata em situações de stress biótico e abiótico. Dessa forma, assumir que a expressão diferencial de ERV-K seja de alguma forma uma resposta do organismo a situações de stress (como transformação celular e invasão do organismo por um retrovírus exógeno) parece um possível cenário, que certamente merece ser melhor investigado.

5.6 Considerações finais: Interação entre ERV e hospedeiros

5.6.1 Coevolução?

Por invadirem o hospedeiro e utilizar sua maquinaria para amplificação, elementos transponíveis (TEs) foram descritos como sequências parasitas no final da década de 70, (Doolittle e Sapienza, 1980). Entretanto, coevolução entre parasita e hospedeiro é reconhecida como um importante mecanismo de geração de diversidade (Thompson, 1994; Buckling e Hodgson, 2007). Em sistemas modelo, a invasão de um organismo por um patógeno contribui com a evolução do sistema de defesa do hospedeiro, que por sua vez

direciona a evolução do parasita para um aumento da sua virulência (Thompson, 1994). Esse tipo de relação, conhecido como coevolução antagonista, é reconhecido por impactar na manutenção da reprodução sexual e na dinâmica populacional do hospedeiro e do parasita (Hamilton *et al*., 1990; Thompson, 1998). Porém, se tratando de retrovírus endógenos, essa bidirecionalidade é um pouco diferente. Devido a integração no genoma hospedeiro e posterior transmissão vertical, o potencial patogênico de um ERV tende a diminuir ao longo do tempo, caso contrário ele é purgado. Este processo dinâmico coevolutivo acontece desde a integração do elemento no genoma hospedeiro, porém, a relação entre eles tende a mudar ao longo do tempo, de deletéria a adaptativa. Assim, embora invasão, integração e fixação sejam, até certo ponto, processos estocásticos, que dependem de diversas variáveis, a "domesticação" ao longo do tempo dos elementos fixados, como forma de equilíbrio entre parasita e hospedeiro, é uma conseqüência real (Volff, 2006; Rouzic *et al*., 2007). De fato, a utilização de material genético (regulatório e codante) de ERVs (domesticação) pelo hospedeiro pode ser observada em diversas situações: LTRs provirais funcionando como promotor alternativo de genes do hospedeiros (Medstrand *et al*., 2000; Landry *et al*., 2002; Landry *et al*., 2003), epítopos de ERVs recrutando células T em situações de infecções exógenas (Garrison *et al*., 2007) e em proteínas provirais participando de diferenciação de tecido placentário (Mallet *et al.,* 2004) .

5.6.2 Reprodução sexuada e evolução de TEs

A interação entre parasitas e hospedeiros é proposta como uma das razões que favorece a reprodução sexuada sobre a assexuada, a despeito do custo duplo do sexo (Hamilton *et al*., 1990). Uma vez que proporciona a combinação de diferentes alelos, acelerando o aparecimento de variabilidade, a reprodução sexuada pode ser vista como uma adaptação dos organismos para aumentar a resistência a parasitas (Hamilton *et al*., 1990). Dessa forma, se elementos transponíveis são como parasitas (ao menos logo após a invasão do genoma hospedeiro), a reprodução sexuada deveria atuar também no controle e eliminação desses elementos. Entretanto, devido ao grande número de TEs em todos os genomas e, principalmente de retrovírus endógenos em vertebrados, isso não parece ocorrer.

A grande tolerância a TEs em organismos que se reproduzem sexuadamente, embora paradoxal, pode ser vista como outra vantagem que o sexo permite. Segundo Hamilton *et al*., (1990), uma das principais vantagens do sexo seria a maior permissividade a manutenção de alelos deletérios no genoma. Isso porque, a vantagem do ganho de material genético (TE)

potencialmente útil em longo prazo, compensaria o seu caráter deletério. Ao final, os novos genes ou promotores podem se tornar parte da malha gênica do hospedeiro, como de fato ocorreu em muitos exemplos anteriormente discutidos.

Existem diversas famílias de transposons, retrotransposons e ERVs no genoma dos diferentes organismos. Aparentemente, os hospedeiros são 'seletivos' `a presença de certos tipos de TEs em detrimento a outros, possivelmente devido à fatores genéticos e populacionais. Os resultados obtidos neste trabalho sugerem que a dinâmica evolutiva de ERV-K, especificamente, não depende de um único fator, sendo a demografia dos hospedeiros aparentemente um ponto decisivo na fixação ou perda de elementos. A utilização de promotores e genes de ERV-K por parte do hospedeiro parece ser uma consequência inevitável, e que certamente teve um papel crucial na evolução dos seus genomas.

6 CONCLUSÃO

Os provírus da família K estão presentes nos genomas de primatas do Velho mundo, integrados há aproximadamente 35 milhões de anos. Fez parte deste trabalho encontrar, caracterizar e mapear genomas completos de ERV-K em alguns destes primatas. Análises destes provírus sugeriram que a fixação e perda de ERV-K no genoma dos hospedeiros depende, primariamente, de fatores demográficos e populacionais dos hospedeiros, como expansões e gargalos de garrafa sofridos ao longo da sua evolução.

Além de prover adicionais evidências a respeito da existência de membros ativos e recentemente integrados no genoma humano, os resultados sugeriram que há provirus recentes também no genoma de orangotangos. Isso porém, parece não ter ocorrido nos genomas de macaco rhesus.

Embora a atividade de ERV-K em humanos já tenha sido descrita em diversas situações, este trabalho foi inédito ao associar os transcritos obtidos em outros trabalhos aos respectivos provírus. Nesse aspecto, foi surpreendente o achado de que não só os elementos mais recentes são expressos, mas também alguns dos mais antigos elementos. Mais interessante ainda, a atividade pareceu ser característica apenas dos elementos pertencentes a um determinado Cluster (Grupo O/N). A expressão dos elementos foi associada à regiões promotoras da LTR proviral, e não exatamente a integridade de seus genes. Esse resultado pode ter duas implicações: (*i*) Como o espalhamento de ERV-K no genoma hospedeiro não requer integridade total do genoma proviral, é possível que a complementação em *trans* seja um importante mecanismo de aumento de número de cópias, e (*ii*) a atividade diferencial entre provírus do Grupo O/N e Grupo I, além das diferenças entre transcritos obtidos de pacientes com diferentes patologias, sugere que a transativação dos ERV-K é específica, e pode estar relacionada a patologias distintas.

REFERÊNCIAS

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. J Mol Biol. 1990;215:403-10.

Anderson ML, Lindeskog M, Medstrand P, Westley B, May F, Blomberg J. Diversity of human endogenous retrovirus class II-like sequences. J Gen Virol. 1999;80:255–60.

Armbruester V, Sauter M, Krautkraemer E, Meese E, Kleiman A, Best B, Roemer K, Mueller-Lantzsch N. A novel gene from the human endogenous retrovirus K expressed in transformed cells. Clin Cancer Res.2002;8:1800-07.

Armbruester V, Sauter M, Roemer K, Best B, Hahn S, Nty A, et al. Np9 protein of human endogenous retrovirus K interacts with ligand of numb protein X. J Virol. 2004;78:10310-19.

Baltimore D. Viral RNA-dependent DNA Polymerase: RNA-dependent DNA Polymerase in Virions of RNA Tumour Viruses. Nature. 1970;226:1209–11.

Barbulescu M, Turner G, Seaman MI, Deinard AS, Kidd KK, Lenz J.Many human endogenous retrovirus K (HERV-K) proviruses are unique to humans. Curr Biol. 1999;9: 861-66.

Bejerano G, Lowe CB, Ahituv N, King B, Siepel A, Salama SR, et al. A distal enhancer and an ultraconserved exon are derived from a novel retroposon. Nature. 2006;441:87-90.

Belshaw R, Pereira V, Katzourakis A, Talbot G, Paces J, Burt A, Tristem M. Long-term reinfection of the human genome by endogenous retroviruses. Proc Natl Acad Sci. 2004; 6:4894-99.

Belshaw R, Katzourakis A. BlastAlign: a program that uses blast to align problematic nucleotide sequences. Bioinformatics. 2005;21:122-3.

Blomberg J, Benachenhou F, Blikstad V, Sperber G, Mayer J. Classification and Nomenclature of Endogenous Retroviral Sequences (ERVs). Problems and Recommendations. Gene. In Press 2009.

De acordo com:

International Comitee of Medical Journal Editors. Uniform Requirements for Manuscripts Subitted to Biomedical Journal: Sample References. Avaliable from: http://www.icmje.org [2007 May 22].

Blond JL, Beseme F, Duret L, Bouton O, Bedin F, Perron H, Mandrand B, Mallet F. Molecular characterization and placental expression of HERV-W, a new human endogenous retrovirus family. J Virol. 1999;73:1175–85.

Blond JL, Lavillette D, Cheynet V, Bouton O, Oriol G, Chapel-Fernandes C, et al. An envelope glycoprotein of the human endogenous retrovirus HERV-W is expressed in the human placenta and fuses cells expressing the type D mammalian retrovirus receptor. J. Virol. 2000;74:3321-29.

Boeke JD, Stoye JP. Retrotransposons, endogenous retroviruses and the evolution of retroelements. In: Coffin JM, Hughes SH, Varmus EH, editors. Retroviruses. New York: Cold Spring Harbor Laboratory Press; 1997. p. 343–435.

Bogerd HP, Wiegand HL, Hulme AE, Garcia-Perez JL, O'Shea KS, Moran JV, Cullen BR. Cellular inhibitors of long interspersed element 1 and Alu retrotransposition. Proc Natl Acad Sci USA. 2006;103: 8780-05.

Boller K, Konig H, Sauter M, Mueller-Lantzsch N, Lower R, Lower J, Kurth R.Evidence that HERV-K is the endogenous retrovirus sequence that codes for the human teratocarcinoma-derived retrovirus HTDV. Virology. 1993;196:349–53.

Brandle U, Ono H, Vincek V, Klein D, Golubic M, Grahovac B, Klein J. Trans-species evolution of Mhc-DRB haplotype polymorphism in primates: organization of DRB genes in the chimpanzee. Immunogenetics.1992;36:39–48.

Britten RJ, McCormack TJ, Mears TL, Davidson EH. Gypsy/Ty3-class retrotransposons integrated in the DNA of herring, tunicate, and echinoderms. J Mol Evol.1995;40:13-24.

Brookfield JFY. Models of spread of transposable elements containment of transposable element copy number. Genet Res. 1996;67:199-209.

Bucheton A. The relationship between the flamenco gene and gypsy in Drosophyla; how to tame a retrovirus. Trend Genet. 1995;224:33-39.

Buzdin A. HumanSpecific Endogenous Retroviruses. TSWJ. 2007; 7:1848–68.

Buzdin A, Kovalskaya-Alexandrova E, Gogvadze E, Sverdlov E. At Least 50% of Human-Specific HERV-K (HML-2) Long Terminal Repeats Serve In Vivo as Active Promoters for Host Nonrepetitive DNA Transcription. J Virol. 2006;80:10752-62.

Capy P, Buzdin C, Higuet D, Langin T. Classification of transposable elements. In Capy P, Buzdin C, Higuet D, Langin T, editors. Dynamic and evolution of transposable elements. New York: Spring Verlag; 1998. p. 155-170.

Callahan R, Chiu IM, Wong JF, Tronick SR, Roe BA, Aaronson SA, Schlom J. A new class of endogenous human retroviral genomes. Science. 1985;7:1208-11.

Camian JH, Sokal RR. A method for deducing branching sequences in phylogeny. Evolution. 1965; 19: 311-326.

Cavalli-Sforza LL, Edwards AWF. Phylogenetic analysis: models and estimation procedures. Am. J. Hum. Genet. 1967;19:122–257.

Christensen T. Association of human endogenous retroviruses with multiple sclerosis and possible interactions with herpes viruses. Rev Med Virol. 2005;15:179–211.

Coffin JM, Hughes SH, Varmus EH, editors. Retrovíruses. Plainview (NY): Cold Spring Harbor Laboratory Press; 1997

Cohen CJ, Lock WM, Mager DL. Endogenous retroviral LTRs as promoters for human genes: A critical assessment. Gene. In press 2009.

Collins TM, Wimberger P H, Naylor G N P. Compositional bias, character-state bias, and character-state reconstructionusing parsimony. Syst Biol. 1994:43:482–496.

Contreras-Galindo R, González M, Almodovar-Camacho S, González-Ramírez S, Lorenzo E, Yamamura Y. A new Real-Time-RT-PCR for quantitation of human endogenous retroviruses type K (HERV-K) RNA load in plasma samples: increased HERV-K RNA titers in HIV-1 patients with HAART non-suppressive regimens. J Virol Methods. 2006a;136:51-57.

Contreras-Gallindo R, Kaplan MH, Markovitz DM, Lorenzo E, Yamamura Y.. Detection of HERV-K (HML-2) viral RNA in Plasma of HIV type 1 Infected individuals. Aids Res Hum Evol. 2006b;22:979-84.

Contreras-Galindo R, Kaplan MH, Leissner P, Verjat T, Ferlenghi I, Bagnoli F, Giusti F, Dosik MH, Hayes DF, Gitlin SD, Markovitz DM. Human Endogenous Retrovirus-K (HML-2) Elements in the Plasma of People with Lymphoma and Breast Cancer. J Virol. 2008;82 (19):9329-36.

Costas J. Evolutionary Dynamics of the Human Endogenous Retrovirus Family HERV-K Inferred from Full-Length Proviral Genomes. J Mol Evol. 2001;53:237-43.

Cullen BR. Role and Mechanism of action of the APOBEC3 family of antiretroviral resistent factors. J Virol. 2006;80:1067-76.

Dangel AW, Mendoza AR, Baker BJ, Daniel CM, Carroll MC, Wu LC, Yu CY. The dichotomous size variation of human complement C4 genes is mediated by a novel family of endogenous retroviruses, which also establishes species-specific genomic patterns among Old World primates. Immunogenetics. 1994;40(6):425–36.

Dangel AW, Baker BJ, Mendoza AR, Yu CY. Complement component C4 gene intron 9 as a phylogenetic marker for primates: long terminal repeats of the endogenous retrovirus ERV-K(C4) are a molecular clock of evolution. Immunogenetics. 1995;42:41–52.

Deceliere G, Charles S, Biemont C. The dynamics of transposable elements in structured populations. Genetics. 2005; 169:467–74.

Denne M, Sauter M, Armbruester V, Licht JD, Roemer K, Mueller-Lantzsch N. Physical and Functional Interactions of Human Endogenous Retrovirus Proteins Np9 and Rec with the Promyelocytic Leukemia Zinc Finger Protein. J Virol. 2007; 81:5607–16.

Di Cristofano A, Strazzullo M, Parisi T, La Mantia G. Mobilization of an ERV9 human endogenous retroviral element during primate evolution. Virology. 1995;213:271–75.

Doolittle W, Sapienza C. Selfish genes, the phenotype paradigm and genome evolution. Nature. 1980; 284: 601–03.

Dunn CA, van de Lagemaat LN, Baillie GJ, Mager DL. Endogenous retrovirus long terminal repeats as ready-to-use mobile promoters: the case of primate beta3GAL-T5. Gene. 2005; 364:2-12.

Edgar RC. MUSCLE: multiple sequence alignment with high accuracy and high throughput, Nucleic Acids Res. 2004;32(5),1792-97.

Edwards AWF . Likelihood. Maryland:The Johns Hopkins University Press;1992.

Edwards AWF, Cavalli-Sforza L. The reconstruction of evolution. Ann Hum Gen. 1963; 27:105.

Efron B. Bootstrap methods: another look at the jackknife. Ann Statist. 1979;7:1-26.

Farris JS. Hennig86 version 1.5 manual; software and MSDOS program. 1988.

Farris JS. Phylogenetic analysis under Dollo's Law. Syst Zool. 1977;26: 77-88.

Faulkner GJ, Kimura Y, Daub CO, Wani S, Plessy C, Irvine KM, et al. The regulated retrotransposon transcriptome of mammalian cells. Nat Genetics. 2009;41:563-71.

Felsenstein J. Evolutionary trees from DNA sequences: a maximum likelihood approach. J Mol Evol. 1981; 17:368–376.

Felsenstein J. PHYLIP - Phylogeny Inference Package (Version 3.2). Cladistics. 1989; 5:164-66.

Felsenstein J. Phylogenies from molecular sequences: inference and reliability. Annu Rev Genet. 1988; 22:521–65.

Felsenstein J. Inferring Phylogenies. 2nd ed. Sunderland, Massachussets: Sinauer Associates; 2004.

Finnegan DJ. Eukaryotic transposable elements and genome evolution. Trends Genet. 1989;5(4):103-7.

Finnegan DJ. Transposable elements. In: indsley DL, Zimm GG, editors The genome of Drosophila melanogaster. San Diego: Academic Press; 1992; vol. 2, p. 1096–107.

Fischer A, Wiebe V, Paabo S, Przeworski M. Evidence for a complex demographic history of chimpanzees. Mol Biol Evol. 2004;21:799-808.

Fitch WM, Margoliash E. Construction of phylogenetic trees. Science. 1967;155:279-84.

Frank O, Giehl M, Zheng C, Hehlmann R, Leib-Mösch C, Seifarth W. Human Endogenous Retrovirus Expression Profiles in Samples from Brains of Patients with Schizophrenia and Bipolar Disorders. J Virol. 2005;79(17):10890–901.

Frazer KA, Chen X, Hinds DA, Pant PV, Patil N, Cox DR. Genomic DNA insertions and deletions occur frequently between humans and nonhuman primates. Genome Res. 2003; 13(3):341-6.

Garrison KE, Jones RB, Meiklejohn DA, Anwar N, Ndhlovu LC, Chapman JM, et al.T Cell Responses to Human Endogenous Retroviruses in HIV-1 Infection. PLoS Path. 2007;3:1617-27.

Glazko GV, Nei M. Estimation of divergence times for major lineages of primate species. Mol Biol Evol. 2003;20(3):424-34.

Goodchild NL, Wilkinson DA, Mager DL. Recent evolutionary expansion of a subfamily of RTVL-H human endogenous retrovirus-like elements. Virology. 1993;196:778.

Greenwood AD, Stengel A, Erfle V, Seifarth W, Leib-Mösch C. The distribution of pol containing human endogenous retroviruses in non-human primates. Virology. 2005;334(2):203-13.

Han K, Konkel MK, Xing J, Wang H, Lee J, Meyer TJ, et al. Mobile DNA in Old World monkeys: a glimpse through the rhesus macaque genome. Science. 2007;316:238-40.

Harris JW, Stocker H. Maximum Likelihood Method. In: Handbook of Mathematics and Computational Science. New York: Springer-Verlag; 1998. p. 824.

Hasegawa M, Kishino H, Yano T. Daying of the human-ape splitting by a molecular clock of mitochondrial DNA. J Mol Evol.1985;22:160-74.

Hastings WK. Monte Carlo Sampling Methods Using Markov Chains and Their Applications. Biometrika. 1970; 57: 97-109.

Henning, W. Grundzüge einer theorie der phylogenestichen systematik. Berlin: Deutscher Zentralverlag; 1950.

Henning W. Phylogenetic systematics. Urbana: University of Illinois Press; 1966.

Herbst H, Sauter M, Mueller-Lantzsch N. Expression of human endogenous retrovirus K elements in germ cell and trophoblastic tumors. Am J Pathol. 1996;149:1727–35.

Huelsenbeck JP. The robustness of two phylogenetic methods: four-taxon simulations reveal a slight superiority of maximum likelihood over neighbor joining. Mol Biol Evol. 1995;12:843-9.

Huelsenbeck JP. Is the Felsenstein zone a fly trap? Syst Biol. 1997;46:69-74.

Huelsenbeck JP, Rannala B, Larget B. A Bayesian framework for the analysis of cospeciation. Evolution. 2000;54:352-64.

Hudson RR. Properties of a neutral allele model with intragenic recombination. Theor Pop Biol. 1983; 23:183–201.

Hua-Van A, Le Rouzic A, Maisonhaute C, Capy P. Abundance, distribution and dynamics of retrotransposable elements and transposons: similarities and differences. Cytogenet Genome Res. 2005;110(1-4):426-40.

Johnson WE, Coffin JM. Constructing primate phylogenies from ancient retrovirus sequences. Proc Natl Acad Sci U S A. 1999;31:10254-60.

Jones S, Martin RD, Pilbeam DR. The Cambridge Encyclopedia of Human evolution. 2nd ed. Cambridge: Cambridge University Press; 1994. 524 p.

Jordan IK, Rogozin IB, Glazko GV, Koonin EV. Origin of a substantial fraction of human regulatory sequences from transposable elements. Trends Genet. 2003;19:68–72.

Jukes TH, Cantor CR. Evolution of protein molecules. In Munro HN, editor. Mammalian protein metabolism. New York: Academic Press; 1969. p. 21-123.

Kaessmann H, Wiebe V, Weiss G, Pääbo S. Great ape DNA sequences reveal a reduced diversity and an expansion in humans. Nat Genetics.2003;27:155–156.

Katzourakis A, Rambaut A, Pybus OG. The evolutionary dynamics of endogenous retroviruses. Trends Microbiol. 2005;13(10):463-8.

Katzourakis A, Tristem M, Pybus OG, Gifford RJ. Discovery and analysis of the first endogenous lentivirus. Proc Natl Acad Sci U S A. 2007;104:6261-5.

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

Kingman JFC. The coalescent. Stochast Proc Appl. 1982;13:235–48.

Kingman JFC. Origins of the Coalescent: 1974–1982. Genetics. 2000;156:1461-63.

Kluge AG, Farris JS. Quantitative phyletics and the evolution of anurans. Syst Zool. 1969; 18: 1-32.

Kuhner MK, Felsenstein J. A simulation comparison of phylogeny algorithms under equal and unequal evolutionary rates. Mol Biol Evol. 1994; 11:459-68.

Kulski JK, Gaudieri S, Inoko H, Dawkins RL. Comparison Between Two Human Endogenous Retrovirus (HERV)-Rich Regions Within the Major Histocompatibility Complex. J Mol Evol.1999;48:675-83.

Kulski JK, Shigenari A, Shiina T, Ota M, Hosomichi K, James I, Inoko H. Human Endogenous Retrovirus (HERVK9) Structural Polymorphism with Haplotypic HLA-A Allelic Associations. Genetics. 2008;182(1):445-57.

Kurth R, Lower R, Lower J, Harzmann R, Pfeiffer R, Schmidt CG, Fogh J, Frank H. Oncornavirus synthesis in human teratocarcinoma cultures and an increased antiviral immune reactivity in corresponding patients. In: Essex M, Todaro G, zur Hausen H editors. Viruses in naturally occurring cancers. Cold Spring Harbor (NY): Cold Spring Harbor Laboratory; 1980. p. 835–46.

Kwun HJ, Han HJ, Lee WJ, Kim HS, Jang KL. Transactivation of the human endogenous retrovirus K long terminal repeat by herpes simplex virus type 1 immediate early protein 0. Virus Res. 2008;86(1-2):93-100.

Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, Baldwin J, et al. Initial sequencing and analysis of the human genome. Nature. 2001;409 (6822);860-921.

Landry JR, Rouhi A, Medstrand P, Mager DL. The Opitz syndrome gene Mid1 is transcribed from a human endogenous retroviral promoter. Mol Biol Evol. 2002;19:1934-42.

Landry JR, Mager DL. Functional analysis of the endogenous retroviral promoter of the human endothelin B receptor gene. J Virol . 2003;13: 7459–66.

Langley CH, Brookfield JFY, Kaplan N. Transposable elements in Mendelian populations. I Theory. Genetics. 1983; 104:457-71.

Larget B, Simon DL. Markov Chain Monte Carlo Algorithms for the Bayesian Analysis of Phylogenetic Trees. Mol. Biol. Evol. 1999;16:750-59.

Lee J, Han K, Meyer TJ, Kim H-S, Batzer MA. Chromosomal Inversions between Human and Chimpanzee Lineages Caused by Retrotransposons. PLoS One. 2008;3(12):e4047.

Leib-Mosch C, Seifarth W. Evolution and biological significance of human retroelements. Virus Genes. 1996;11:133–45.

Leib-Mosch C, Brack-Werner R, Werner T, Bachmann M, Faff O, Erfle V, Hehlmann R. Endogenous retrohuman DNA. Cancer Res. 1990;50:5636–42.

Lerat E, Brunet F, Bazin C, Capy P. Is the evolution of transposable elements modular? Genetica. 1999;107:15-25.

Lewin R. Patterns in Evoluiton: The New Molecular View. New York: Scientific American Library;1997.

Lewis PO. Phylogenetic systematics turns over a new leaf. Trends Ecol Evol. 2001;16:30-37.

Li W-H, Graur D. Fundamentals of molecular evolution. 2nd ed. Sunderland: Sinauer Associates; 1991. 481 p.

Lower R, Boller K, Hasenmaier B, Korbmacher C, Müller-Lantzsch N, Löwer J, Kurth R. Identification of human endogenous retroviruses with complex mRNA expression and particle formation. Proc Natl Acad Sci USA. 1993;90(10):4480-84.

Lower R, Lower J, Kurth R. The viruses in all of us: Characteristics and biological significance of human endogenous retrovirus sequences. Proc Natl Acad Sci USA 1996;93(11):5177–84.

Lower R, Thelen K, Hasenmaier B, Knobetal M, Kurth R, Lower J. Molecular biology of the human endogenous retrovirus family HERV-K: transactivation by viral proteins. International Conference on AIDS,Vancouver, Canada; 1996; vol 11, p. 60. Abstract n.1035.

Lower R, Tönjes RR, Korbmacher C, Kurth R, Löwer J. Identification of a Rev-related protein by analysis of spliced transcripts of the human endogenous retroviruses HTDV/HERV-K. J Virol. 1995;69:141-9.

Macfarlane C, Simmonds P. Allelic variation of HERVK (HML-2) endogenous retroviral elements in human populations. J Mol Evol. 2004;59: 642-56.

Mack M, Bender K, Schneider PM. Detection of retroviral antisense transcripts and promoter activity of the HERV-K(C4) insertion in the MHC class III region. Immunogenetics. 2004;56 (5):321–32.

Maddison DR, Maddison WP. MacClade 4: Analysis of Parsimony and Character Evolution, 4.06th ed. Sunderland, MA: Sinauer Associates; 2003.

Magistrelli C, Samoilova E, Agarwal RK, Banki K, Ferrante P, Vladutiu A, Phillips PE, Perl A. Polymorphic genotypes of the HRES-1 human endogenous retrovirus locus correlate with systemic lupus erythematosus and autoreactivity. Immunogenetics. 1999;49(10)829–34.

Magness CL, Fellin PC, Thomas MJ, Korth MJ, Agy MB, Proll SC, et al. Analysis of the Macaca mulatta transcriptome and the sequence divergence between Macaca and human. Genome Biol, 2006;6:R60.

Marguerat S, Wang WY, Todd JA, Conrad B. Association of human endogenous retrovirus K-18 polymorphisms with type 1 diabetes. Diabetes. 2004;53(3):852–4.

Marques AC, Dupanloup I, Vinckenbosch N, Reymond A, Kaessmann H. Emergence of young human genes after a burst of retroposition in primates. PLoS Biol. 2005;3:e357.

Mayer J, Meese EU. The human endogenous retrovirus family HERV-K(HML-3). Genomics. 2002;80:331-43.

Mayer J, Meese E, Muller-Lantzsch N. Human Endogenous Retrovirus K Homologous Sequences and Their Coding Capacity in Old World Primates. J Virol. 1998;72:1870-75.

McCarthy EM, McDonald JF. LTR_STRUC: a novel search and identification program for LTR retrotransposons. Bioinformatics. 2003;19:362-67.

McClintock B. The origin and behavior of mutable loci in maize. Proc Natl Acad Sci USA. 1950;36 (6):344–55.

McDonald JF. Transposable elements: possible catalysts of organismic evolution. Tree. 1995;3:123-6.

Medstrand P, Blomberg J. Characterization of novel reverse transcriptase encoding human endogenous retroviral sequences similar to type A and type B retroviruses: differential transcription in normal human tissues. J Virol. 1993;67:6778–87.

Medstrand P, Landry JR, Mager DL. Long terminal repeats are used as alternative promoters for the endothelin B receptor and apolipoprotein C-I genes in humans. J Biol Chem. 2001;19:1896-903.

Medstrand P, van de Lagemaat L, Dunn C, Landry J, Svenback D, Mager D. Impact of transposable elements on the evolution of mammalian gene regulation. Cytogenet Genome Res. 2005;110(1-4):342–52.

Medstrand P, van de Lagemaat LN, Mager DL. Retroelement distributions in the human genome: variations associated with age and proximity to genes. Genome Res. 2002;12: 1483–95.

Metropolis N, Rosenbluth AW, Rosenbluth MN, Teller AH, Teller E. Equations of State Calculations by Fast Computing Machines. J of Chem Physics. 1953; 21: 1087–92.

Mi S, Lee X, Li X, Veldman GM, Finnerty H, Racie L, LaVallie E, Tang XY, Edouard P, Howes S, Keith JC Jr, McCoy JM. Syncytin is a captive retroviral envelope protein involved in human placental morphogenesis. Nature. 2000;403:785–789.

Michener CD, Sokal RR. A quantitative approach to a problem in classification. Evolution. 1957;11:130-162.

Minghetti PP, Dugaiczyk A. The emergence of new DNA repeats and the divergence of primates. Proc Natl Acad Sci USA. 1993;90:1872-76.

Morgan D, Brodsky I. Human endogenous retrovirus (HERV-K) particles in megakaryocytes cultured from essential thrombocythemia peripheral blood stem cells. Exp Hematol. 2004;32: 520-25.

Moyes DL, Martin A, Sawcer S, Temperton N, Worthington J, Griffiths DJ, Venables PJ. The distribution of the endogenous retroviruses HERV-K113 and HERV-K115 in health and disease. Genomics. 2005;86:337-41.

Mueller LD, Ayala FJ. Estimation and interpretation of genetic distance in empirical studies. Genet Res.1982; 40:127-37.

Nei M, Kumar S. Molecular evolution and phylogenetics. New York: Oxford University Press; 2000.

Noor MAF, Chang A. Evolutionary Genetics: Jumping into a New Species. Curr Biol. 2006;16:R890-92.

OhAinle M, Kerns JA, Li MM, Malik HS, Emerman M. Antiretroelement activity of APOBEC3H was lost twice in recent human evolution. Cell Host Microbe. 2008;4:249-59.

Ono M, Yasunaga T, Miyata T, Ushikubo H. Nucleotide sequence of human endogenous retrovirus genome related to the mouse mammary tumor virus genome. JVirol. 1986; 60:589–98.

Piriyapongsa J, Marino-Ramirez L, Jordan IK. Origin and evolution of human microRNAs from transposable elements. Genetics. 2007;176:1323–37.

Polavarapu N, Bowen NJ, McDonald JF. Identification, characterization and comparative genomics of chimpanzee endogenous retroviruses. Genome Biol. 2006a;7: R51.

Polavarapu N, Bowen NJ, McDonald JF. Newly Identified Families of Human Endogenous Retroviruses. J Virol. 2006b; 80(9):4640–42.

Promislow DEL, Jordan IK, McDonald JF. Genomic demography: a life-history analysis of transposable element evolution. Proc R Soc Lond. 1999;266:1555-60.

Rannala B, Yang Z. Probability distribution of molecular evolutionary trees: a new method of phylogenetic inference. J Mol Evol. 1996;43:304–11.

Ray DA, Xing J, Salem AH, Batzer MA.SINEs of a nearly perfect character. Syst Biol. 2006; 55:928-35.

Romano CM, de Melo FL, Corsini MA, Holmes EC, Zanotto PM. Demographic histories of ERV-K in humans, chimpanzees and rhesus monkeys. PLoS One. 2007;2:e1026.

Romano CM, Ramalho RF, Zanotto PM. Tempo and mode of ERV-K evolution in human and chimpanzee genomes. Arch Virol. 2006;151:2215-28.

Romano CM, Zanotto PM, Holmes EC. Bayesian coalescent analysis reveals a high rate of molecular evolution in GB virus C. J Mol Evol. 2008;66:292-97.

Saitou N, Nei M. The neighbor-joining method: a new method for reconstructing phylogenetic trees. Mol Biol Evol. 1987;4 (4):406-25.

Salemi M, Vandamme AM. The Phylogenetic Handbook: A Practical Approach to DNA and Protein Phylogeny. Cambridge: Cambridge University Press; 2003. 430 p.

Sawyer SL, Emerman M, Malik HS. Ancient Adaptive Evolution of the Primate Antiviral DNA-Editing Enzyme APOBEC3G . Plos Biol. 2004;2:1278-85.

Schwartz A, Chan DC, Brown LG, Alagappan R, Pettay D, Disteche C, et al. Reconstructing hominid Y evolution: X-homologous block, created by X-Y transposition, was disrupted by Yp inversion through LINE-LINE recombination. Hum Mol Genet. 1998;7(1):1-11.

Seifarth W, Spiess B, Zeilfelder U, Speth C, Hehlmann R, Leib-Mosch C. Assessment of retroviral activity using a universal retrovirus chip. J Virol Methods. 2003;112:79–91.

Seifarth W, Frank O, Zeilfelder U, Spiess B, Greenwood AD, Hehlmann R, Leib-Mösch C.Comprehensive analysis of human endogenous retrovirus transcriptional activity in human tissues with a retrovirus-specific microarray. J Virol. 2005;79(1):341-52.

Seino S, Bell GI, Li W-H. Sequences of Primate Insulin Genes Support the Hypothesis of a Slower Rate of Molecular Evolution in Humans and Apes than in Monkeys. Mol Biol Evol. 1992;9:193-203.

Serafino A, Balestrieri E, Pierimarchi P, Matteucci C, Moroni G, Oricchiob E, et al. The activation of human endogenous retrovirus K (HERV-K) is implicated in melanoma cell malignant transformation. Exp Cell Res. 2009;315:849-62.

Sergei L, Pond K, Frost SDW. Datamonkey: rapid detection of selective pressure on individual sites of codon alignments. Bioinformatics. 2005;21(10):2531-33.

Sheehy AM, Gaddis NC, Choi JD, Malim MH. Isolation of a human gene that inhibits HIV-1 infection and is suppressed by the viral Vif protein. Nature. 2002;418:646-50.

Smalheiser NR, Torvik VI. Mammalian microRNAs derived from genomic repeats. Trends Genet. 2005;21:322–26.

Smit AF. Interspersed repeats and other mementos of transposable elements in mammalian genomes. Curr Opin Genet Dev. 19999:657–63.

Steel MA, Hendy MD, Penny D. Loss of information in genetic distances. Nature. 1988; 336:118.

Steinhuber S, Brack M, Hunsmann G, Schwelberger H, Dierich MP, Vogetseder W. Distribution of human endogenous retrovirus HERV-K genomes in humans and different primates. Hum Genet. 1995; 96(2):188-92.

Steiper ME. Population history, biogeography, and taxonomy of orangutans (Genus: *Pongo*) based on a population genetic meta-analysis of multiple loci. J Hum Evol. 2006;50:509-22.

Steiper ME, Young NM, Sukarna TY. Genomic data support the hominoid slowdown and an Early Oligocene estimate for the hominoid–cercopithecoid divergence. Proc Natl Acad Sci USA. 2004;49:17021-26.

Stengel A, Roos C, Hunsmann G, Seifarth W, Leib-Mosch C, Greenwood AD. Expression Profiles of Endogenous Retroviruses in Old World Monkeys. J Virol. 2006;80;4415–21.

Stremlau M, Owen CM, Perron MJ, Kiessling M, Autissier P, Sodroski J. The cytoplasmic body component TRIM5a restricts HIV-1 infection in Old World monkeys. Nature. 2004; 427:848–53.

Suchard MA, Weiss RE, Sinsheimer JS. Bayesian selection of continuous-time Markov chain evolutionary models. Mol Biol Evol.2001;18:1001-1013.

Sverdlov ED. Retroviruses and primate evolution. BioEssays. 2000;22:161-71.

Swofford DL, Olsen GJ. Phylogeny reconstruction. In: Hillis DM, Moritz C, editors. Molecular Systematics. Sunderland, Massachusett: Sinauer Associates;1990. p 411-501.

Swofford DL, Maddison, WP. Reconstructing Ancestral Character States Under Wagner Parsimony. Math Biosc. 1987; 87: 199-229.

Swofford DL. PAUP*. Phylogenetic Analysis Using Parsimony (*and other methods). Version 4. Sunderland (MA): Sinauer Associates; 2000.

Tajima F. Evolutionary relationship of DNA sequences in finite populations. Genetics. 1983;105:437–60.

Takahata N, Satta Y. Evolution of the primate lineage leading to modern humans: Phylogenetic and demographic inferences from DNA sequences. Proc Natl Acad Sci. 1997;94:4811-15.

Tamura K, Nei M. Estimation of the number of nucleotide substitutions in the control region of mitochondrial DNA in humans and chimpanzees. Mol Biol Evol. 1993;10:512-526.

Tanaka S, Ikeda H, Otsuka N, Yamamoto Y, Sugaya T, Yoshiki T. Tissue Specific High Level Expression of a Full Length Human Endogenous Retrovirus Genome Transgene, HERV-R, Under Control of its Own Promoter in Rats. Transgenic Res. 2003;12(3):319-28.

Tavare S. Some probabilistic and statistical problems in the analysis of DNA sequences. Lec Math Life Sci. 1986;17:57-86.

Temin HM. Viral RNA-dependent DNA Polymerase: RNA-dependent DNA Polymerase in Virions of RNA Tumour Viruses. Nature. 1970;226:1211-13.

Teng B, Burant CF, Davidson NO. Molecular cloning of an apolipotprotein B messenger RNA editing protein. Science.1993;260:1816-19.

Thompson JN. Rapid evolution as an ecological process. Trends Ecol Evol. 1998;13:329–32.

Thorne JL, Kishino H, Painter IS. Estimating the rate of evolution of the rate of molecular evolution. Mol Biol Evol. 1998;15:1647-57.

Tuffley C, Steel M. Links between maximum likelihood and maximum parsimony under a simple model of site substitution. Bull Math Biol. 1997:59:581–607.

Turner G, Barbulescu M, Su M, Jensen-Seaman MI, Kidd KK, Lenz J. Insertional polymorphisms of full-length endogenous retroviruses in humans. Curr Biol. 2001;1:1531–35.

van de Lagemaat LN, Landry JR, Mager DL, Medstrand P. Transposable elements in mammals promote regulatory variation and diversification of genes with specialized functions. Trends Genet. 2003;19:530–36.

van de Lagemaat LN, Medstrand P, Mager DL. Multiple effects govern endogenous retrovirus survival patterns in human gene introns. Genome Biol. 2006;7:R86.

Voisset C, Bouton O, Bedin F, Duret L, Mandrand B, Mallet F, Paranhos-Baccala G. Chromosomal distribution and coding capacity of the human endogenous retrovirus HERV-W family. AIDS Res Hum Retroviruses. 2000;16:731–40.

Volff J-N. Turning junk into gold: domestication of transposable elements and the creation of new genes in eukaryotes. BioEssays . 2006;28:913-22.

Voytas DF, Konieczny A, Cummings MP, Ausubel FM. Copia-like retrotransposons are ubiquitous among plants. Proc Natl Acad Sci USA. 1992;89:7124-28.

Wagner WH. Problems in the classification of ferns. In: Bailey D, editors. Recent advances in Botany. Toronto: University of Toronto Press; 1961.p 841-844.

Wang-Johanning F, Frost AR, Johanning GL, Khazaeli MB, LoBuglio AF, Shaw DR, Strong TV. Expression of human endogenous retrovirus K envelope transcripts in human breast cancer. Clin Cancer Res. 2001;**7**:1553-60.

Wiley EO. Phylogenetic Systematics. New York: John Wiley; 1981.

Watkins WS, Ricker CE, Bamshad MJ, Carroll ML, Nguyen SV, et al. Patterns of ancestral human diversity: An analysis of alu-insertion and restriction-site polymorphisms. Am J Hum Genet. 2001;68:738-52.

Whitelaw E, Martin DIK. Retrotransposons as epigenetic mediators of phenotypic variation in mammals. Nat Gen. 2001;27:361-65.

Whitlock MC. Fixation Probability and Time in Subdivided Populations. Genetics. 2003;164:767–779.

Wicker T, Sabot F, Hua-Van A, Bennetzen JL, Capy P, Chalhoub B. A unified classification system for eukaryotic transposable elements. Nat Rev Genet. 2007;8:973-82. Review.

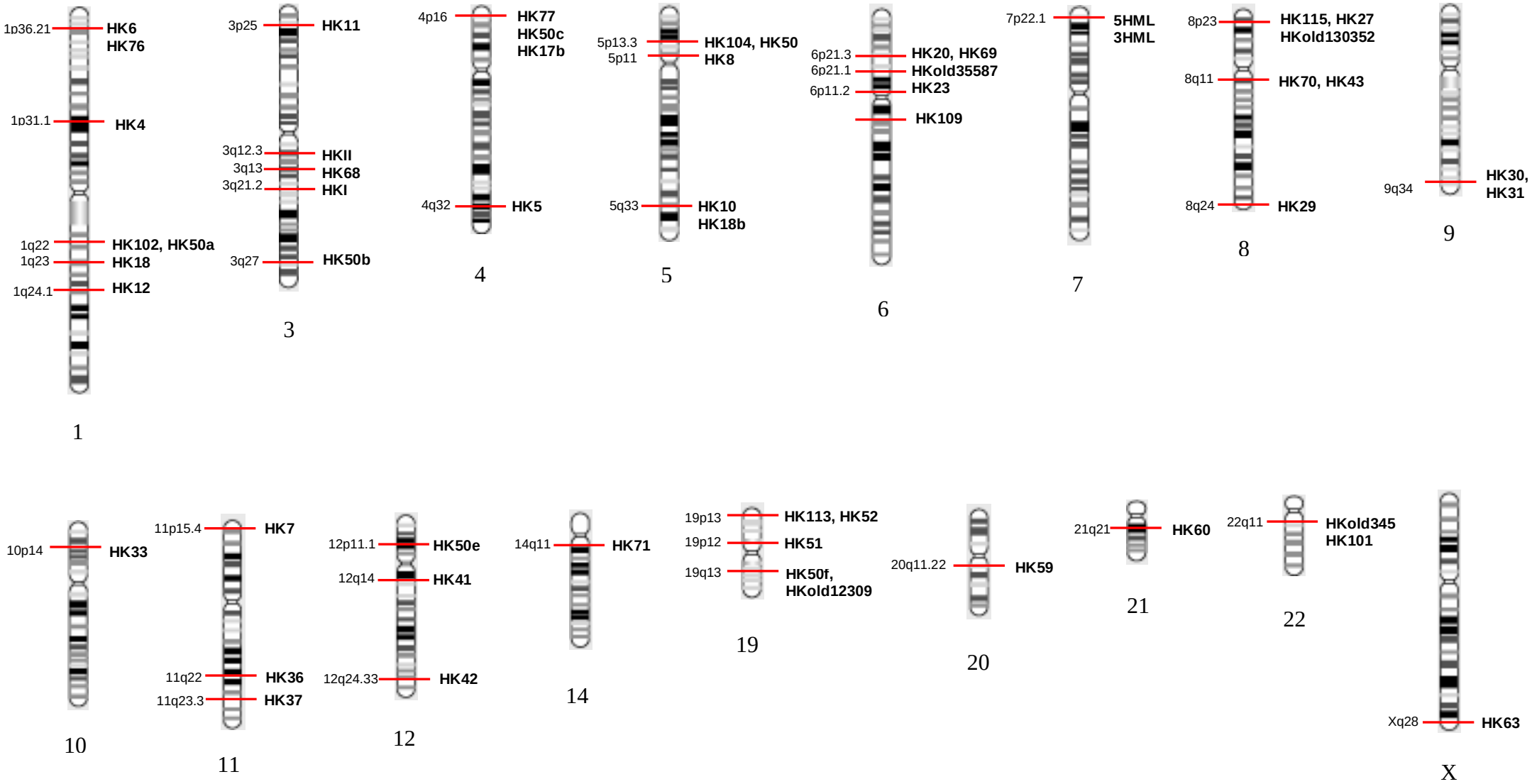
Xiong Y, Eickbush T. Origin and evolution of retroelements based upon their reverse transcriptase sequences. EMBO J. 1990;9:353-362.

Zwickl DJ. Genetic algorithm approaches for the phylogenetic analysis of large biological sequence datasets under the maximum likelihood criterion [Ph.D. Thesis]. Texas, Austin: The University of Texas; 2006.

ANEXOS

ANEXO A

Mapa de localização cromossômica dos HERV-K no genoma humano



Arch Virol (2006)
DOI 10.1007/s00705-006-0792-1

Archives of Virology
Printed in Austria

Tempo and mode of ERV-K evolution in human and chimpanzee genomes

C. M. Romano, R. F. Ramalho, and **P. M. de A. Zanotto**

Laboratory of Molecular Evolution and Bioinformatics, Department of Microbiology, Biomedical Sciences Institute – ICB II, University of São Paulo, São Paulo, Brazil

Received October 31, 2005; accepted April 24, 2006
Published online July 10, 2006

Summary. Several families of endogenous retrovirus (ERV) exist in copious numbers in the genomes of primate species. Therefore, we undertook a systematic search for endogenous retrovirus sequences from the ERV-K family, comparing across both human (*Homo sapiens*) and chimpanzee (*Pan troglodytes*) genomes. Using conserved motifs of the ERV-K as query we identified and characterized 76 complete ERV-K elements, 54 in human (HERV-K), 34 of which were described previously, and 21 in the chimpanzee (CERV-K). Phylogenetic analysis using coding regions and LTRs showed the existence of two main branches. Group I was the most heterogeneous and had an average integration time of 18.3 MYBP (million years before present), using rates ranging from 1.5 to 4.0×10^{-9} s/s/y (substitution per site per year). Group O/N integrated around 19.4 MYBP and nested Group N integrated about 14 MYBP. We found evidence for strong positive selection on the *gag*, *pol* and *env* coding regions and for A/T hypermutation. Our data suggest that the endogenous elements were possibly involved in chromosomal rearrangements and retained a great deal of information from their active stage, most likely as a consequence of host interactions. This study also contributes to the annotation effort of both human and chimpanzee genomes.

Introduction

Primate genomes carry thousands of copies of retroelements at different levels of integrity, such as retrotransposons and ERVs, which are remains of ancient and recent viral infection events in the germ line cells and subsequent vertical transmission [6]. The human genome has around 98,000 endogenous retroviruses (family *Retroviridae*) and thousands of solo long terminal repeats (LTR) [15, 30] (*i.e.*, 8% of the human genome). This may be explained by: (i) horizontal transmission when replicating as retroviral infectious agents, (ii) active retrotransposition in

cis, or by complementation in *trans* [5], (iii) replication using LINE machinery or, (iv) duplication of chromosomal loci during chromosome rearrangement events [6]. The ERV-K family has genomes of about 9.4 Kb, with 5′ and 3′ LTRs around 1000 bp [6, 20], found to be active in both human and chimpanzee lineages [13, 22, 28, 36], causing inversions, translocations, duplications, insertions and deletions (indels) [5, 14, 20]. Additionally, LTRs, acting as enhancers, could be important during evolution given their potential role in key cellular processes, from tissue differentiation to immune-modulation [10, 38]. Most ERV families differentiated after the separation of Old and New World monkeys at 30 to 45 million years before present (MYBP) [12, 19, 23]. In this study, we did a search for complete ERV-K proviral sequences in the human and chimpanzee genome database available on-line. Herein, we present a phylogenetic analysis on the tempo, mode of evolution and speciation of ERV-K in hominids.

Methods

ERV-K genome screening and annotation

With ERVFinder (available by request from the authors), both build 35 of the complete human (NCBI, http://www.ncbi.nih.gov [15]) and build 1 from the chimpanzee genomes (Chimpanzee Genome Sequencing Consortium [40]) were screened for complete ERV-K genomes using the consensus sequences for typical *gag*, *pol*, and *env*, in conjunction with the Blastn program [1] (default parameters), BLAT programs [17], Octopus [8], NCBI-Fastacmd [1], EMBOSS-Extractseq [33], and Blixem [37]. We also confirmed the chromosomal localization of human and chimpanzee ERV-K using the NCBI MapViewer and BLAT. We annotated the putative ERV-K genomes and identified typical ERV-K protein motifs using the Artemis software [34] and RPS-Blast [2], using HERV-K18 and -K10 as references (data available at the website http://www.lemb.icb.usp.br/hervs-k).

Genomic alignment and phylogenetic trees

Retroviral sequences were aligned with BlastAlign [4] using optimized parameter values. We also aligned ERV-K to the available genomic coding sequences of four simian exogenous retroviruses (SRV1, SRV2, SMRV-HLB, and Mason-Pfizer Monkey Virus) as outgroups, in order to orient cladogenetic events in time. The best evolutionary models and trees were found with the Model Test (V. 3.06) program [31], the parallel version of Tree-Puzzle [35] and PAUP 4.0b10 [39]. Patristic distances and dates were calculated with PatDist (Perl script available from the authors upon request), which uses branch lengths (and its standard errors) from trees calculated with PAUP. Patristic distances were averaged following the coalescent pattern of tree components.

Dating of integration events

An estimate on the time (T) of provirus integration can be obtained by using the relation $T = K/2r$ [21], where K is the distance between the 5′ and the 3′ LTR and r is the host mutation rate, ranging from 1.5 to 5×10^{-9} s/s/y (substitution per site per year) [6, 41]. We determined r from K values obtained by aligning 1000 bp of ERV-K flanks of human and chimpanzee, assuming their split at 6.5 MYBP [16]. LTRs were aligned and considered as independent taxa. Pairwise distances among 5′–3′ LTRs were calculated with the Mega 2 software [18] using the Tamura-Nei model with the optimal value for the shape parameter (α). The corrected genetic

distances between the 5′ and 3′ LTR of each genome were then used for dating integration events.

Selection detection

Evolutionary rates at different codon positions were estimated as average differences in genetic distances among the three positions, and distance matrices were calculated with PAUP 4.0b10 using the optimal evolutionary model determined by maximum likelihood. We also investigated selective pressures by calculating the synonymous (*dS*) and non-synonymous (*dN*) rates of change using several pairwise and codon-based methods implemented in Mega 2 and CODEML program in PAML v. 3.14 [43].

Results and discussion

Finding ERV-K genomes

Several molecular phylogenetic [5, 7, 24, 27, 42, 44] and functional studies using molecular approaches [3, 26, 29, 32, 36] have been done with HERV-K. Nevertheless, given the availability of human and chimpanzee genomes, we conducted a systematic computer search for complete ERV-K genomes on both genomes and found twenty-one full-length proviruses in the chimpanzee genome (CERV-K), and fifty-five in human (HERV-K). Table 1 shows: (i) nomenclatures used for both HERV-K and CERV-K, (ii) their chromosomal localization, (iii) the human clone in the genome containing it, (iv) ERV duplication found in the human genome, (v) the counterparts of HERV-K proviruses in chimpanzee (*i.e.*, CERV-K), and (vi) the absolute time of provirus integration in MYBP (million years before present) and the standard error estimated from LTR 5′ and 3′ genetic distance. Thirty-four out of fifty-five human ERV-K proviral sequences were present in the HERVd database. DotPlots comparing ERV-K genomes and pictograms showing human chromosomal localization are available at: (supplementary material at http://www.lemb.icb.usp.br/hervs-k). We found that several human ERV-K had counterparts in the chimpanzee genome (Fig. 1), which was confirmed by inspection of their flanking regions. The ERV-K genomes found in this study were fully annotated and are available at the GenBank (NCBI) under accession numbers DQ112093–DQ112156.

Phylogenetic analysis and genomic structural change

Maximum likelihood phylogenetic trees were inferred from (i) the complete genome alignment, (ii) only gene coding regions (*partgen*) containing *gag*, *pro*, and *pol* genes, and (iii) LTRs. A phylogenetic tree for complete genomes rooted at the SRV outgroup (Fig. 1a and b), pointed to the existence of at least two main groups, named Group O and I. Group O nested a relatively recent burst of cladogenesis, named Sub-group N. This result was confirmed with UPGMA clusters, using both *partgen* and complete ERV-K genomes (for which there was no SRV outgroup). The low bootstrap support below 100% for basal nodes may be a consequence of the poor placement of a root based on highly divergent SRV

Table 1. Nomenclature utilized and chromosomal localization of human and chimpanzee endogenous retrovirus. Duplicated HERVs, orthologous counterparts on both genomes and the absolute integration time estimated from 5′ and 3′ genetic distances are indicated

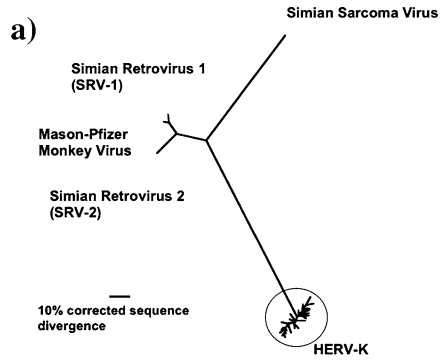
Provirus	Chromosome	ID (GenBank)	HERV-K duplicates	Chimpanzee orthologues	Age	
					MYBP	[s.e.]
Group N						
HERV-KI	3q21.2	AC092903			2.75	1.125
HERV-K11	3p25	AC018809			7.75	1.125
HERV-K113	19p13.11	AY037928			0.375	0.250
HERV-K10	5q33	AC016577			0	0
HERV-K33	10p24	AL392086		CKOLD12309	4.125	0.875
HERV-K36	11q22	AP000776			0.5	0.250
HERV-KII	3q12.3	AC084198			2.576	0.606
HERV-K37	11q23.3	AP002954			0.5	0.250
HERV-K41	12q14	AC025420			0.5	0.250
HERV-K5	4q32	AC106872			4.625	0.750
HERV-K8	5p11	AC125750			6.75	1
HERV-K102	1q22	AC044819	HK50a		0.25	0.250
HERV-K18	1q23	AL121985		CK60	5.25	0.875
HERV-K4	1p31	AC093156			0.125	0.125
HERV-K109	6q14.1	AL590785			0.375	0.250
HERV-K51	19p12	AC011467			7.625	1.250
HERV-K60	21q21	AF240627			0.375	0.250
HERV-K101	22q11	AC007326			0.25	0.125
HERV-K50d	5p13	AC025757			2.125	0.5
HERV-K68	3q13.2	AC078785			0	0
5HMLHOM 3HML2HOM	7p22.1	AC072054			0.25	0.125
HERV-K104	5p13.3	AC116309	HK50d		2.125	0.875
HERV-K115	8p23.1	AC134684	HK109		0.75	0.5
HERV-K50a	1q22	AL353807			0.25	0.25
HERV-K50b	3q27	AC099661/ AC133473			0.375	0.25
CERV-K60	1				7.5	1.375
CERV-K101	Y	AC0147136			2	0.6
CERV-K10	1				5.152	1.061
CERV-KOLD12309	8				1.2	0.6
CERV-K102	10				0	0
CERV-K100	Y				10	3.2
Group O						
HERV-KOLD12309	19q13.12	AC012309	HK50f	CK59	15.152	1.818
HERV-K50f	19q13	AC010632		CK59	15.152	1.818
HERV-K69	6p21.3	AL671879	HK20	CK20	17	2.4
HERV-K20	6p21.3	AL121932/ AL390196		CK20	17.2	2.4

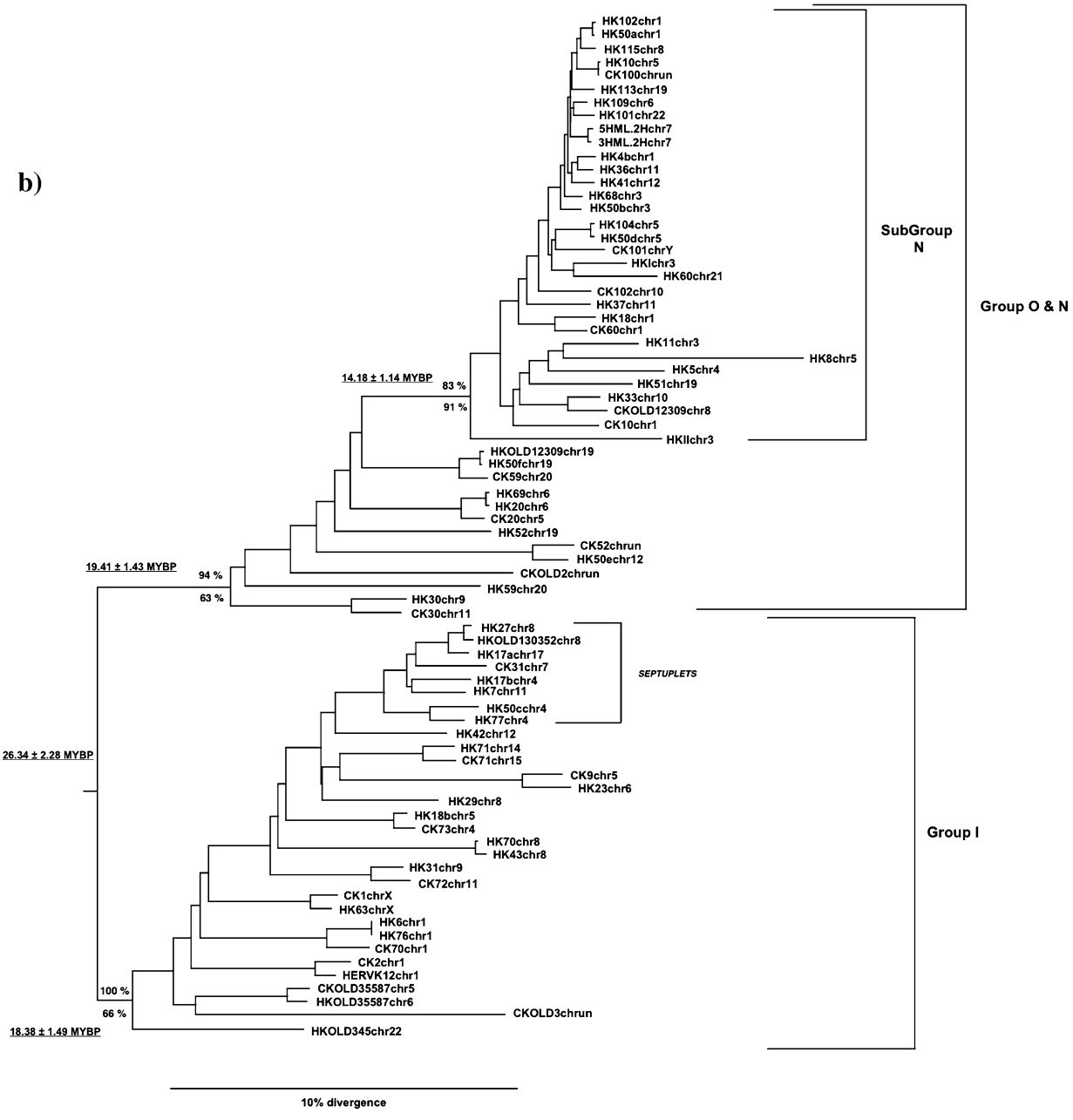
(*continued*)

Table 1 (*continued*)

Provirus	Chromosome	ID (GenBank)	HERV-K duplicates	Chimpanzee orthologues	Age	
					MYBP	[s.e.]
Group O						
HERV-K52	19p13	AC078899			26	2
HERV-K50e	12p11.1	AC144535		CK52	40	3.8
HERV-K30	9q34	AL355987		CK30	13.6	1.8
HERV-K59	20q11.22	AL031668			9.848	1.515
CERV-K59	20				8.03	1.515
CERV-K20	5				14.2	2.2
CERV-K52	Not found				42	4
CERV-KOLD2	Not found				48	4.8
CERV-K30	11				13	1.8
Group I						
HERV-K17a*	17				9.091	1.364
HERV-K17b*	4p16	AC108519			11.667	1.515
HERV-KOLD130352*	8p23	AC130352			9.848	1.515
HERV-K27*	8p23	AC068020			10.303	1.515
HERV-K50c*	4p16.1	AC105916			9.848	0.197
HERV-K7*	11p15.4	AC127526			9.697	1.364
HERV-K77*	4p16.3	AC116562			7.727	1.212
HERV-K71	14q11.2	AL136419		CK71	7.576	1.364
HERV-K42	12q24.33	AC026786			3.939	0.758
HERV-K23	6p11.2	AL590227		CK9	6.4	1.2
HERV-K29	8q24	AF235103			8.182	1.212
HERV-K31	9q34.11	AL441992		CK72	7.273	1.061
HERV-K70	8q11	AC113134	HK43		16.667	1.515
HERV-K43	8q11.2	AC113134			16.667	2.424
HERV-K18b	5q33	AC112175		CK73	7.121	1.061
HERV-K63	Xq28	AF277315		CK1	14.333	2.333
HERV-K6	1p36.21	AL603890	HK76	CK70	15.152	1.515
HERV-K76	1p36.21	AL365443		CK70	15.152	1.515
HERV-KOLD35587	6p21.1	AL035587		CKOLD35587	0	0
HERV-K12	1q24.1	AL611962		CK2	22	3
HERV-KOLD345	22q11	AP000345			18	4
CERV-K31*	7				11.97	1.515
CERV-K71	15				18.939	2.273
CERV-K9	5_randon				5.8	1
CERV-K72	11				6.667	1.061
CERV-KOLD3	Not found	AC144385			30	2.8
CERV-K73	4				6.515	1.061
CERV-K1	X	AC144385			12.333	2
CERV-K70	1_randon				18.485	1.97
CERV-KOLD35587	5				6.2	1
CERV-K2	1				22.333	3

*The human segmental duplications (setuplets) and their chimpanzee counterpart


a)
Simian Sarcoma Virus
Simian Retrovirus 1
(SRV-1)
Mason-Pfizer
Monkey Virus
Simian Retrovirus 2
(SRV-2)
10% corrected sequence
divergence
HERV-K


b)
HK102chr1
HK50achr1
HK115chr8
HK10chr5
CK100chrun
HK113chr19
HK109chr6
HK101chr22
5HML.2Hchr7
3HML.2Hchr7
HK4bchr1
HK36chr11
HK41chr12
HK68chr3
HK50bchr3
HK104chr5
HK50dchr5
CK101chrY
HKIchr3
HK60chr21
CK102chr10
HK37chr11
HK18chr1
CK60chr1
HK11chr3
HK8chr5
HK5chr4
HK51chr19
HK33chr10
CKOLD12309chr8
CK10chr1
HKIIchr3
14.18 ± 1.14 MYBP
83 %
91 %
HKOLD12309chr19
HK50fchr19
CK59chr20
HK69chr6
HK20chr6
CK20chr5
HK52chr19
CK52chrun
HK50echr12
CKOLD2chrun
HK59chr20
19.41 ± 1.43 MYBP
94 %
63 %
HK30chr9
CK30chr11
HK27chr8
HKOLD130352chr8
HK17achr17
CK31chr7
HK17bchr4
HK7chr11
HK50cchr4
HK77chr4
HK42chr12
SEPTUPLETS
26.34 ± 2.28 MYBP
HK71chr14
CK71chr15
CK9chr5
HK23chr6
HK29chr8
HK18bchr5
CK73chr4
HK70chr8
HK43chr8
HK31chr9
CK72chr11
CK1chrX
HK63chrX
HK6chr1
HK76chr1
CK70chr1
CK2chr1
HERVK12chr1
100 %
66 %
CKOLD35587chr5
HKOLD35587chr6
CKOLD3chrun
18.38 ± 1.49 MYBP
HKOLD345chr22
10% divergence
SubGroup
N
Group O & N
Group I

outgroups. Nevertheless, the tree for the LTRs (see supplementary material) had congruent adjacency patterns to that in Fig. 1b.

The evolutionary history of ERV-K

It is difficult to ascertain exactly when the split of major groups took place, due to possible differences in mutation rates between exogenous and endogenous retroviruses and different mutational pressures at different loci. Nevertheless, older ERV-K lineages found in the human genome (mainly from Group O and I), may reflect events that took place in ancestral primates [14, 32]. Most of the Group O and I ERV-K found in this work had human and chimpanzee counterparts, with an integration time from Early Miocene (20.6 MYBP for the Group O) up to Late Miocene (11.7 MYBP for the Group I). A recent cladogenetic burst at 8.4 MYBP gave rise to Group N, found in human and chimpanzee, but the newest lineages of Group N appeared around 100–500 thousand years ago, at the time of the emergence of modern humans (*e.g.*, HERV-K68) in Africa [16]. In several instances, LTR sequences had a higher rate of change than the rest of the genome (see the K value and the rates obtained for each provirus at supplementary material at http://www.lemb.icb.usp.br/hervs-k). We could not explain these results by the simple fact that we coded a single shared character at each gap in the LTR alignments in order to account for distance underestimation. Possibly, enhancer and promoter elements in LTRs could alter the transcription patterns of neighboring host genes [20]. As a consequence of that, a higher rate of nucleotide substitution in the LTR could lead to its inactivation, counteracting its deleterious effects.

Several instances of ERV-K duplications possibly took place along ancestral hominoid lineages, generating six HERV-K pairs (Table 1, Fig. 1). For each pair we noticed that the 5′ LTR of each provirus was more similar to the 5′ LTR of its counterpart than to its own 3′ LTR, as indicated by the LTR phylogenetic tree (not shown). Likewise, the 3′ LTR was more similar to the 3′ LTR of its counterpart than to its own 5′ LTR. Moreover, each pair of proviruses was in the same chromosome (Table 1), suggestive of a typical chromosomal duplication event. Because of the size of the human genome, and due to the random nature of retroviral integration [38], the finding of sister ERV taxons at nearby loci would be better explained by a chromosomal duplication rather than by replicative transposition. Nevertheless, it is difficult to estimate which provirus originated from which, since it is expected that both proviruses were identical at the time of duplication. Therefore, the data in Table 1 may provide misleading estimates on integration dates of duplicated elements. In sum, we may only estimate an absolute time of integration and perhaps

Fig. 1. Composite maximum likelihood phylogenetic tree for ERV-K. **a**) Tree rooted with the four exogenous simian retrovirus (SRV) outgroups using the *partgen* dataset. **b**) ERV-K phylogeny inferred from complete genomes of 76 ERV-K. The monophyletic groups I and O-N are shown in **b**. Levels of support and average integration date of each group are shown above (quartet-method with Tree-Puzzle) and below (maximum likelihood with PAUP) branches near relevant nodes

speculate on duplication times under the circumstances. We also found seven ERV-K ("septuplets", Fig. 1) in a cluster suggestive of seven independent host chromosome duplication events. This indicated that the integration of the ancestral element took place before the origin of *Homo* and *Pan*. This hypothesis was supported by our molecular phylogenies, since the septuplets' emergence was estimated at 12.78 ± 3.9 MYBP using LTR distances. Five of these genomes were reported before [5]. The 5′ LTR HERV-K115 (known to be active [3]), was more similar to both, 5′ and 3′ HERV-K109 LTRs than to its own 3′ LTR and their flanking regions were distinct. Moreover, the integration of HERV-K115 estimated from LTRs was about 750 thousand years and HERV-K109 at around 375 thousand years ago (Table 1). Therefore, it is likely that HERV-K109 may have originated by replicative transposition of HERV-K115 (or from an undetected closely related provirus).

Impact of ERV-K random integration on host evolution

Since their separation from a common ancestor, both lineages leading to humans and chimpanzee experienced chromosomal rearrangements [25]. A possible example may involve CERV-K1 and its orthologue HERV-K63, which is at the border of a duplicated inversion of a 30.2-Kbp genomic segment only in the human X chromosome (Fig. 2). Moreover, by aligning 27 Kbp from chimpanzee and its

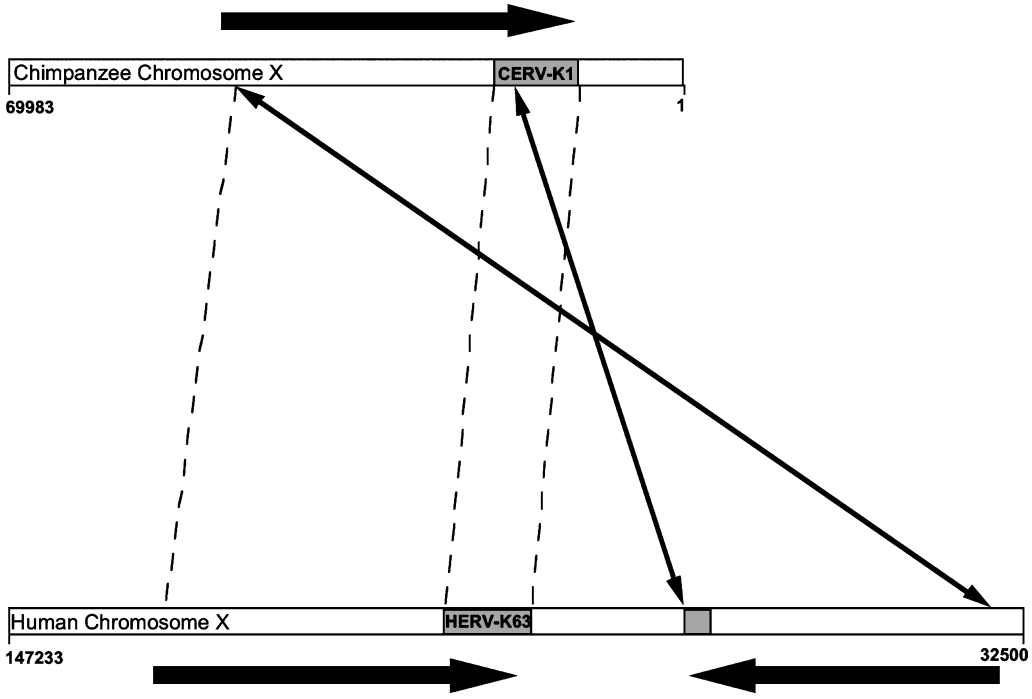


Fig. 2. Human and chimpanzee orthologous loci at chromosome X. Possibly, the replicative transposition of the 5′ CERV-K1 LTR was involved in the duplication and inversion of a 30.2-Kbp-long fragment of the human chromosome X. CERV-K1 and HERV-K63 were found to be orthologues integrated at the same region of chromosome X in both host species

equivalent in the human genome, we found that the duplication in the human genome took place approximately 500 thousand years ago.

AT composition bias as further evidence for host-virus interactions

ERV-K had a highly biased nucleotide composition, with an abundance of A and T in the viral positive strand (around 58% for both human and chimpanzee elements, and 63% for *partgen*). The exogenous counterparts of ERV-K, SRV-1 and MMTV, also have a biased AT-rich composition [46]. Interestingly, the increase of AT follows a preferential utilization of AT-rich third codon positions, with a four-fold increase in G → A and C → T transitions. The detection of a strong AT bias at the genomic coding strand does agree with the notion that those elements had a conflicting interaction with the host cells. Possibly, the AT bias is a consequence of most strategies to control the proliferation of virus and transposable elements, including gene methylation and siRNA-mediated co-suppression. Recently, the capacity of the host-encoded APOBEC nucleic-acid-editing enzyme in inducing lethal mutations in some endogenous retroviruses has been shown [9].

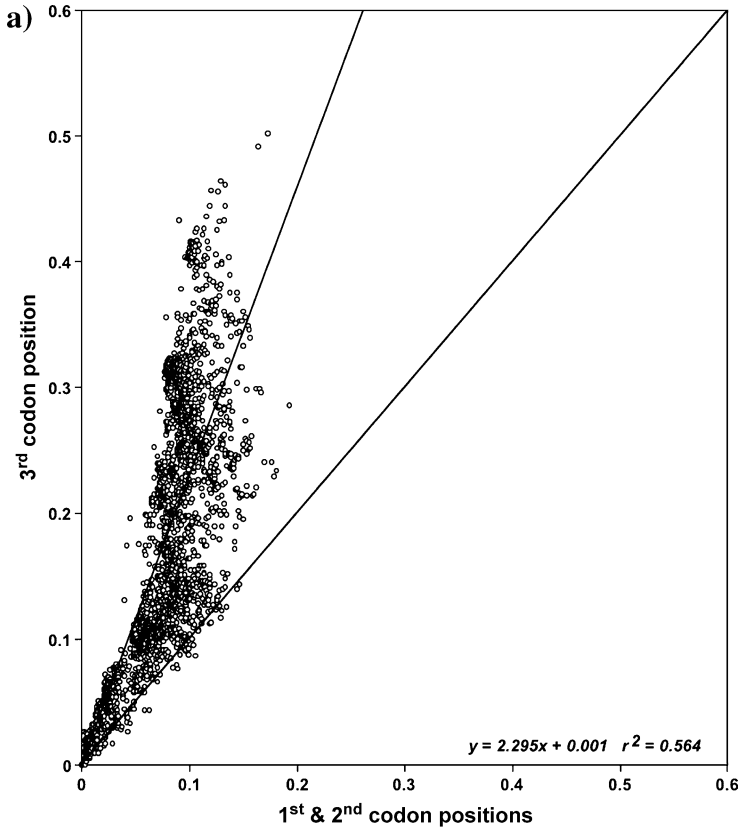
Our results are consistent with the hypermutation hypothesis as suggested by Esnault [9], and may support the notion that APOBEC could constitute a cellular defense mechanism against endogenous and exogenous retroviruses. This could be the case if uncontrolled transposition of endogenous retroviruses is deleterious for the host, and as a response, the APOBEC could inactivate ERVs by interfering with their propagation.

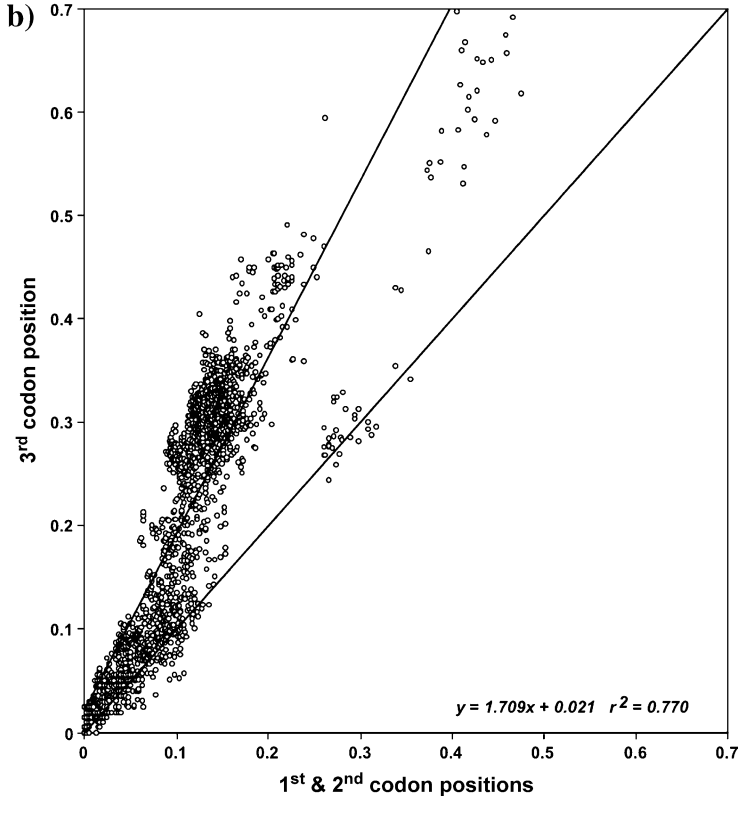
Host range expansion of ERV-K during primate evolution?

In order to test the monophily of ERVs, we generated a rooted tree for a 714-bp-(with gaps) long dataset including shared *gag* sequences from ERV-K from Groups O, N, I, and three Old World monkey sequences (one from vervet monkey *Cercopithecus aethiops* and 2 from rhesus monkey, *Macaca fascicularis*) [27]. The maximum likelihood tree obtained (not shown, but congruent to that in Fig. 1) indicated that *gag* regions from vervets and rhesus were sister taxa, branching within Group O (supplementary material at http://www.lemb.icb.usp.br/hervs-k). This reinforced the notion that the *gag* sequences from Old World monkey are more related to those belonging to Group O than to Group I or N. Using patristic distance estimate from a *gag* tree and a rate of 3.3×10^{-9} s/s/y, the split of ERV-K leading to the lineage infecting great apes and hominoids from the one leading to Old World monkeys appears to have occurred at 28.81 ± 6.45 MYBP ($K = 0.06580 \pm 0.01446$). These estimates are compatible with the argument of co-divergence of ERV-K and primate host species.

Evidence for selection in ERV proviral sequences

Evidence for negative (purifying) selection in HERV-W genomes has been associated with proviral activity, with possible implications in several pathologies, such as neuropsychiatric disorders and cancer [11]. Therefore, we calculated pairwise


a)
0.6
0.5
0.4
0.3
0.2
0.1
0
3rd codon position
y = 2.295x + 0.001 r2 = 0.564
0
0.1
0.2
0.3
0.4
0.5
0.6
1st & 2nd codon positions


b)
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
3rd codon position
y = 1.709x + 0.021 r2 = 0.770
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
1st & 2nd codon positions

distances using the *partgen* and *p36* (envelope) datasets. The regression slopes (Fig. 3a and b) indicated that the observed rates of change at the first and second positions were lower compared to the 3rd codon position for all endogenous retrovirus groups, indicative of functional constraints possibly associated with purifying selection. Lower rates of synonymous (*dS*) compared to non-synonymous (*dN*) changes at synonymous and non-synonymous sites, respectively, indicated purifying selection at the coding regions of all 3 groups (O, N and I), for human and chimpanzee. We also used a codon-based likelihood method implemented in the CODEML program (PAML, version 3.14). Nine codons in the human *p36* gene, two of which are also found in the chimpanzee, had high (>99%) posterior probabilities of being under positive selection. Most of the codons under positive selection were found in Group I, one specifically in Group N, another in Group O, but just one was found in both lineages of the three groups. In the *gag* and *pol* regions, 25 codons were positively selected. Twenty-tree were found only in human proviral genes, and two codons were found in both HERV and CERV-K. Of these, two codons were under positive selection in the *gag* (p24) domain, two in the RNase-H domain, one in the G-patch (glycine-rich nucleic-acid-binding) domain, only in HERV. Moreover, the RVT (RNA-dependent DNA polymerase) domain had seven codons, including two in CERV. Moreover, most parsimonious reconstructions (MPRs) of the *env* gene codons (data not shown) indicated that sites under positive selection were synapomorphies of old groups of lineages from Groups O and I, and that most of these sites underwent reversals or changes under positive selection afterwards. This is suggestive of serially incorporated adaptive changes common in exogenous viruses (*i.e.*, escape mutations) [46] associated with horizontal transmission events. Interestingly, of the 207 codons of *p36*, 13 (6.2%) had high posterior probability (>99.9%) of being under positive selection. On the other hand, fourteen (1%) codon sites out of 1288 in the *gag*, *pro*, and *pol* genes were under positive selection. The six-fold excess in the number of positively selected sites does support the notion that the *p36* product could be a target for immune neutralization, or other types of functional adaptations, like in HIV-1 [45]. In this regard, an antigen presented in samples of cutaneous and ocular melanoma appears to be expressed due to a frame shift of the gene coding for p36 [14].

Finally, a higher positive selection for Group N than for Group O is in accordance with the notion that "newer" viruses should present a stronger purifying selection signal, as shown before for both the HML-2 [5, 7] and the HERV-E,

Fig. 3. Excess of substitutions at the 3rd compared to the 1st and 2nd codon position of the ERV-K coding region (*partgen* dataset, including *gag* and *pro-pol* domains) (**a**), and *p36* (envelope) (**b**). Although the scattering of points above the diagonal provide evidence for purifying selection at the coding positions for both coding domains, some highly divergent *env* genes appear to show significant increase in the relative rate of coding position changes as indicated by points falling below the diagonal. Neutrally evolving genes or pseudo-genes would fall at the diagonal representing $dS/dN = 1$

F, H, K, R, S, T, W families [5]. This suggests that re-infection as an important mechanism for ERV radiation from replication-competent elements in the germ line, such as HERV-K113, -K115, -K10, etc. Moreover, we argue that integrated ERV-K elements carried information of continuous selective pressure from the time when they were actively replicating (indicated by overall purifying selection) and negotiating their survival around the host immune system (as shown by the presence of a few strongly positively selected sites and a strong A/T bias). Phylogenetic studies of RT and *env* genes of the baboon endogenous retroviruses (BaEV) in particular [42] and among retroelements in general have provided further evidence for rampant cross-species transmission. Possibly, ERV-K had a significant past history as an exogenous retrovirus, actively infecting divergent host species, often crossing species barriers like SIVSM from sooty mangabeys moving into humans as HIV-2 and SIVMAC from rhesus monkeys and, most notably the SIV from chimpanzee moving into humans as HIV-1.

Acknowledgments

The VGDN program FAPESP under project number 00/04205-6 funded this research. PMAZ holds a CNPq PQ Research Scholarship.

References

1. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ (1990) Basic local alignment search tool. J Mol Biol 215: 403–410
2. Altschul SF, Thomas LM, Alejandro AS, Jinghui Z, Zheng Z, Webb M, David JL (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res 25: 3389–3402
3. Barbulesco M, Turner G, Seaman IM, Deinard SA, Kidd KK, Lenz J (1999) Many human endogenous retrovirus K (HERV-K) proviruses are unique to humans. Curr Biol 9: 861–868
4. Belshaw R, Katzourakis A (2005) BlastAlign: a program that uses blast to align problematic nucleotide sequences. Bioinformatics 21: 122–123
5. Belshaw R, Pereira V, Katzourakis A, Talbot G, Paces J, Burt A, Tristem M (2004) Long-term reinfection of the human genome by endogenous retroviruses. Proc Natl Acad Sci 101: 4894–4899
6. Boeke JD, Stoye JP (1997) Retrotransposons, endogenous retroviruses and the evolution of retroelements. In: Coffin JM, Hughes SH, Varmus EH (eds) Retroviruses. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY, pp 343–435
7. Costas J (2001) Evolutionary dynamics of the human endogenous retrovirus family HERV-K inferred from full-Length poviral genomes. Mol Evol 53: 237–243
8. Durand P, Canard L, Mornon J (1997) Visual BLAST and Visual FASTA: graphic workbenches for interactive analysis of full BLAST and FASTA outputs under Microsoft Windows 95/NT. Comput Appl Biosci 13: 407–413
9. Esnault C, Heidmann O, Delebecque F, Dewannieux M, Ribet D, Hance AJ, Heidmann T, Schwartz O (2005) APOBEC3G cytidine deaminase inhibits retrotransposition of endogenous retroviruses. Nature 433: 430–433
10. Feuchter AE, Mager DL (1990) Functional heterogeneity of a large family of human LTR-like promoters and enhancers. Nucleic Acids Res 11: 1261–1270

11. Gifford R, Tristem M (2003) The evolution, distribution and diversity of endogenous retroviruses. Virus Genes 26: 291–315
12. Greenwood AD, Stengel A, Erfle V, Seifarth W, Leib-Mosch C (2005) The distribution of pol containing human endogenous retroviruses in non-human primates. Virology 334: 203–213
13. Griffiths JD (2001) Endogenous retroviruses in the human genome sequence. Genome Biol 6: 1017.1–1017.5
14. Hughes JF, Coffin JM (2001) Evidence for genomic rearrangements mediated by human endogenous retroviruses during primate evolution. Nature 29: 487–489
15. [IHGSC] International Human Genome Sequencing Consortium (2001) Initial sequence and analysis of the human genome. Nature 409: 860–921
16. Jones S, Martin RD, Pilbean DR (1994) The Cambridge encyclopedia of human evolution, 1st edn. Cambridge University Press, Cambridge
17. Kent WJ (2002) BLAT- the BLAST-like alignment tool. Genome Res 12: 656–664
18. Kumar S, Tamura K, Jakobsen IB, Nei M (2001) MEGA2: Molecular evolutionary genetics analysis software. Bioinformatics 17: 1244–1245
19. Lebedev YB, Belonovitch OS, Zybrova NV, Khil PP, Kurdyukov SG, Vinogradova TV, Hunsmann G, Sverdlov ED (2000) Differences in HERV-K LTR insertions in orthologous loci of humans and great apes. Gene 18: 265–277
20. Leib-Mosch C, Haltmeier M, Werner T, Geigl EM, Brack-Werner R, Francke U, Erfle V, Hehlmann R (1993) Genomic distribution and transcription of solitary HERV-K LTRs. Genomics 18: 261–269
21. Li W-H, Graur D (1991) Rates and patterns of nucleotide substitution. In: Fundamentals of molecular evolution, Chapter 4. Sinauer Associates, Sunderland
22. Lower R, Boller K, Hasenmaier B, Korbmacher C, Muller-Lantzsch N, Lower J, Kurth R (1993) Identification of human endogenous retroviruses with complex mRNA expression and particle formation. Proc Natl Acad Sci 90: 4480–4484
23. Lower R, Lower J, Kurth R (1996) The viruses in all of us: characteristics and biological significance of human endogenous retrovirus sequences. Proc Natl Acad Sci 93: 5177–5184
24. Malik HS, Henikoff S, Eickbush TH (2000) Poised for contagion: evolutionary origins of the infectious abilities of invertebrate retroviruses. Genome Res 10: 1307–1318
25. Marquès-Bonet T, Cáceres M, Bertranpetit J, Preuss TM, Thomas JW, Navarro A (2004) Chromosomal rearrangements and the genomic distribution of gene-expression divergence in humans and chimpanzees. Trends Genet 20: 524–529
26. Mayer J, Meese EU (2002) The human endogenous retrovirus family HERV-K (HML-3). Genomics 80: 331–343
27. Mayer J, Meese E, Mueller-Lantzsch N (1998) Human endogenous retrovirus K homologous sequences and their coding capacity in old world primates. J Virol 72: 1870–1875
28. Medstrand P, Mager DL (1998) Human-specific integrations of the HERV-K endogenous retrovirus family. J Virol 72: 9782–9787
29. Ono M, Yasunaga T, Miyata T, Ushikubo H (1986) Nucleotide sequence of human endogenous retrovirus genome related to the mouse mammary tumor virus genome. J Virol 60: 589–598
30. Paces J, Pavlícek A, Paces V (2002) HERVd: database of human endogenous retroviruses. Nucleic Acids Res 30: 205–206
31. Posada D, Crandall KA (1998) MODELTEST: testing the model of DNA substitution. Bioinformatics 14: 817–818

32. Reus K, Mayer J, Sauter M, Zischler H, Muller-Lantzsch N, Meese E (2001) HERV-K(OLD): ancestor sequences of the human endogenous retrovirus family HERV-K(HML-2). J Virol 75: 8917–8926
33. Rice P, Longden I, Bleasby A (2000) EMBOSS: The European molecular biology open software suite. Trends Genet 16: 276–277
34. Rutherford K, Parkhill J, Crook J, Horsnell T, Rice P, Rajandream M-A, Barrell B (2000) Artemis: sequence visualization and annotation. Bioinformatics 16: 944–945
35. Schmidt HA, Strimmer K, Vingron M, von Haeseler A (2002) TREE-PUZZLE: maximum likelihood phylogenetic analysis using quartets and parallel computing. Bioinformatics 18: 502–504
36. Simpson RG, Patience C, Lower R, Tonjes RR, Moore MDH, Weiss AR, Boyd TM (1996) Endogenous D-Type (HERV-K) Related sequences are packaged into retroviral particles in the placenta and posses open reading frames for reverse transcriptase. Virology 222: 451–456
37. Sonnhammer E, Durbin R (1994) A workbench for large scale sequence homology analysis. Comput Appl Biosci 10: 301–307
38. Sverdlov ED (2000) Retroviruses and primate evolution. Bio Essays 22: 161–171
39. Swofford DL (2002) PAUP*. Phylogenetic analysis using parsimony (* and other methods). Version 4. Sinauer Associates, Sunderland
40. The Chimpanzee Sequencing and Analysis Consortium (2005) Initial sequence of the chimpanzee genome and comparison with the human genome. Nature 437: 69–87
41. Turner G, Barbulescu M, Su M, Jensen-Seaman IM, Kidd KK, Lenz J (2001) Insertional polymorphisms of full-length endogenous retrovirus in humans. Curr Biol 11: 1531–1535
42. van der Kuyl CA, Mang R, Dekker JT, Goudsmit J (1997) Complete nucleotide sequence of simian endogenous type-D retrovirus with intact genome organization: evidence for ancestry to simian retrovirus and baboon endogenous virus. J Virol 71: 3666–3676
43. Yang Z (1997) PAML: a program package for phylogenetic analysis by maximum likelihood. Comput Appl Biosci 13: 555–556
44. Zanotto PMA, Gibbs MJ, Gould EA, Holmes EC (1996) A revaluation of the higher taxonomy of viruses based on RNA polymerases. J Virol 70: 6083–6096
45. Zanotto PMA, Kallas EG, de Souza RF, Holmes EC (1999) Genealogical evidence for positive selection in the *nef* gene of HIV-1. Genetics 153: 1077–1089
46. Zsíros J, Jebbink MF, Lukashov VV, Voúte PA, Berkhout B (1999) Biased nucleotide composition of the genome of HERV-K related endogenous retroviruses and its evolutionary implications. J Mol Evol 48: 102–111

Author's address: Paolo M. de A. Zanotto, Laboratory of Molecular Evolution and Bioinformatics, Department of Microbiology, Biomedical Sciences Institute – ICB II, University of São Paulo – USP, Av. Prof. Lineu Prestes, 1734, Sao Paulo, SP, 05508-000 Brazil; e-mail: pzanotto@usp.br

OPEN ACCESS Freely available online

PLoS one

Demographic Histories of ERV-K in Humans, Chimpanzees and Rhesus Monkeys

Camila M. Romano[1], Fernando L. de Melo[1], Marco Aurelio B. Corsini[1], Edward C. Holmes[2,3], Paolo M. de A. Zanotto[1]*

1 Laboratory of Molecular Evolution and Bioinformatics, Department of Microbiology, Biomedical Sciences Institute–ICBII, University of São Paulo, Brazil, **2** Mueller Laboratory, Department of Biology, Center for Infectious Disease Dynamics, The Pennsylvania State University, University Park, Pennsylvania, United States of America, **3** Fogarty International Center, National Institutes of Health, Bethesda, Maryland, United States of America

We detected 19 complete endogenous retroviruses of the K family in the genome of rhesus monkey (*Macaca mulatta;* RhERV-K) and 12 full length elements in the genome of the common chimpanzee (*Pan troglodytes*; CERV-K). These sequences were compared with 55 human HERV-K and 20 CERV-K reported previously, producing a total data set of 106 full-length ERV-K genomes. Overall, 61% of the human elements compared to 21% of the chimpanzee and 47% of rhesus elements had estimated integration times less than 4.5 million years before present (MYBP), with an average integration times of 7.8 MYBP, 13.4 MYBP and 10.3 MYBP for HERV-K, CERV-K and RhERV-K, respectively. By excluding those ERV-K sequences generated by chromosomal duplication, we used 63 of the 106 elements to compare the population dynamics of ERV-K among species. This analysis indicated that both HERV-K and RhERV-K had similar demographic histories, including markedly smaller effective population sizes, compared to CERV-K. We propose that these differing ERV-K dynamics reflect underlying differences in the evolutionary ecology of the host species, such that host ecology and demography represent important determinants of ERV-K dynamics.

Citation: Romano CM, de Melo FL, Corsini MAB, Holmes EC, Zanotto PM (2007) Demographic Histories of ERV-K in Humans, Chimpanzees and Rhesus Monkeys. PLoS ONE 2(10): e1026. doi:10.1371/journal.pone.0001026

INTRODUCTION

A considerable proportion (~45%) of the primate genome consists of copies of mobile genetic elements [1]. These elements are divided into two classes based on their mechanism of mobilization: those involving an RNA intermediate, or those that transpose via DNA excision and reintegration into the host genome (transposons). The via-RNA elements (Class I) are represented by retrotransposons and endogenous retroviruses (ERVs). ERVs are relics of ancient viral infection events in the germ line, followed by long-term vertical transmission. They can increase in copy number by means of active replication (in *cis* or in *trans*) or by chromosomal duplication [2], and represent about 3% of all transposable elements (TE) related sequences. Proviral activity may occur over long periods of time until they become inactivated by loss of promoter functionality due to host chromosome rearrangements, insertions, deletions or point mutations. Because the LTRs (long terminal repeats) of proviruses carry transcriptional regulatory elements, such as promoters and enhancers, its likely that the insertion of a provirus, or only its LTRs, near genes or regulatory regions will be detrimental to host fitness [3–5].

The human ERV-K (HERV-K) family includes some of the most active retroviral elements in human genome [6,7]. Although most of the proviral copies of ERV-K in the genome are inactive, some show evidence of past positive selection at the *env* gene [8,9]. ERVs, as well the other retroelements, can invade the host genome due to transposition bursts [10], counteracted by host-driven excision and purging [11,12]. This dynamical process plays an important role in the evolution of host genomes as a consequence of the rearrangement, transduction and inactivation of genes [13,14]. In the absence of any host selection pressure to inhibit the fixation and replication, ERV copy number could increase to extreme levels [15,16]. However, the preferential integration of LTR elements in gene-poor regions and in an antisense orientation suggests that these elements are routinely purged from gene-rich regions by purifying selection [17,18], which is perhaps a major force restricting ERV copy number. Consequently, determining the mechanisms of transposition control, inactivation and purging are central to the understanding of proviral dynamics in the host genome [5,12,16,19,20].

To explore the evolutionary dynamics of ERVs in more detail, we determined the demographic history of ERV-K in three primates: human (*Homo sapiens*), common chimpanzee (*Pan troglodytes*) and rhesus monkey (*Macaca mulatta*). Our findings suggest that host population size and ecology plays a major role in shaping patterns of ERV-K evolution in primates.

RESULTS

ERV-K Characterization and Phylogeny

Nineteen complete proviruses, designated RhERV-K, were found in the rhesus monkey (*Macaca mulatta*) draft assembly genome (Text S1). Similarly, 12 new elements in *Pan troglodytes* (CERV-K) genome were found (Text S1) and compared to 20 CERV and 55 human HERV-K previously reported, producing a total of 106 ERV-K genomes. Three RhERV-K proviruses had almost identical LTR, indicative of recent integration and therefore of possible recent activity. Conversely, RhERV-K19 had highly divergent 5′ and 3′ LTR that could not be aligned due to several insertion-deletion events (indels), indicating that the estimated

Academic Editor: Jean Carr, Institute of Human Virology, United States of America

Received August 17, 2007; **Accepted** September 21, 2007; **Published** October 10, 2007

Funding: This research was funded by VGDN program FAPESP (00/04205-6). CMR and FLM hold a CAPES doctorate fellowship (DO) and PMAZ holds a CNPq PQ Research Scholarship.

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

* **To whom correspondence should be addressed.** E-mail: pzanotto@usp.br.

integration time of about 46 MYBP (see below) may be misleading. As no RhERV-K orthologue was closely related to those in either the chimpanzee or human genomes, all RhERV-K proviruses appear to have arisen by active transposition rather than chromosomal duplication. In contrast, *Pan* and *Homo* share several ERV-K, and exhibit many closely related elements that most likely originated by chromosomal duplications and rearrangement events (*e.g.*, CERV-K32, CERV-K31, CERV-K34; CERV-K26, 27 and 28 on the Y chromosome).

A phylogenetic tree (Fig. 1) for a 4130 bp alignment from the conserved domains (the Partial data set) shared by 106 ERV-K genomes, had a topology congruent to those obtained previously for both ERV-K genomic fragments [21] and complete genomes [9]. To facilitate data presentation, tree components involving two or more adjacent lineages in the same host, were collapsed and were indicated as colored wedges in Figure 1. Human and chimpanzee appear to share a large number of ERV-K as indicated by at least 18 *Pan-Homo* sister taxa pairs at the tips of the

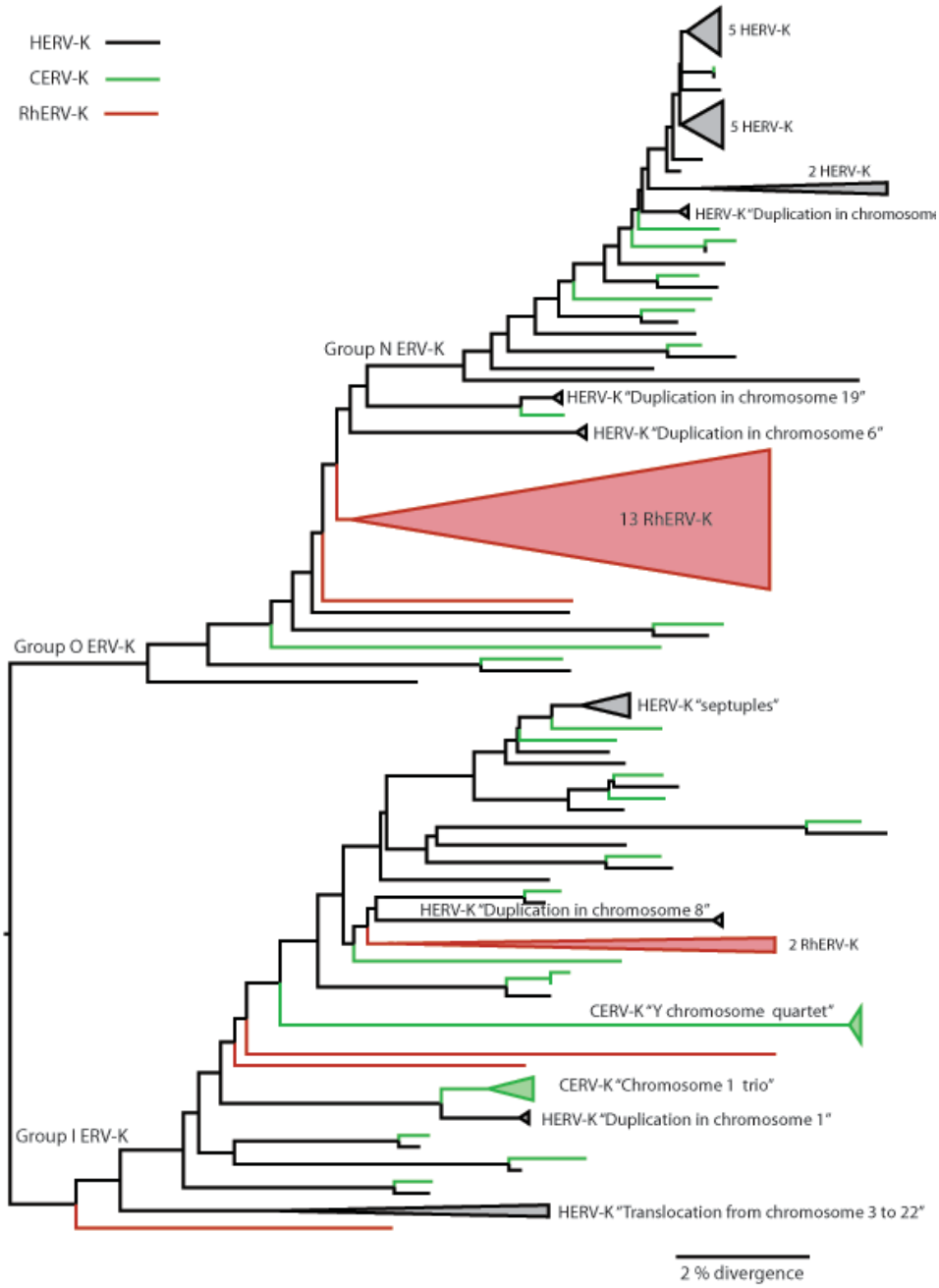


Figure 1. Maximum likelihood tree for 106 ERV-K genomes. ML tree for 4130 bp of shared (Partial) sequences from ERV-K genomes of human (*Homo sapiens*) (55 sequences), common chimpanzee (*Pan troglodytes*) (32 sequences) and, rhesus monkey (*Macaca mulatta*) (19 sequences). Thirteen RhERV-K (shown as a collapsed red wedge in the tree) arise from a single ancient branch in Group O, while four other deep lineages radiate independently from within Group I. No RhERV-K was observed in Group N. The HERV-K, CERV-K and RhERV-K elements are shown by black, green and red branches, respectively. Duplications of the same provirus appear in colored collapsed wedges.
doi:10.1371/journal.pone.0001026.g001

tree. Interestingly, 13 RhERV-K clustered in a distinct group, radiating within Group O [9], represented by the largest wedge in Figure 1. The other six RhERV-K genomes fell in four distinct lineages within Group I. None of the six lineages of RhERV-K shared recent orthologues with *Homo* or *Pan*, and only three (RhERV-K3, RhERV-K8 and RhERV-K19) were possibly integrated into the common ancestor of all three primates. This notion was further supported by the fact that no traces of ERV-K were found in the orthologous chromosomal regions in human and chimpanzee, where we would expect to find the descendents of RhERV-K3 and the eight ERVs that predate the separation of all three lineages. Conversely, fragments of LTR and *gag* sequences were found on chromosome 9 of both human and chimpanzee at the integration site of RhERV-K19, suggesting that they the ERV-K viruses have been purged from these genomes.

ERV-K Population Dynamics

Bayesian skyline plots, reflecting changes in effective population size through time, were inferred for 31 HERV-K found in *Homo sapiens* (Figure 2a), 21 CERV-K found in *Pan troglodytes* (Figure 2b) and 19 RhERV-K found in *Macaca mulatta* (Figure 2c). The high ESS values (near 1000) indicated that the sample sizes, although small, were sufficient for convergence during parameter estimation. Strikingly different plots were seen in the three species, and with a particularly complex dynamic in humans, although both *Homo* and rhesus ERV-K experienced an initial burst in ERV copy number followed by a significant reduction in the number of complete proviruses after 20 MYBP. In contrast, CERV-K experienced an apparently flat dynamic after a significantly (around ten-fold) higher growth in numbers up until 15 MYBP, and had very much larger effective population sizes than the other two species. Finally, and perhaps most notable of all, during the last 5 MY there was an increase in ERV-K numbers in the human genome, possibly caused by the radiation of the newer human elements (Group N) [9].

One possible reason for differences in the dynamics observed is heterogeneity in evolutionary rate among the primate hosts. In particular, it has been established that the rate of evolution in humans suffered a slowdown relative to that of the chimpanzee [22,23], with an approximately two-fold reduction in evolutionary rate relative to Old World monkeys and chimpanzee [24]. Therefore, based on previous estimates on the differences among substitution rates for the species considered here [22–25], we repeated our analysis of population dynamics using dates of integration based on rates of 5.94×10^{-9} s/s/y for CERV-K and 6.93×10^{-9} s/s/y for RhERV-K (with human still at 3.3×10^{-9} s/s/y). The comparisons shown in Figure 3 clearly indicate that the differences in population dynamics are not changed qualitatively by host rate heterogeneity. Hence, these results indicate that the evolutionary rate of the host genome is a less important determinant of differences among ERV-Ks than host population dynamics.

DISCUSSION

ERV-K in primates

Herein, we described several new complete ERV-K elements in the genomes of the common chimpanzee (*Pan troglodytes*) and rhesus monkey (*Macaca mulatta*) and compared them to those found in humans. We show, for the first time, that the demographic history of the host may be a major factor determining the dynamics of an endogenous retrovirus. Despite the draft quality of the rhesus genome assembly, we found many complete proviruses that have a marked similarity in their fluctuating demographic history to that of humans, with both these species distinct from that observed in the chimpanzee (Figure 3). In particular, we found a distinct group of 13 RhERV-K, which diverged around 12 MYBP that were absent in both humans and chimpanzees. Moreover, there was no evidence of RhERV-K amplification caused by chromosomal duplication. On the other hand, both *Homo* and *Pan* had many closely related ERV-K, some of which had several duplicated counterparts. Important differences between CERV-K and HERV-K were also evident. For example, four CERV-K where found on the Y chromosome, three of which were found within an apparently low complexity repeat region, as a consequence of DNA duplication (*i.e.*, CERV-K "Y chromosome quartet" in Figure 1). Interestingly, the human Y chromosome has the same repeat region without traces of retrovirus integration, suggesting that elements have been purged along the human lineage.

Demography and Dynamics of ERV-K

The Bayesian skyline plots revealed fluctuating ERV-K population sizes in all three primate species, although with a relatively large sampling error (Figures 2 and 3). Although HERV-K and RhERV-K had similarly complex skyline plots, it is striking that the latter exhibited a signal of rapid population growth up until 25 MYBP, coinciding with both fossil and molecular data for the radiation of the *Cercopithecidae*. Conversely, the signal for the initial burst for HERV-K and CERV-K occurred at approximately 17–18 MYBP, followed by a reduction of the number copy of the elements, first in *Homo* and then in *Pan*. This growth signature, common to all three primates, may reflect some of the shared history of ERV-K colonization of Catarrhines from the Oligocene (30 MYBP) to Miocene (20 MYBP).

The rate of retrovirus-driven transposition and excision is evidently insufficient to explain their permanence and integrity. Since, in finite populations, size fluctuations have a drastic impact on genome architecture, ERV-K numbers in time must ultimately depend on host population dynamics [26]. Nevertheless, the mechanisms of purging [5], reduction of transposition efficiency by APOBEC [20], excision [11] and stabilization under weak selection [16], or the balance between host migration rates and ERV-K transposition rates [12], as well as synergistic epistasis among integrated ERV-K [27], may have played a role in preventing the continued growth of the three ERV populations towards the present from 10 to 20 MYBP. The loss of cladogenetic signal from older ERV-K lineages could therefore be a consequence of a strong host-driven purging that is more evident in the *Homo* and rhesus lineages. This agrees with our finding that 61% percent of the human elements compared to 21% of the chimpanzee and 47% of rhesus had estimated integration times less than 4.5 MYBP.

Since all partial sequences we dismissed were likely generated by incomplete purging events it is evident that our approach has underestimated the loss of ERV proviruses. Nevertheless, by investigating complete genomes were able to estimate integration times, which is only possible when both LTRs are present. The Bayesian skyline plot for HERV-K showed a conspicuous population bottleneck in the last 17 MY, comprising a significant reduction in complete proviral numbers up until 4MYBP, after which a cladogenetic burst within ERVs from Group N [9] took place. This population bottleneck could indicate a recent loss of ancient signal in the hominids, since the difference in the skyline signatures predates the split of *Homo* and *Pan*. Possibly, bottlenecks since the Plio-Pleistocene may have played an important role, facilitating both the loss of unfixed alleles and the fixation of deleterious ones by genetic drift [28], and which could help explain the observed complex dynamics of HERV-K. Intriguingly, the time frame for a "re-colonization" of the hominids by Group N HERV-K at around 1.5 MYBP coincides with the emergence of human-specific life history traits [23], such as increased generation time.

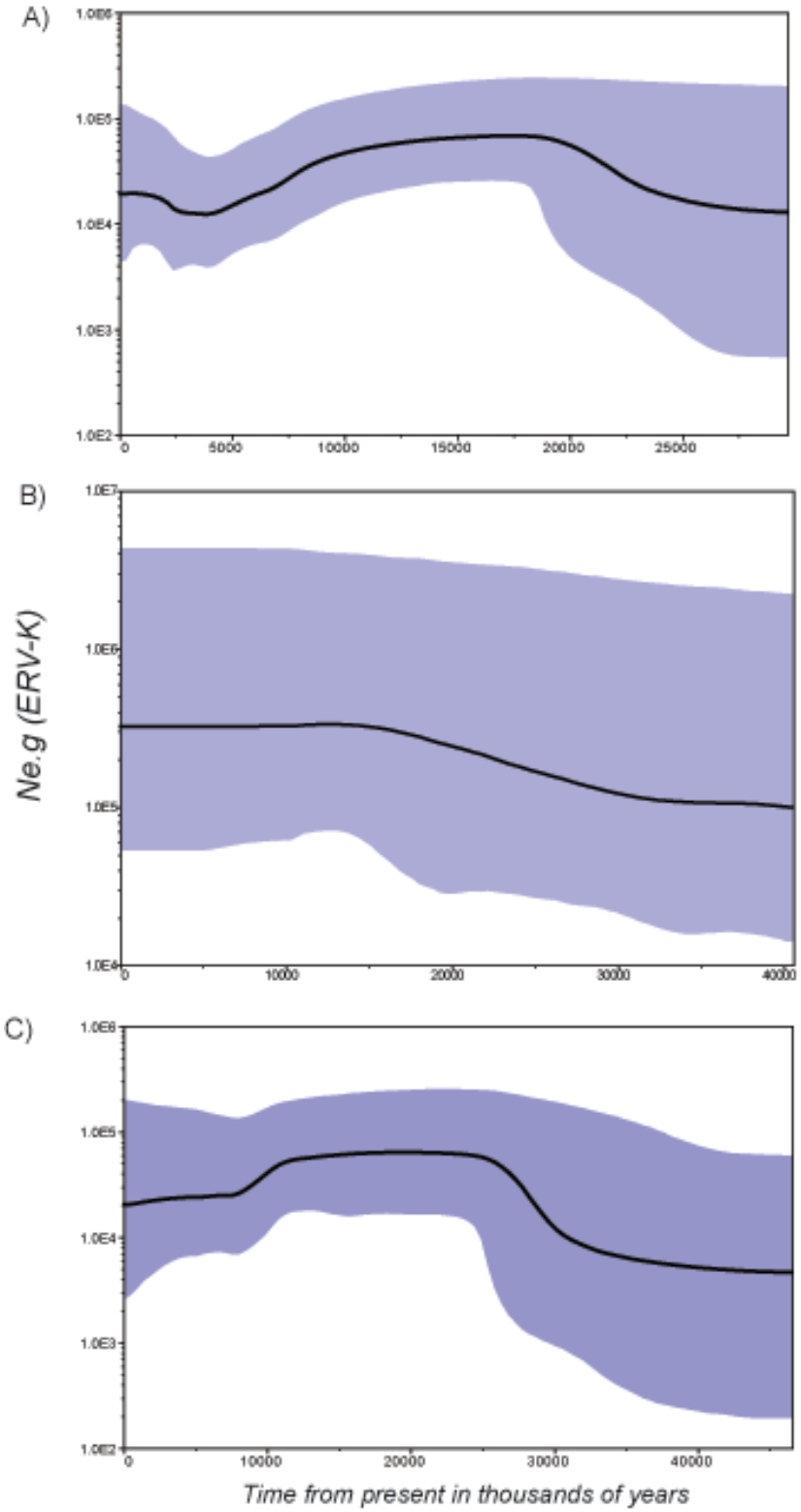


Figure 2. Bayesian skyline plots of three primate ERV-K. A) human (*Homo sapiens*) ERV-K (HERV-K), **B)** common chimpanzee (*Pan troglodytes*) ERV-K (CERV-K) and, **C)** rhesus monkey (*Macaca mulatta*) ERV-K (RhERV-K). Time is presented in million years from the present and effective population sizes multiplied by the generation time (*Ne*.g) are presented in a logarithmic scale on the y-axis. The bold line represents the median estimate for each species while the 95% HPDs (reflecting statistical uncertainty) are shaded. Integration times for all ERV-K were estimated using a rate of 3.3×10^{-9} substitutions per site per year (s/s/y).
doi:10.1371/journal.pone.0001026.g002

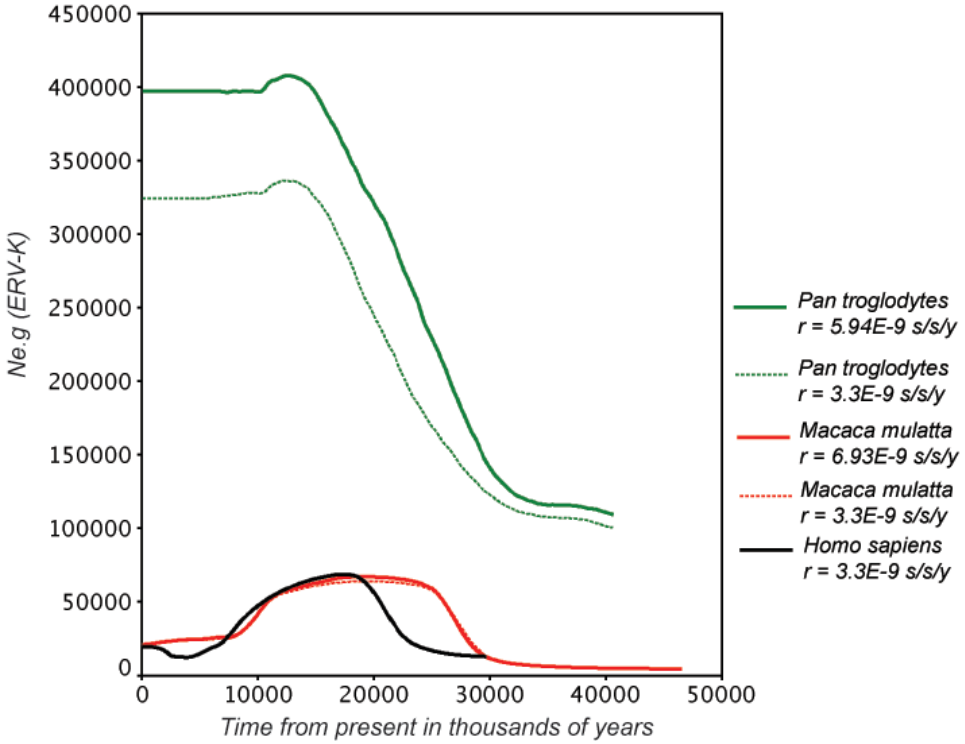


Figure 3. Comparative population dynamic of ERV-K. The figure shows the superimposed median values of *Ne*.g through time taken from the Bayesian skyline plots for the primate species. Time is presented in million years from the present and effective population time generation time (*Ne*.g) sizes are given in a linear scale without the 95% HPD values shown in
doi:10.1371/journal.pone.0001026.g003

Unlike the single extant species of the genus *Homo*, the genus *Macaca* is represented by a large number of species (19) despite being a relative young clade [29,30]. *Macaca mulatta* originated from a *fascicularis*-like ancestor around 2.5 MYBP and became widely distributed within a relatively short period, from western India to the eastern coast of China. The strong decrease in RhERV-K population size (Figure 3) coincided with the emergence of the genus *Macaca* around 10 MYBP, which is one of the most specious groups among *Cercopithecidae* [29]. The impact of the intense cladogenesis in *Cercopithecidae* on RhERV-K dynamics remains to be addressed. Nevertheless, the elevated dispersal of both *Homo* and *Macaca* compared to *Pan* may be an important factor that could explain the similarities in the demographic histories of HERV-K and RhERV-K.

Unlike HERV-K and RhERV-K, the chimpanzee ERV-K demographic signal was characterized by a far larger effective population size. Assuming that host dynamics impacts on ERV-K numbers, the recent flat curve of *Pan* skyline after 6 MYBP agrees with the lack of evidence for severe bottlenecks in the *Pan* lineage and a 3.2 times larger effective ancestral population size [31]. The latter could have facilitated the maintenance of a higher number of integrated elements observed in the chimpanzee genome, because of a weaker effect of genetic drift, although the wide HPD values caution against over-interpretation.

METHODS

ERV Screening, Phylogenetic Inference and Sequence Analysis

We screened the genomes of *Pan troglodytes* (build 2 v.1) and the *Macaca mulatta* draft assembly (v.1) by BLAT search [32] using complete ERV-K genomes as a query. This analysis revealed 116 complete retroviral genome sequences, 78 of which were previously reported and are deposited in GenBank as DQ112093-DQ112156. These sequences were then aligned with both MUSCLE [33] and BlastAlign [34]. To minimize systematic errors caused by insertion/deletion events (indels), for which there is no adequate model of evolution, we also constructed a 4130 bp data set using gene coding regions only (designated as the 'Partial' data set from now on). Maximum likelihood (ML) trees of these data were then inferred by PAUP v.4.0b [35], using the TVM+γ evolutionary model as determined by MODELTEST 3.7 [36]. Tree topologies were evaluated from an initial neighbor joining tree (NJ), using a heuristic search approach that implemented successively branch-swapping methods: (*i*) tree bisection-reconnection (TBR) branch-swapping, (*ii*) subtree pruning-regrafting (SPR) and, (*iii*) nearest-neighbor interchange (NNI). The integration time (T) of each provirus was estimated using the relation $T = d/2r$, where d is the genetic distance between 5′ and 3′ LTR and r is the rate of nucleotide substitution per site. Errors in T where assumed to be the transformed values of the standard errors for d estimations. Because rates of substitution for ERVs can range from $1.5–5\times10^{-9}$ substitutions per site per year (s/s/y), [2,37] we used an average rate of 3.3×10^{-9}. Finally, pairwise distances among ERVs were calculated using Tamura-Nei model available in MEGA2 [38].

Population Dynamics

For this analysis we constructed a smaller 2530 bp region from the Partial dataset that contained those nucleotide sites shared by all proviruses. Proviruses that were both sister taxa (*i.e.* adjacent in the phylogenetic tree) and had similar flanking regions up to10 kb away from the insertion locus, were excluded from the demographic analyses as they most likely to have arisen by

chromosomal duplication. Following this screening, 19 RhERV-K, 21 CERV-K and 31 HERV-K sequences were available for analysis. Rates of nucleotide substitution per site, the time to the Most Recent Common Ancestor TMRCA and the demographic history of each ERV-K group (*Homo*, *P. troglodytes* and *M. mulatta*) were estimated using a Bayesian Markov Chain Monte Carlo (MCMC) method available in the BEAST package [39]. For this analysis, dates of integration based on LTR distances were used as "sampling dates" since, once integrated, ERV-K proviruses would behave as if they were "frozen" in the genome and so evolve at rates equivalent to those of host DNA. Such LTR-based "sampling dating" is justified since the differences in the rates of evolution of exogenous retroviruses are six orders of magnitude higher than those of their endogenous ("frozen") counterparts. Because LTR comparisons indicate that ERV-K have been integrating into primate DNA for at least 40 million years, the assumption that all ERV-K were sampled today would entail a far greater systematic error. To infer population dynamics of the different primate ERV-K we fitted sequence data to the demographic models available in the Bayesian coalescent method in BEAST. In particular we used the Bayesian skyline plot to depict changes in effective population size through time ($N_e.g$, where N_e is the effective population size and g the generation time). For this analysis we used the HKY+γ model of nucleotide substitution under the assumption of a relaxed (uncorrelated exponential) molecular clock. The HKY+γ was consistently the best-supported model in MODELTEST when the data from each species were analyzed separately. In all cases chain lengths of 40–50 million were sufficient to obtain Effective Sample Sizes (ESS) greater than 100.

SUPPORTING INFORMATION

Text S1 GenBank information
Found at: doi:10.1371/journal.pone.0001026.s001 (0.03 MB DOC)

ACKNOWLEDGMENTS

Author Contributions

Conceived and designed the experiments: PZ. Performed the experiments: CR. Analyzed the data: EH PZ CR Fd MC. Contributed reagents/materials/analysis tools: MC. Wrote the paper: EH PZ CR.

REFERENCES

1. Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, et al. (2001) Initial sequencing and analysis of the human genome. Nature 409: 860–921.
2. Boeke JD, Stoye JP (1997) Retrotransposons, endogenous retroviruses and the evolution of retroelements. In: Coffin JM, Hughes SH, Varmus EH (eds) Retroviruses. Cold Spring Harbor Laboratory Press: Cold Spring Harbor, NY. pp 343–435.
3. Schulte AM, Lai S, Kurtz A, Czubayko F, Riegel AT, Wellstein A (1996) Human trophoblast and choriocarcinoma expression of the growth factor pleiotrophin attributable to germ-line insertion of an endogenous retrovirus. Proc Natl Acad Sci U S A 93: 14759–14764.
4. Nuzhdin SV (1999) Sure facts, speculations, and open questions about the evolution of transposable element copy number. Genetica. 107: 129–137.
5. Le Rouzic A, Capy P (2005) The first steps of transposable elements invasion: parasitic strategy vs. genetic drift. Genetics 169: 1033–1043.
6. Medstrand P, Mager DL (1998) Human-specific integrations of the HERV-K endogenous retrovirus family. J Virol 72: 9782–9787.
7. Lower R, Boller K, Hasenmaier B, Korbmacher C, Muller-Lantzsch N, Lower J, Kurth R (1993) Identification of human endogenous retroviruses with complex mRNA expression and particle formation. Proc Natl Acad Sci USA 90: 4480–4484.
8. Belshaw R, Pereira V, Katzourakis A, Talbot G, Paces J, Burt A, Tristem M (2004) Longterm reinfection of the human genome by endogenous retroviruses. Proc Natl Acad Sci USA 101: 4894–4899.
9. Romano CM, Ramalho RF, Zanotto PM (2006) Tempo and mode of ERV-K evolution in human and chimpanzee genomes. Arch Virol 151: 2215–2228.
10. Wisotzkey RG, Felger I, Hunt JA (1997) Biogeographic analysis of the Uhu and LOA elements in the Hawaiian Drosophila. Chromosoma 106: 465–477.
11. Promislow DE, Jordan IK, McDonald JF (1999) Genomic demography: a life-history analysis of transposable element evolution. Proc Biol Sci 266: 1555–1560.
12. Deceliere G, Charles S, Biemont C (2005) The dynamics of transposable elements in structured populations. Genetics 169: 467–474.
13. John B, Miklos G (1988) The Eukaryote Genome in Development and Evolution. London: Allen&Unwin. 416 p.
14. Crombach A, Hogeweg P (2007) Chromosome Rearrangements and the Evolution of Genome Structuring and Adaptability. Mol Biol Evol 24: 1130–1139.
15. Doolittle WF, Sapienza C (1980) Selfish genes, the phenotype paradigm and genome evolution. Nature 284: 601–603.
16. Tsitrone A, Charles S, Biemont C (1999) Dynamics of transposable elements under the selection model. Genet Res 74: 159–164.
17. Medstrand P, van de Lagemaat LN, Mager DL (2002) Retroelement distributions in the human genome: variations associated with age and proximity to genes. Genome Res 12: 1483–1495.
18. Smit AFA (1993) Identification of a new, abundant superfamily of mammalian LTR-transposons. Nucleic Acids Res 21: 1863–1872.
19. Ohta T (1986) Population genetics of an expanding family of mobile genetic elements. Genetics 113: 145–159.
20. Sawyer LS, Emerman M, Malik HS (2004) Ancient Adaptive Evolution of the Primate Antiviral DNA-Editing Enzyme APOBEC3G. Plos Biol 2: 1278–1285.
21. Belshaw R, Dawson AL, Woolven-Allen J, Redding J, Burt A, Tristem M (2005) Genomewide screening reveals high levels of insertional polymorphism in the human endogenous retrovirus family HERV-K (HML2): implications for present-day activity. J. Virol 79: 12507–12514.
22. Li WH, Ellsworth DL, Krushkal J, Chang BH, Hewett-Emmett D (1996) Rates of nucleotide substitution in primates and rodents and the generation-time effect hypothesis. Mol Phylogenet Evol 5: 182–187.
23. Elango N, Thomas JW, NISC Comparative Sequencing program, Yi, SV (2006) Variable molecular clocks in hominoids. Proc Natl Acad Sci USA 103: 1370–1375.
24. Martin AP, Palumbi SR (1993) Body size, metabolic rate, generation time, and the molecular clock. Proc Natl Acad Sci USA 90: 4087–4091.
25. Seino S, Bell GI, Li WH (1992) Sequences of primate insulin genes support the hypothesis of a slower rate of molecular evolution in humans and apes than in monkeys. Mol Biol Evol 9: 193–203.
26. Gherman A, Chen PE, Teslovich T, Stankiewicz P, Withers M, et al. (2007) Population Bottlenecks as a Potential Major Shaping Force of Human Genome Architecture. PLoS Genetics In press..
27. Barton NH, Charlesworth B (1998) Why sex and recombination? Science 25: 1986–1990.
28. Whitlock MC (2003) Fixation probability and time in subdivided populations. Genetics 164: 767–779.
29. Brandon-Jones D, Eudey AA, Geissmann T, Groves CP, Melnick DJ, et al. (2004) Asian primate classifcation. Int J Primatol 25: 97–164.
30. Purvis A, Nee S, Harvey PH (1995) Macroevolutionary inferences from primate phylogeny. Proc Biol Sci 260: 329–333.
31. Kaessmann H, Wiebe V, Weiss G, Paabo S (2001) Great ape DNA sequences reveal a reduced diversity and an expansion in humans. Nat Genet 27: 155–156.
32. Kent WJ (2002) BLAT- the BLAST-like alignment tool. Genome Res 12: 656–664.
33. Edgar RC (2004) MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res 32: 1792–1797.
34. Belshaw R, Katzourakis A (2005) BlastAlign: a program that uses blast to align problematic nucleotide sequences. Bioinformatics 21: 122–123.
35. Swofford DL (2002) PAUP*: Phylogenetic analysis using parsimony (and other methods) 4.0. Sunderland (MA) Sinauer Associates.
36. Posada D, Crandall KA (1998) MODELTEST: testing the model of DNA substitution. Bioinformatics 14: 817–818.
37. Johnson WE, Coffin JM (1999) Constructing primate phylogenies from ancient retrovirus sequences. Proc Natl Acad Sci USA 96: 10254–10260.
38. Kumar S, Tamura K, Jakobsen IB, Nei M (2001) MEGA2: Molecular evolutionary genetics analysis software. Bioinformatics 17: 1244–1245.
39. Drummond AJ, Rambaut A (2003) BEAST v1.0. Available: http://evolve.zoo.ox.ac.uk/beast/.

ANEXO D

BSL de PongERV-K

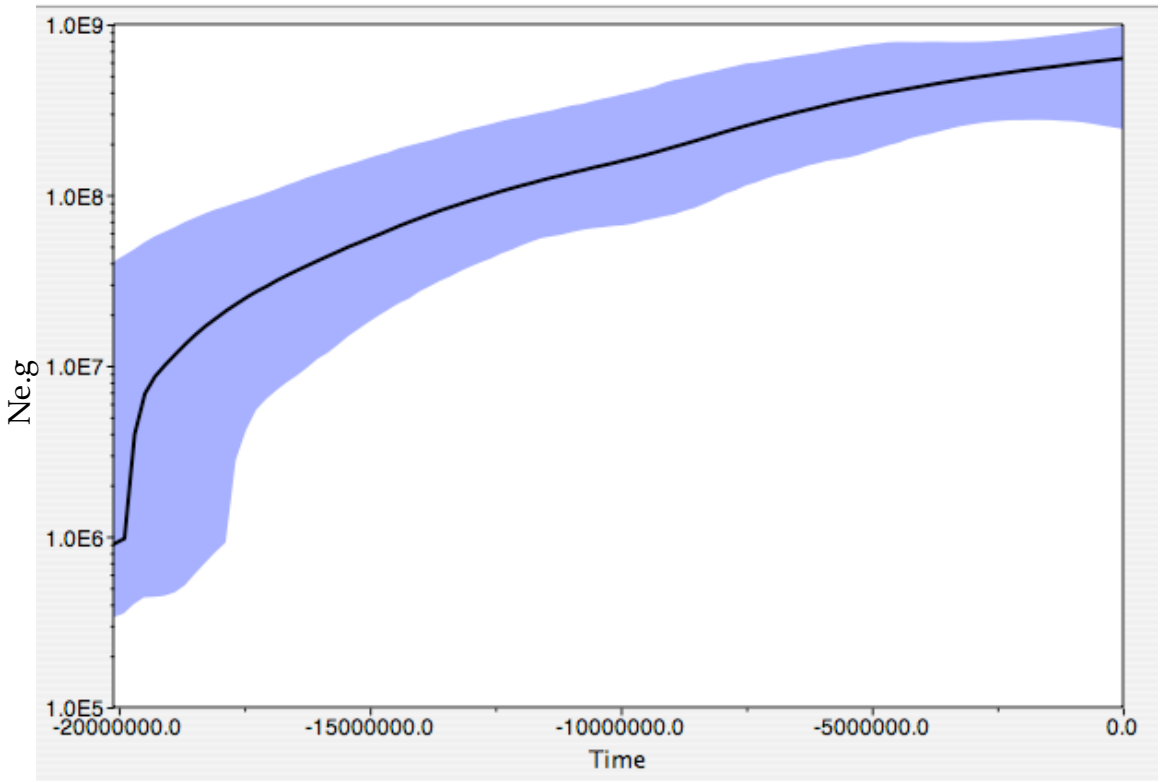


Figura D1. Bayesian Skyline plot de ERV-K de orangotango. O eixo X representa o tempo em milhões de anos (onde 0 é o presente). O eixo Y representa o tamanho da população expresso em Ne.g (população efetiva multiplicada por tempo de geração), em escala logaritmica. A linha central indica a mediana e a margem azul representa o HPD de 95% (high posterior probability), mostrando os valores máximo e mínimo de Ne.g. Notar que desde o período de integração, os ERV-K de orangotango não experimentaram diminuição do seu tamanho efetivo.

Recently integrated HERV-K proviruses with conserved promoter elements are preferentially transactivated upon HIV-1 infection and in cancer cells

Running title: Diferential transactivation of HERV-K

Camila Malta Romano[1,2], Fernando Lucas de Melo[1] and Paolo Marinho de Andrade Zanotto[1, †]

[1] *Laboratory of Molecular Evolution and Bioinformatics, Department of Microbiology, Biomedical Sciences Institute – ICBII. Av Prof. Lineu Prestes, 1374, Butanta, University of São Paulo, SP. Brazil. Tel. 55 11 30917290*

[2] *Laboratory of Virology, Department of Infectious Diseases, Institute of Tropical Medicine, University of São Paulo - School of Medicine, São Paulo, SP. Brazil. Tel. 55 11 30617020*

[†]Corresponding author:

Paolo M. de A. Zanotto

Summary

Human endogenous retroviruses of the K family (HERV-K) are the most recently integrated viruses in the primate genome and, unlike all other human endogenous retroviruses, are still capable to produce viral particles. HERV-K proviral mRNA has been detected in high levels in blood from HIV-1 infected patients and in several types of cancer cells, suggesting HERV-K transcription enhancement under these conditions. However, despite the recurrent association between HERV-K expression and distinct pathological manifestations, there is no effort in order to associate which proviruses are expressed in such conditions. The aim of this work was to identify the proviral source of each HERV-K found to be transcript and the level of its promoter conservation. Using maximum likelihood, proviral transcripts were compared to fifty-five complete HERV-K genomes. We found that only a restrict group of HERV-K, with conserved promoter regions in their LTR, appears to be transcribed during HIV-1 infection and in cancer cells. Moreover, 80% (12: 15) of the transcripts from cancer tissues belonged to type 1 HERV-K, compared to 56% (13:23) detected in plasma of HIV-1 carriers, suggesting selective transactivation in tumor cells. Our results may have implications on the elucidation of the biological roles of HERV expression under pathological conditions.

Key words HERV-K - HIV-1 - cancer cells – transactivation - promoter

Table 1. Promoter integrity of Group 1 and 2 HERV-K proviruses and matches to cellular transcripts. HERV-K type (1, 2 or not determined *n.d.*), absolute integration time in million years, presence of transcripts in HIV-1 infected patients and cancer cells and LTR regions integrity are shown.

Provirus		**Age**	**Transcripts**		**LTR Integrity**				
Group 1	**Type**	**MY**	**HIV**	**Cancer**	**Enh**	**UP**	**BS**	**TATA**	**Inr**
HERV-K68	1	0.125	●	●	+	+	+	+	+
HERV-K4	1	0.125	●	●	+	+	+	+	+
HERV-K10	1	0.25	●	●	+	+	+	+	+
HERV-K102/-50a	1	0.3	●	●	+	+	+	+	+
HML-2	2	0.3	●		+	+	+	+	+
HERV-K113	2	0.37			+	+	+	+	+
HERV-K101	1	0.37	●	●	+	+	+	+	+
HERV-K50b	1	0.37	●	●	+	+	+	+	+
HERV-K60	1	0.375	●	●	+	+	+	+	+
HERV-K36	2	0.5	●	●	+	+	+	+	+
HERV-K41	2	0.5	●		+	+	+	+	+
HERV-K37	1	0.62	●	●	+	+	+	+	+
HERV-K109/-K115	2	1.0	●	●	+	+	+	+	+
HERV-K104/-K50d	2	2.3			+	+	+	+	+
HERV-KII	1	2.4	●	●	+	+	+	+	+
HERV-KI	2	2.75	●		+	+	+	+	+
HERV-K33	2	4.1	●		+	+	+	+	+
HERV-K18	1	5.0	●	●	+	+	+	+	+
HERV-K11	1	6.75	●	●	-	+	+	+	+
HERV-K12309/-K50f	2	7.5	●	●	+	+	-	+	+
HERV-K51	1	7.6	●		-	+	+	+	+
HERV-K59	2	9.5			-	+	-	+	-
HERV-K5	n.d.	13.8			-	-	+	+	+
HERV-K30	2	13.8			-	+	-	+	-
HERV-K8	1	15.3	●	●	-	-	-	+	+
HERV-KOLD345	2	18.2	●		-	+	-	-	-
HERV-K20/-K69	2	18.5	●		-	+	-	+	+
HERV-K52	2	22	●		-	+	-	+	+
HERV-K50e	2	39.8			-	-	-	+	+

Table 1. Continuation.

Provirus		Age	Transcripts		LTR Integrity				
Group 2	**Type**	**MY**	**HIV**	**Cancer**	**Enh**	**UP**	**BS**	**TATA**	**Inr**
HERV-KOLD35587	2	0			-	-	-	+	-
HERV-K42	2	3.1			-	-	-	+	-
HERV-K23	n.d.	6			-	+	-	+	-
HERV-K18b	2	7.1			-	+	-	+	-
HERV-K31	2	7.4			-	-	-	+	-
HERV-K71	n.d.	7.5			-	-	-	+	-
HERV-K77	2	7.8			-	-	-	+	-
HERV-K29	2	8.4			-	+	-	+	-
HERV-K17a	2	9	●		-	-	-	+	-
HERV-K130352	2	9.8			-	-	-	+	-
HERV-K50c	2	9.8			-	-	-	+	-
HERV-K7	2	10			-	-	-	+	-
HERV-K27	2	10.3			-	-	-	+	-
HERV-K17b	2	10.7			-	-	-	+	-
HERV-K30	n.d.	13.8			-	+	-	+	-
HERV-K63	2	14.3			-	+	-	+	-
HERV-K43/K70	2	17.2			-	+	-	+	-
HERV-K6/-K76	2	18			-	-	-	+	-
HERV-K12	2	22			-	-	-	+	-

* To minimize ambiguity, proviruses originated by recent duplication events were grouped and were considered both as putative sources of cellular transcripts.

Figure 1. HERV-K Maximum a posteriori (MAP) tree. The black branches correspond to HERV-K matching with transcripts and the gray branches show non-active elements. The presence (●) or absence (O) of each promoter region by element is also showed, and they are depicted in the same order as they appear in table 1. Cladogenetic events through time were scaled in million years from past to present (thin vertical lines). At the top of the trees are depicted the Bayesian skyline plots (BSLs) of each cluster, describing the fluctuation in the number of elements through time. The bold line represents the median estimate for each group, and the thin lines indicate 95% of the statistical uncertainty density (95% HPD).

Figure 2. Relation between HERV-K proviral conservation and LTR integrity. The graph shows the correlation between the proviral substitution accumulation (X axis) and its LTR integrity (Y axis). The proviral conservation were evaluated by the genetic distance from the tip to the root, and the LTR integrity was classified in five levels, were each of the five promoter regions was considered as a character. The HERV-K from Cluster 1 are represented by the blue dots and the elements from cluster 2 are the red dots.

Introduction

The human genome harbors a large number of endogenous retroviral sequences (HERVs) at different levels of integrity, which are remnants of ancient viral infections in the ancestral germ-line cells. During evolutionary time, the proviral copy number increased due to recurrent infection events, retrotransposition and chromosomal duplication. It has been proposed that HERVs proliferate actively within the host genome until undergoing inactivation by mutations or, alternatively, by being purged by recombination (Boeke and Stoye 1997). HERV-K includes the youngest and most active HERV family that can still produce viral particles (Lower et al. 1993). They share homologues in all Old Word monkeys but not in New Word monkeys, which suggests that their association with primates traces back to a time after the separation of Old World and New Word monkeys (Steinhuber *et al.*, 1995). Recently we mapped 55 human elements, 21 chimpanzee (*Pan troglodytes*) and 19 rhesus (*Macaca mulatta*) ERV-K full-length proviruses (Romano et al. 2006; Romano et al. 2007). They were classified in two main groups: Cluster 1, also called Group Old/New by comprise the oldest and the newer elements and, Cluster 2 or Group Intermediate (I), which includes the proviruses with intermediate integration time. Possible beneficial roles have been postulated for HERV because of their maintenance in host genomes. Actually, a HERV-coded protein named syncytin, is capable to mediate the formation of syncytiotrophoblasts in human placental morphogenesis (Mi et al. 2000). On a similar vein, several LTRs were found to act as an alternative host gene promoter (Medstrand *et al*., 2001; Landry *et al*., 2002; Landry *et al*., 2003). Nevertheless, HERVs overexpression have also been linked to several diseases, including neurological disorders, diabetes and cancer (Margerat et al., 2004; Leib-Mosh et al. 1990; Herbst et al. 1999; Wang-Johanning et al. 2001; Contreras-Galindo et al. 2008). Recently, HERV-K transcripts were found in plasma of HIV-1 infected patients, but not in HCV infected or healthy individuals, but a real association of HERV overexpression with AIDS progression is not well established yet (Contreras-Gallindo, et al.2006a; Contreras-Gallindo et al. 2006b). A recent work indicated that HERV super expression under HIV-1 infection results in stimulation of a HERV-specific CD8+ T cell response, suggesting that the super expression of endogenous retrovirus may play a role on the immune response against HIV-1 infection (Garrison *et al*., 2007).

Despite the recurrent detection of HERV-K transcripts in different tissues, and their putative biological implications no direct association of these transcripts to their proviral sources was attempted so far. To address this issue, HERV-K transcripts from *gag*, *pol* and *env* genes amplified from cells of HIV infected patients (Contreras-Gallindo et al. 2006a) and tumoral tissues (Contreras-Galindo et al. 2008) were used as queries during *in silico* searches against our full-length HERV-K genomes database (Romano et al. 2006). We then ranked HERV-K promoters into different levels of sequence conservation based on the presence or absence of known promoter elements and motifs and also associated them to the putative proviral transcription detection. Our data strongly suggests that only a specific group of HERV-K, most of them integrated after the divergence of human and chimpanzee, and keeping conserved promoter elements, is capable to be trans-activated in such conditions.

Methods

Datasets and phylogenetic inferences. We constructed datasets from alignments of *gag*, *pol* and *env* genes comprising known HERV-K proviruses and transcripts obtained from HIV-1 infected patients (Contreras-Galindo et al. 2006a). Also, an additional dataset containing *env* sequences from proviruses and transcripts from tumoral tissues (Contreras-Galindo et al. 2008) was also analyzed. The datasets were aligned

using Muscle v.3.52 (Edgar 2004) and phylogenetic trees were reconstructed with GARLI v.0.951 (Genetic Algorithm for Rapid Likelihood Inference), which employs a more comprehensive branch-swapping procedure and optimizes the evolutionary model iteratively during the search (Zwick 2006). The maximum likelihood (ML) method was used because it uses models of sequence evolution that correct for superimposed replacement events while estimating genetic distances. A maximum of 2% distance was chosen to discriminate among mRNA proviral sources, since it covers: (*i*) nucleotide misincorporation by the RNA-dependent DNA-polymerase, (*ii*) error introduced from sequencing, (*iii*) polymorphisms in the human genome and, (*iv*) polymorphism in the proviral populations spreading among humans (Deb et al. 1998; Zhao et al. 2000). In some cases, it was impossible to determine the real source of a particular gene transcript coming from distinct proviral sequences that were identical at those loci (*i.e.* duplicated HERV-K). For this reason, we considered duplicated near-identical proviruses as a single putative source, which were grouped together in Table1.

Population Dynamic. The demographic history of HERV-K Cluster 1 and 2 was estimated using a Bayesian Markov Chain Monte Carlo (MCMC) method available in the BEAST package (Drummond & Rambaut 2007). The Bayesian skyline model was used to describe changes in the composite parameter *Ne*.*g* through time under a relaxed molecular clock (*Ne* being a surrogate for effective population size and *g* generation time scaled in millions of years before the present (MYBP). Since we can estimate the time of integration of each sequence based on the genetic distance between its LTRs, the MCMC estimates were made more informative by using heterochronous sequences (Romano et al. 2007).

Evaluation of LTR integrity. To evaluate the extent to which LTR conservation in the full-length proviruses correlated with the amount of mRNA clustering, we searched for previously described conserved regulatory motifs in the viral promoter regions. We then analyzed the integrity of: (*i*) putative binding site (BS) for the transcription factor OTF-2, (*ii*) the TATA box, (*iii*) the transcription initiation site (Inr) located at the transcription start site, (*iv*) the putative upstream promoter (UP) located in the regulatory region (R) and, (*v*) the enhancer site for cellular transcription factor YY 1 that is located in the U3 region (Kovalskaya et al. 1991; Seto et al. 1991; Knossl et al. 1999).

After integration and under neutral selection, it is expected that LTRs and proviral genes accumulate substitution at the same rate. However, if the activity of a proviral promoter is deleterious for the host, we may expect the provirus to be purged. Alternatively, the provirus inactivation can occur by an excess of substitution in the promoter region. Therefore, we constructed a graph to correlate the LTR integrity to the level of proviral genes substitution. The LTR integrity was evaluate in five degrees (1 to 5) according the presence or absence of the five described promoter regions analyzed. The proviral conservation was measured as distance (*k*) of the provirus from the root of the tree in a ML reconstruction using the TreeStat (http://evolve.zoo.ox.ac.uk/software/treestat) program.

Results and Discussion

HERV-K proviral transcripts in HIV-1 infected patients and cancer cells. Our ML trees indicated that several full length proviruses appear to be capable of being transcribed in both, HIV-1 infected patients and cancer cells (Table 1 and black branches in the trees of the Figure 1). Table 1 shows that 80% (23:29) of the proviruses from Cluster 1 had at least one gene (*gag*, *pol* or *env*) related to transcripts in HIV-1 patients and 15 (51%) of the proviruses, all of them expressed also in HIV carriers, clustered to transcripts obtained from cancer cells.

Although only envelope HERV-K sequences were available from cancer cells, the higher number of HIV related transcripts cannot be attributed for the different data availability, since the number of envelope sequences amplified from both conditions were roughly the same (280 HERV-K envelope from HIV carriers and 254 from cancer cells). Most important, all the eight HERV-K expressed exclusively in HIV plasma were also detected by envelope matches, and five of them clustered to a minimum of six transcripts. Consequently, if these proviruses were expressed in cancer cells, they should have been detected, even if in lower levels.

Cluster 2 had one of its 19 proviruses clustering with transcripts expressed in HIV-1 patients but none with cancer-cell material. Importantly, the proviral activity of the two clusters was significantly different (χ^2 = 10.0, p<0.001, d.f. = 1), suggesting functional distinctions. The BSLs reconstructed for each cluster (depicted in the top of the trees in Figure 1) showed that, while Cluster 2 was active only until 15 million years ago, the Cluster 1 experienced a burst very close to the present. Apparently, both clusters include proviruses that where replication competent in the past (also evidenced intense cladogenesis between 30 and 10 million years ago), but only the cluster 1 kept the activity until the present.

The data also suggested that some HERV-K loci are more transcriptionally active than others. HERV-K101, K102 and K18, clustered to the highest number of transcripts, from both HIV and cancer. Interestingly, the K18 has an integration time around five million years, but clustered to more transcripts than several recent integrated elements. Actually, we found that not only recent integrated proviruses clustered to transcripts, but also some of those integrated before the *Homo-Pan* split (*e.g.* HERV-K20, K69, K52, Kold345, K8, K51, Kold12309 and K50f from Cluster 1 and K17a from Cluster 2), indicating that older elements are also capable to be transcribed. Our findings that several ancient integrated elements are still producing mRNA suggest that the integration time is not critical for activity.

It was intriguingly however, that one of the most recently integrated elements, HERV-K113, did not match to any transcript. However, because HERV-K113 is a polymorphic provirus, present in 5 -30% in human population (Turner et al. 2001), it is possible that the transcripts analyzed may not have included that locus.

It was also found several transcripts that did not match to any known HERV-K. One may argue that these transcripts were generated by non-full length HERV-K, which were not included in this study. Alternatively, as already noticed before (Flockerzi et al., 2007; Laufer et al., 2009), it is possible that some sequences analyzed may be generated by recombination between individual transcripts, generated *in vitro* during reverse transcription. However, since more than a half of envelope transcripts did not mach to any known HERV-K, we argue that the observed results is a consequence of both, non-full length originated transcripts and *in vitro* recombinant generated.

Although our results showed that only proviruses from Cluster 1 appear to be capable to generate transcripts, it is important to mention that promoter activity is tissue specific (Ruda et al., 2004). Consequently, since we analyzed transcripts obtained only from HIV and cancer patients, it is possible that different results could be obtained by analyzing transcripts coming from other conditions. However, for any type of tissue, proviral transactivation needs the integrity of promoter regions enclosed in their LTRs. Because of that, we assessed the integrity of 5' proviral LTR.

LTR integrity and transcript detection. Overall, the integrity of proviral promoter elements correlated well with the presence of transcripts. Figure 1 and Table 1show a higher number of conserved elements in the LTRs of Cluster 1 compared to Cluster 2 (χ^2 = 21.03, p<0.001, d.f. = 1). Only the TATA box was well conserved in all HERV-K

analyzed, with the exception of the KOLD345. Both the Inr motif, which assists in the formation of the RNA polymerase II transcription-complex, and the upstream promoter site were conserved in roughly all members of the Cluster 1. Crucially, most of the proviruses lacking one of these regions had no transcripts detected in both types of tissues, suggesting that they may be important to LTR promoter activity. On the other hand, the binding site (BS) region for the cofactor OTF-2 and enhancer for transcription factor YY1 (Yoza and Roeder 1990) appeared not to be strongly associated with promoter transactivation in these tissues, as suggested by their absence in older proviruses clustering with transcripts. Nevertheless, because the LTR transactivation is tissue-specific (Ruda et al. 2004), we cannot exclude the possibility that other promoter elements that we are unaware of, may promote HERV-K transcription under specific conditions. As a matter of fact, some minor transcripts initiated before the mapped transcription start site were detected in testicular tissues by RT-PCR experiments, but the alternative start site could not to be determined (Kovalskaya et al., 2006).

It was unexpected that proviruses belonging Cluster 2 have less conserved promoter elements than those from Cluster 1, since the average proviral integration time is nearly the same between them. Crucially, Figure 2 shows that, differently from Cluster 1, proviruses from Cluster 2 lost the promoter regions faster than expected by neutrality, since the LTR integrity did not correlated to their proviral integrity. This reinforces the notion of the biological implications of the activity of specific proviruses only.

Implications of HERV-K activity in cancer and HIV-1 infection. Two types of HERV-K, are present in the human genome, differentiated by a 292 bp fragment deleted between *pol* and *env* in type 1. In this work, we detected transcripts originated from type 1 and type 2 HERV-K (both of which are present in Cluster 1 and 2) in both cancer cells and HIV-1 carriers. Interestingly, we found that 80% (12:15) of the active proviruses in cancer cells were type 1, compared to 56% (13:23) in HIV carriers. Type 1 proviruses encode the Np9 oncoprotein that is expressed only in carcinomas and tumor cell lines (Ambruester et al. 2002) and is capable of interfering with the cellular-differentiation pathway (Ambruester et al. 2004). Differently from type 1, type 2 HERV-K encodes Rec, a HIV-1-similar protein, preferentially expressed in tumors in the germ line (Yang et al. 1999). Due to the detection of high levels of Np9 but very low levels of Rec transcripts in distinct tumor cells (Ambruester et al. 2002), it has been suggested a possible role of Np9 in tumorogenesis. Nevertheless, the expression of only type 1 proviruses in cancer cells was unexpected, since the promoter elements in the LTR of the two types were very similar (Table 1) and did not display any feature that can clearly distinguish them. Moreover, in a LTR phylogenetic tree, there is no segregation by type, (data not showed), probably due to convergence (Sverdlov 2000). Actually, some evidence indicates that differential methylation, which tends to vary according to chromosomal position, may influence selectively HERV-K expression levels in several cancer cells (Gotzinger et al. 1996). Nevertheless, differential methylation alone seems insufficient to explain the puzzling exclusive activity of type 1 proviruses in cancer cells.

Although several lines of evidence have indicated a potential role of HERV-K expression in tumorogenesis and other diseases, it is still unclear whether HERV-encoded proteins play a causative role in pathogenesis or, alternatively, if the enhancement of proviral activity is a consequence of cytopathology. Yet, one could hypothesize that HERV overexpression could have benefic roles for the host under such conditions. For example, the HERV-K env protein was found to be a strong tumor antigen capable to elicit a host antibody response in seminoma patients (Sauter et al. 1995). In addition, recent evidence indicated that

HERV-specific CD8+ T lymphocytes kill cells expressing their cognate peptide (Garrinson et al. 2008). As a consequence, we claim for the importance to investigate the impact of HERV-K on the HIV-1 replication and viral load and, on the proliferation of cancer cells as well.

In sum, we showed for the first time that several HERV-K loci are differentially transactivated in cancer tissues and plasma of HIV-1 infected individuals. The HERV-K expression appears to be dependent on the integrity of specific promoter regions enclosed in the 5' proviral LTR, which were maintained only by selected proviruses. By associate the HERV-K transcripts to their putative source, our results may help to clarify the potential role of HERV transactivation in pathological conditions.

Acknowledgments This research was funded by VGDN program FAPESP under project number 00/04205-6.

References

Armbruester, V., Sauter, M., Krautkraemer, E., Meese, E., Kleiman, A., Best, B., Roemer, K. & Mueller-Lantzsch, N. (2002). A novel gene from the human endogenous retrovirus K expressed in transformed cells. *Clin Cancer Res* **8**,1800-1807.

Armbruester, V., Sauter, M., Roemer, K., Best, B., Hahn, S., Nty, A., Schmid, A., Philipp, S., Mueller, A. & Mueller-Lantzsch, N. (2004). Np9 protein of human endogenous retrovirus K interacts with ligand of numb protein X. *J Virol* **78,**10310-10319.

Badenhoop, K., Tonjes, R.R., Rau, H., Donner, H., Rieker, W., Braun, J., Herwing, J., Mytilineos J., Kurth, R. & Usadei, K.H. (1996). Endogenous retroviral long terminal repeats of the HLA-DQ region are associated with susceptibility to insulin-dependent diabetes mellitus. *Hum Immunol* **50,**103–110.

Boeke, J.D. & Stoye, J.P. (1997). Retrotransposons, endogenous retroviruses and
the evolution of retroelements. In *Retroviruses*. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, pp. 343–435. Edited by J. H. Coffin, S.H. Hughes & E.H. Varmus. NY.

Contreras-Galindo, R., González, M., Almodovar-Camacho, S., González-Ramírez, S., Lorenzo, E. & Yamamura, Y. (2006a). A new Real-Time-RT-PCR for quantitation of human endogenous retroviruses type K (HERV-K) RNA load in plasma samples: increased HERV-K RNA titers in HIV-1 patients with HAART non-suppressive regimens. *J Virol Methods* **136,**51-57.

Contreras-Gallindo, R., Kaplan, M.H., Markovitz, D.M., Lorenzo, E. & Yamamura, Y. (2006b). Detection of HERV-K (HML-2) viral RNA in Plasma of HIV type 1 Infected individuals. *Aids Res Hum Evol* **22,** 979-984.

Contreras-Galindo, R., Kaplan, M.H., Leissner, P., Verjat, T., Ferlenghi, I., Bagnoli, F., Giusti, F., Dosik, M.H., Hayes, D.F., Gitlin, S.D. & Markovitz, D.M. (2008). Human Endogenous Retrovirus-K (HML-2) Elements in the Plasma of People with Lymphoma and Breast Cancer. *J Virol* **82**,9329-9336.

Deb, P., Klempan, A., O'Reilly, R.L. & Singh, S.M. (1998). A single-primer PCR-based retroviral-related DNA polymorphism shared by two distinct human populations. *Genome* **41**, 662-668.

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

Edgar, R.C. (2004). MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res* **32,**1792-1797.

Garrison, K.E., Jones, R.B., Meiklejohn, D.A., Anwar, N., Ndhlovu, L.C., Chapman, J.M., Erickson, A.L., Agrawal, A., Spotts, G., & other authors (2007). T Cell Responses to Human Endogenous Retroviruses in HIV-1 Infection. *PLoS Pathog* **3,** 1617-1627.

Gotzinger, N., Sauter, M., Roemer, K. & Mueller-Lantzsch N. (1996). Regulation of human endogenous retrovirus-K Gag expression in teratocarcinoma cell lines and human tumours *J Gen Virol* **77,** 2983-2990.

Herbst, H., Kuhler-Obbarius, C., Lauke, H., Sauter, M., Mueller- Lantzsch, N., Harms, D. & Loning, T. (1999). Human endogenous retrovirus (HERV)-K transcripts in gonadoblastomas and gonadoblastoma-

derived germ cell tumors. *Virchows Arch* **434**, 11–15.

Jern, P., Sperber, G.O. & Blomberg, J. (2006). Divergent Patterns of Recent Retroviral Integrations in the Human and Chimpanzee Genomes: Probable Transmissions between Other Primates and Chimpanzees *J Virol* **80,** 1367–1375.

Johnson, W.E. & Coffin, J.M. **(1999)** Constructing primate phylogenies from ancient retrovirus sequences. *Proc Natl Acad Sci USA* **31,**10254-10260.

Kovalskaya, E., Buzdin, A., Gogvadze, E., Vinogradova, T. & Sverdlov, E. (1991). Functional human endogenous retroviral LTR transcription start sites are located between the R and U5 regions. *Virology* **346**, 373-378.

Knossl, M., Lower, R. & Lower, J. (1999). Expression of the Human Endogenous Retrovirus HTDV/HERV-K Is Enhanced by Cellular Transcription Factor YY1. *J Virol* **73**, 1254-1261.

Landry, J.R., Rouhi, A., Medstrand, P. & Mager, D.L. (2002). The Opitz syndrome gene Mid1 is transcribed from a human endogenous retroviral promoter. *Mol Biol Evol* **19,**1934-1942.

Landry, J.R. & Mager, D.L. (2003). Functional analysis of the endogenous retroviral promoter of the human endothelin B receptor gene. *J Virol* . **13,** 7459–7466.

Leib-Mosch, C., Brack-Werner, R., Werner, T., Bachmann, M., Faff, O., Erfle, V. & Hehlmann, R. (1990). Endogenous retrohuman DNA. Cancer Research **50,** 5636–5642.

Löwer, R., Boller, K., Hasenmaier, B., Korbmacher, C., Müller-Lantzsch, N., Löwer, J. & Kurth, R. (1993). Identification of human endogenous retroviruses with complex mRNA expression and particle formation. *Proc Natl Acad Sci USA* **15,** 4480-4484.

Maddison, W.P. & Maddison, D.R. (1989). Interactive Analysis of Phylogeny and Character Evolution Using the Computer Program MacClade. Folia Primatol **53,**190-202.

Marguerat, S., Wang, W.Y., Todd, J.A. & Conrad, B. (2004) Association of human endogenous retrovirus K-18 polymorphisms with type 1 diabetes. *Diabetes* **53,** 852–85456.

Medstrand, P., Landry, J.R. & Mager, D.L. (2001). Long terminal repeats are used as alternative promoters for the endothelin B receptor and apolipoprotein C-I genes in humans. *J Biol Chem* **19**,1896-1903.

Mi, S., Lee, X., Li, X., Veldman, G.M., Finnerty, H., Racie, L., Allie, E., Tang, X.Y., Eduard P., & other authors. (2000). Syncytin is a captive retroviral envelope protein involved in human placental morphogenesis. *Nature* **403**, 785–789.

Romano, C.M., Ramalho, R.F. & Zanotto, P.M.A. (2006). Tempo and mode of ERV-K evolution in human and chimpanzee genomes. *Arch Virol* **151**, 2215-2228.

Romano, C.M., Melo, F.L., Corsini, M.A.B., Holmes, E.C. & Zanotto, P.M.A. (2007). Demographic histories of ERV-K in humans, chimpanzees and rhesus monkeys. *PLoS One* **10**, e1026.

Roelofs, H., van Gurp, R.J.H.L.M., Oosterhuis, J.W. & Looijenga, L.H.J. (1998). Detection of Human Endogenous Retrovirus Type K-Specific Transcripts in Testicular Parenchyma and Testicular Germ Cell Tumors of Adolescents and Adults. Clinical and Biological Implications. *Am J Pathol*. **153,** 1277-1282.

Ruda, V.M., Akopov, S.B., Trubetskoy, D.O., Manuylov, N.L., Vetchinova, A.S., Zavalova, L.L., Nikolaev, L.G. & Sverdlov, E.D. (2004). Tissue specificity of enhancer and promoter activities of a HERV-K (HML-2) LTR. *Virus Res* **104**, 11-16.

Sauter, M., Schommer, S., Kremmer, E., Remberger, K., Dolken, G., Lemm, I., Buck, M., Best, B., Neumann-Haefelin, D. & Mueller-Lantzsch, N. (1995). Human endogenous retrovirus K10: expression of Gag protein and detection of antibodies in patients with seminomas. *J Virol* **69,** 414 – 421.

Seto, E., Shi, Y. & Shenk, T. (1991). YY1 is an initiator sequence binding protein that directs and activates transcription in vitro. *Nature* **354**, 241-245.

Steinhuber, S., Brack, M., Hunsmann, G., Schwelberger, H., Dierich, M.P. & Vogetseder, W. (1995) Distribution of human endogenous retrovirus HERV-K genomes in humans and different primates. *Hum Genet* **96,**188-92.

Sverdlov, E.D. (2000). Retroviruses and primate evolution. *BioEssays*. **22,** 161-171.

Turner, G., Barbulescu, M., Su, M., Jensen-Seaman, M. I., Kidd, K.K. & Lenz, J. (2001). Insertional polymorphisms of full-length endogenous retroviruses in humans. *Curr Biol* **11,** 1531-1535.

Wang-Johanning, F., Frost, A.R., Johanning, G.L., Khazaeli, M.B., LoBuglio, A.F., Shaw, D.R. & Strong, T.V. (2001). Expression of human endogenous retrovirus K envelope transcripts in human breast cancer. *Clin Cancer Res* **7,** 1553-1560.

Yang, J., Bogerd, H.P., Peng, S., Wiegand, H., Truant, R. & Cullen, B.R. (1999). An ancient family of human endogenous retroviruses encodes a functional homolog of the HIV-1 Rev protein. *Proc Natl Acad Sci USA* **96**: 13404-13408.

Yoza, B.K. & Roeder, R.G. (1990). Identification of a novel factor that interacts with an immunoglobulin heavy-chain promoter and stimulates transcription in conjunction with the lymphoid cell-specific factor OTF2. *Mol Cell Biol* **10,** 2145-2153.

Zhao, Z., Jin, L., Fu, Y., Ramsay, M., Jenkins, T., Leskinen, E., Pamilo, P., Trexler, M., Patthy, L., & other authors (2000). Worldwide DNA sequence variation .in a 10-kilobase noncoding region on human chromosome 22. *Proc Natl Acad Sci USA* **97,** 11354-11358.

Zwickl, D.J. (2006). Genetic algorithm approaches for the phylogenetic analysis of large biological sequence datasets under the maximum likelihood criterion. Ph.D. thesis, The University of Texas at Austin. Available from http://www.zo.utexas.edu /faculty/antisense/garli/Garli.html