FUNDAÇÃO OSWALDO CRUZ
INSTITUTO GONÇALO MONIZ

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

DISSERTAÇÃO DE MESTRADO

ESTUDO DA DIVERSIDADE GENÔMICA DO HTLV-1 EM PORTADORES DO VÍRUS COM DIFERENTES CONDIÇÕES CLÍNICAS

MELINA MOSQUERA NAVARRO BORBA

Salvador - Bahia
2020

FUNDAÇÃO OSWALDO CRUZ

INSTITUTO GONÇALO MONIZ

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

ESTUDO DA DIVERSIDADE GENÔMICA DO HTLV-1 EM PORTADORES DO VÍRUS COM DIFERENTES CONDIÇÕES CLÍNICAS

MELINA MOSQUERA NAVARRO BORBA

Orientadora: Dra. Maria Lourdes Farre Vallve
Co-orientadora: Dra. Fernanda Khouri Barreto

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

Salvador - Bahia
2020

Ficha Catalográfica elaborada pela Biblioteca do
Instituto Gonçalo Moniz / FIOCRUZ - Salvador - Bahia.

Borba, Melina Mosquera Navarro.
B726e Estudo da diversidade genômica do HTLV-1 em portadores do vírus com diferentes condições clínicas. / Melina Mosquera Navarro Borba. - 2020.
124 f. : il. ; 30 cm.

Orientador: Prof. Dra. Maria Lourdes Farre Vallve, Laboratório de Patologia Experimental.
Dissertação (Mestrado em Biotecnologia em Saúde e Medicina Investigativa) - Instituto Gonçalo Moniz, Fundação Oswaldo Cruz, 2020.

1. HTLV-1. 2. Infecções. 3. Células clonais. 4. Genômica. I. Título.

CDU 616.98

"ESTUDO DA DIVERSIDADE GENÔMICA DO HTLV-1 EM PORTADORES DO VÍRUS COM DIFERENTES CONDIÇÕES CLÍNICAS".

MELINA MOSQUERA NAVARRO BORBA

FOLHA DE APROVAÇÃO

Salvador, 18 de dezembro de 2020.

COMISSÃO EXAMINADORA

Dra. Marta Giovanetti
Professora Colaboradora
UFMG

Dr. Pedro Dantas Oliveira
Professor Adjunto
UFS

Dr. Pablo Ivan Pereira Ramos
Pesquisador
IGM/FIOCRUZ

FONTES DE FINANCIAMENTO

Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) – Código de Financiamento 001
Fundação de Amparo à Pesquisa do Estado da Bahia (FAPESB)
Conselho Nacional de Desenvolvimento Científico e Tecnológico
Instituto Gonçalo Moniz – Fiocruz Bahia

AGRADECIMENTOS

Primeiramente, agradeço à Deus por ter me permitido chegar até aqui! Gratidão ao universo.

Aos meus pais, Dolores e Raimundo, por sempre acreditarem em mim e não medirem esforços para sempre me apoiarem. Meus exemplos de força e coragem, minha base e meus portos seguros.

Ao meu irmão, Diego Mosquera, por estar ao meu lado e por sempre me apoiar. Obrigada!

Às minhas primas e família, por estarem sempre comigo, acreditando em mim.

Ao meu namorado, Iago Ribeiro, por toda força e paciência, obrigada por ter acredito em mim!

Às minhas amigas de alma, vocês fizeram parte disso, fizeram toda diferença!

À minha orientadora, Lourdes Farre, por todas oportunidades e confiança.

À minha co-orientadora, Fernanda Khouri, que está comigo há tantos anos, gratidão pela oportunidade de ter trabalho com você! Obrigada por toda confiança e paciência.

Ao Instiuto Gonçalo Moniz, por toda estrtura e investimento.

Ao LAPEX, por toda estrutura, e aos amigos que lá fiz e ajudaram na realização desse trabalho.

A todos os técnicos das plataformas e aos colaboradores da biblioteca, limpeza, segurança e manutenção do IGM pelo carinho e excelente serviço prestado à instituição e aos estudantes.

Ao Programa de Pós-graduação em Biotecnologia em Saúde e Medicina Investigativa.

À FAPESB, pela disponibilização da bolsa de estudos, essencial para a conclusão do trabalho.

BORBA, Melina Mosquera Navarro. Estudo da diversidade genômica do HTLV-1 em portadores do vírus com diferentes condições clínicas. 2020. 124 f. Dissertação (Mestrado em Biotecnologia em Saúde e Medicina Investigativa) – Fundação Oswaldo Cruz, Instituto Gonçalo Moniz, Salvador, 2020.

RESUMO

INTRODUÇÃO: O vírus linfotrópico de células T humanas tipo 1 (HTLV-1) foi o primeiro retrovírus humano descrito e estima-se que aproximadamente 5 à 10 milhões de pessoas são infectadas pelo HTLV em todo o mundo. O HTLV-1 é o agente etiológico da paraparesia espástica tropical/mielopatia associada ao HTLV (HAM/TSP), Leucemia/Linfoma de Células T do Adulto (ATLL), Dermatite Infecciosa associada ao HTLV-1 (DIH), entre outras condições clinicas. Sabe-se que os indivíduos infectados pelo HTLV-1 podem desenvolver as patologias associadas ao vírus ou serem considerados portadores assintomáticos (AC). Os fatores que determinam o perfil clínico em um indivíduo infectado ainda não foram estabelecidos. Acredita-se na possível influência do modo de transmissão, da carga proviral, além de fatores genéticos do hospedeiro e do vírus. Considerando que as células infectadas circulantes em um indivíduo portador não são idênticas, existindo diferentes grupos de células infectadas (diferentes clones), a clonalidade dessas células infectadas é um parâmetro que precisa ser considerado também no estudo da diversidade genômica do HTLV-1. Apesar de ser uma infecção endêmica, os indivíduos infectados permanecem sem opções terapêuticas eficazes. **OBJETIVO:** Avaliar a diversidade genética do HTLV-1 em portadores do vírus com diferentes condições clínicas. **MATERIAL E MÉTODOS:** Inicialmente foi avaliado o grau de conservação genética da ORF-I em 32 sequências provenientes de portadores com diferentes perfis clínicos de e os resultados foram comparados com 2.406 sequências da ORF-I disponíveis no GenBank. Em seguida, 242 sequências de genoma completo do HTLV-1 disponíveis no GenBank foram submetidas à análises moleculares e análise de Machine Learning em busca de associações entre mutações e sintomatologia. Ainda, foram desenhados um conjunto de 29 iniciadores para sequenciamento do genoma completo do HTLV-1 utilizando o sequenciamento baseado em nanoporos. **RESULTADOS:** No primeiro trabalho os dados demonstraram uma baixa diversidade genética da região ORF do HTLV-1, confirmando a estabilidade genômica do provirus. Em seguida, evidenciamos uma correlação entre mutações no gene LTR e na região pX que podem estar associados a indivíduos sintomáticos. **CONCLUSÃO:** Esse trabalho possibilitou a disponibilização no GenBank de sequências da ORF-I do HTLV-1 provenientes de pacientes com DIH, ATLL, HAM/TSP e pacientes assintomáticos, além de evidenciar possíveis alvos para desenvolvimento da vacina.

Palavras-chave: HTLV-1; Mutações; Clonalidade.

BORBA, Melina Mosquera Navarro. Estudo da diversidade genômica do HTLV-1 em portadores do vírus com diferentes condições clínicas. 2020. 124 f. Dissertação (Mestrado em Biotecnologia em Saúde e Medicina Investigativa) – Fundação Oswaldo Cruz, Instituto Gonçalo Moniz, Salvador, 2020.

ABSTRACT

INTRODUCTION: The human T-cell lymphotropic virus type 1 (HTLV-1) was the first human retrovirus described. It is estimated that approximately 5 to 10 million people are infected with HTLV worldwide. HTLV-1 is the etiologic agent of tropical spastic paraparesis / HTLV-associated myelopathy (HAM / TSP), Adult T Cell Leukemia / Lymphoma (ATLL), Infectious Dermatitis HTLV-1 associated (IDH), among other clinical conditions. It is known that part of the individuals infected with HTLV-1 remain asymptomatic throughout life. The factors thatrelated with the manifestation of an HTLV-1 associated diseases are not well stablished. The transmission routes, proviral load levels, in addition to genetic factors of the carrier and/or the virus may be involved. As circulating HTLV-1 infected cells are not identical and different clones can be detected, the clonality of infected cells should be a relevant parameter to be considered in the study of the HTLV-1 genomic diversity. Despite HTLV-1 infection is endemic, no vaccine strategies are available to fight against the virus nor effective therapeutic options are available for HTVL-1 associated diseases. **OBJECTIVE:** To evaluate the genetic diversity of HTLV-1 in carriers with different clinical conditions. **MATERIAL AND METHODS:** The degree of genetic conservation of ORF-I HTLV-1 genetic region was evaluated in 32 sequences from carriers with different clinical status and was compared with 2,406 ORF-I sequences available in the GenBank. Additionally, 77 complete HTLV-1 genomes available in the GenBank were subjected to molecular and Machine Learning analysis in to investigate the association between mutations and clinical manifestations. A set of 29 pair of primers were designed in order to sequence the complete HTLV-1 genome with a nanosequencing platform. **RESULTS:** In the first study, a low genetic diversity in the ORF region was observed, indicating genomic stability of provirus. The obtained sequences were published in the GenBank. Additionally, a correlation between mutations in the LTR gene and in the pX region that may be associated with symptomatic individuals was observed. **CONCLUSION:** With this study, we contributed with the availability of sequences of the HTLV-1 ORF region in the Gen Bank obtained from patients with IDH, ATLL, HAM/TSP and asymptomatic patients as well as to identify possible targets for vaccine development against HTLV-1.

Keywords: HTLV-1; Mutations; Clonality.

LISTA DE FIGURAS

LISTA DE ABREVIATURAS E SIGLAS

HTLV	Vírus Linfotrópico de células T Humanas
ATLL	Leucemia/Linfoma de células T do Adulto
HAM/TSP	Paraparesia Espástica Tropical/Mielopatia Associada ao HTLV
DIH	Dermatite Infecciosa associada ao HTLV-1
PTLV	Vírus Linfotrópico de células T de Primatas
ECM	Matriz Extracelular
CA	Capsídeo
DNA	Ácido Desoxirribonucleico
HIV	Vírus da Imunodeficiência Humana
HLA	Antígeno Leucocitário Humano
ILPCR	Reação em Cadeia da Polimerase Longo e Invertido
PCR	Reação em Cadeia da Polimerase
IN	Integrase
MA	Matriz
LTR	Repetição Terminal Longa
MTOC	Centro de Organização Microtubular
ONT	Oxford Nanopore Technologies
ORF	Open Reading Frame
pb	pares de base
PR	Protease
RNA	Ácido Ribonucleico

SI	Sinapse Imunológica
SNV	Variantes de Nucleotídeos Ùnicos
SU	Proteína de Superfície
TM	Proteína de Transmembrana
RT	Transcriptase Reversa
VS	Sinapse Virológica
SGS	Sequenciamento por Síntese
PCMB	Células Mononucleares do Sangue Periférico
HBZ	HTLV-1 bZIP factor gene

SUMÁRIO

1 INTRODUÇÃO

O vírus linfotrópico de células T humanas tipo 1 (HTLV-1) foi o primeiro retrovírus humano descrito (POIESZ et al., 1980). Estima-se que aproximadamente 5 a 10 milhões de pessoas são infectadas pelo HTLV em todo o mundo (GESSAIN e CASSAR, 2012). No Brasil, em Salvador-Bahia, a soroprevalência na população geral estima-se em 1,8 %, o que corresponderia a cerca de 40 a 50 mil indivíduos infectados na cidade (DOURADO et al., 2003). O HTLV-1 é o agente etiológico da paraparesia espástica tropical/mielopatia associada ao HTLV (HAM/TSP), uma doença neurológica crônico-degenerativa. Essa doença atinge o sistema nervoso central causando, principalmente, o aumento da espasticidade dos membros inferiores (GESSAIN et al., 1985; OSAME et al., 1986). Este vírus também está associado a doenças de natureza neoplásica como a Leucemia/Linfoma de Células T do Adulto (ATLL) (YOSHIDA, MIYOSHI e HINUMA, 1982) e de natureza inflamatória como a dermatite infecciosa associada ao HTLV-1 (DIH) (LA GRENADE, 1996).

Sabe-se que parte dos indivíduos infectados pelo HTLV-1 permanecem assintomáticos ao longo da vida e as pesquisas em relação aos fatores que determinam o desenvolvimento de uma das doenças associadas ao vírus, em um indivíduo infectado, ainda não são conclusivas. Acredita-se na possível influência do modo de transmissão, da carga proviral, tipo e magnitude da resposta imune do hospedeiro contra os antígenos do HTLV-1, além de fatores genéticos individuais como polimorfismos em genes de HLA (Antígeno Leucocitário Humano) e genes envolvidos na resposta imune (MARTINS e STANCIOLI, 2006). Acredita-se, ainda, na possível influência da localização da integração viral. Estudos demonstraram que a integração viral no genoma da célula hospedeira não é totalmente aleatória, cada retrovírus parece possuir alvos preferenciais no genoma humano (MITCHELL et al., 2004; GILLET et al., 2013).

Estudos demonstram que mutações em genes específicos do HTLV-1 também podem estar relacionadas ao desfecho da infecção. Algumas mutações naturais localizadas na ORF (Open Reading Frame)-I podem influenciar a carga proviral e a manifestação clínica de HAM/TSP (BARRETO et al., 2016). Foi demonstrado que a presença de SNV (Variantes de Nucleotídeos Únicos) na ORF-I, uma região do genoma viral localizada entre o gene *env* e a extremidade 3', podem alterar as

quantidades relativas de expressão das proteínas virais p12 e de p8 (PISE-MASISON et al., 2014), que estão relacionadas à evasão do sistema imune (EDWARDS et al., 2011).

O entendimento e identificação dos fatores que levam um indivíduo infectado a permanecer assintomático ou desenvolver alguma doença relacionada ao vírus é de extrema relevância, principalmente para o acompanhamento clínico dos portadores assintomáticos que podem apresentar um desses fatores e, então, poderiam ter o diagnóstico o mais precoce possível. Por outro lado, também são tópicos importantes para contribuir com o desenvolvimento de vacinas e terapias específicas, uma vez que os portadores assintomáticos permanecem sem tratamento e os pacientes que desenvolvem alguma das manifestações clínicas são tratados de maneira paliativa. O fato de não haver uma terapia específica e eficaz contra a infecção pelo HTLV-1 e suas enfermidades compromete a erradicação deste vírus.

Apesar da importância clínica e epidemiológica do HTLV-1, há um número limitado de genomas completos disponíveis, cerca de 0,12 genomas completos por 10.000 indivíduos infectados. Além disso, há uma escassez de estudos relacionados às sequências (totais e parciais) disponíveis nos bancos de dados públicos. Dessa forma, esse projeto visa caracterizar genomas do HTLV-1 de pacientes com HTLV-1 com diferentes condições clínicas.

2 REVISÃO DE LITERATURA

2.1 VÍRUS LINFOTRÓPICO DE CÉLULAS T HUMANAS TIPO-1

O HTLV-1 está classificado dentro da família Retroviridae, subfamília Orthoretrovirinae e gênero *Deltaretrovirus*. O vírus linfotrópico de células T humanas tipo-1 (HTLV-1) foi o primeiro retrovírus humano descrito (POIESZ et al., 1980). Em 1982, o HTLV-2 foi descoberto (KALYANARAMAN et al., 1982) e alguns anos depois, os subtipos 3 e 4 (CALATTINI et al., 2005). O HTLV-1 é o agente etiológico da paraparesia espástica tropical/mielopatia associada ao HTLV (HAM/TSP), uma doença neurológica crônico-degenerativa que atinge o sistema nervoso central causando, principalmente, o aumento da espasticidade dos membros inferiores (GESSAIN et al., 1985; OSAME et al., 1986). Este vírus também está associado a doenças de natureza neoplásica como a Leucemia/Linfoma de Células T do Adulto (ATLL) (YOSHIDA, MIYOSHI e HINUMA, 1982) e de natureza inflamatória como a dermatite infecciosa associada ao HTLV-1 (DIH) (LA GRENADE, 1996).

Observa-se que pacientes com DIH podem manifestar HAM/TSP ainda na infância ou adolescência (PRIMO et al., 2005) e ainda, evoluir para ATLL (HANCHARD et al., 1991; BITTENCOURT et al., 2001). Já foi descrito uma estreita associação entre essas patologias (BITTENCOURT e OLIVEIRA, 2005), mas ainda é desconhecido os fatores envolvidos no processo.

A ATLL pode, ainda, ser classificado em 5 formas clínicas: aguda, crônica, linfomatosa, smoldering (indolente) e tumoral primária de pele (BITTENCOURT et al., 2007; SHIMOYAMA, 1991). Nas duas primeiras, observa-se a presença de linfocitose, enquanto as outras formas clínicas não apresenta. (SHIMOYAMA, 1991).

2.2 ORIGEM DO HTLV

Estudos demonstraram, por análises de filogenética, que a região da África Central seria o ponto de origem para o ancestral comum dos PTLVs (Vírus Linfotrópicos de Células T de Primatas), devido ao fato de ser o único continente com a presença de todos os PTLVs (VAN DOOREN et al., 1998; VERDONCK et al., 2007).

A hipótese mais aceita, então, sugere que o vírus seguiu para os países do Novo Mundo (Caribe, Estados Unidos e América do Sul) por negros africanos, durante o trafico de escravos, e ao Japão, durante os séculos XVI e XVII (GALLO, SLISKI e WONG-STAAL, 1983; GESSAIN, GALLO e FRANCHINI, 1992; CATALAN-SOARES, PROIETTI, CARNEIRO-PROIETTI, 2001).

2.3 INFECÇÃO DO HTLV PELO MUNDO: DISTRIBUIÇÃO GEOGRÁFICA

Estima-se que aproximadamente 5-10 milhões de pessoas estejam infectadas pelo HTLV-1 no mundo e a infecção restringe-se a regiões que constituem áreas endêmicas como o sudoeste do Japão, África Ocidental e Central, Caribe, América do Sul, Caribe e Sudoeste dos EUA (Figura 1) (GESSAIN e CASSAR, 2012).

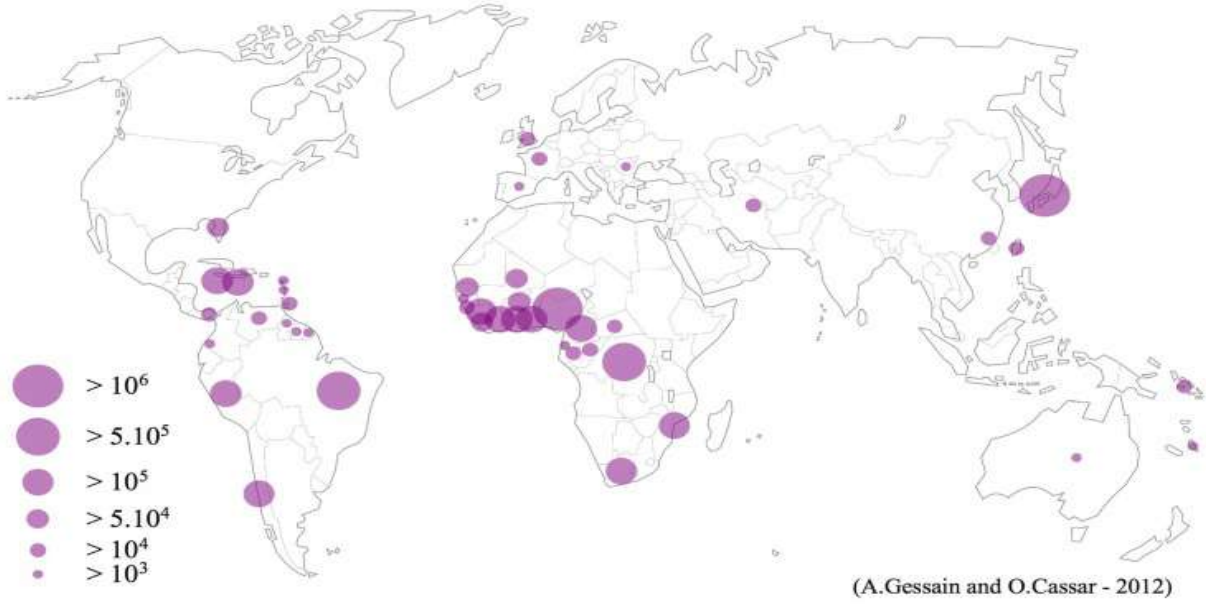


Figura 1: Distribuição, em número de casos (indicados pelos círculos de diferentes tamanhos) da infecção pelo HTLV no mundo. Adaptado de GESSAIN e CASSAR, 2012.

No Brasil se estima que aproximadamente 2,5 milhões de pessoas são portadoras do HTLV-1 (CARNEIRO-PROIETTI et al., 2002). No nosso país, a maior parte de estudos de soroprevalência foram realizados em bancos de sangue (Figura 2) e mostraram estados com uma maior prevalência do que outros, como a Bahia. Em

Salvador foi estudada a soroprevalência de infecção pelo HTLV-1 no marco do projeto Bahia Azul e foi considerada uma prevalência em torno de 1,8% da população geral.

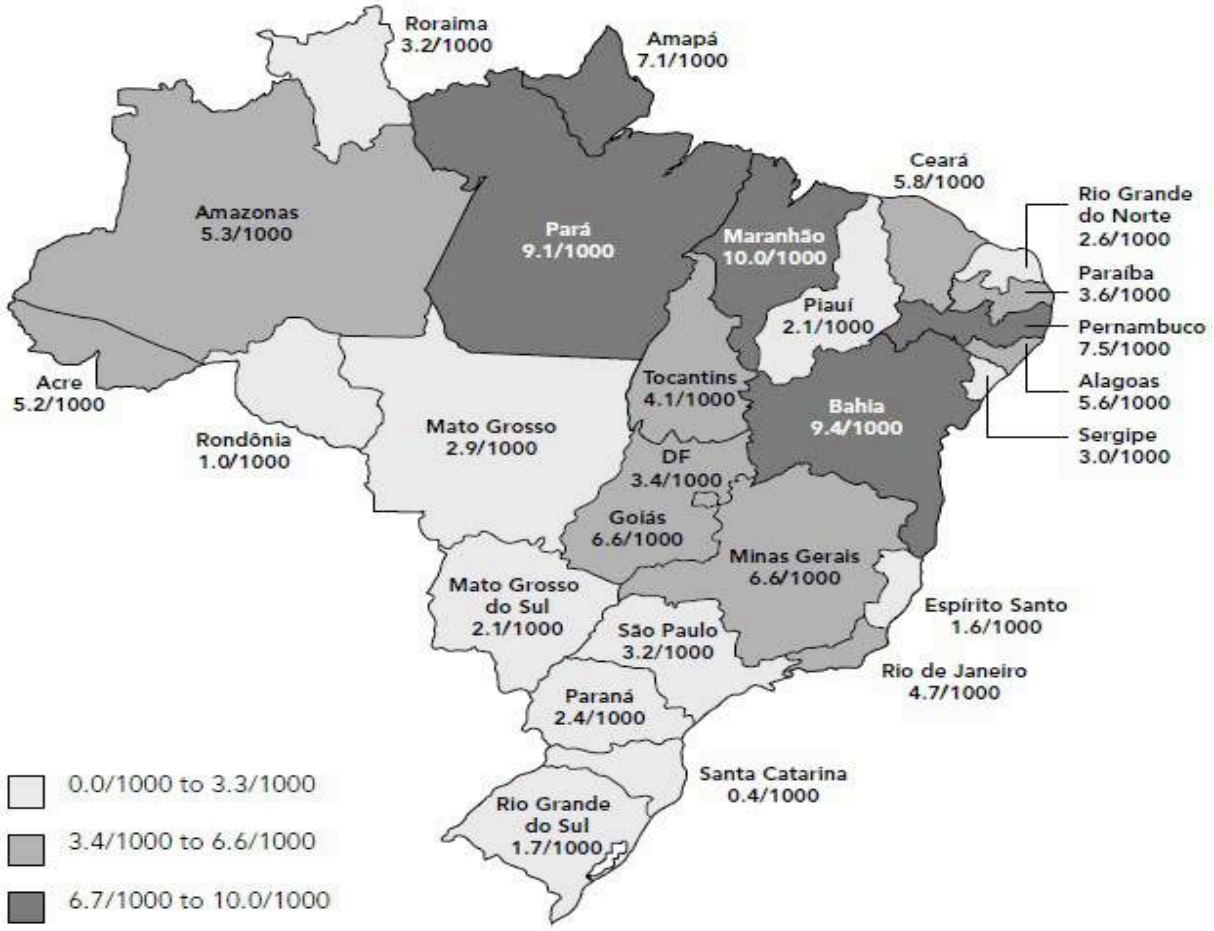


Figura 2: Mapa da soroprevalência para o HTLV em doadores de sangue no Brasil. Adaptado de CATALAN-SOARES et al., 2005.

2.4 PRINCIPAIS VIAS DE TRANSMISÃO DO HTLV-1

A transmissão do HTLV-1 ocorre predominantemente por três vias: vertical (mãe para filho), principalmente por intermédio da amamentação, onde a probabilidade de transmissão materno-infantil é de 10-30%, e ocorrendo também por via transplacentária durante o parto (TAJIMA e ITO, 1992); horizontal, através da relação sexual com parceiro portador sem proteção, com transmissão mais eficiente do homem para a mulher; e parenteral, por transfusão sanguínea ou

compartilhamento de instrumentos perfuro-cortantes contaminados principalmente entre os usuários de drogas injetáveis (ZALA et al., 1994).

2.5 ESTRUTURA MORFOLÓGICA E GENÔMICA VIRAL

O HTLV-1 possui uma estrutura morfológica e genômica similar aos outros retrovírus (figura 3). É formado por um envelope composto por duas proteínas: proteína de superfície (SU), a gp46, e uma proteína transmembrana (TM), a gp21. Além do envelope, contém a matriz viral (MA) e o capsídeo (CA), que envolve o material genético viral, composto por duas fitas de ácido ribonucleico (RNA) com polaridade positiva (BURKE et al., 1997). Há também outras proteínas importantes para a infecção viral no interior do capsídeo como a protease funcional (PR), a transcriptase reversa (TR) que converte a molécula de RNA em DNA, e a integrase (IN) que promove a inserção do DNA proviral no genoma da célula hospedeira (VERDONK et al, 2007; GOFF, 2001).

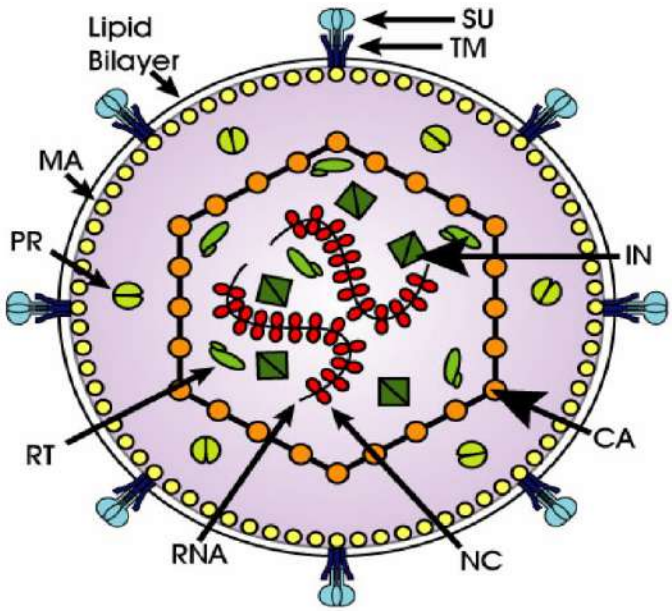


Figura 3: Estrutura morfológica do HTLV-1. Adaptado de VAN DOOREN, 2005.

Como se observa na figura 4, o DNA proviral possui em torno de 9032 pares

de bases (pb) e é formado pelos genes estruturais e enzimáticos retrovirais: *gag, pro, pol e env*. O genoma viral é flanqueado, nas duas extremidades, por sequência de repetição terminal longa (LTR), essencial à integração do provírus no genoma do hospedeiro (GREEN e CHEN, 2001). Estudos demonstraram que a transcrição viral é controlada por algumas proteínas (fatores de transcrição) que se ligam em regiões presentes na região promotora (GAUDRAY et al., 2002).

As proteínas p19 - matriz (MA), p24 – capsídeo (CA) e p15 – nucleocapsídeo (NC) são produtos (proteínas maduras) da clivagem realizada pela protease viral (produto do gene *pro*) na proteína precursora *gag* (sintetizada pelo gene *gag*). As enzimas TR e IN, por sua vez são um produto do gene *pol.*

O gene *env* codifica inicialmente uma proteína precursora que ao ser clivada gera as proteínas TM e SU. Essas proteínas estão fortemente ligadas à respostas humoral e celular, devido a sua localização na superfície da estrutura do vírus (DELAMARRE et al., 1996).

Além destas regiões, existe uma região localizada entre *env* e LTR da extremidade 3', conhecida como pX, que é particular do HTLV-1, que contém quatro quadros de leitura abertos parcialmente sobrepostos (ORF I, II, III e IV). A ORF-I codifica a proteína p12 que, posteriormente, é clivada em p8 (KORALNIK et al., 1992). As proteínas da ORF-I são capazes de interagir com proteínas celulares e de modular diferentes funções da célula do hospedeiro como a proliferação e a sinalização de células T, promovendo a evasão do HTLV-1 ao sistema imunológico, a transmissão do vírus para células vizinhas e uma infecção persistente no hospedeiro. A ORF-II produz as proteínas acessórias p13 e p30, envolvidas na regulação da transcrição gênica (BINDHU, NAIR e LAIRMORE, 2004). Já as ORFs III e IV produzem proteínas regulatórias: Rex (p27) e Tax (p40). Rex é essencial para regulação da expressão génica (YOUNIS e GREEN, 2005) e Tax para replicação viral (PROIETTI, 2006). Ainda na região pX encontramos o gene HTLV-1 bZIP factor gene (HBZ) envolvido na regulação da transcrição gênica e aumento da proliferação de células T (MESNARD, BARBEAU e DEVAUX, 2006).

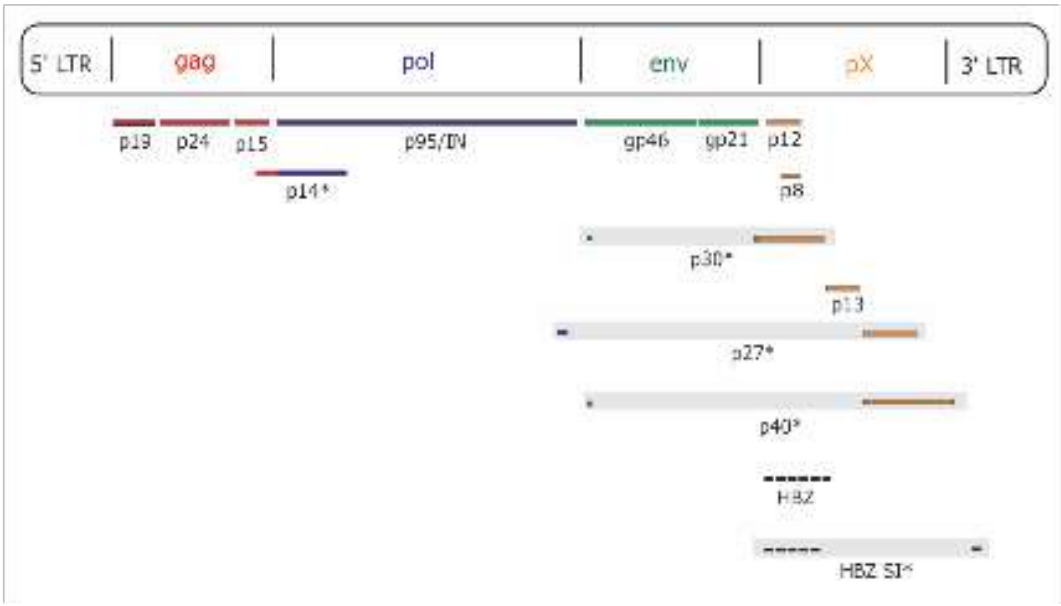


Figura 4: Estrutura genômica do HTLV-1. Adaptado de BARRETO et al., 2017.

2.6 EPIDEMIOLOGIA MOLECULAR DO HTLV-1

Os estudos aprofundados sobre as regiões que compõem o provírus permitiram estabelecer a epidemiologia molecular do HTLV-1. Ao analisarem os genes *env* e LTR do DNA proviral e identificar alterações nucleotídicas, foi possível classificar o HTLV-1 em 7 subtipos distintos, baseado nessas diferenças encontradas. O primeiro subtipo “a” ou, também denominado, Cosmopolita (SEIKI, HATTORI e YOSHIDA, 1982), como o nome já sugere, pode ser encontrado em quase todo o mundo. Os outros subtipos, se localizam em regiões específicas (como demonstra a figura 5), como o “b” ou Central Africano, “c” ou da Melanésia (GESSAIN et al, 1991; BASTIAN et al, 1993); “d” ,subtipo isolado de pigmeus em Camarões e no Gabão (CHEN et al, 1995; MAHIEUX et al, 1997);“e” e “f”, isolados de pigmeus na República Democrática do Congo e de um indivíduo do Gabão, respectivamente (SALEMI et al, 1998) e um novo subtipo isolado em Camarões, na África Central, denominado “g” (WOLFE et al, 2005).

Figura 5: Distribuição dos subtipos do HTLV-1 no mundo. Adaptado de ARAÚJO, 2012.

2.7 PATOGÊNESE DO HTLV-1

2.7.1 Propagação do HTLV no Organismo

Quando nos referimos ao HTLV-1, as células T CD4 + são o alvo principal e preferencial para infecção por HTLV-1 in vivo, embora também possa infectar linhagens de células dendríticas e mieloides (JONES et al., 2008). Logo após a infecção primária, pelas vias de transmissão mencionadas anteriormente, o HTLV-1 tenta se expandir colonizando novos alvos. O vírus utiliza diversos mecanismos como transferência célula a célula, transcrição reversa do RNA viral, integração do provírus no cromossomo, expressão de proteínas virais e brotamento de novos vírions para se expandir. Mas, as principais vias utilizadas envolvem a divisão mitótica de uma célula contendo um pró-vírus integrado (expansão clonal) e a sinapse virológica (VS).

No plasma de pacientes infectados pelo HTLV-1 comumente encontramos baixa ou até nenhuma presença de partículas virais. Os vírions do HTLV-1 são pouco infecciosos e a infecção mediada por células é muito mais eficiente do que a infecção

livre de células (FAN et al., 1992), mas mesmo que amostras demonstrem quantidades indetectáveis das partículas virais, estudos demonstraram que células infectadas pelo HTLV-1 apresentam replicação viral e expressão de antígenos (HANON et al, 2000). O vírus utiliza, como uma das principais vias de replicação, a expansão clonal de células infectadas, via mitose (WATTEL et al., 1995), não sendo necessário a utilização da transcriptase reversa. A não utilização dessa enzima contribui para sua maior estabilidade genômica, quando comparamos aos outros retrovírus.

Além disso, o vírus pode induzir o contato de células infectadas com células não infectadas, onde acontece a transferência de material viral, caracterizando a sinapse virológica (BANGHAM et al., 2003). Quando uma célula infectada entra em contato com uma célula não infectada, o centro de organização microtubular (MTOC) polariza-se na junção célula-célula, promovendo o acúmulo de proteínas Gag e o material genético do vírus nessas regiões, permitindo a passagem de partículas virais entre elas (IGAKURA et al., 2003; MATSUOKA e JEANG, 2007). Vale ressaltar a diferença entre a VS e a Sinapse Imunológica (SI): na VS, o citoesqueleto da célula infectada direciona-se para a célula-alvo (ainda não infectada) (NEJMEDDINE; BANGHAM, 2010). Ainda na VS há também transmissão de material viral por biofilmes virais (parecidos com os biofilmes bacterianos). O biofilme é composto por componentes da matriz extracelular (ECM) e lectinas celulares e funciona como uma espécie de proteção ao vírus do reconhecimento imunológico, enquanto partículas virais são montadas e transferidas para células-alvos (JIN, SHERER e MOTHES, 2010). Outra estratégia é a transmissão dos vírions, partícula viral infectante, por contato célula-célula que pode ocorrer, também, por meio de conduítes celulares, induzidos pela proteína viral p8, a qual favorece a persistência do HTLV-1 no hospedeiro (PROOYEN et al., 2010).

Estudos demonstraram que a integração viral no genoma da célula hospedeira não é totalmente aleatória, cada retrovírus parece possuir alvos preferenciais no genoma humano (MITCHELL et al, 2004). Existem alguns fatores que pode influenciar a preferência de cada retrovírus como as propriedades da integrase viral (LEWINSKI et al, 2006), proteínas celulares (BUSHMAN, 2003), a estrutura da cromatina no local

da integração (PRUSS et al, 1994; WANG et al, 2007) e as proteínas que auxiliam na adesão ao DNA (AL-MAWSAWI et al, 2006). O HIV, por exemplo, possui um viés em direção aos genes ligados às ilhas CpG e aos sítios de iniciação de transcrição dos genes (CIUFFI et al, 2006). Quando falamos sobre o HTLV-1, pouco é conhecido sobre esse viés, mas estudos demonstraram que, in vivo, há fatores que possivelmente influenciam para distribuição genômica dos sítios de integração proviral em infecção persistente pelo HTLV-1 (MEEKINGS et al, 2008).

Como uma resposta imune antiviral também é iniciada rapidamente, a eficácia do ciclo infeccioso é severamente atenuada logo após a infecção (CARPENTIER et al., 2015). Como observado na figura 6, o ciclo infeccioso inicia-se com a adsorção do mesmo, utilizando alguns receptores de membrana celular como GLUT1 (transportador de glicose). Após a fusão das membranas e a internalização da partícula viral, acontece a ação da enzima transcriptase reversa que consiste na transcrição do RNA em DNA. E então, este DNA é inserido no genoma da célula hospedeira pela integrasse. Após a ligação do DNA viral no DNA cromossômico, o mesmo é chamado de DNA proviral (PINON et al, 2003; MANEL, 2005; GHEZ et al, 2006; JONES et al, 2011).

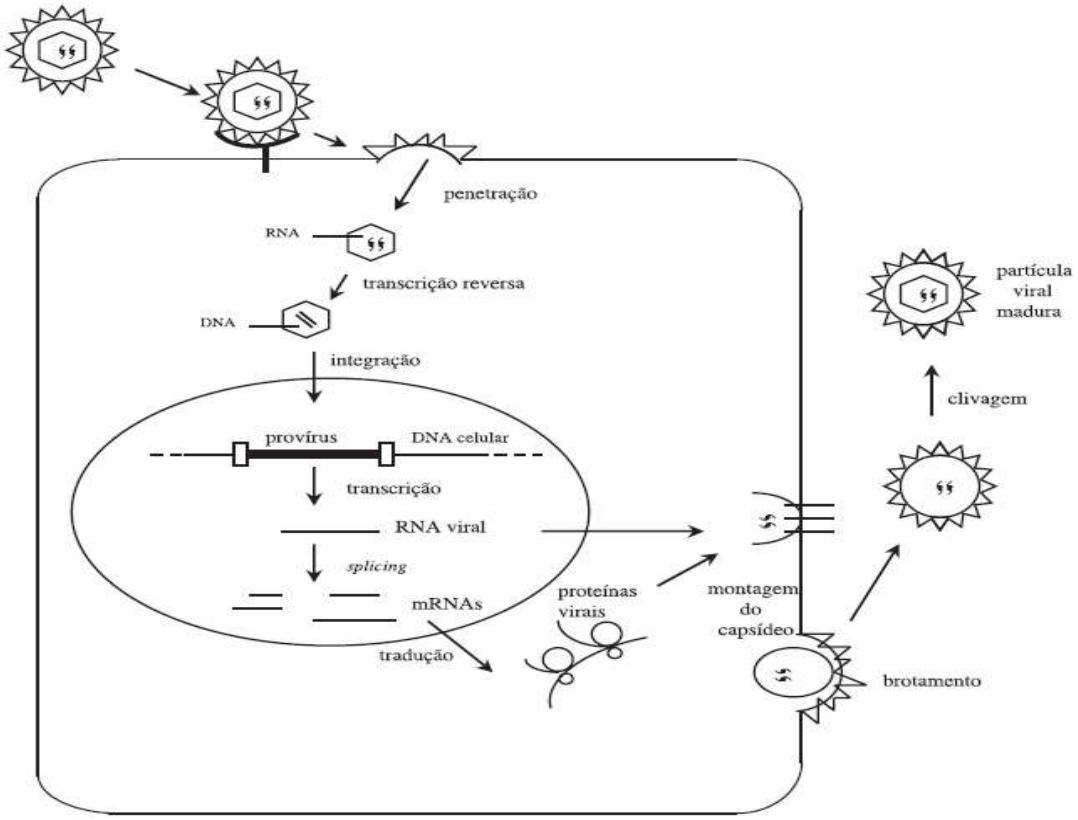


Figura 6: Ciclo de replicação clássica dos retrovírus. Adaptado de PROIETTI, 2006.

2.8 FISIOPATOLOGIA DAS DOENÇAS ASSOCIADAS AO HTLV-1

O contato inicial com o HTLV-1 ocorre principalmente por meio da amamentação, relação sexual e transfusão de sangue (GONÇALVES et al., 2010). Após um longo período de exposição a infecção costuma aparecer os sintomas (PROIETTI et al., 2005). Há no organismo a presença de múltiplas populações de células infectadas, cada uma dessas populações contêm o DNA proviral integrado em um local diferente do genoma do hospedeiro, caracterizando uma população policlonal.

Diante da contínua disseminação viral no organismo, há aumento de células infectadas no organismo, esse aumento de células infectadas é medido através da carga proviral no organismo. Um modelo de replicação viral por contato célula a célula nos estágios iniciais da infecção, seguido por uma proliferação clonal sustentada contrabalançando a resposta imune do hospedeiro. Ciclos repetitivos de expressão

viral seguidos de silenciamento transcricional desafiam continuamente a resposta imune, iniciando assim a inflamação e, em última análise, levando a HAM/TSP (CARPENTIER et al., 2015b). Em pacientes com DIH e HAM/TSP encontra-se uma população policlonal de linfócitos T infectados (oriunda da proliferação de diferentes clones infectados) e um padrão de integração policlonal (policlonalidade dos linfócitos) (GILLET et al, 2011) (Figura 7). Os pacientes com DIH tipicamente têm uma alta carga proviral de HTLV-1, comparável aos pacientes com HAM / TSP (PRIMO et al., 2009) (BATISTA et al., 2019). Na ATLL, uma das múltiplas populações existentes pode adquirir vantagem proliferativa por acúmulo de alterações genéticas por um processo ainda não esclarecido, levando à proliferação monoclonal exacerbada com um único local de integração do DNA proviral no genoma do hospedeiro (Figura 7). Nessa patologia há a indução da proliferação das células e inibição da apoptose, levando ao aumento do número de células que está relacionado com a elevação da carga proviral (MATSUOKA, 2005). A patogênese da ATLL é pouco esclarecida, mas sabe-se que múltiplos fatores podem estar relacionados com o desenvolvimento de ATLL: aqueles associados ao vírus, como carga proviral elevada, e aqueles associados ao portador (MORTREUX, GABET e WATTEL, 2003).

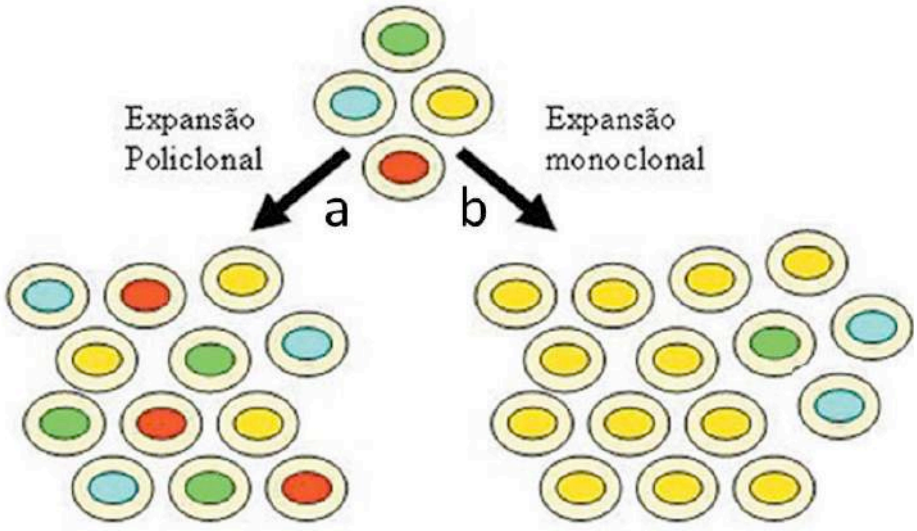


Figura 7: Padrões da expansão clonal de linfócitos T infectados. Adaptado de FARRE et al, 2008.

2.9 ESTRATÉGIAS PARA ANÁLISE DA INTEGRAÇÃO VIRAL

A primeira estratégia desenvolvida foi baseada na técnica do Southern Blot e visava realizar diagnóstico molecular da ATLL. Aproveitando que o DNA proviral não tem sítio de restrição para a enzima *EcoRI*, o DNA extraído do PBMC obtido de um portador do HTLV-1 era digerido com esta enzima de restrição e posteriormente submetido a técnica de Southern blot utilizando uma sonda para o HTLV-1, o que possibilitava detectar na membrana aqueles fragmentos de DNA humano que continham o DNA proviral inserido. Caso houvesse uma única banda, indicava que existia um clone predominante (possivelmente neoplásico) o que era indicativo de ATLL. Nos portadores assintomáticos, o padrão obtido no Southern Blot era de um conjunto de bandas tipo “smear”, o que indicava a presencia de múltiplos clones. Esta técnica, além de ser demorada e requerer uma quantidade considerável de DNA de partida, o que não sempre está disponível, não identifica o local onde o DNA proviral está inserido e deste modo também não identifica o clone predominante, dificultando o acompanhamento da progressão clinica dos pacientes de ATLL ou possíveis recaídas terapêuticas.

Na técnica do PCR invertido e longo (ILPCR) se pretende identificar, por amplificação por PCR e sequenciamento, as sequências genômicas humanas nas que o DNA proviral se inseriu. Como estas sequências são desconhecidas, o passo inicial nesta estratégia é inverter a sequência onde o provírus está integrado no meio de uma sequência humana desconhecida, para que esta sequência desconhecida esteja flanqueada por sequencia viral conhecida (conforme está demonstrado na figura 8) (ETOH et al, 1997). Para isto, se inicia com a digestão do DNA extraído das células infectadas pela *EcoRI* (não digere regiões provirais do HTLV-1) e, por meio da T4 DNA ligase, gerar uma autoligação. Posteriormente, a região pX do HTLV-1 sofrerá digestão pela enzima *MluI.* O objetivo desta ultima digestão é abrir a estrutura circular que foi gerada durante a ligação de maneira que se obtenha um fragmento com as sequencias virais conhecidas nos extremos e as sequencias humanas desconhecidas no interior. Deste modo, é possível amplificar as sequências humanas desconhecidas

utilizando oligos desenhados com as sequências virais conhecidas. Finalmente, os amplicons obtidos são sequenciados com uma plataforma Sanger e identificados utilizando a ferramenta BLAST (https://blast.ncbi.nlm.nih.gov/Blast.cgi). Por utilizar o sequenciamento Sanger, e permitir uma única leitura de sequência, esta estratégia tem a importante limitação de que somente é informativa e proporciona bons resultados nos casos em que existe um clone predominante que apresentou proliferação exacerbada, o que acontece nos casos de ATLL, principalmente naqueles com formas agudas e crônicas que presentam carga proviral muito alta e marcada linfocitose. Porém, os resultados não são tão satisfatórios para a evolução de pacientes de ATLL com formas que se caracterizam por não ter linfocitose como as formas smoldering, limfomatosa e tumoral primaria de pele, assim como para estudar a clonalidade em pacientes de DIH, HAM/TSP ou portadores assintomáticos, o que seria muito importante para o acompanhamento clinico destes portadores e a detecção precoce de possível evolução para ATLL.

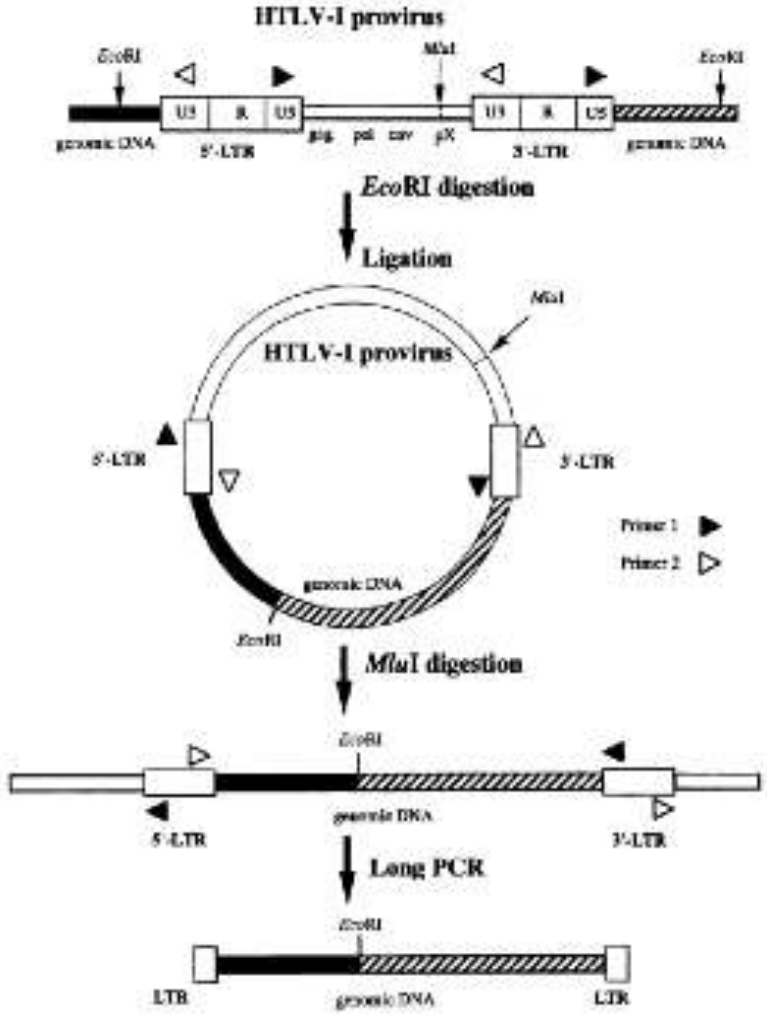


Figura 8: Demonstração em forma de esquema da técnica de Bangham**.** Adaptado de ETOH et al, 1997.

Em 2011 foi publicada outra estratégia para a avaliação da integração viral baseada no sequenciamento de genoma completo pela tecnologia de sequenciamento de nova geração utilizando a plataforma HiSEq da Illumina (GILLET et al., 2011). Nesta estratégia, e de maneira resumida, inicialmente era fragmentado (por métodos mecânicos ou enzimáticos) o DNA extraído das PBMC obtidas de portadores do HTLV-1 com diferentes condições clinicas, sequenciado todo o genoma da célula hospedeira e finalmente, por técnicas de bioinformática, eram selecionados aqueles reads continham sequencias virais e avaliadas as sequencias humanas contiguas para determinar o local de integração. Esta abordagem permitia um estudo

muito aprofundado dos clones existentes sem uma etapa de amplificação de sequencias, o que evita o viés favorecendo a detecção de algumas sequencias em detrimento de outras. Deste modo permitiu fazer uma avaliação muito acurada de todos os locais de integração e da abundância dos clones presentes na amostra assim como das possíveis alterações ocorridas nas regiões genômicas humanas interrompidas como impacto da integração do provirus (GILLET et al., 2013). Esta estratégia foi desenvolvida para fins principalmente de pesquisa e o custo econômico do procedimento e a complexidade técnica e de análise de dados limitam sua utilização para fins diagnósticos. Há geração da informação do genoma total da célula hospedeira, mas se utiliza somente a informação da região onde o provírus se inseriu. Por outro lado, a utilização da plataforma Illumina para o sequenciamento comportava uma limitação do tamanho dos fragmentos que podiam ser sequenciados a 300 bp, o que dificulta muito a montagem do provirus do HTLV-1 que tem aproximadamente 9kb. Nesta situação, somente é possível chegar a uma sequência única do provirus total, não sendo possível relacionar a presença de SNV no provirus com um local de integração proviral e, por tanto, associar o SNV ao clone.

2.10 NOVA METODOLOGIA DE SEQUENCIAMENTO

O primeiro método de sequenciamento de DNA surgiu em 1975, desenvolvido por Sanger (SANGER e COULSON, 1975), constituindo método de terminação de cadeia. Seguido por outra metodologia baseada em sequenciamento químico, de Maxam e Gilbert em 1977 (MAXAM e GILBERT, 1977). Esses métodos formaram os sequenciamentos de primeira geração. Em 2005 surgiu, então, a segunda geração com o pirosequenciamento. E, também, o sequenciamento por síntese (SGS). Essas metodologias possuíam alto rendimento e baixo custo, permitindo uma ampla utilização. Apesar disso, geravam reads curtos, ocasionando montagens bastante fragmentadas, principalmente quando nos referimos a genoma maiores. Com o surgimento dessa demanda, começou-se o aparecimento da terceira geração. Essa geração além de produzir reads mais longos, permitiu que o sequenciamento fosse realizado em tempo real, sem a necessidade da utilização do PCR (diminuindo viés)

e, consequentemente, redução de tempo de sequenciamento (SCHADT, TURNER e KASARSKIS, 2010).

Em 2014 surgiu uma nova tecnologia de sequenciamento oferecida pela ONT (Oxford Nanopre Technologies) que possui, entre outras vantagens, sequenciamento de cadeias de nucleotídeos muito mais longas que outras plataformas disponíveis como Illumina ou Ion Torrent (Thermo Fisher). Apesar de apresentar uma maior taxa de erro, é uma tecnologia com o custo muito menor, sem a necessidade de infraestrutura complexa e alto conhecimento no processo de produção da biblioteca.
O dispositivo utilizado para o sequenciamento é denominado de MinION. Pequeno, leve e o principal: um dispositivo portátil. Possui um software chamado de MinKNOW que permite-o realizar desde a aquisição até a análise de dados (em tempo real) (LU, GIORDANO e NING, 2016). Esse software é executado no computador host ao qual o MinION está conectado. O MinKNOW realiza várias tarefas principais como aquisição de dados, análise e feedback em tempo real.

Assim como a tecnologia de SGS, a preparação da biblioteca é necessária. Essa preparação consiste em etapas como reparo das extremidades em fragmentos cortados, adição de uma base “A” na extremidade 3’ (cauda dA), ligação do adaptador e, para finalizar, purificação para remover nucleotídeos e enzimas.

O dispositivo possui uma flow cell, como podemos observar na Figura 9, contém uma membrana cheia de nanoporos. Quando a molécula de DNA está atravessando o poro, uma primeira proteína transforma a dupla hélice do DNA em duas fitas. A segunda proteína forma um poro na membrana e mantém a molécula adaptadora no poro, a qual consegue manter as bases na passagem por tempo suficiente para que possam ser eletronicamente identificadas. As bases ao passarem pelo poro induzem uma alteração da corrente iônica presente na membrana, a qual é medida e caracterizada. Cada base causa uma alteração distinta, com intensidades diferentes, tornando possível identifica-las. Esses dados são repassados para o software que é responsável pela aquisição e análise dos dados.

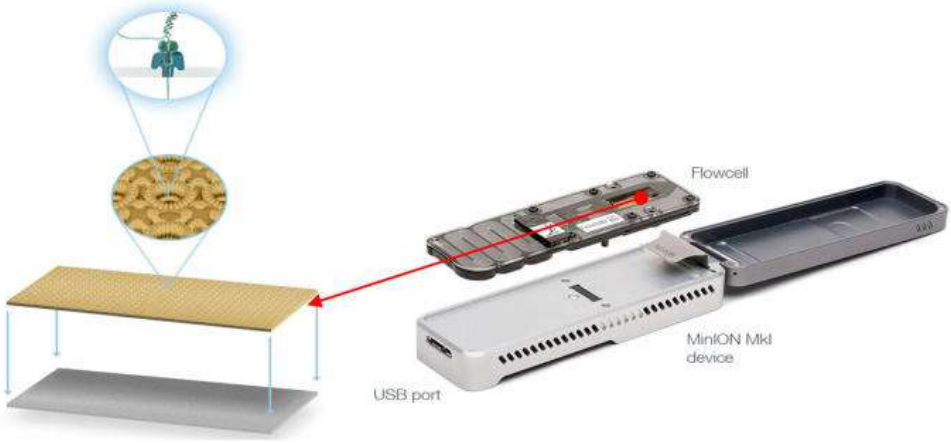


Figura 9: Dispositivo de sequenciamento MinION. Adaptado de LU, GIORDANO e NING, 2016.

3 OBJETIVOS

3.1 GERAL

Avaliar a diversidade genética do HTLV-1 em portadores do vírus com diferentes condições clínicas.

3.2 ESPECÍFICOS

(I) Analisar a diversidade genética da ORF-I do HTLV-1 em pacientes com diferentes condições clínicas;

(II) Identificar SNV no genoma completo do HTLV-1 que possam estar relacionados ao desfecho clínico do paciente em sequencias já publicadas;

(III) Desenhar uma metodologia para o sequenciamento baseado em nanoporos do genoma completo do HTLV-1.

4 RESULTADOS

Os resultados obtidos neste trabalho sobre o sequenciamento e caracterização do genoma completo do HTLV-1 de portadores do vírus com diferentes condições clínicas foram organizados em 3 seções.

Na seção 1 estão os resultados da avaliação da diversidade genética da ORF-I do HTLV-1 em pacientes com diferentes perfis clínicos com o artigo já publicado e intitulado "*Assessment of genetic diversity of HTLV-1 ORF-I sequences collected from patients with different clinical profiles*". **(doi: 10.1089/AID.2019.0127)**

Na seção 2 encontra-se o resultado da caracterização do genoma completo do HTLV-1 de pacientes com diferentes condições clínicas com o artigo, já submetido na revista Virus Research, intitulado *"Analysis of HTLV-1 complete genomes from patients with different clinical outcomes".*

Na seção 3 está o desenho dos primers para sequenciamento do genoma completo do HTLV-1 utilizando o dispositivo portátil MinION.

4.1 SEÇÃO 1: ASSESSMENT OF GENETIC DIVERSITY OF HTLV-1 ORF-I SEQUENCES COLLECTED FROM PATIENTS WITH DIFFERENT CLINICAL PROFILES

AIDS RESEARCH AND HUMAN RETROVIRUSES
Volume 00, Number 00, 2019

DOI: 10.1089/aid.2019.0127

Assessment of Genetic Diversity of HTLV-1 ORF-I Sequences Collected from Patients with Different Clinical Profiles

Melina Mosquera Navarro Borba,[1] Lourdes Farre,[1,2] Achilea Lisboa Bittencourt,[3] Maria Fernanda de Castro-Amarante,[1] Bernardo Galvão-Castro,[4] Luciane Amorim Santos,[1,4] Thessika Hialla Almeida Araújo,[1,4] Luiz Carlos Junior Alcantara,[5] and Fernanda Khouri Barreto[5,6]

Abstract

The human T cell lymphotropic virus type 1 (HTLV-1) infects 5 to 10 million individuals and remains without specific treatment. This retrovirus genome is composed of the genes gag, pol, env, and a region known as pX. This region contains four open reading frames (ORFs) that encode specific proteins. The ORF-I produces the protein p12 and its cleavage product, p8. In this study, we analyzed the genetic diversity of 32 ORF-I sequences from patients with different clinical profiles. Seven amino acid changes with frequency over 5% were identified: G29S, P34L, L55F, F61L, S63P, F78L, and S91P. The identification of regions where the posttranslational sites were identified showed a high identity among the sequences and the amino acid changes exclusive of specific clinical profile were found in less than 5% of the samples. We compare the findings with 2.406 sequences available in GenBank. The low overall genetic diversity found suggested that this region could be used in the HTLV-1 vaccine development.

Keywords: HTLV-1, ORF-I, mutations

THE HUMAN T cell lymphotropic virus type 1 (HTLV-1) was the first described human retrovirus.[1] It is estimated that 5–10 million people are infected with HTLV-1 in the world, and although this infection is endemic in different geographic regions, it still remains without effective therapeutic methods.[2] It is known that most patients infected with HTLV-1 do not develop clinical manifestations, but this retrovirus is the etiologic agent of infective dermatitis associated to HTLV-1 (IDH), HTLV-1-associated myelopathy/tropical spastic paraparesis (HAM/TSP), and adult T cell leukemia/lymphoma (ATLL), among others.[3–5] The major barriers for the development of HTLV-1 therapeutic vaccine is the comprehension why some individuals develop pathological processes while others remain asymptomatic and what is the best way to prevent viral persistence and infectivity.

A recent research demonstrated that the persistence of HTLV-1 infection is influenced by the expression of p12 and p8 proteins, encoded from the open reading frame (ORF)-I of the pX gene region. This study suggests that some natural ORF-I mutations alter the expression of the p12 and p8 proteins and that equivalent concentrations of both are necessary to prevent recognition and lysis of HTLV-1 infected cells by cytotoxic T cells.[6] We previously suggested that some of these natural ORF-I mutations might influence the proviral load and clinical manifestation of HAM/TSP.[7]

Considering the influence of the HTLV-1 ORF-I expression on the course of infection, this study aims to evaluate whether this region could be used as a target for the development of a therapeutic vaccine through the analysis of ORF-I genetic diversity.

In the first stage, we analyzed samples from 32 patients with defined clinical profile: 6 from patients with HAM/TSP, 6 from ATLL patients, 14 from asymptomatic patients, and 6 samples from patients with IDH. The clinical classification was carried out by medical experts according to World Health Organization (WHO). All samples were

[1]Fundação Oswaldo Cruz, Salvador, Brazil.
[2]Catalan Institute of Oncology, Bellvitge Biomedical Research Institute, Barcelona, Spain.
[3]Universidade Federal da Bahia, Salvador, Brazil.
[4]Escola Bahiana de Medicina e Saúde Pública, Salvador, Brazil.
[5]Fundação Oswaldo Cruz, Rio de Janeiro, Brazil.
[6]Universidade Federal da Bahia, Instituto Multidisciplinar em Saúde, Vitória da Conquista, Brazil.

TABLE 1. FREQUENCY OF MAJOR OPEN READING FRAME-I NATURAL MUTATIONS AND THEIR RESPECTIVE MOTIF

Mutation	*IDH* (n=6)	*HAM/TSP* (n=6)	*ATLL* (n=6)	*Asymptomatic* (n=14)	*Motif*
G29S	—	1	0	2	Transmembrane domain
P34L[a,b]	2	5	1	7	SH3 binding domain
L55F	2	0	0	0	Transmembrane domain
F61L	—	1	0	1	Transmembrane domain
S63P	6	5	6	11	Transmembrane domain
F78L	—	—	1	2	—
S91P	6	5	6	11	SH3 binding domain

[a]Mutation able to change the chemical physical profile.
[b]$p = 0.047$ between IDH and HAM/TSP profiles.
ATLL, adult T cell leukemia/lymphoma; HAM/TSP, HTLV-1-associated myelopathy/tropical spastic paraparesis; IDH, infective dermatitis associated to HTLV-1; SH3, Src homology 3.

anonymized and informed consent was written and obtained from each subject. These samples were obtained from Center of HTLV-1 Bahia School of Medicine and Public Health—Salvador, Bahia, Hospital Complex Prof. Edgar Santos—Salvador, Bahia and from Hemocentro—Ribeirão Preto, São Paulo, and the research was approved by the Ethics Committee of the Centro de Pesquisa Gonçalo Moniz/Fiocruz (N^0 377/2012).

In the second step, a search of the HTLV-1 ORF-1 sequences available on GenBank was performed to compare the results obtained from the sequences of our patients with other available sequences.

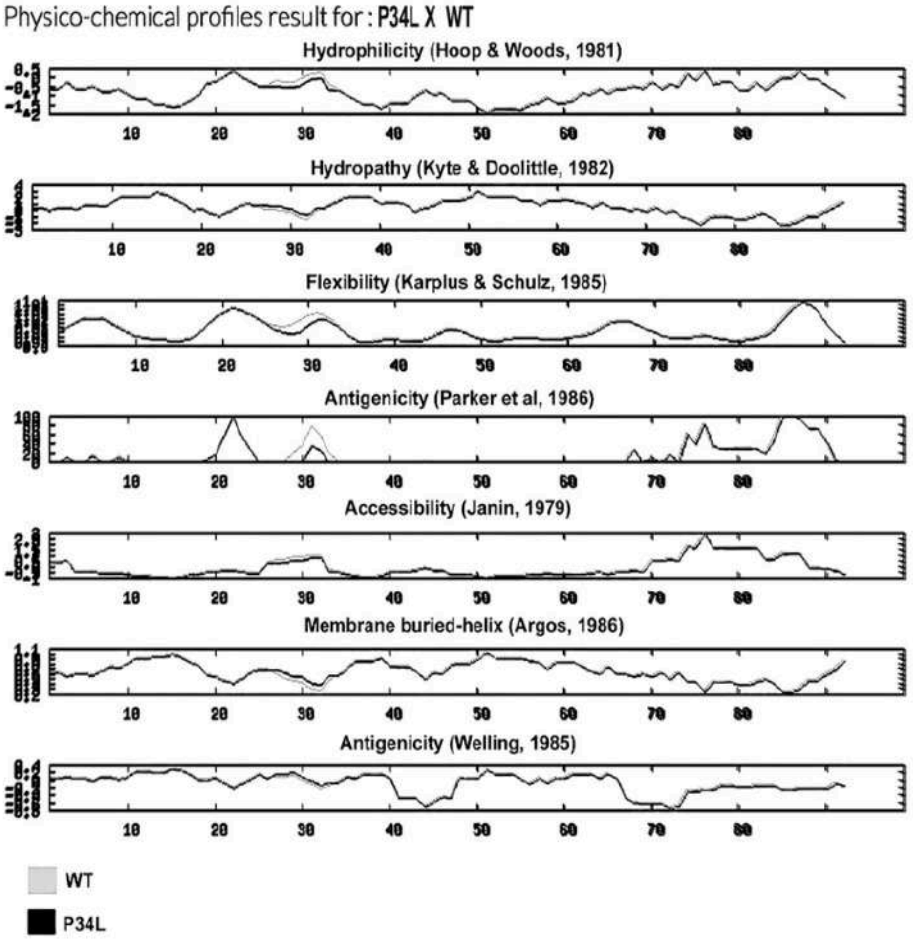


FIG. 1. Physicochemical analysis of P34L mutation versus wild type. The graphs are organized as follows: Hydrophilicity; Hydropathy; Flexibility; Antigenicity; Accessibility; Membrane-buried helix; Antigenicity.

TABLE 2. GENETIC DISTANCES IN HUMAN T CELL LYMPHOTROPIC VIRUS TYPE 1 OPEN READING FRAME-I SEQUENCES FROM PATIENTS WITH DIFFERENT CLINICAL PROFILES

	IDH	*HAM/TSP*	*ATLL*	*Asymptomatic*
IDH	0.006	0.007	0.006	0.007
HAM/TSP	0.007	0.007	0.007	0.008
ATLL	0.006	0.007	0.006	0.007
Asymptomatic	0.007	0.008	0.007	0.008

The peripheral blood mononuclear cells were obtained and DNA was extracted using spin column DNA extraction system (QIAamp DNA Mini Kit; Qiagen). The samples were used for amplification of ORF-I through polymerase chain reaction (PCR), as follows: denaturation (94°C, 3 min) annealing (94°C, 15 s), 65°C (45 s), 72°C (1 min), 35 times of cycle, and a final extension of 72°C for 8 min, with the primers 24^{+} (5′CGTATCGCCTCCCTCGCGCCATCAGAGTATGCTGCCCAGAACAG3′) and 27^{-} (5′CTATGCGCCTTGCCAGCCCGCTCAGGGTTCCATGTATCCATTTCGGA3′). The amplicons were purified using PureLink PCR Purification Kit (Thermo Fisher Scientific) and sequenced in an ABI Prism 3100 DNA Sequencer (Applied Biosystems Inc., Foster City, CA) using Taq FS Dye (Applied Biosystems) terminator cycle sequencing with the same PCR primers.

The files from the 32 sequences were trimmed, manually edited, and aligned to the HTLV-1 reference sequence ATK-1 (J02029) to generate the consensus sequence of each patient. The final dataset was first submitted to a search for the major natural amino acid changes, identified in at least of 5% of the sequences. Then the minor mutations, found in less than 5% of the sequences, were identified. All these analyses were done with Geneious R6 software.[8] The statistical analyses were performed using Fisher's exact test and a *p*-value lower than 0.05 was considered statistically significant. Then, we compared our results with ORF-I sequences available in the GenBank. This GenBank dataset was composed of 2.406 sequences, with 1.399 sequences from patients with HAM/TSP, 57 from ATLL patients, 945 from asymptomatic patients and 5 sequences from patients with IDH.

To perform the molecular analysis of the mutations identified, physicochemical analysis was carried out using Network Protein Sequence Analysis (NPS@) (http://npsa-pbil.ibcp.fr/) and the potential protein domain analysis was executed using the PROSITE tool.[9,10] The genetic distances within and between the sequences from patients with different clinical profiles were measured using the Mega 6.0 program.[11]

Seven natural amino acid changes with frequency over 5% were identified within the dataset: G29S, P34L, L55F, F61L, S63P, F78L, and S91P. Among them, five were located in specific motifs and were previously described as mutations that influence the expression profile of the HTLV-1 ORF-I protein product: G29S, P34L, F61L, S63P, and S91P.[6] The L55F (found only in sequences of patients with IDH) and F78L mutations were not described yet. Among the seven mutations identified, only P34L was found with a statistically significant difference in the frequency within the IDH and HAM/TSP groups ($p = 0.047$) (Table 1). Analysis performed with the GenBank available sequences reinforces these data, exception of the L55F and F78L mutations, found at low frequency.

The wild-type and mutated sequences were submitted to physicochemical analysis and only the P34L mutation was able to alter protein profile. The NPS@ analysis suggested that the ORF-I product with a leucine in 34 position was less hydrophilic, flexible, and antigenic than the wild type. The accessibility was also decreased, while the hydropathy and membrane-buried helix profile were slightly increased (Fig. 1).

To identify if these amino acid mutations were able to create or abrogate potential protein domains, we submitted the 32 ORF-I sequences to the scan PROSITE tool and no changes were observed. All sequences have a casein kinase II phosphorylation site at the 23–26 position and a protein kinase C phosphorylation site at the 75–77 position, which were not altered by the mutations (data not shown).

The analysis of ORF-I sequences revealed 10 mutations found in less than 5% of the samples. Despite being in low frequency, all these mutations have an important characteristic: they are observed only in specific clinical profiles. Six amino acid changes were detected in samples from asymptomatic patients (S7G, P45L, S69G, P73S, R82*, and A96V), two mutations were exclusive of HAM/TSP sequences (C39Y and P86S), while L5I and F84L mutation were identified only in ATLL samples.

In the GenBank dataset, S7G and A96V mutations and the P86S mutation were also found only in asymptomatic individuals and in HAM/TSP patients, respectively.

The overall diversity between sequences from patients with HAM/TSP, ATLL, IDH, and asymptomatic was 0.007, and the genetic distance values within and between the different clinical profiles are described in Table 2. The low overall genetic diversity found corroborates the fact that the HTLV-1 genome exhibits relatively few sequence variations and that the development of a therapeutic vaccine is possible. However, studies demonstrated that the induction of HTLV-1 protective immune response is not so simple.[12–15] Here, we suggest that a therapeutic vaccine may be a better alternative and the HTLV-1 ORF-I is a good target for the development of this vaccine. More analyses involving sequences from patient with others HTLV-1 pathologies can provide more information about the ORF-I genetic diversity and these data can be used for a design of HTLV-1 vaccine.

Acknowledgments

The authors are grateful to all participating donors, the professionals of the centers of care and the sequencing platform of FIOCRUZ / IGM. This work was supported by Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq 400900/2013-0, CNPq 150892/2018-7).

Availability of Supporting Data

All sequences are available in the GenBank database (accession numbers MF158987-MF159019).

Author Disclosure Statement

No competing financial interests exist.

References

1. Poiesz BJ, Ruscetti FW, Gazdar AF, Bunn PA, Minna JD, Gallo RC: Detection and isolation of type C retrovirus particles from fresh and cultured lymphocytes of a patient

with cutaneous T-cell lymphoma. Proc Natl Acad Sci U S A 1980;77:7415–7419.
2. Gessain A, Cassar O: Epidemiological aspects and world distribution of HTLV-1 infection. Front Microbiol 2012;3:388.
3. La Grenade L, Manns A, Fletcher V, *et al.*: Clinical, pathologic, and immunologic features of human T-lymphotrophic virus type I-associated infective dermatitis in children. Arch Dermatol 1998;134:439–444.
4. Gessain A, Barin F, Vernant JC, *et al.*: Antibodies to human T-lymphotropic virus type-I in patients with tropical spastic paraparesis. Lancet Lond Engl 1985;2:407–410.
5. Yoshida M, Miyoshi I, Hinuma Y: Isolation and characterization of retrovirus from cell lines of human adult T-cell leukemia and its implication in the disease. Proc Natl Acad Sci U S A 1982;79:2031–2035.
6. Pise-Masison CA, de Castro-Amarante MF, Enose-Akahata Y, *et al.*: Co-dependence of HTLV-1 p12 and p8 functions in virus persistence. PLoS Pathog. 2014;10:e1004454. [cited July 10, 2015].
7. Barreto FK, Khouri R, Rego FF de A, *et al.*: Analyses of HTLV-1 sequences suggest interaction between ORF-I mutations and HAM/TSP outcome. Infect Genet Evol 2016;45:420–425.
8. Kearse M, Moir R, Wilson A, *et al.*: Geneious Basic: An integrated and extendable desktop software platform for the organization and analysis of sequence data. Bioinforma Oxf Engl 2012;28:1647–1649.
9. Combet C, Blanchet C, Geourjon C, Deléage G: NPS@: Network protein sequence analysis. Trends Biochem Sci 2000;25:147–150.
10. Sigrist CJA, de Castro E, Cerutti L, *et al.*: New and continuing developments at PROSITE. Nucleic Acids Res 2013;41:D344–D347.
11. Tamura K, Stecher G, Peterson D, Filipski A, Kumar S: MEGA6: Molecular Evolutionary Genetics Analysis version 6.0. Mol Biol Evol 2013;30:2725–2729.
12. Bomford R, Kazanji M, De Thé G: Vaccine against human T cell leukemia-lymphoma virus type I: Progress and prospects. AIDS Res Hum Retroviruses 1996;12:403–405.
13. de Thé G, Bomford R: An HTLV-I vaccine: Why, how, for whom? AIDS Res Hum Retroviruses 1993;9:381–386.
14. Rodríguez SM, Florins A, Gillet N, *et al.*: Preventive and therapeutic strategies for bovine leukemia virus: Lessons for HTLV. Viruses 2011;3:1210–1248.
15. Mahieux R: A vaccine against HTLV-1 HBZ makes sense. Blood 2015;126:1052–1053.

Address correspondence to:
Fernanda Khouri Barreto, PhD
Instituto Multidisciplinar em Saúde
Universidade Federal da Bahia
Campus Anísio Teixeira
Rua Hormindo Barros, 58, Quadra 17, Lote 58
Bairro Candeias: 45.029-094
Vitória da Conquista
Brazil

E-mail: fernanda.khouri@hotmail.com

4.2 SEÇÃO 2: ANALYSIS OF HTLV-1 COMPLETE GENOMES FROM PATIENTS WITH DIFFERENT CLINICAL OUTCOMES

Analysis of HTLV-1 complete genomes from patients with different clinical outcomes

Melina Mosquera Navarro Borba[1], Luciane Amorim Santos[1,2,3,4], Reinaldo Conceição Neto[5], Felipe de Oliveira Andrade[6], Álvaro Salgado[7], Filipe Ferreira de Almeida Rego[3], Lourdes Farre[8], Fernanda Khouri Barreto[6,*]

[1] Instituto Gonçalo Moniz, Salvador, Brazil

[2] Escola Bahiana de Medicina e Saúde Pública, Salvador, Brazil

[3] Universidade Católica do Salvador, Salvador, Brazil

[4] Programa de Pós-graduação em Ciências da Saúde, Faculdade de Medicina da Bahia, Universidade Federal da Bahia, Salvador, Brazil

[5] Faculdade Cruzeiro do Sul, Vitória da Conquista, Brazil

[6] Universidade Federal da Bahia, Vitória da Conquista, Brazil

[7] Universidade Federal de Minas Gerais, Belo Horizonte, Brazil

[8] Institut Català d'Oncologia, Institut d'Investigació Biomédica de Bellvitge, Barcelona, Spain

*Correspondence: Federal University of Bahia, Multidisciplinary Health Institute, Campus Anísio Teixeira. Rua Hormindo Barros, 58, lot 58. Bairro Candeias: 45.029-094. Vitória da Conquista, Brazil. Email: fernanda.khouri@hotmail.com

Abstract

The Human T-lymphotropic virus type 1 (HTLV-1) is the etiologic agent of adult T-cell leukemia/lymphoma (ATLL), HTLV-1-associated myelopathy/tropical spastic paraparesis (HAM/TSP), and HTLV-1-associated infectious dermatitis (IDH), among other diseases. It is estimated that 5 to 10 million people are infected with HTLV-1 worldwide. Although HTLV-1 was the first human retrovirus described, the reasons why some carriers remain asymptomatic while others develop pathologies associated with the virus have not been fully established. The present study aims to identify nucleotide variations in the proviral genome that might be related with the different clinical conditions associated to the virus. For this purpose, 242 complete HTLV-1 genomes available in GenBank were downloaded and subjected to *in silico* analyzes to identify mutations and their possible impacts on the course of HTLV-1 infections. We selected 22 mutations, such as P34L (found in the p12 protein, that seems to play a role in clinical profile development) and the 8452 variation (found in the LTR, that results in the abrogation of the site for SP1 transcription factor and also seems to be important in patients' clinical condition). The present study suggests that, despite the low genetic diversity among HTLV-1 genomes, some mutations have the potential to alter the conformation of viral proteins that are important for infection outcomes. Therefore, further functional studies should be performed to assess the impact of these variations on the pathogenesis and on the development of clinical manifestations associated with HTLV-1.

Keywords: HTLV-1; complete genome; mutations; machine learning.

1. Introduction

The Human T-lymphotropic virus type 1 (HTLV-1) was the first human retrovirus described (Poiesz et al., 1980). It is estimated that 5 to 10 million people are infected with HTLV-1 worldwide, highlighting some geographic regions that have high endemic rates such as Southwest Japan, West and Central Africa, South America, the Caribbean, and Southwest USA (Gessain and Cassar, 2012). This retrovirus is the etiologic agent of adult T-cell leukemia/lymphoma (ATLL), HTLV-1-associated myelopathy/tropical spastic paraparesis (HAM/TSP), and HTLV-1-associated infectious dermatitis (IDH), among other clinical conditions (Gessain et al., 1985; Grenade et al., 1998; Yoshida et al., 1982).

Despite HTLV-1 infection being associated with the development of these pathologies, there are infected individuals who do not show symptoms associated with the virus or have nonspecific symptoms throughout life and are, therefore, classified as asymptomatic carriers of HTLV-1 (AC) (Saito et al., 2012). The reasons that lead some infected individuals to remain asymptomatic carriers while others develop pathologies associated with the virus are still unclear. In this context, there are hypotheses that suggest that genetic factors related to the host, as well as viral genetic variations, may influence the outcome of HTLV-1 infections (Araujo and Silva, 2006).

The HTLV-1 genome is composed of gag, pol, and env genes, flanked by two long terminal repeats (LTR), and the pX region that is situated between the env gene and the 3' end. The pX region has four open reading frames (ORFs) that encode accessory proteins, such as p12 (ORF-I), p13, p30 (ORF-II), and the regulatory proteins Rex (ORF-III) and Tax (ORF-IV) (Matsuura et al., 2010). The pX region also has the HTLV-1 bZIP factor gene (HBZ), which encodes HBZ protein and also participates in the regulation of genetic transcription and increased proliferation of infected T-cells (Satou et al., 2006; Arnold et al., 2006). This virus is classified into seven subtypes according to the env and LTR nucleotide variations: subtypes "a to g", where the subtype “a” (cosmopolitan) is further subdivided into five subgroups (1aA-1aE) (Hahn et al., 1984; Gessain et al., 1991; Salemi et al., 1998; Miura et al., 1994; Chen et al., 1995; Wolfe et al., 2005; Ehrlich et al., 1992; Gessain et al., 1992).

We have previously reported that nucleotide variations present in the ORF-I of the provirus can interfere with the proviral load and clinical manifestations of HAM/TSP (Barreto et al., 2016). Other studies have reported that mutations in the gene encoding the HTLV-1

surface glycoprotein induce specific neutralizing antibodies (Blanchard et al., 1999), and that the presence of the N93D mutation in gp46 protein may be related to symptomatic clinical status (Socorro de Almeida Viana et al., 2018). Also, a study has shown asymptomatic individuals with a higher proviral load in the presence of the G232A mutation in the LTR region (Neto et al., 2011).

In order to contribute and better understand the factors that lead an asymptomatic carrier to a certain clinical condition, the present study aims to identify possible contributions of nucleotide variations in the proviral genome for the development of HTLV-1-related diseases.

2. Materials and Methods

The molecular analysis was conducted by downloading all the complete provirus genomes of HTLV 1 (n= 242) with their respective clinical and epidemiological data, available in the GenBank database until August 17, 2020, using the search algorithm *(“Complete sequence” OR "Complete genome" OR "Whole Genome" AND "HTLV-1" OR "T-human Cell Lymphotropic Virus I")*. For the formation of this dataset, the following inclusion criteria were applied: (i) sequences of complete HTLV-1 provirus genomes; and (ii) HTLV-1 complete genome sequences not obtained from cell lines. The exclusion criterium was: sequences of HTLV-1 complete genomes that were not composed by all the coding or non-coding genome’s regions (gag, pol, env, pX, and LTR regions).

All genomes were aligned in the Geneious Prime® 2020.2.3 program using the MAFFT alignment software (Kearse et al., 2012) and were then submitted to subtyping and to the analysis of genetic distances. Considering that this study aims to identify nucleotide variations in the viral genome that can be related to different clinical conditions, mutations specific to each subtype could be a bias in our analysis. As a result, only the subtype 1a subgroup A sequences were analyzed for complete molecular characterization. The other subtypes were not included due to the lack of available epidemiological and clinical data associated with the sequences.

The molecular characterization of the subtype 1a subgroup A sequences were carried out in stages. First, a search for mutations was made and then these variations were assayed by machine learning analyzes to rank them according to their importance for clinical manifestations. Then, the mutations that showed frequencies higher than 5% and were ranked

by machine learning analyzes were selected for evaluation of their possible impact on physicochemical profile, post-translational modification sites, and binding sites for transcription factors.

2.1. Subtyping and Genetic Distances

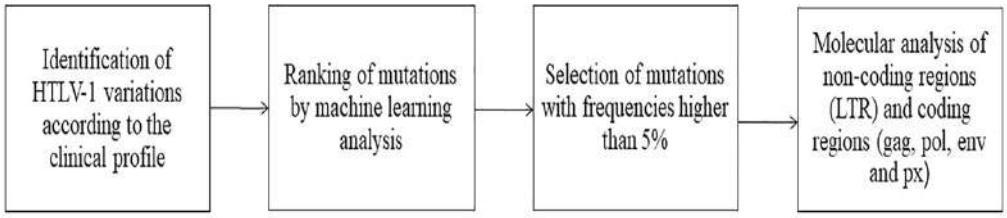
The LTR region was used for subtyping, along with reference sequences for each subtype and subgroup previously described in the literature: (a) subtype 1a subgroup A: DQ005552-DQ005555, Y16476, Y16477, Y16479, Y16480, Y16482, DQ005564-DQ005567, EF672333-EF672336, FJ853491, GU225731, and GU225732; (b) subtype 1a subgroup B: J02029, M37299, Y16484, and Y16487; (c) subtype 1a subgroup C: AY342303 and AY342304; (d) subtype 1a subgroup D: U12804-U12806; (e) subtype 1a subgroup E: AF054627 and Y16481; (f) subtype 1b: AY818425 and L76309; (g) subtype 1c: L02534; (h) subtype 1d: L76312; and (i) subtype 1e: Y17014.

The maximum likelihood phylogenetic tree was constructed using the IQ-TREE web server, under the TN+F+G4 (Tamura Nei with Empirical base frequencies and gamma distribution) nucleotide substitution model inferred by ModelFinder and implemented in the IQ-TREE (Trifinopoulos et al., 2016). Ultrafast bootstrap analysis with 1000 replicates was used to calculate the statistical support of the tree branches, graphically visualized in the Figtree v1.4.4 program (http: // tree.bio.ed.ac.uk/software/figtree/). The Genetic distances were measured within and between the different subtypes using the MEGA 10.1 software (Kumar et al., 2018).

2.2. Molecular characterization of subtype 1a subgroup A sequences

The identification of the variations in the complete HTLV-1 genomes was performed through the Geneious Prime® 2020.2.3 program using, as a reference, a consensus sequence generated from all genomes of subgroup 1aA. For annotation of positions, this consensus sequence was aligned according to the ATK-1 sequence (J02029) (Barreto et al., 2017). The workflow described in the methodology is represented schematically in Figure 1.

Figure 1. Schematic figure of the workflow of molecular characterization analyzes.



2.2.1 Machine Learning (ML)

The dataset was divided into three groups according to the clinical profile (ATLL, HAM/TSP, and IDH) and each group was compared to asymptomatic carrier sequences as control. The "one-hot encoding" binarizing scheme to convert the categorical genomic dataset to a numerical format appropriate for ML application was applied. To adjust the machine learning model, two different ML models, the XGBoost and random forest, were used for each dataset (Chen and Guestrin, 2016; Friedman, 2000; Schapire, 2003; Breiman, 2001). Model parameters were adjusted in a k-fold cross validation scheme, and model performance was evaluated on a test dataset previously withheld. Model results were interpreted using SHAP framework (Lundberg and Lee, 2017), which is a mathematical method used to interpret ML models once they have been adjusted on a dataset. It is based on game theory, and identifies which attributes (mutations, in this particular case) are more relevant to the ML model as it performs its classification between each group analyzed. In other words, it highlights which mutations, among all present in the dataset, are more relevant for discriminating between the groups being compared. The results of this analysis appear according to the respective importance of the mutations identified for the clinical profile. The mutations identified as important were presented through a rank, in order: from the most important mutation to the least important.

Finally, a Decision tree J48, a supervised method, was used for clinical manifestation classification (Decision tree J48). For this analysis, we considered variables such as mutations of the genome, geographic location, age, and gender data to construct the Decision tree. To

execute and validate the J48 decision tree approach, we used the data mining suite Weka, and accuracy greater than 70% was considered significant (Witten and Frank, 2002).

2.2.2 Molecular Analysis

For the molecular analysis, we considered those mutations that met the following criteria: (i) showed frequencies higher than 5%; (ii) appeared in the rank of machine learning analyzes; and (iii) generated an exchange of amino acids. To identify the possible impact of the selected variations in the protein structure, a physicochemical analysis was performed using Network Protein Sequence Analysis (NPS@) (http://npsa-pbil.ibcp.fr/) and the search for post-translational modifications sites was conducted utilizing the GeneDoc10 software and the Prosite tool (Combet et al., 2000; Sigrist et al., 2013). GraphPad Prism 8.4.2 (GraphPad Software, San Diego, California, USA) was used to plot the graphs of the physicochemical profile of proteins.

The selected mutations in the LTR region were subjected to an analysis for *in silico* identification of the transcription factor binding sites using the TFSCAN plugin – TRANSFAC database, from the Geneious software. In order to consider a motif, we used, as minimum limit, the size of four base pairs, and incompatibility was not allowed. After identifying the binding sites for transcription factors, we verified whether the selected mutations led to the creation or abrogation of the sites.

3. Results

3.1 Selection of complete HTLV-1 sequences

Our dataset consisted of 242 complete sequences of the HTLV-1 genome. Metadata included: geographic region, clinical profile, and sequencing technology. These data are organized in Table I.

Table I: Metadata of the complete HTLV-1 genome sequences available in GenBank database on August 17th.

Sequence data	Number of Sequences (n=242)

Geographic Region (Continents)	
Asia	184
North America	3
Central America	2
South America	34
Africa	9
Europe	2
Oceania	6
Not available	*2*
Clinical Profile	
ATLL[a]	26
HAM/TSP[b]	43
IDH[c]	5
Asymptomatic carriers	17
Not available	*151*
Sequencing Methodology	
Ion Torrent	31
Illumina Technology (Miseq platform)	37
Sanger	163
Not available	*11*

[a]ATLL: adult T-cell leukemia/lymphoma;

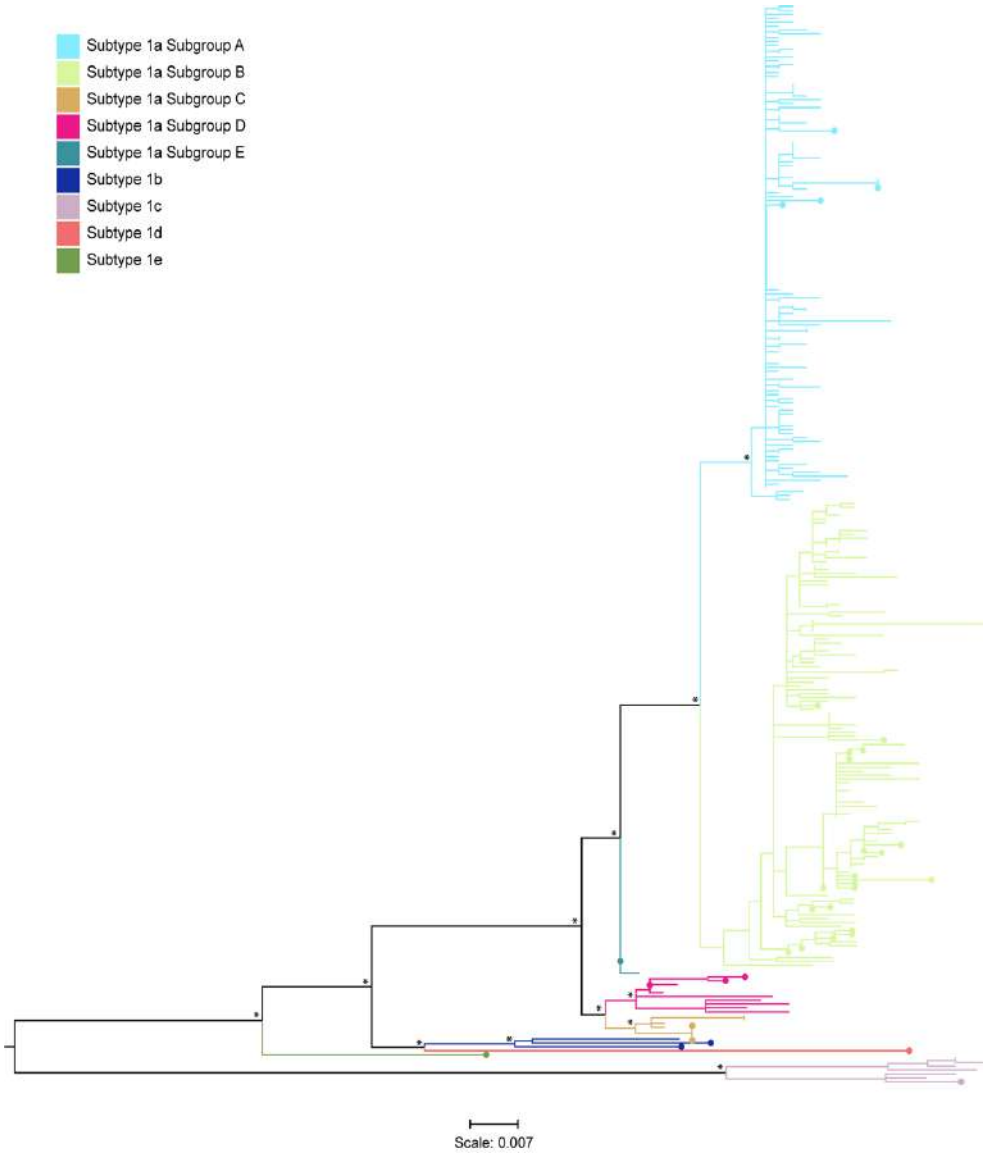
[b]HAM/TSP: HTLV-1-associated myelopathy/tropical spastic paraparesis;

[c]IDH: infective dermatitis associated with HTLV-1.

3.2 Subtyping

The sequences were classified by subtype and subgroups according to phylogenetic reconstruction. A total of 235 sequences were classified as subtype 1a, of which 99 were from subgroup 1aA, 125 from subgroup 1aB, four from subgroup 1aC, seven from subgroup 1aD, one from subtype 1b, and six from subtype 1c (Figure 2).

Figure 2. Phylogenetic tree with 242 complete genomes downloaded from GenBank on August 17th, with the reference sequences for each subtype and subgroup. The access numbers of the reference sequence are listed in the methodology, section 2.1. In addition, the respective access numbers of the dataset used can be found in supplementary table 1. Sequence from Subtype 1c was used as an outgroup to root the tree. The branches with bootstrap support >90% are indicated with one asterisk (*).



The analysis of genetic distance showed a global genetic distance of 0.0123 considering all 242 sequences. Successively, the distance between each subtype and subgroup were calculated and the minimum variation observed was 0.0034 (within subgroup 1aB), while the

maximum was 0.0875 (subgroup 1aB vs subtype 1c). (Supplementary material section – Table S2).

3.3 Molecular analysis of subtype 1a subgroup A sequences

The molecular analysis was performed with the subgroup 1aA dataset because it was the only one with clinical information associated to the sequences in 77 out of a total of 99. Subgroup 1aB included a significant number of sequences, but unfortunately only six had clinical information. As mentioned in the materials and methods section, we performed the molecular analysis by subtype to avoid bias caused by specific subtype mutations.

The dataset of the 1aA 77 complete genome sequences was subjected to molecular characterization and machine learning analysis. Of the 77, 16 sequences corresponded to ATLL, 5 to IDH, 42 to HAM/TSP, and 14 to AC. Considering the selection criteria for molecular analyzes, 22 mutations were evaluated: 11 were located in the promoter regions (LTR), shown in table II, and 11 in the coding regions, including the gag, pol, env, and pX regions, shown in table III.

In the LTR region, we observed that a variation of the nucleotide at position 677 was able to create a new transcription factor binding site (TFBS) for c-Myb, while the mutations that occurred at positions 125, 174, and 8452 caused the abrogation of the TFBS for SP1. Interestingly, these mutations were identified in the machine learning rank analysis as specific to a given clinical manifestation. Machine learning rank consists of an order of importance for each mutation in relation to the clinical profile classification (ATLL, HAM/TSP, and IDH).

The A125G mutation was the twenty-first in the IDH rank, while the G174A was the seventh in the ATLL classification. The variation at the 8452 position was ranked for both ATLL and IDH classification. Although changes at positions 501, 550, 631, 8828, and 8879 did not generate changes in TFBS, they were also identified as important mutations for a given clinical profile (Table II).

Table II. Mutations found in the LTR region and their impacts on the binding site for transcription factors.

Nucleotide Position	Nucleotide Change	Frequency (%)	Machine Learning Rank	TFBS[d] Criation	TFBS Abrogation

			Asymptomatic vs. ATLL[a]	vs. HAM[b]	vs. IDH[c]		
89	C -> A	47.3	-	-	23	-	-
125	A -> G	15.8	-	-	21	-	SP1[e]
174	G -> A	14.3	7	-	-	-	SP1
501	T -> C	36.4	-	3	-	-	-
550	A -> G	36.4	-	6	-	-	-
631	C -> T	15.6	-	9	-	-	-
677	G -> A	35.1	-	12	8	c-Myb	-
768	T -> G	33.8	3	4	2	-	-
8452	G -> A	15.6	17	-	16	-	SP1
8828	A -> G	36.4	16	-	-	-	-
8879	C -> T	9.1	-	15	-	-	-

[a]ATLL: adult T-cell leukemia/lymphoma; [b]HAM/TSP: HTLV-1-associated myelopathy/tropical spastic paraparesis; [c]IDH: infective dermatitis associated with HTLV-1, [d]TFBS: binding sites for transcription factors; [e]SP1: specificity protein 1.

The possible impacts of mutations present in the genes responsible for encoding structure and accessory proteins were also analyzed. The alteration observed at the 2693 position, located in the region that codifies the p14 protein, resulted in a change in G244D at protein level and caused the abrogation of a myristylation site. This alteration occupied the second position in the rank of importance yielded by the machine learning analyzes for the group of patients with HAM/TSP. In addition, the 6957 mutations, affecting the sequence that codifies the p30 (sense) and the HBZ (antisense) proteins, also caused the abrogation of post-translational modification sites corresponding to an amidation site and a PKC_Phosphorylation site, respectively (Table III).

Table III. Mutations found in genes that encode structural and accessory HTLV-1 proteins and their impacts.

Nucleotide Position	Region	CD[a]	Nucleotide Change	Frequency (%)	Amino acid Change	Machine Learning Rank: Asymptomatic vs. ATLL[b]	vs. HAM[c]	vs. IDH[d]	PTMS[e] Abrogation
1863	gag	p15	A -> C	7.8	Q3P	18	-	-	-
2693	pol	p14	G -> A	39.0	G244D	-	2	-	Myristylation
5366	env	gp46	T -> C	36.4	L55P	22	11	-	-
5394	env	gp46	G -> C	7.8	Q64H	26	-	-	-
5416	env	gp46	A -> G	39.0	S72G	-	7	-	-
5941	env	gp46	G -> A	14.3	V247I	21	-	20	-
6957	pX	HBZ	C -> T	18.2	G90R	19	13	-	Amidation
6957	pX	p30	C -> T	18.2	R37C	19	13	-	PKC_Phosphorylation
6957	pX	p12	C -> T	18.2	P34L	19	13	-	-

7181	pX	HBZ	T -> C	35.1	K169R	24	-	-	-
7974	pX	p40	C -> T	19.5	P136S	20	-	4	-

[a]CD: Coding region

[b]ATLL: adult T-cell leukemia/lymphoma;

[c]HAM/TSP: HTLV-1-associated myelopathy/tropical spastic paraparesis;

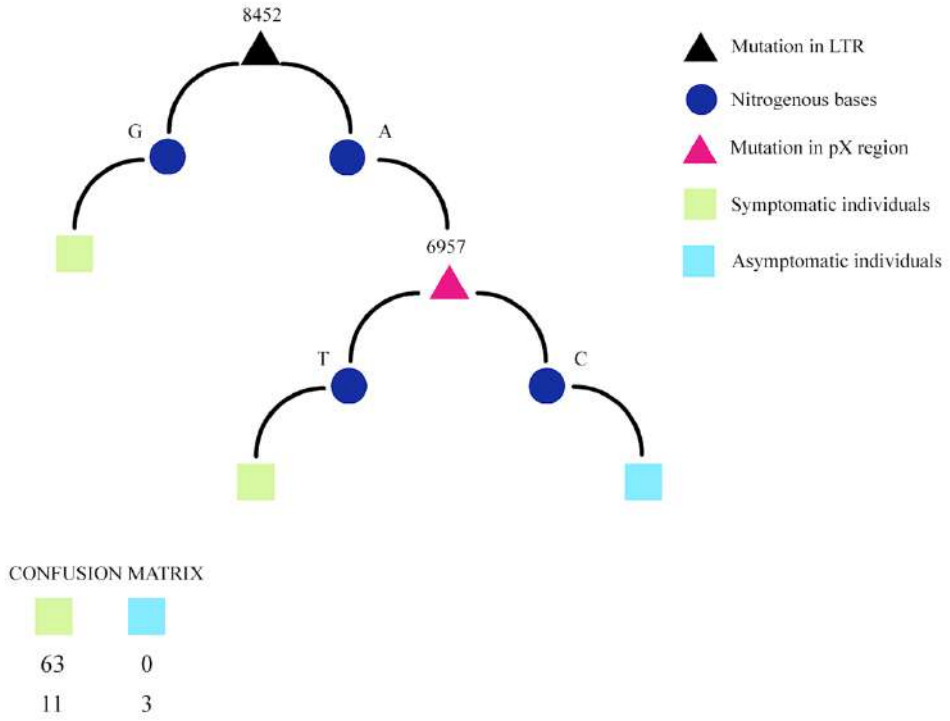
[d]IDH: infective dermatitis associated with HTLV-1.

[e]PTMS: sites of post-translational modifications.

To identify whether the observed amino acid variations were able to alter the physicochemical profile of proteins, the wild-type and mutated sequences were submitted to structural prediction. The evaluated aspects were: hydrophilicity, hydropathy, flexibility, accessibility, membrane buried-helix, and antigenicity. As expected, the NPS@ analysis suggested that all mutations were able to alter protein profiles (Supplementary material section, Figure S1-S8).

Finally, to identify possible correlation between mutations and disease development, we generated a decision tree using the data mining suite Weka. The analysis suggested that the presence of a guanine at 8452 position of the LTR region may be related to the development of symptoms (symptomatic individuals). Similarly, the presence of an adenine at the same LTR position in combination with a thymine at 6957 of the pX region seem to be also associated with symptom development. On the other hand, provirus sequences with this adenine at 8452 position and cytosine at 6957 position resulted in asymptomatic individuals. Although the tree is 85.7% accurate, the confusion matrix showed a high error rate in the classification of asymptomatic patients (Figure 3).

Figure 3. Analysis of interactions among mutations in LTR and pX regions, and clinical condition.



4. Discussion

The HTLV-1 infection has a heterogeneous distribution, affecting 5 to 10 million people worldwide (Gessain and Cassar, 2012). Despite that, this infection is a neglected public health condition, and some questions remain unanswered, such as a full understanding of the variations in the virus genome and their relationship with infection outcomes. Here, we analyzed 242 HTLV-1 complete genome sequences and performed a molecular characterization in 77 subtype 1a subgroup A sequences.

It is known that the provirus transcription is regulated by factors such as the specificity of the protein 1 factor (SP1) that connects to specific sites (Fauquenoy et al., 2017). Some mutations found in the HTLV-1 LTR region (A125G, G174A, and G8452A) caused the deletion of these sites and this may be related to alterations of the HTLV-1 transcription genes. In addition, the machine learning analyzes suggest that the variation at 8452 position, which results in SP1 site abrogation, may play a role in the infection outcome, and this may be related

to the HBZ protein function. The HBZ protein is transcribed in the antisense direction by a promoter present in the 3' LTR and is constantly expressed by HTLV-1 infected cells (Baratella et al., 2017). This constant expression is related to inducing the proliferation of infected cells and inhibiting apoptosis, thus contributing to cell immortalization, maintenance, and multiplication (Mitobe et al., 2015; Satou et al., 2006). Previous studies have shown that some sites for SP1 binding are critical for HBZ activity and to positively regulate antisense LTR (Yoshida et. al, 2008; Fauquenoy et. al, 2017).

It is important to note that the hbz expression also correlates to the Tax protein (Gaudray et al., 2002). The mutation described here may act in the negative regulation of the expression of the hbz gene, in the absence of the Tax protein, suggesting its action during viral latency (Okumura et al., 1996). One study demonstrated a possible correlation between the development of ATLL, reduced expression of Tax (genetics), and epigenetic changes (Takeda et al., 2004). In fact, the mutation at position 8452 of the 3' LTR region appears in the machine learning rank for the IDH and ATLL classification, and was also reported as a mutation with a possible correlation with a variation at position 6957 of the pX region, important for the clinical manifestation. Another nucleotide change, at position 677, caused the creation of a TFBS for binding factor c-Myb, considered as a proto-oncogene (Lipsick and Wang, 1999). Studies have shown that the Tax protein inhibits c-Myb-dependent transcription through activation of the NF-κB pathway (Nicot et al., 2001).

In addition to the control of viral transcription, there are mechanisms for cell regulation used by some pathogens (e.g. HTLV-1), such as post-translational modifications that are characterized by chemical modifications in post-translational modification sites (PTMS) located in the amino acid sequence of the protein (Dissinger et al., 2014). Some mutations in coding regions caused the abrogation of PTMS and, therefore, may have led to changes in the structure of protein products and, consequently, in the physicochemical profile.

Mutations G90R (HBZ) and G244D (p14) generated an abrogation of amidation and myristylation sites, respectively, which could cause structural changes. In addition, the exchange of an apolar amino acid (glycine) for a positively charged amino acid (arginine) or negatively charged one (aspartate) can directly influence the structure and function of the protein. The NPS@ analysis suggests that this exchange results in proteins with characteristics of greater hydrophilicity and antigenicity than the wild sequence, but with less hydropathy. In

the analysis of machine learning, the mutation found in protein p14 showed great importance (ranked second) in the clinical manifestation of HAM/TSP, when compared to asymptomatic carriers. In contrast, the change in amino acid R37C and in protein p30 excluded a site that would receive a phosphate group by the phosphokinase C enzyme, and generated a more hydrophilic protein product with less hydropathy and no changes in antigenicity.

As well as the HBZ, p14, and p30 variations, the mutations found in the env region seem to result in protein alteration. The S72G and V247I mutations (gp46) have been previously described and our data corroborate that changes at position 247 appear to influence the physicochemical profile of the protein generated, making it less antigenic when compared to the wild-type sequence. Besides being identified in individuals with high proviral load and appearing to cause changes in the secondary structure of the protein (Mota-Miranda et al., 2013), the S72G mutation was also ranked seventh in importance (machine learning rank) in the HAM/TSP clinical profile.

The analyzes of the physicochemical profile of the proteins generated also demonstrated changes in the p12 protein caused by the presence of the P34L mutation. Although the mutation influenced the generation of a less hydrophilic and antigenic protein, the hydropathy increased. This mutation has been described in other studies and also appeared in the decision tree as an important mutation in the correlation between mutations, individuals with high proviral load, and the development of HAM/TSP (Barreto et al., 2016; Borba et al., 2019). It is important to highlight that in these studies, some other ORF-I amino acid changes that here did not show a frequency higher than 5% and, therefore was not able to be selected and participate in the next analyzes, were analyzed. Barreto et al. (2016) performed the characterization of fifteen mutations that influence the expression profile of the ORF-I protein product, as previously described by Pise-Masison et al. (2014), while Borba et al. (2019) analyzed the genetic diversity of 32 ORF-I sequences and found seven amino acid changes with frequency over 5%, including P34L mutation. In this sense, we emphasize that this mutation needs to be better evaluated and tested as a biomarker for HTLV-1 progression or as a possible target for the development of a therapeutic vaccine.

Although the decision tree shows possible correlations between mutations in LTR and pX regions with the development of a symptomatic clinical profile, the analyzes of mutations in the asymptomatic group has a relevant error rate (Figure 3). This data can lead us to question

whether individuals classified as asymptomatic carriers are truly asymptomatic or have nonspecific symptoms and, consequently, the decision tree recognizes them as sick. Despite this, it demonstrates that the presence of mutations in the provirus alone is not sufficient to determine a clinical manifestation.

Individuals infected with HTLV-1 still remain without an effective therapeutic regimen and, in view of this scenario, it is necessary to emphasize the importance of not only generating new viral sequences, but also making available all clinical and epidemiological data about the sample. These data can contribute to studies like the present one which attempts to identify the factors that influence the development of pathologies associated with the virus. As we can see in Table I, of the 242 complete virus genome sequences available in GenBank, only 91 contained information about the patient's clinical profile. The information is even more scarce when it refers to other clinical data, such as proviral load, age, gender, and the different clinical forms of ATLL. The absence of complete data results in a small data set, impacting our machine learning analysis.

Still regarding these factors, the clonality of the infected cells is relevant in HTLV-1 infections, since the clonal profile of ATLL patients who have a predominance of a clone, is different from patients with IDH, HAM/TSP as well as asymptomatic carriers who have a polyclonal profile (Bangham et al., 2014; Wattel et al., 1995). The sequencing methodologies presented in table I do not consider the profile of the clonality involved, due to the high fragmentation of the proviral DNA and, consequently, an assembly of reads from different mixed clones. In addition, the use of conventional primers generates sequences that are not truly complete genomes because they do not have the start part of the 5' end and the end part of the 3' end of the promoter region. It is a limitation that can be overcome by encouraging the use of other techniques that promote increased sensitivity in the detection of proviral sequences (Katsuya et al., 2019).

In addition to the low number of full HTLV-1 genome sequences available, there are even fewer sequences of subtypes 1b and 1c than 1a, as we can see in Figure 2. A dataset consisting of a variety of sequences from each subtype would generate more complete results. A greater variety of sequences from the other subtypes would be extremely relevant for studies like this, mainly due to the low genetic diversity among them. These data provide us with advantageous information for the development of vaccines and drugs.

5. Conclusions

The present study found low genetic diversity among HTLV-1 subtypes and possible correlations between mutations and disease development. Some mutations in the viral genome lead to changes in protein physicochemical profile and may be related to clinical manifestations. At the same time, our analysis emphasizes the importance of expanding the identification of other factors, such as those related to the host. In addition, we highlight the importance of providing clinical and epidemiological data with viral genome sequences during submission to the databases. We believe that these data can be used for the development of specific therapies and prophylactic vaccines to fight HTLV-1.

Funding

This research was funded by Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq), N. 421342/2018-8 and Fundação de Amparo à Pesquisa do Estado da Bahia **(**FAPESB), N. BOL0157/2019.

Conflicts of Interest

The authors declare no conflict of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript; or in the decision to publish the results.

Acknowledgment

We would like to thank Daisy Glass for revising the English text.

CRediT authorship contribution statement

Melina M. N. Borba: Conceptualization, Methodology, Formal Analysis, Investigation, Writing – Original Draft and Visualization. **Luciane A. Santos:** Methodology, Data Curation and Writing – Review & Editing. **Reinaldo C. Neto:** Formal Analysis, Writing – Review & Editing and Visualization. **Felipe de O. Andrade:** Formal Analysis, Writing – Review & Editing and Visualization. **Álvaro Salgado:** Methodology, Data Curation and Writing – Review & Editing. **Filipe Ferreira de A. Rego:** Methodology, Formal Analysis, Writing –

Review & Editing. **Lourdes Farre:** Conceptualization and Writing – Review & Editing. **Fernanda Khouri Barreto:** Conceptualization, Methodology, Formal Analysis, Writing – Review & Editing and Supervision.

References

Araujo, A., Silva, M., 2006. The HTLV-1 neurological complex. Lancet Neurol 5, 1068–1076. https://doi.org/10.1016/S1474-4422(06)70628-7

Arnold, J., Yamamoto, B., Li, M., Phipps, A.J., Younis, I., Lairmore, M.D., Green, P.L., 2006. Enhancement of infectivity and persistence in vivo by HBZ, a natural antisense coded protein of HTLV-1. Blood 107, 3976–3982. https://doi.org/10.1182/blood-2005-11-4551.

Bangham, C.R.M., Cook, L.B., Melamed, A., 2014. HTLV-1 clonality in adult T-cell leukaemia and non-malignant HTLV-1 infection. Semin Cancer Biol 26, 89–98. https://doi.org/10.1016/j.semcancer.2013.11.003

Baratella, M., Forlani, G., Accolla, R.S., 2017. HTLV-1 HBZ Viral Protein: A Key Player in HTLV-1 Mediated Diseases. Frontiers in Microbiology 8, 2615. https://doi.org/10.3389/fmicb.2017.02615

Barreto, F.K., Araújo, T.H., Rego, F.F.A., Alcantara, L.C., 2017. A Fully Annotated Genome Sequence of Human T-Cell Lymphotropic Virus Type 1 (HTLV-1) 1, 3.

Barreto, F.K., Khouri, R., de Almeida Rego, F.F., Santos, L.A., de Castro-Amarante, M.F., Bialuk, I., Pise-Masison, C.A., Galvão-Castro, B., Gessain, A., Jacobson, S., Franchini, G., Alcantara, L.C., 2016. Analyses of HTLV-1 sequences suggest interaction between ORF-I mutations and HAM/TSP outcome. Infect Genet Evol 45, 420–425. https://doi.org/10.1016/j.meegid.2016.08.020

Blanchard, S., Astier-Gin, T., Tallet, B., Moynet, D., Londos-Gagliardi, D., Guillemain, B., 1999. Amino Acid Changes at Positions 173 and 187 in the Human T-Cell Leukemia Virus Type 1 Surface Glycoprotein Induce Specific Neutralizing Antibodies. J Virol 73, 9369–9376.

Borba, M.M.N., Farre, L., Bittencourt, A.L., Castro-Amarante, M.F. de, Galvão-Castro, B., Santos, L.A., Araújo, T.H.A., Alcantara, L.C.J., Barreto, F.K., 2019. Assessment of Genetic Diversity of HTLV-1 ORF-I Sequences Collected from Patients with Different

Clinical Profiles. AIDS Res. Hum. Retroviruses 35, 881–884. https://doi.org/10.1089/AID.2019.0127

Breiman, L., 2001. Random Forests. Machine Learning 45, 5–32. https://doi.org/10.1023/A:1010933404324

Chen, J., Zekeng, L., Yamashita, M., Takehisa, J., Miura, T., Ido, E., Mboudjeka, I., Tsague, J.-M., Hayami, M., Kaptue, L., 1995. HTLV Type I Isolated from a Pygmy in Cameroon Is Related to but Distinct from the Known Central African Type. AIDS Research and Human Retroviruses 11, 1529–1531. https://doi.org/10.1089/aid.1995.11.1529

Chen, T., Guestrin, C., 2016. XGBoost: A Scalable Tree Boosting System, in: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. Presented at the KDD '16: The 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, ACM, San Francisco California USA, pp. 785–794. https://doi.org/10.1145/2939672.2939785

Combet, C., Blanchet, C., Geourjon, C., Deléage, G., 2000. NPS@: network protein sequence analysis. Trends Biochem. Sci. 25, 147–150. https://doi.org/10.1016/s0968-0004(99)01540-6

Dissinger, N., Shkriabai, N., Hess, S., Al-Saleem, J., Kvaratskhelia, M., Green, P.L., 2014. Identification and Characterization of HTLV-1 HBZ Post-Translational Modifications. PLoS One 9. https://doi.org/10.1371/journal.pone.0112762

Ehrlich, G.D., Andrews, J., Sherman, M.P., Greenberg, S.J., Poiesz, B.J., 1992. DNA sequence analysis of the gene encoding the HTLV-I p21e transmembrane protein reveals inter- and intraisolate genetic heterogeneity. Virology 186, 619–627. https://doi.org/10.1016/0042-6822(92)90028-n

Fauquenoy, S., Robette, G., Kula, A., Vanhulle, C., Bouchat, S., Delacourt, N., Rodari, A., Marban, C., Schwartz, C., Burny, A., Rohr, O., Van Driessche, B., Van Lint, C., 2017. Repression of Human T-lymphotropic virus type 1 Long Terminal Repeat sense transcription by Sp1 recruitment to novel Sp1 binding sites. Scientific Reports 7, 43221. https://doi.org/10.1038/srep43221

Friedman, J., Hastie, T., Tibshirani, R., 2000. Additive logistic regression: a statistical view of boosting (With discussion and a rejoinder by the authors). Annals of Statistics, v. 28, n. 2, p. 337–407.

Gaudray, G., Gachon, F., Basbous, J., Biard-Piechaczyk, M., Devaux, C., Mesnard, J.-M., 2002. The complementary strand of the human T-cell leukemia virus type 1 RNA genome encodes a bZIP transcription factor that down-regulates viral transcription. J. Virol. 76, 12813–12822. https://doi.org/10.1128/jvi.76.24.12813-12822.2002

Gessain, A., Barin, F., Vernant, J.C., Gout, O., Maurs, L., Calender, A., de Thé, G., 1985. Antibodies to human T-lymphotropic virus type-I in patients with tropical spastic paraparesis. Lancet 2, 407–410. https://doi.org/10.1016/s0140-6736(85)92734-5

Gessain, A., Cassar, O., 2012. Epidemiological Aspects and World Distribution of HTLV-1 Infection. Front Microbiol 3. https://doi.org/10.3389/fmicb.2012.00388

Gessain, A., Gallo, R.C., Franchini, G., 1992. Low degree of human T-cell leukemia/lymphoma virus type I genetic drift in vivo as a means of monitoring viral transmission and movement of ancient human populations. J Virol 66, 2288–2295.

Gessain, A., Yanagihara, R., Franchini, G., Garruto, R.M., Jenkins, C.L., Ajdukiewicz, A.B., Gallo, R.C., Gajdusek, D.C., 1991. Highly Divergent Molecular Variants of Human T-Lymphotropic Virus Type I from Isolated Populations in Papua New Guinea and the Solomon Islands. Proceedings of the National Academy of Sciences of the United States of America 88, 7694–7698.

Grenade, L.L., Manns, A., Fletcher, V., Carberry, C., Hanchard, B., Maloney, E.M., Cranston, B., Williams, N.P., Wilks, R., Kang, E.C., Blattner, W.A., 1998. Clinical, Pathologic, and Immunologic Features of Human T-Lymphotrophic Virus Type I–Associated Infective Dermatitis in Children. Arch Dermatol 134, 439–444. https://doi.org/10.1001/archderm.134.4.439

Hahn, B.H., Shaw, G.M., Popovic, M., Lo Monico, A., Gallo, R.C., Wong-Staal, F., 1984. Molecular cloning and analysis of a new variant of human T-cell leukemia virus (HTLV-ib) from an African patient with adult T-cell leukemia-lymphoma. Int. J. Cancer 34, 613–618. https://doi.org/10.1002/ijc.2910340505

Katsuya, H., Islam, S., Tan, B.J.Y., Ito, J., Miyazato, P., Matsuo, M., Inada, Y., Iwase, S.C., Uchiyama, Y., Hata, H., Sato, T., Yagishita, N., Araya, N., Ueno, T., Nosaka, K., Tokunaga, M., Yamagishi, M., Watanabe, T., Uchimaru, K., Fujisawa, J.-I., Utsunomiya, A., Yamano, Y., Satou, Y., 2019. The Nature of the HTLV-1 Provirus in

Naturally Infected Individuals Analyzed by the Viral DNA-Capture-Seq Approach. Cell Rep 29, 724-735.e4. https://doi.org/10.1016/j.celrep.2019.09.016

Kearse, M., Moir, R., Wilson, A., Stones-Havas, S., Cheung, M., Sturrock, S., Buxton, S., Cooper, A., Markowitz, S., Duran, C., Thierer, T., Ashton, B., Meintjes, P., Drummond, A., 2012. Geneious Basic: an integrated and extendable desktop software platform for the organization and analysis of sequence data. Bioinformatics 28, 1647–1649. https://doi.org/10.1093/bioinformatics/bts199

Kumar, S., Stecher, G., Li, M., Knyaz, C., Tamura, K., 2018. MEGA X: Molecular Evolutionary Genetics Analysis across Computing Platforms. Mol Biol Evol 35, 1547–1549. https://doi.org/10.1093/molbev/msy096

Lipsick, J.S., Wang, D.M., 1999. Transformation by v-Myb. Oncogene 18, 3047–3055. https://doi.org/10.1038/sj.onc.1202745

Lundberg, S.M., Lee, S.-I., 2017. A Unified Approach to Interpreting Model Predictions, in: Guyon, I., Luxburg, U.V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., Garnett, R. (Eds.), Advances in Neural Information Processing Systems 30. Curran Associates, Inc., pp. 4765–4774.

Matsuura, E., Yamano, Y., Jacobson, S., 2010. Neuroimmunity of HTLV-I Infection. J Neuroimmune Pharmacol 5, 310–325. https://doi.org/10.1007/s11481-010-9216-9.

Mitobe, Y., Yasunaga, J., Furuta, R., Matsuoka, M., 2015. HTLV-1 bZIP Factor RNA and Protein Impart Distinct Functions on T-cell Proliferation and Survival. Cancer Res. 75, 4143–4152. https://doi.org/10.1158/0008-5472.CAN-15-0942

Miura, T., Fukunaga, T., Igarashi, T., Yamashita, M., Ido, E., Funahashi, S., Ishida, T., Washio, K., Ueda, S., Hashimoto, K., 1994. Phylogenetic subtypes of human T-lymphotropic virus type I and their relations to the anthropological background. Proc Natl Acad Sci U S A 91, 1124–1127.

Mota-Miranda, A.C.A., Barreto, F.K., Amarante, M.F.C., Batista, E., Monteiro-Cunha, J.P., Farre, L., Galvão-Castro, B., Alcantara, L.C.J., 2013. Molecular characterization of HTLV-1 gp46 glycoprotein from health carriers and HAM/TSP infected individuals. Virol J 10, 75. https://doi.org/10.1186/1743-422X-10-75

Neto, W.K., Da-Costa, A.C., de Oliveira, A.C.S., Martinez, V.P., Nukui, Y., Sabino, E.C., Sanabani, S.S., 2011. Correlation between LTR point mutations and proviral load levels

among human T cell lymphotropic virus type 1 (HTLV-1) asymptomatic carriers. Virology Journal 8, 535. https://doi.org/10.1186/1743-422X-8-535

Nicot, C., Mahieux, R., Pise-Masison, C., Brady, J., Gessain, A., Yamaoka, S., Franchini, G., 2001. Human T-Cell Lymphotropic Virus Type 1 Tax Represses c-Myb-Dependent Transcription through Activation of the NF-κB Pathway and Modulation of Coactivator Usage. Mol Cell Biol 21, 7391–7402. https://doi.org/10.1128/MCB.21.21.7391-7402.2001

Okumura, K., Sakaguchi, G., Takagi, S., Naito, K., Mimori, T., Igarashi, H., 1996. Sp1 family proteins recognize the U5 repressive element of the long terminal repeat of human T cell leukemia virus type I through binding to the CACCC core motif. J. Biol. Chem. 271, 12944–12950. https://doi.org/10.1074/jbc.271.22.12944

Pise-Masison, C.A., de Castro-Amarante, M.F., Enose-Akahata, Y., Buchmann, R.C., Fenizia, C., Washington Parks, R., Edwards, D., Fiocchi, M., Alcantara, L.C., Bialuk, I., Graham, J., Walser, J.-C., McKinnon, K., Galvão-Castro, B., Gessain, A., Venzon, D., Jacobson, S., Franchini, G., 2014. Co-dependence of HTLV-1 p12 and p8 Functions in Virus Persistence. PLoS Pathog 10. https://doi.org/10.1371/journal.ppat.1004454

Poiesz, B.J., Ruscetti, F.W., Gazdar, A.F., Bunn, P.A., Minna, J.D., Gallo, R.C., 1980. Detection and isolation of type C retrovirus particles from fresh and cultured lymphocytes of a patient with cutaneous T-cell lymphoma. Proc Natl Acad Sci U S A 77, 7415–7419.

Saito, M., Jain, P., Tsukasaki, K., Bangham, C.R.M., 2012. HTLV-1 Infection and Its Associated Diseases [WWW Document]. Leukemia Research and Treatment. https://doi.org/10.1155/2012/123637

Salemi, M., Van Dooren, S., Audenaert, E., Delaporte, E., Goubau, P., Desmyter, J., Vandamme, A.-M., 1998. Two New Human T-Lymphotropic Virus Type I Phylogenetic Subtypes in Seroindeterminates, a Mbuti Pygmy and a Gabonese, Have Closest Relatives among African STLV-I Strains. Virology 246, 277–287. https://doi.org/10.1006/viro.1998.9215

Satou, Y., Yasunaga, J., Yoshida, M., Matsuoka, M., 2006. HTLV-I basic leucine zipper factor gene mRNA supports proliferation of adult T cell leukemia cells. Proc Natl Acad Sci U S A 103, 720–725. https://doi.org/10.1073/pnas.0507631103

Schapire, R.E., 2003. The Boosting Approach to Machine Learning: An Overview, in: Denison, D.D., Hansen, M.H., Holmes, C.C., Mallick, B., Yu, B. (Eds.), Nonlinear Estimation and Classification, Lecture Notes in Statistics. Springer New York, New York, NY, pp. 149–171. https://doi.org/10.1007/978-0-387-21579-2_9

Sigrist, C.J.A., de Castro, E., Cerutti, L., Cuche, B.A., Hulo, N., Bridge, A., Bougueleret, L., Xenarios, I., 2013. New and continuing developments at PROSITE. Nucleic Acids Res 41, D344–D347. https://doi.org/10.1093/nar/gks1067

Socorro de Almeida Viana, M. de N. do, Santos Nobre, A.F., Costa Jr, E., Silva, I.C., Pinheiro, B.T., Pereira, C.C.C., de Souza Canto Ferreira, L., de Almeida, D.S., de Araújo, M.W.L., da Silva Borges, M., da Costa, C.A., Ishikawa, E.A.Y., Ferrari, S.F., Silva de Sousa, M., 2018. Stability of the HTLV-1 glycoprotein 46 (gp46) gene in an endemic region of the Brazilian Amazon and the presence of a significant mutation (N93D) in symptomatic patients. Virol J 15. https://doi.org/10.1186/s12985-018-0984-9

Takeda, S., Maeda, M., Morikawa, S., Taniguchi, Y., Yasunaga, J., Nosaka, K., Tanaka, Y., Matsuoka, M., 2004. Genetic and epigenetic inactivation of tax gene in adult T-cell leukemia cells. International Journal of Cancer 109, 559–567. https://doi.org/10.1002/ijc.20007

Trifinopoulos, J., Nguyen, L.-T., von Haeseler, A., Minh, B.Q., 2016. W-IQ-TREE: a fast online phylogenetic tool for maximum likelihood analysis. Nucleic Acids Res 44, W232–W235. https://doi.org/10.1093/nar/gkw256

Wattel, E., Vartanian, J.P., Pannetier, C., Wain-Hobson, S., 1995. Clonal expansion of human T-cell leukemia virus type I-infected cells in asymptomatic and symptomatic carriers without malignancy. J Virol 69, 2863–2868.

Witten, I.H., Frank, E., 2002. Data mining: practical machine learning tools and techniques with Java implementations. SIGMOD Rec. 31, 76–77. https://doi.org/10.1145/507338.507355

Wolfe, N.D., Heneine, W., Carr, J.K., Garcia, A.D., Shanmugam, V., Tamoufe, U., Torimiro, J.N., Prosser, A.T., LeBreton, M., Mpoudi-Ngole, E., McCutchan, F.E., Birx, D.L., Folks, T.M., Burke, D.S., Switzer, W.M., 2005. Emergence of unique primate T-lymphotropic viruses among central African bushmeat hunters. Proc Natl Acad Sci U S A 102, 7994–7999. https://doi.org/10.1073/pnas.0501734102

Yoshida, M., Miyoshi, I., Hinuma, Y., 1982. Isolation and characterization of retrovirus from cell lines of human adult T-cell leukemia and its implication in the disease. Proc. Natl. Acad. Sci. U.S.A. 79, 2031–2035. https://doi.org/10.1073/pnas.79.6.2031.

Yoshida, M., Satou, Y., Yasunaga, J.-I., Fujisawa, J.-I., Matsuoka, M., 2008. Transcriptional control of spliced and unspliced human T-cell leukemia virus type 1 bZIP factor (HBZ) gene. J Virol 82, 9359–9368. https://doi.org/10.1128/JVI.00242-08

Supplementary Materials:

Table S1: Clinical and epidemiological data from HTLV-1 sequences collected on GenBank

Accession Number	Geographic Location	Subtype	Clinical Profile	Age	Gender	Sequencing Methodology
LC192505	Japan	1aA	Asymptomatic	-	-	Miseq
LC192504	Japan	1aA	Asymptomatic	-	-	Miseq
LC192503	Japan	1aA	Asymptomatic	-	-	Miseq
LC192502	Japan	1aA	Asymptomatic	-	-	Miseq
LC192501	Japan	1aB	Asymptomatic	-	-	Miseq
LC192500	Japan	1aA	Asymptomatic	-	-	Miseq
KY007273	Brazil	1aA	Asymptomatic	36	Female	Ion Torrent
KY007271	Brazil	1aA	Asymptomatic	59	Female	Ion Torrent
KY007269	Brazil	1aA	Asymptomatic	52	Male	Ion Torrent
KY007262	Brazil	1aA	Asymptomatic	46	Male	Ion Torrent
KY007260	Brazil	1aA	Asymptomatic	50	Female	Ion Torrent
KY007258	Brazil	1aA	Asymptomatic	45	Female	Ion Torrent
KY007257	Brazil	1aA	Asymptomatic	44	Female	Ion Torrent
KY007256	Brazil	1aA	Asymptomatic	46	Female	Ion Torrent
KY007250	Brazil	1aA	Asymptomatic	50	Female	Ion Torrent
MN781156	Guinea	1aD	Asymptomatic	52	Female	Sanger
MN781155	Africa	1aD	Asymptomatic	29	Male	Sanger
KC807984	China	1aA	ATLL	-	-	-
D13784	Central America	1aC	ATLL	-	-	-
KY007251	Brazil	1aA	ATLL	46	Female	Ion Torrent
KY007249	Brazil	1aA	ATLL	20	Male	Ion Torrent
KY007248	Brazil	1aA	ATLL	52	Male	Ion Torrent
KY007247	Brazil	1aA	ATLL	34	Female	Ion Torrent
KY007246	Brazil	1aA	ATLL	40	Female	Ion Torrent

KY007245	Brazil	1aA	ATLL	42	Male	Ion Torrent
KY007244	Brazil	1aA	ATLL	35	Male	Ion Torrent
AB513134	Japan	1aB	ATLL	-	-	-
J02029	Japan	1aB	ATLL	-	-	Sanger
MH399767	Iran	1aA	ATLL	-	-	Sanger
LC484862	-	1aB	ATLL	-	-	-
MH399769	Iran	1aA	ATLL	-	-	Sanger
LC378575	Japan	1aA	ATLL	-	-	Sanger
LC183873	Japan	1aA	ATLL	-	-	Sanger
AF042071	Germany	1aA	ATLL	-	-	-
MH399768	Iran	1aA	ATLL	-	-	Sanger
MH395864	Iran	1aA	ATLL	-	-	Sanger
MH392265	Iran	1aA	ATLL	-	-	Sanger
MN781154	Africa	1aC	ATLL	45	Male	Sanger
MN781153	Africa	1aD	ATLL	54	Male	Sanger
MN781152	Cote d'Ivoire - Africa	1aC	ATLL	59	Male	Sanger
MN781150	Mauritania - Africa	1aD	ATLL	25	Male	Sanger
MN781149	Mali - Africa	1aD	ATLL	45	Male	Sanger
U19949	Japan	1aB	ATLL	-	-	Sanger
KY007261	Brazil	1aA	DIH	12	Female	Ion Torrent
KY007255	Brazil	1aA	DIH	14	Male	Ion Torrent
KY007254	Brazil	1aA	DIH	19	Female	Ion Torrent
KY007253	Brazil	1aA	DIH	10	Male	Ion Torrent
KY007252	Brazil	1aA	DIH	21	Male	Ion Torrent
LC192536	Japan	1aA	HAM/TSP	-	-	Miseq
LC192535	Japan	1aA	HAM/TSP	-	-	Miseq

LC192534	Japan	1aA	HAM/TSP	-	-	Miseq
LC192533	Japan	1aA	HAM/TSP	-	-	Miseq
LC192532	Japan	1aA	HAM/TSP	-	-	Miseq
LC192531	Japan	1aA	HAM/TSP	-	-	Miseq
LC192530	Japan	1aA	HAM/TSP	-	-	Miseq
LC192529	Japan	1aA	HAM/TSP	-	-	Miseq
LC192528	Japan	1aA	HAM/TSP	-	-	Miseq
LC192527	Japan	1aA	HAM/TSP	-	-	Miseq
LC192526	Japan	1aA	HAM/TSP	-	-	Miseq
LC192525	Japan	1aA	HAM/TSP	-	-	Miseq
LC192524	Japan	1aA	HAM/TSP	-	-	Miseq
LC192523	Japan	1aA	HAM/TSP	-	-	Miseq
LC192522	Japan	1aA	HAM/TSP	-	-	Miseq
LC192521	Japan	1aA	HAM/TSP	-	-	Miseq
LC192520	Japan	1aA	HAM/TSP	-	-	Miseq
LC192519	Japan	1aA	HAM/TSP	-	-	Miseq
LC192518	Japan	1aA	HAM/TSP	-	-	Miseq
LC192517	Japan	1aA	HAM/TSP	-	-	Miseq
LC192516	Japan	1aA	HAM/TSP	-	-	Miseq
LC192515	Japan	1aA	HAM/TSP	-	-	Miseq
LC192514	Japan	1aA	HAM/TSP	-	-	Miseq
LC192513	Japan	1aA	HAM/TSP	-	-	Miseq
LC192612	Japan	1aA	HAM/TSP	-	-	Miseq
LC192511	Japan	1aA	HAM/TSP	-	-	Miseq
LC192510	Japan	1aA	HAM/TSP	-	-	Miseq
LC192509	Japan	1aA	HAM/TSP	-	-	Miseq
LC192508	Japan	1aA	HAM/TSP	-	-	Miseq
LC192507	Japan	1aA	HAM/TSP	-	-	Miseq
LC192506	Japan	1aA	HAM/TSP	-	-	Miseq
KY007274	Brazil	1aA	HAM/TSP	52	Female	Ion Torrent

KY007272	Brazil	1aA	HAM/TSP	59	Female	Ion Torrent
KY007270	Brazil	1aA	HAM/TSP	42	Female	Ion Torrent
KY007268	Brazil	1aA	HAM/TSP	38	Female	Ion Torrent
KY007267	Brazil	1aA	HAM/TSP	51	Female	Ion Torrent
KY007266	Brazil	1aA	HAM/TSP	51	Male	Ion Torrent
KY007265	Brazil	1aA	HAM/TSP	64	Male	Ion Torrent
KY007264	Brazil	1aA	HAM/TSP	67	Male	Ion Torrent
KY007263	Brazil	1aA	HAM/TSP	50	Female	Ion Torrent
KY007259	Brazil	1aA	HAM/TSP	50	Female	Ion Torrent
M86840	Caribbean	1aB	HAM/TSP	-	-	-
L36905	France	1aA	HAM/TSP	-	-	Sanger
KF242506	Australia	1C	-	29	Female	Sanger
KF242505	Australia	1C	-	60	Female	Sanger
JX891479	Australia	1C	-	59	Female	Sanger
JX891478	Australia	1C	-	67	Male	Sanger
JX507077	Brazil	1b	-	-	-	Sanger
HQ606138	Canada	1aA	-	-	-	-
AY563954	Brazil	1aA	-	-	-	Sanger
AY563953	Brazil	1aA	-	-	-	Sanger
KX905203	Australia	1C	-	73	Female	Sanger
KX905202	Vanuatu	1C	-	76	Female	Sanger
LC183874	Japan	1aB	-	-	-	-
LC183872	Japan	1aA	-	-	-	-
LC210071	Japan	1aB	-	-	-	Sanger
LC210070	Japan	1aB	-	-	-	Sanger
LC210069	Japan	1aB	-	-	-	Sanger
LC210068	Japan	1aB	-	-	-	Sanger
LC210067	Japan	1aA	-	-	-	Sanger
LC210066	Japan	1aB	-	-	-	Sanger
LC210065	Japan	1aB	-	-	-	Sanger

LC210064	Japan	1aA	-	-	-	Sanger
LC210063	Japan	1aB	-	-	-	Sanger
LC210062	Japan	1aB	-	-	-	Sanger
LC210061	Japan	1aB	-	-	-	Sanger
LC210060	Japan	1aB	-	-	-	Sanger
LC210059	Japan	1aB	-	-	-	Sanger
LC210058	Japan	1aB	-	-	-	Sanger
LC210057	Japan	1aB	-	-	-	Sanger
LC210056	Japan	1aB	-	-	-	Sanger
LC210055	Japan	1aB	-	-	-	Sanger
LC210054	Japan	1aB	-	-	-	Sanger
LC210053	Japan	1aB	-	-	-	Sanger
LC210052	Japan	1aA	-	-	-	Sanger
LC210051	Japan	1aA	-	-	-	Sanger
LC210050	Japan	1aA	-	-	-	Sanger
LC210049	Japan	1aA	-	-	-	Sanger
LC210048	Japan	1aA	-	-	-	Sanger
LC210047	Japan	1aB	-	-	-	Sanger
LC210046	Japan	1aB	-	-	-	Sanger
LC210045	Japan	1aA	-	-	-	Sanger
LC210044	Japan	1aA	-	-	-	Sanger
LC210043	Japan	1aB	-	-	-	Sanger
LC210042	Japan	1aB	-	-	-	Sanger
LC210041	Japan	1aB	-	-	-	Sanger
LC210040	Japan	1aB	-	-	-	Sanger
LC210039	Japan	1aB	-	-	-	Sanger
LC210038	Japan	1aB	-	-	-	Sanger
LC210037	Japan	1aB	-	-	-	Sanger
LC210036	Japan	1aB	-	-	-	Sanger
LC210035	Japan	1aB	-	-	-	Sanger

LC210034	Japan	1aB	-	-	-	Sanger
LC210033	Japan	1aB	-	-	-	Sanger
LC210032	Japan	1aB	-	-	-	Sanger
LC210031	Japan	1aB	-	-	-	Sanger
LC210030	Japan	1aB	-	-	-	Sanger
LC210029	Japan	1aB	-	-	-	Sanger
LC210028	Japan	1aB	-	-	-	Sanger
LC210027	Japan	1aB	-	-	-	Sanger
LC210026	Japan	1aB	-	-	-	Sanger
LC210025	Japan	1aB	-	-	-	Sanger
LC210024	Japan	1aB	-	-	-	Sanger
LC210023	Japan	1aB	-	-	-	Sanger
LC210022	Japan	1aB	-	-	-	Sanger
LC210021	Japan	1aB	-	-	-	Sanger
LC210020	Japan	1aB	-	-	-	Sanger
LC210019	Japan	1aB	-	-	-	Sanger
LC210018	Japan	1aB	-	-	-	Sanger
LC210017	Japan	1aB	-	-	-	Sanger
LC210016	Japan	1aB	-	-	-	Sanger
LC210015	Japan	1aB	-	-	-	Sanger
LC210014	Japan	1aB	-	-	-	Sanger
LC210013	Japan	1aB	-	-	-	Sanger
LC210012	Japan	1aB	-	-	-	Sanger
LC210011	Japan	1aB	-	-	-	Sanger
LC210010	Japan	1aA	-	-	-	Sanger
LC210009	Japan	1aB	-	-	-	Sanger
LC210008	Japan	1aB	-	-	-	Sanger
LC210007	Japan	1aB	-	-	-	Sanger
LC210006	Japan	1aB	-	-	-	Sanger
LC210005	Japan	1aB	-	-	-	Sanger

LC210004	Japan	1aB	-	-	-	Sanger
LC210003	Japan	1aB	-	-	-	Sanger
LC210002	Japan	1aB	-	-	-	Sanger
LC210001	Japan	1aB	-	-	-	Sanger
LC210000	Japan	1aB	-	-	-	Sanger
LC209999	Japan	1aB	-	-	-	Sanger
LC209998	Japan	1aB	-	-	-	Sanger
LC209997	Japan	1aB	-	-	-	Sanger
LC209996	Japan	1aB	-	-	-	Sanger
LC209995	Japan	1aB	-	-	-	Sanger
LC209994	Japan	1aB	-	-	-	Sanger
LC209993	Japan	1aB	-	-	-	Sanger
LC209992	Japan	1aB	-	-	-	Sanger
LC209991	Japan	1aB	-	-	-	Sanger
LC209990	Japan	1aB	-	-	-	Sanger
LC209989	Japan	1aB	-	-	-	Sanger
LC209988	Japan	1aB	-	-	-	Sanger
LC209987	Japan	1aB	-	-	-	Sanger
LC209986	Japan	1aB	-	-	-	Sanger
LC209985	Japan	1aB	-	-	-	Sanger
LC209984	Japan	1aB	-	-	-	Sanger
LC209983	Japan	1aB	-	-	-	Sanger
LC209982	Japan	1aB	-	-	-	Sanger
LC209981	Japan	1aB	-	-	-	Sanger
LC209980	Japan	1aB	-	-	-	Sanger
LC209979	Japan	1aB	-	-	-	Sanger
LC209978	Japan	1aB	-	-	-	Sanger
LC209977	Japan	1aB	-	-	-	Sanger
LC209976	Japan	1aB	-	-	-	Sanger
LC209975	Japan	1aB	-	-	-	Sanger

LC209974	Japan	1aB	-	-	-	Sanger
LC209973	Japan	1aB	-	-	-	Sanger
LC209972	Japan	1aB	-	-	-	Sanger
LC209971	Japan	1aB	-	-	-	Sanger
LC209970	Japan	1aB	-	-	-	Sanger
LC209969	Japan	1aB	-	-	-	Sanger
LC209968	Japan	1aB	-	-	-	Sanger
LC209967	Japan	1aB	-	-	-	Sanger
LC209966	Japan	1aB	-	-	-	Sanger
LC209965	Japan	1aB	-	-	-	Sanger
LC209964	Japan	1aB	-	-	-	Sanger
LC209963	Japan	1aB	-	-	-	Sanger
LC209962	Japan	1aB	-	-	-	Sanger
LC209961	Japan	1aB	-	-	-	Sanger
LC209960	Japan	1aB	-	-	-	Sanger
LC209959	Japan	1aB	-	-	-	Sanger
LC209958	Japan	1aB	-	-	-	Sanger
LC192264	Japan	1aB	-	-	-	Sanger
LC192263	Japan	1aB	-	-	-	Sanger
LC192262	Japan	1aA	-	-	-	Sanger
LC192261	Japan	1aB	-	-	-	Sanger
LC192260	Japan	1aB	-	-	-	Sanger
LC192259	Japan	1aB	-	-	-	Sanger
LC192258	Japan	1aA	-	-	-	Sanger
LC192257	Japan	1aB	-	-	-	Sanger
LC192256	Japan	1aA	-	-	-	Sanger
LC192255	Japan	1aB	-	-	-	Sanger
LC192254	Japan	1aB	-	-	-	Sanger
LC185242	Japan	1aB	-	18	Female	Sanger
LC185241	Japan	1aA	-	59	Male	Sanger

LC185240	Japan	1aB	-	20	Male	Sanger
LC185239	Japan	1aB	-	48	Female	Sanger
LC185238	Japan	1aB	-	26	Male	Sanger
LC185237	Japan	1aB	-	58	-	Sanger
LC185136	Japan	1aA	-	47	-	Sanger
LC185235	Japan	1aB	-	46	-	Sanger
KX430031	Africa	1aD	-	-	-	Sanger
KX430030	Africa	1aD	-	-	-	Sanger
HQ606137	Canada	1aA	-	-	-	-
AF033817	-	1aC	-	-	-	Sanger
AF259264	China	1aA	-	-	-	Sanger
AF139170	USA	1aA	-	-	Female	Sanger

Table S2: Genetic distance between and among the different HTLV-1 subtypes.

Subtype	**Subtype**					
	1aB	**1aA**	**1aC**	**1aD**	**1b**	**1c**
1aB	0.0034	0.0103	0.0175	0.0179	0.0336	0.0875
1aA	0.0103	0.0066	0.0180	0.0179	0.0350	0.0874
1aC	0.0175	0.0180	0.0027	0.0133	0.0332	0.0852
1aD	0.0179	0.0179	0.0133	0.0089	0.0338	0.0848
1b	0.0336	0.0350	0.0332	0.0338	n/c[a]	0.0865
1c	0.0875	0.0874	0.0852	0.0848	0.0865	0.0108

[a]n/c = no compute

Figure S1: Analysis of the physicochemical profile of the gp46 protein. The analysis suggests that the L55P mutation may generate an important change in hydrophilicity, hydropathy, antigenicity, accessibility, and membrane buried-helix.

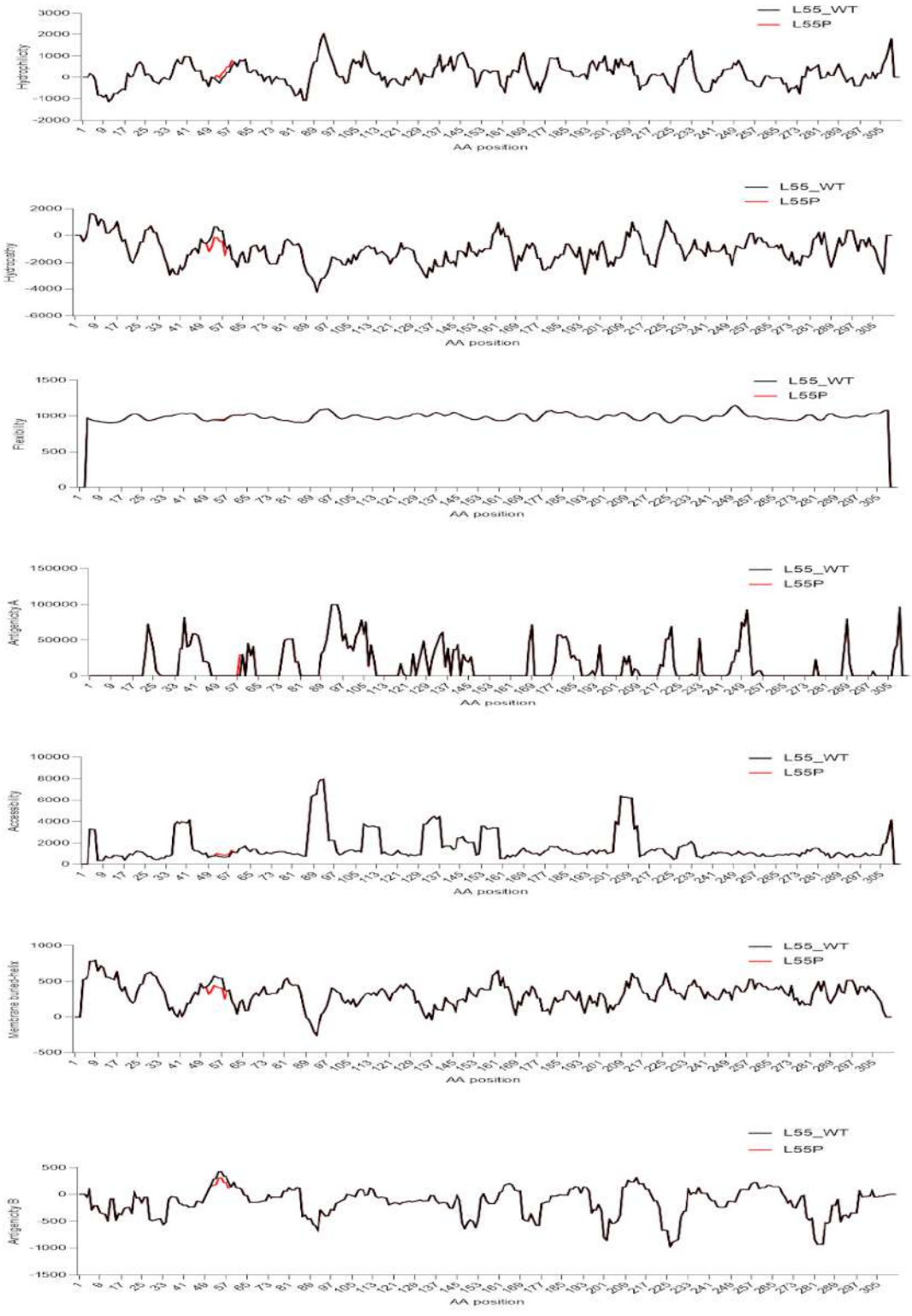

L55_WT
L55P
Hydrophilicity
Hydropathy
Flexibility
Antigenicity A
Accessibility
Membrane buried-helix
Antigenicity B
AA position

Figure S2: Analysis of the physicochemical profile of the gp46 protein. The analysis suggests that the Q64H and S72G mutations may generate an important change in antigenicity, accessibility, and membrane buried-helix. It also suggests that the V247I mutation may generate an important change in antigenicity.

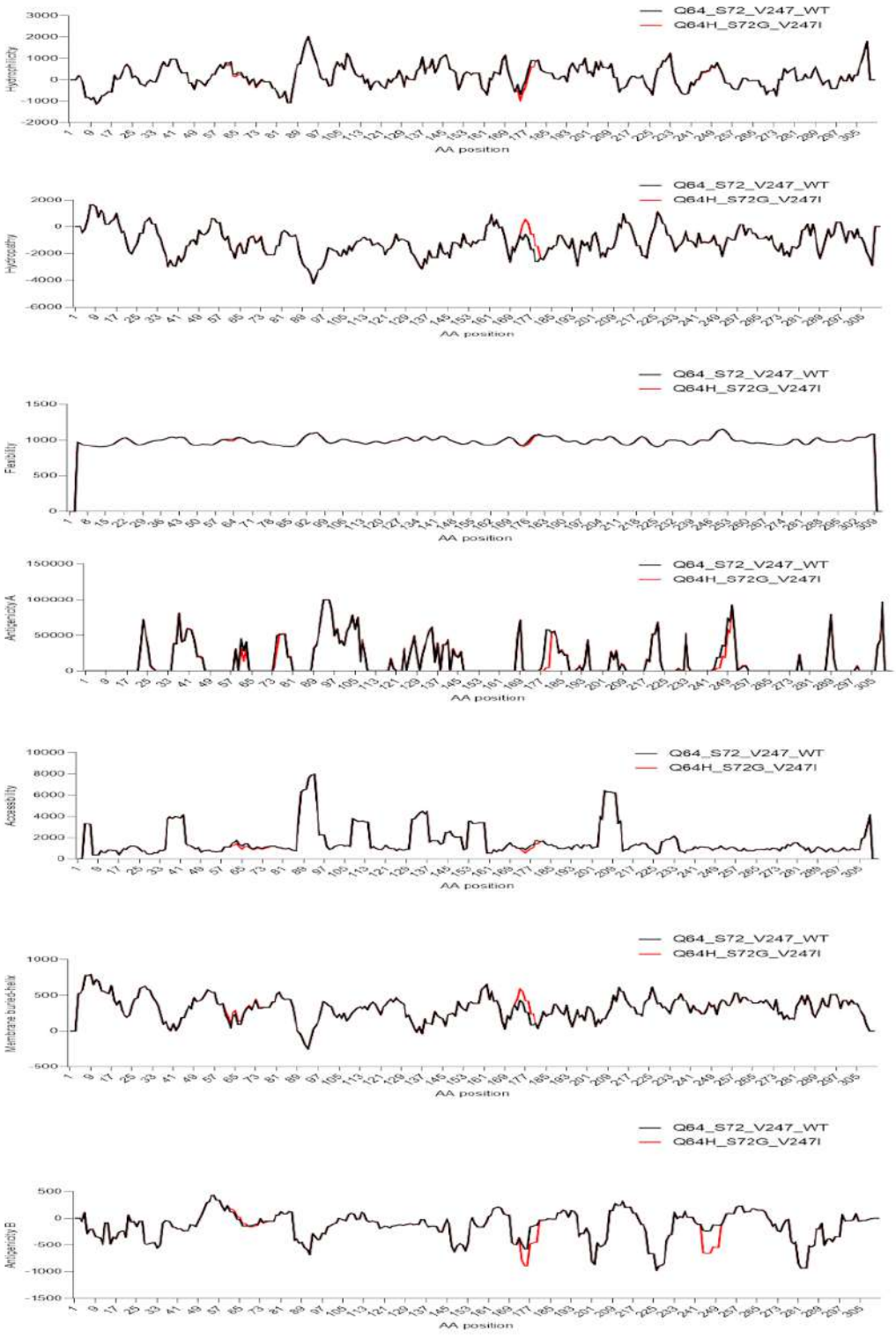
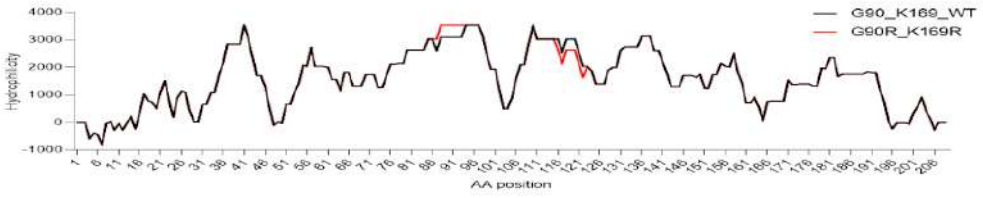
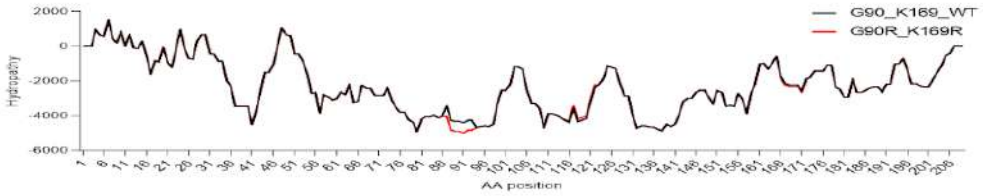
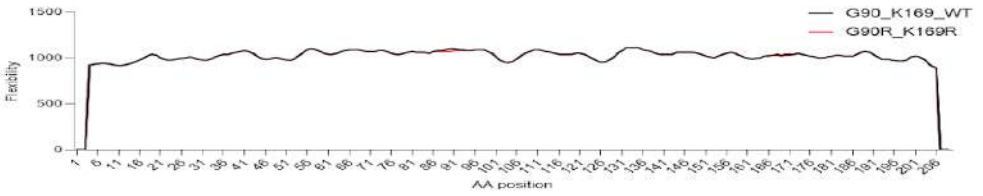
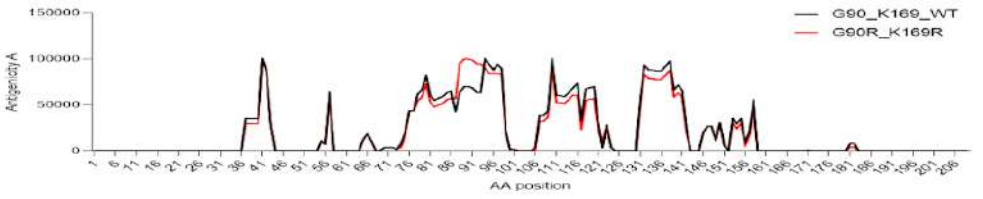

Q64_S72_V247_WT
Q64H_S72G_V247I
Hydrophilicity
Hydropathy
Flexibility
Antigenicity A
Accessibility
Membrane buried helix
Antigenicity B
AA position

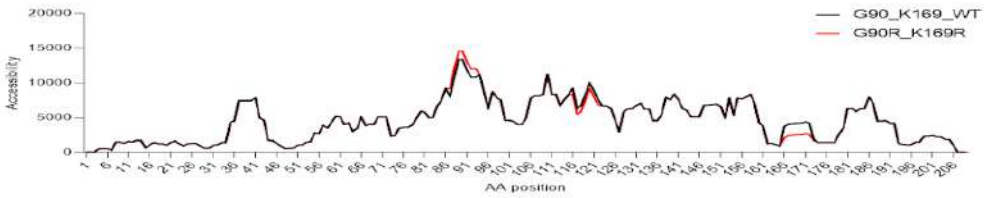
Figure S3: Analysis of the physicochemical profile of the HBZ protein. The analysis suggests that the G90R mutation may generate an important change in hydrophilicity, hydropathy, antigenicity, accessibility, and membrane buried-helix. It also suggests that the K169R mutation may alter accessibility."

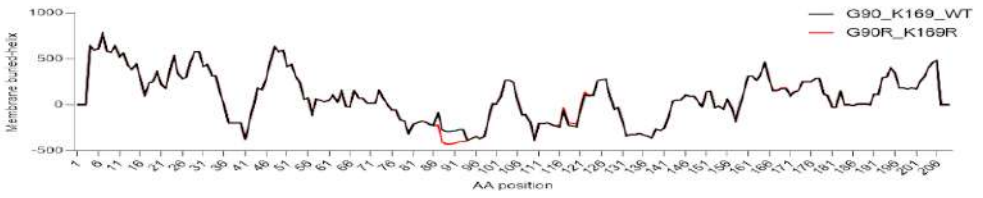

G90_K169_WT
G90R_K169R
Hydrophilicity
AA position


G90_K169_WT
G90R_K169R
Hydropathy
AA position


G90_K169_WT
G90R_K169R
Flexibility
AA position


G90_K169_WT
G90R_K169R
Antigenicity A
AA position


G90_K169_WT
G90R_K169R
Accessibility
AA position


G90_K169_WT
G90R_K169R
Membrane buried helix
AA position

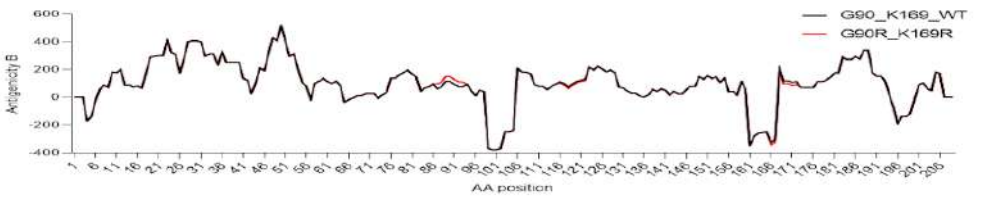

G90_K169_WT
G90R_K169R
Antigenicity B
AA position

Figure S4: Analysis of the physicochemical profile of the p14 protein. The analysis suggests that the G244D mutation may generate an important change in hydrophilicity, hydropathy, antigenicity, accessibility, and membrane buried-helix.

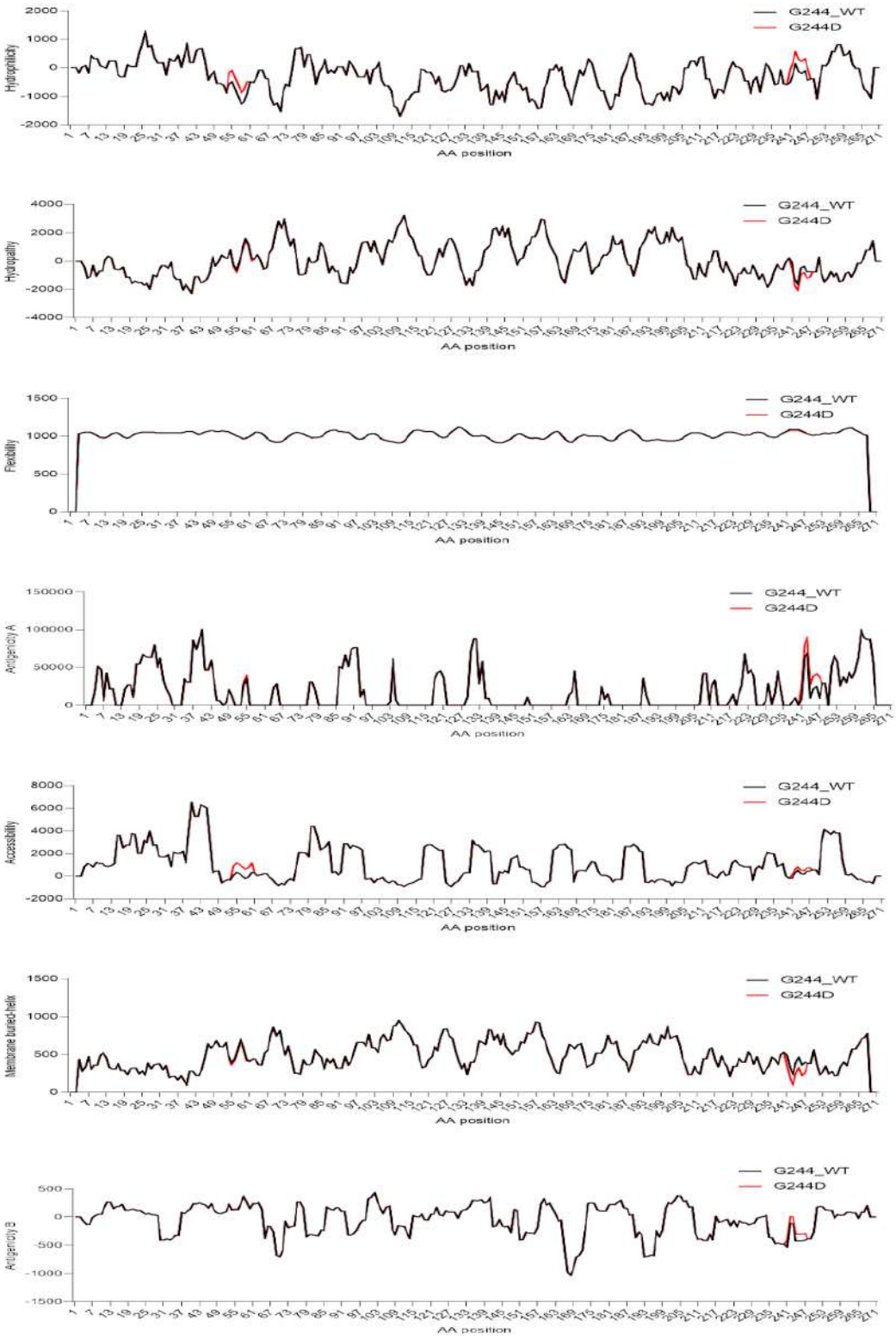

G244_WT
G244D
Hydrophilicity
Hydropathy
Flexibility
Antigenicity A
Accessibility
Membrane buried helix
Antigenicity B
AA position

Figure S5: Analysis of the physicochemical profile of the p15 protein. The analysis suggests that the Q3P mutation may generate an important change in hydropathy, antigenicity, and accessibility.

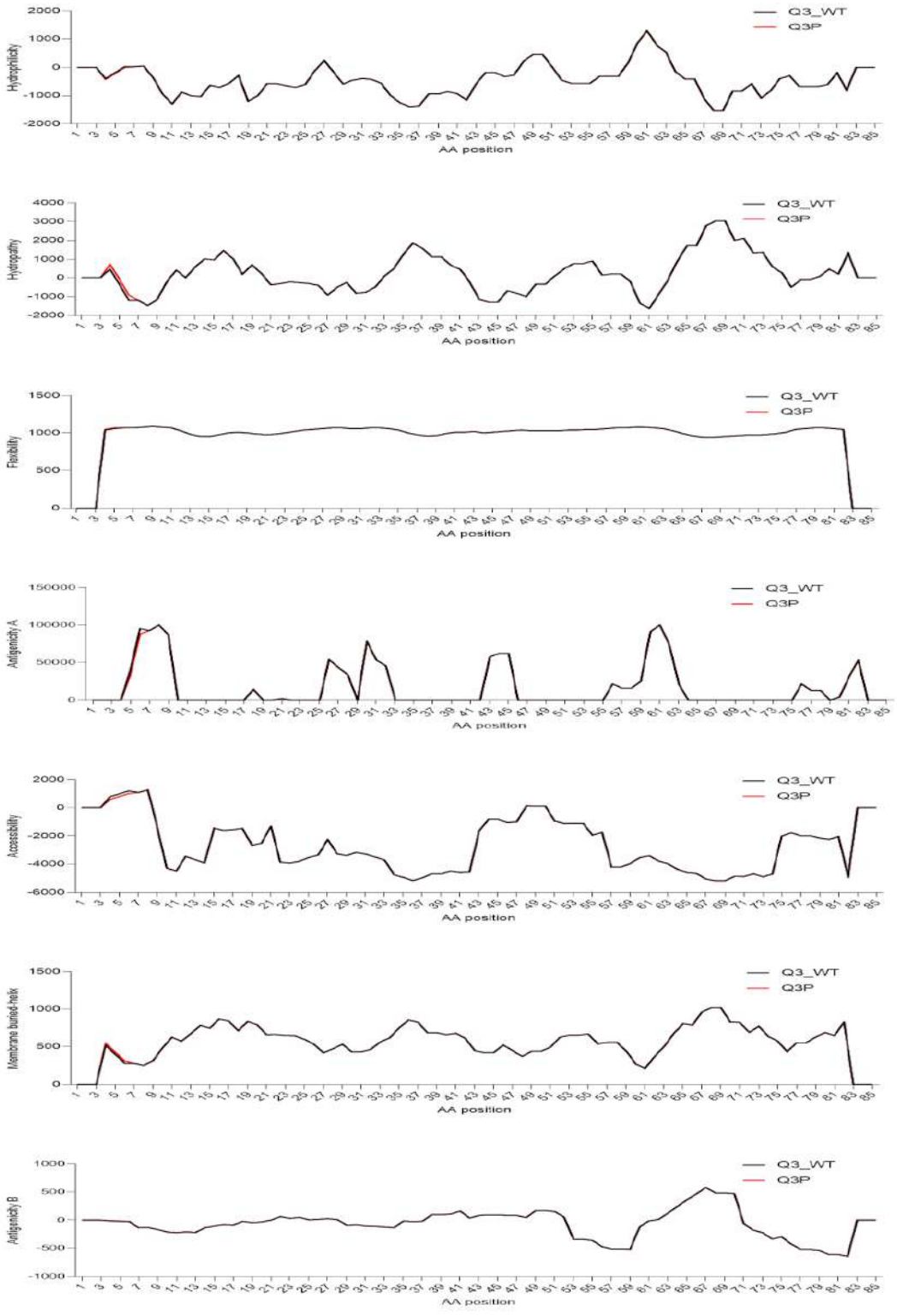

Hydrophilicity
Q3_WT
Q3P
AA position
Hydropathy
Q3_WT
Q3P
AA position
Flexibility
Q3_WT
Q3P
AA position
Antigenicity A
Q3_WT
Q3P
AA position
Accessibility
Q3_WT
Q3P
AA position
Membrane buried-helix
Q3_WT
Q3P
AA position
Antigenicity B
Q3_WT
Q3P
AA position

Figure S6: Analysis of the physicochemical profile of the p30 protein. The analysis suggests that the R37C mutation may generate an important change in hydrophilicity, hydropathy, accessibility, and membrane buried-helix.

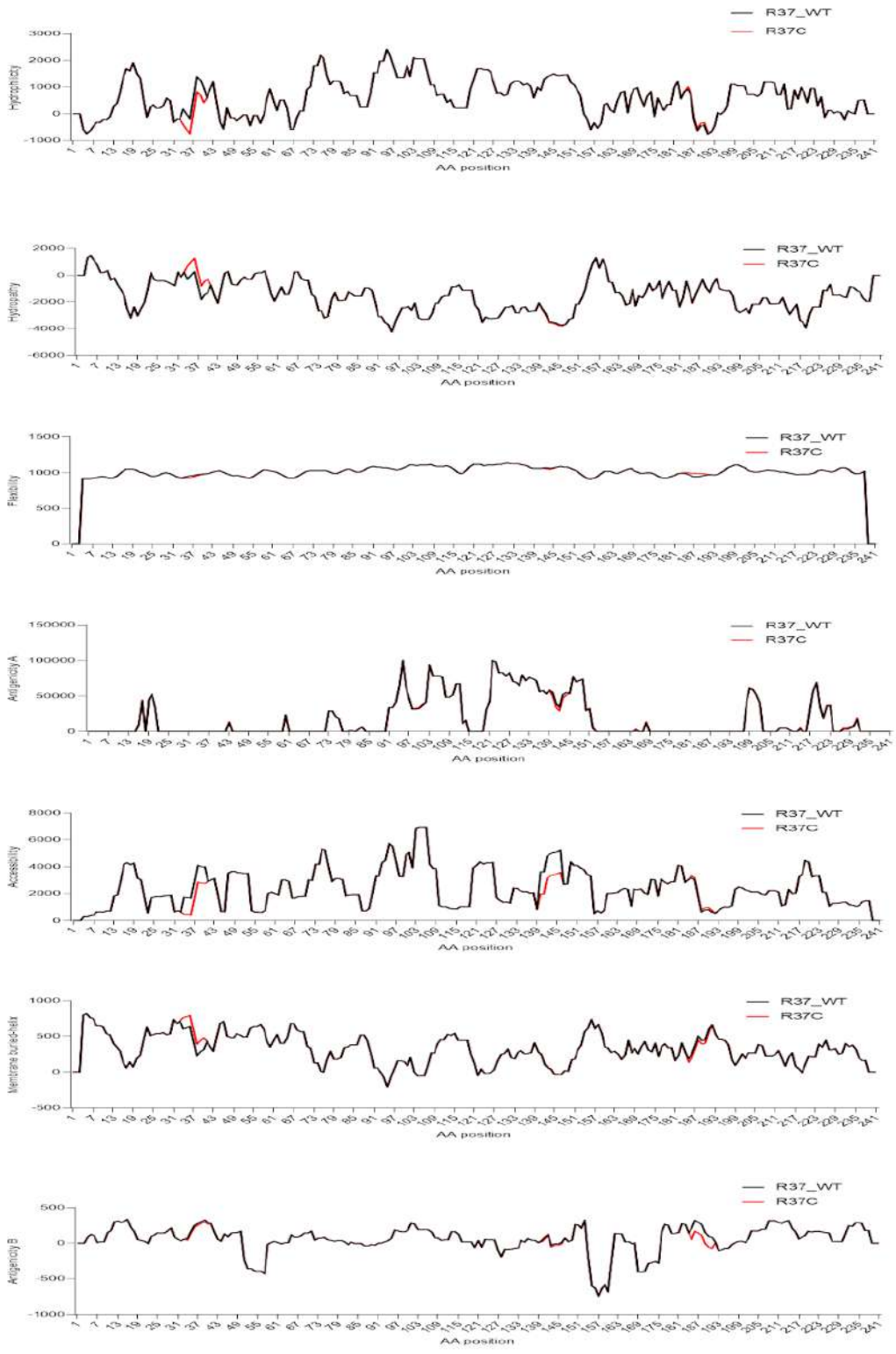


Figure S7: Analysis of the physicochemical profile of the p40 protein. The analysis suggests that the P136S mutation may generate an important change in membrane buried-helix.

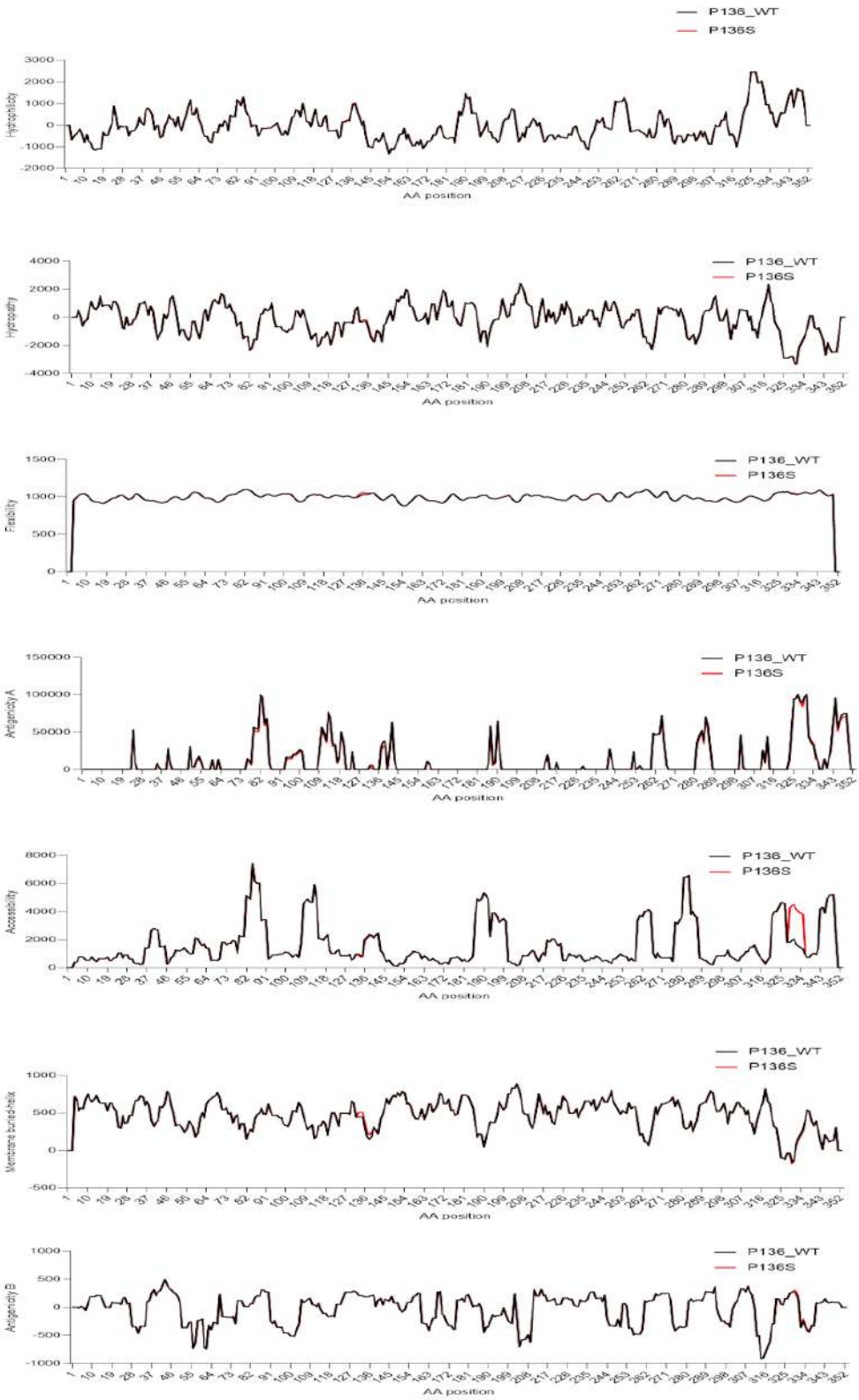
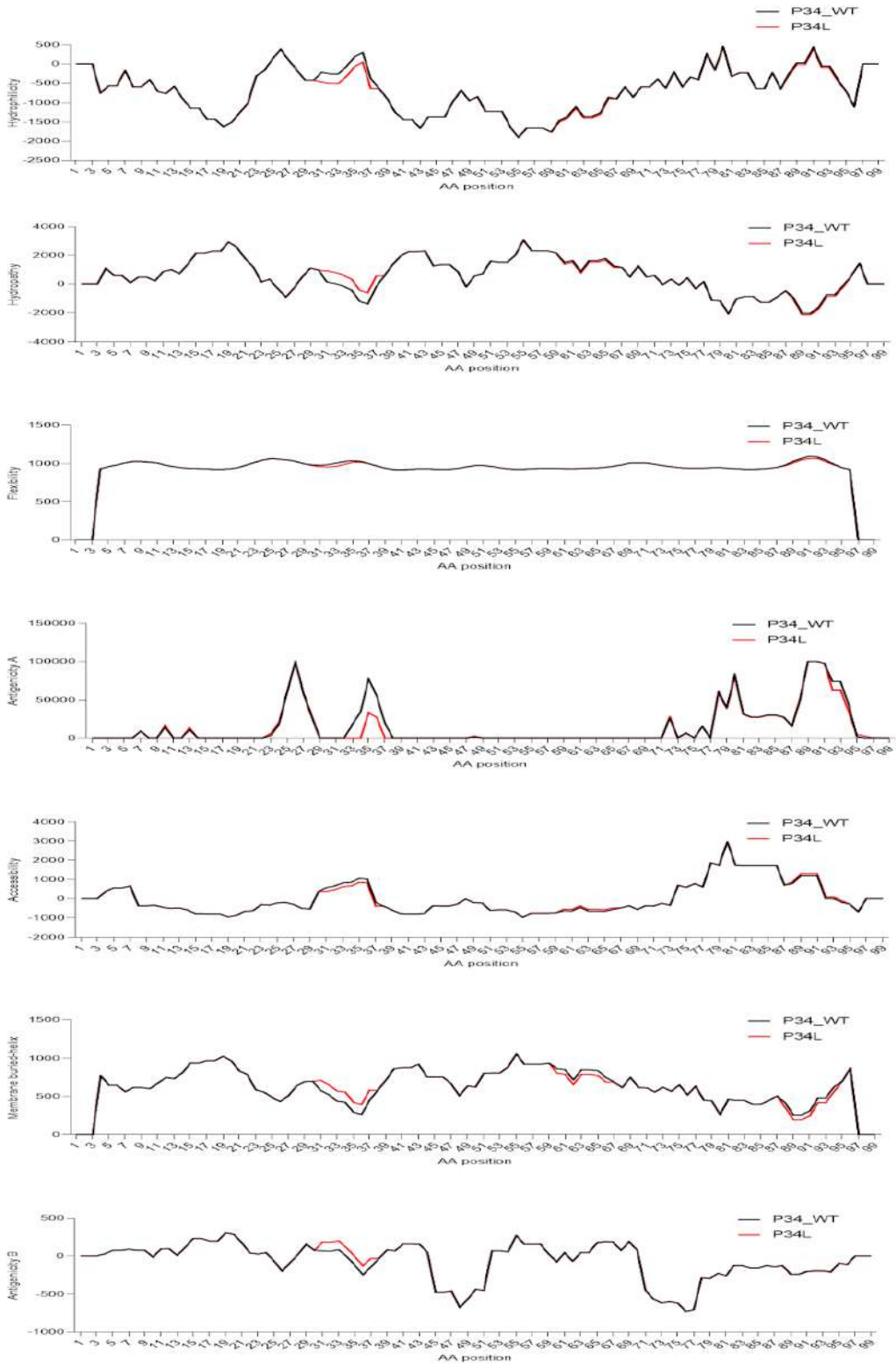

P136_WT
P136S
Hydrophilicity
Hydropathy
Flexibility
Antigenicity A
Accessibility
Membrane buried-helix
Antigenicity B
AA position

Figure S8: Analysis of the physicochemical profile of the p12 protein. The analysis suggests that the P34L mutation may generate an important change in hydrophilicity, hydropathy, antigenicity, flexibility, accessibility, and membrane buried-helix.


P34_WT
P34L
Hydrophilicity
Hydropathy
Flexibility
Antigenicity A
Accessibility
Membrane buried-helix
Antigenicity B
AA position

4.3 SEÇÃO 3: DESENHO DOS PRIMERS PARA SEQUENCIAMENTO BASEANDO EM NANOPOROS DO GENOMA COMPLETO DO HTLV-1 UTILIZANDO O MINION

O desenho dos primers foi baseado em 31 sequências de genoma completo do HTLV-1 geradas pelo grupo e que estão disponíveis no GenBank (KY007244-KY007274), formadas por 8989pb. Após o alinhamento das sequências no programa BioEdit, foram submetidas a uma plataforma de design de primers, a https://primalscheme.com/, que fornece primers multiplex eficientes (QUICK et al, 2017). Foram gerados 29 pares de primers (Figuras 9 e 10) com capacidade de amplificar produtos com tamanho de 400 nucleotídeos. As sequências dos primers gerados foram submetidas à ferramenta BLAST (Basic Local Alignment Search Tool) disponível no site do NCBI (National Center for Biotechnology Information) para confirmação.

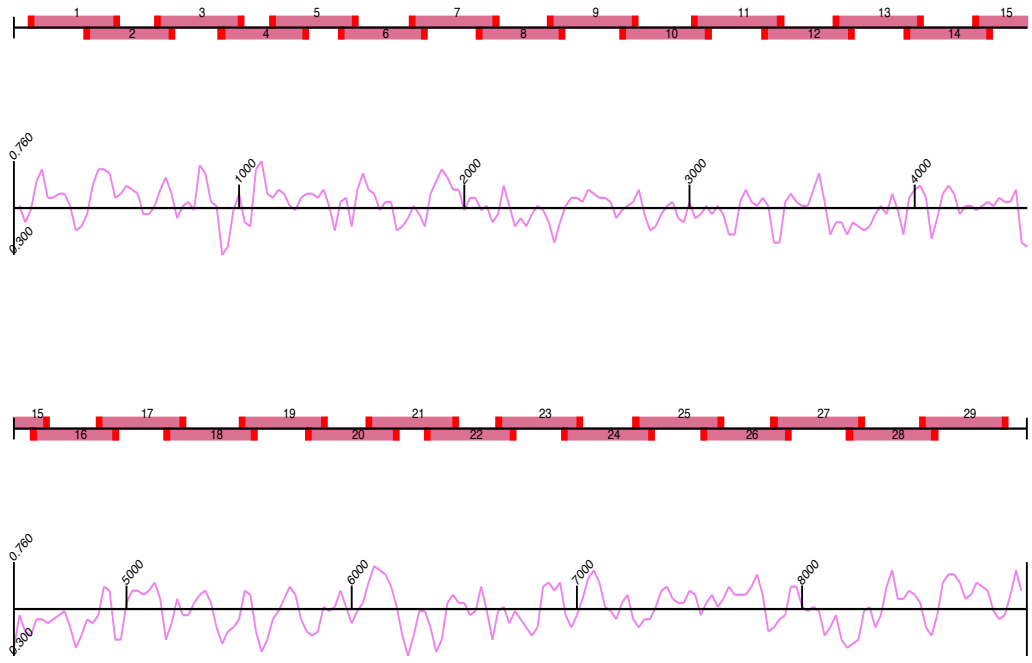


Figura 10: Desenho dos 29 pares de iniciadores para genoma completo do HTLV-1 gerados pelo site https://primalscheme.com/.

HTLV-1.primer

name	pool	seq	size	%gc	tm (use 65)
HTLV-1_1_LEFT	1	GAAAAGGTCAGGGCCCAGACTA	22	54.55	61.87
HTLV-1_1_RIGHT	1	TCTGAACTTACCTAGACGGCGG	22	54.55	61.70
HTLV-1_2_LEFT	2	AAAAGCGTGGAGACAGTTCAGG	22	50.00	61.25
HTLV-1_2_RIGHT	2	CTATAGAATGGGCTGTCGCTGG	22	54.55	60.79
HTLV-1_3_LEFT	1	TTCATTCACGACTAACTGCCGG	22	50.00	61.11
HTLV-1_3_RIGHT	1	CCGGGGTATCCTTTTGGGAGTA	22	54.55	61.41
HTLV-1_4_LEFT	2	ACGATTTCCACCAGTTAAAGAAATTTCT	28	32.14	60.76
HTLV-1_4_RIGHT	2	CAAGCCGGATGGTCTGCATAAA	22	50.00	61.45
HTLV-1_5_LEFT	1	CCCCCTCCTTATGTTGAGCCTA	22	54.55	61.15
HTLV-1_5_RIGHT	1	CAGAGTTGCTGGTATTCTCGCC	22	54.55	61.49
HTLV-1_6_LEFT	2	TTATAACCCCTTAGCCGGTCCC	22	54.55	61.48
HTLV-1_6_RIGHT	2	TTTAGGCGGGACAACTAACACT	22	45.45	59.81
HTLV-1_7_LEFT	1	TCTAGGAGATATGTTGCGGGCT	22	50.00	60.94
HTLV-1_7_RIGHT	1	GGATCTAACGGTATAACTGGCAGA	24	45.83	60.04
HTLV-1_8_LEFT	2	CTCCATAGAGGGGGAGGTTTAAC	23	52.17	59.87
HTLV-1_8_RIGHT	2	ATCACGACCTATGATGGCCCAG	22	54.55	62.12
HTLV-1_9_LEFT	1	TCCGGACAACGCCTATTGTTTT	22	45.45	60.99
HTLV-1_9_RIGHT	1	CCGGGGGAAGATGATGAGAGAT	22	54.55	61.28
HTLV-1_10_LEFT	2	AATGGAACCTGGCGATTCATCC	22	50.00	61.19
HTLV-1_10_RIGHT	2	AGAGTAGTAGTAGGTCCTCATGGG	24	50.00	60.22
HTLV-1_11_LEFT	1	CAAGCTTTCCCCCAATGCACTA	22	50.00	61.33
HTLV-1_11_RIGHT	1	TGCACTAATGATTGAACTTGAGAAGG	26	38.46	60.29
HTLV-1_12_LEFT	2	CAGCCCCTTCACAGTCTCTACT	22	54.55	61.33
HTLV-1_12_RIGHT	2	CTGGGGTGGTCAGATGTTTGAA	22	50.00	60.67
HTLV-1_13_LEFT	1	GCAATCCTATGGGCTACTCTGC	22	54.55	61.05
HTLV-1_13_RIGHT	1	GTTGGAGAGGCCATGCAAAAGT	22	50.00	61.85
HTLV-1_14_LEFT	2	ACAAAGATCATTCCCCCTTCCG	22	50.00	60.54
HTLV-1_14_RIGHT	2	CTTGCAATGTGAGGGCTGTCTG	22	54.55	62.20
HTLV-1_15_LEFT	1	TACTAATCACCCCTGTCCTGCA	22	50.00	60.75
HTLV-1_15_RIGHT	1	CAGGGCCGTTGTCTGTGTTTAT	22	50.00	61.31
HTLV-1_16_LEFT	2	GCTCAGAAGCTATTTCCTCTTTGC	24	45.83	60.52
HTLV-1_16_RIGHT	2	ATGGGTTTGTTTATTGCTGAGGG	23	43.48	59.74

HTLV-1_17_LEFT	1	TTAACCAACTGCCACAAAACCC	22	45.45	60.01
HTLV-1_17_RIGHT	1	TCCAATTGTGAGAGTACAGCAGC	23	47.83	61.12
HTLV-1_18_LEFT	2	GGGTAAGTTTCTCGCCACTTTGA	23	47.83	61.31
HTLV-1_18_RIGHT	2	TCTTGCTGAAACTTCCAGTAGGG	23	47.83	60.50
HTLV-1_19_LEFT	1	CATACCTGGGGTGCCAATCATG	22	54.55	61.52
HTLV-1_19_RIGHT	1	AGTATAGGACGTGCCAAGTGGA	22	50.00	61.07
HTLV-1_20_LEFT	2	TGACCCTTGTCCAGTTAACCCT	22	50.00	61.15
HTLV-1_20_RIGHT	2	TGTAGGAGGCTCTTTCCTGAGG	22	54.55	61.08
HTLV-1_21_LEFT	1	TCACCTGTTCCCACCCTAGGAT	22	54.55	62.46
HTLV-1_21_RIGHT	1	GAGAGGCCAAGGTCCCAGTTAA	22	54.55	61.87
HTLV-1_22_LEFT	2	GGGAGCAAGGAGGATTATGCAA	22	50.00	60.61
HTLV-1_22_RIGHT	2	GAAGTTGCTGCAGGAGAAGGAG	22	54.55	61.37
HTLV-1_23_LEFT	1	ACAATTATTGCAACCACATCGCC	23	43.48	60.93
HTLV-1_23_RIGHT	1	GAGGAGAAGAGGAAGCGGAAAA	22	50.00	60.21
HTLV-1_24_LEFT	2	TGCCTTCTCCTCTTCCTTCCTT	22	50.00	60.95
HTLV-1_24_RIGHT	2	AACACGTAGACTGGGTATCCGA	22	50.00	60.80
HTLV-1_25_LEFT	1	AGCCCTACAGATACAAAGTTAACCA	25	40.00	60.14
HTLV-1_25_RIGHT	1	TGGGTTCCATGTATCCATTTCGG	23	47.83	60.94
HTLV-1_26_LEFT	2	TACCCCGCCAATCACTCATACA	22	50.00	61.41
HTLV-1_26_RIGHT	2	CCTAACAGGCTGGAAAAGGGTG	22	54.55	61.32
HTLV-1_27_LEFT	1	ACTCCTCTATAAAATCTCCCTTACCAC	27	40.74	60.21
HTLV-1_27_RIGHT	1	TGGTCATTGTCATCTGCCTCTTT	23	43.48	60.50
HTLV-1_28_LEFT	2	ACATCTCCTGTTTGAAGAATACACCA	26	38.46	60.91
HTLV-1_28_RIGHT	2	CCACGCTTTTATAGACTCCTGTTAGT	26	42.31	61.07
HTLV-1_29_LEFT	1	CGACAACCCCTCACCTCAAAAA	22	50.00	61.19
HTLV-1_29_RIGHT	1	GAATGAAAGGGAAAGGGGTGGA	22	50.00	60.41

Figura 11: Sequências nucleotídicas dos 29 iniciadores desenhados para sequenciamento do genoma completo do HTLV-1 utilizando o MinION.

Utilizamos a plataforma disponível no site da ThermoFisher (https://www.thermofisher.com/br/en/home/brands/thermo-scientific/molecular-biology/molecular-biology-learning-center/molecular-biology-resource-library/thermo-scientific-web-tools/multiple-primer-analyzer.html) para analisar os iniciadores desenhados para PCR multiplex. Os resultados sugeriram possíveis self-primers e primer cross, mas a alta temperatura de anneling, de 65 à 68^0C, proposta no protocolo para a PCR multiplex (QUICK, 2017), tende a reduzir esse risco.

Esses primers encontram-se em fase de síntese e serão utilizados para sequenciar os genomas completos do HTLV-1 seguindo os protocolos de sequenciamento disponibilizados pela Oxford Nanopore Technologies. Os genomas completos gerados serão caracterizados e disponibilizados em bancos de dados de livre acesso (GenBank), já que os estudos relacionados à caracterização viral e a sua consequente relação com a patogenicidade são importantes para o maior conhecimento científico do assunto e para o desenvolvimento de melhores alternativas para controle da infecção e patogênese viral.

5 DISCUSSÃO

Estima-se que aproximadamente 5 a 10 milhões de pessoas são infectadas pelo HTLV em todo o mundo (GESSAIN e CASSAR, 2012) e é o agente etiológico da paraparesia espástica tropical/mielopatia associada ao HTLV (HAM/TSP) (GESSAIN et al., 1985; OSAME et al., 1986), da Leucemia/Linfoma de Células T do Adulto (ATLL) (YOSHIDA, MIYOSHI e HINUMA, 1982) e da dermatite infecciosa associada ao HTLV-1 (DIH) (LA GRENADE, 1996).

Apesar do HTLV-1 ser considerado um retrovírus estável, devido a sua forma de multiplicação principal ser através da expansão clonal (WATTEL et al., 1995) que dispensa a utilização da enzima transcriptase reversa, nos conduz a pensar que o desenvolvimento de um método terapêutico eficaz e/ou o desenvolvimento de uma vacina seria alcançados mais facilmente, mas os indivíduos infectados pelo vírus ainda permanecem sem tratamento eficaz e vacina. Diante esse fato, as análises da diversidade genética da ORF-I do HTLV-1 em pacientes com diferentes condições clínicas demonstram uma região com baixa diversidade genética e com potencial para ser o alvo de desenvolvimento de uma vacina.

Apesar da importância clínica e epidemiológica do HTLV-1, há um número limitado de genomas completos disponíveis, cerca de 0,12 genomas completos por 10.000 indivíduos infectados. Então, desenhar uma metodologia para o sequenciamento baseado em nanoporos do genoma completo do HTLV-1 facilita a geração de novas sequências e, consequentemente, geração de novos dados a partir das análises realizadas, contribuindo para a erradicação desse vírus.

Além do número limitado de genomas completo disponíveis, há uma escassez de estudos relacionados às sequências (totais e parciais) disponíveis nos bancos de dados públicos. Por isso, buscamos investir em estudos de caracterização molecular em sequências já disponíveis no GenBank, reutilizando dados. Ainda, algumas questões permanecem sem resposta, principalmente a elucidação dos motivos pelos quais alguns indivíduos infectados permanecem como portadores assintomáticos enquanto outros desenvolvem patologias associadas ao vírus. Nesse contexto, existem hipóteses que sugerem que fatores relacionados ao hospedeiro, bem como

variações genéticas virais, podem influenciar no desfecho da infecção pelo HTLV-1. Estudos demonstraram que variações nucleotídicas presentes na ORF-I do vírus podem interferir na carga proviral e na manifestação de HAM/TSP (BARRETO et al., 2016). Outros estudos relataram mutações em outros genes do vírus que podem estar relacionados a indivíduos sintomáticos (SOCORRO DE ALMEIDA VIANA et al., 2018) e que podem estar presente em indivíduos assintomáticos que apresentam uma maior carga proviral (NETO et al., 2011). De acordo com esses fatos, essa dissertação buscou, também, identificar SNVs no genoma completo do HTLV-1 que possam estar relacionados ao desfecho clínico do paciente em sequencias já publicadas. Os dados demonstraram possíveis correlações entre mutações e desenvolvimento de uma sintomatologia. Algumas mutações presentes nos sítios de modificação pós-traducionais podem modificar a conformação proteica e, consequentemente, a função das mesmas. Outra mutação localizada na posição 8452 do gene LTR demonstrou ser uma mutação importante nos grupos DIH e ATLL, além de apresentar uma possível correlação com uma mutação na região pX para o resultado de pacientes sintomáticos. Além disso, pelo fato de termos reunidos sequências disponíveis no GenBank para o estudo, destacamos a importância do fornecimento de dados clínicos e epidemiológicos junto com as sequências virais.

6 CONCLUSÃO

A caracterização do genoma completo do HTLV-1 proveniente de indivíduos com diferentes condições clínicas demonstrou que correlações entre mutações em genes diferentes pode estar relacionado com o desfecho clinico do indivíduo infectado. Além disso, a partir da caracterização molecular da ORF-I proveniente de pacientes com diferentes perfis clínicos foi possível destacar a baixa diversidade genética da ORF-I do HTLV-1 e evidenciar a região como possível alvo de vacinas.

7 APÊNDICE

Revisão Sistemática intitulada como “Application of different sequencing technologies for HTLV-1: a systematic review” já submetido.

Overview of sequencing technology platforms applied to HTLV-1 studies: a systematic review

Felipe de Oliveira Andrade[a], Marina Silveira Cucco[b,c], Melina Mosquera Navarro Borba[b] Reinaldo Conceição Neto[d], Luana Leandro Gois[b,e,f], Filipe Ferreira de Almeida Rego[f], Luciane Amorim Santos[b,c,e,f], Fernanda Khouri Barreto[a,*]

[a] Instituto Multidisciplinar em Saúde, Universidade Federal da Bahia, Vitória da Conquista, Brazil

[b] Instituto Gonçalo Moniz, Fundação Oswaldo Cruz, Salvador, Brazil

[c] Programa de Pós-graduação em Ciências da Saúde, Faculdade de Medicina da Bahia, Universidade Federal da Bahia, Salvador, Brazil

[d] Faculdade Cruzeiro do Sul, Vitória da Conquista, Brazil

[e] Escola Bahiana de Medicina e Saúde Pública, Salvador, Brazil

[f] Universidade Católica do Salvador, Salvador, Brazil

*****Correspondence author: Instituto Multidisciplinar em Saúde, Universidade Federal da Bahia, Campus Anísio Teixeira. Rua Hormindo Barros, 58. Bairro Candeias: 45.029-094. Vitória da Conquista, Brazil. Email: fernanda.khouri@hotmail.com

Abstract

The Human T-lymphotropic virus type 1 (HTLV-1) was the first human retrovirus described. The viral factors involved in the different clinical manifestations of infected individuals are still unknown, and in this sense, sequencing technologies can support the viral genome studies, contributing to a better understanding of infection outcome. Currently, several sequencing technologies are available with different approaches. To understand the methodological advances in the HTLV-1 field, it is necessary to organize a synthesis by a rigorous review. This systematic literature reviews described different technologies used to generate HTLV-1 sequences. The review followed the PRISMA guidelines and the search for articles was performed in PubMed, Lilacs and SciELO databases. From the 358 articles found in search, 60 were selected. The articles showed that, even with the emergence of new sequencing technologies, the traditional Sanger method continues to be the most used methodology for generating HTLV-1 genomes. There are many questions that remain unanswered in the field of HTLV-1 research and this reflects on the low number of studies using Next-Generation Sequencing technologies, which could help address these gaps. The data compiled and analyzed here can help research on HTLV-1, assisting in the choice of sequencing technologies.

Keywords: HTLV-1; sequencing technologies; genome.

1. Introduction

It is estimated that 5-10 million people are infected with Human T-lymphotropic virus type 1 (HTLV-1) worldwide [1, 2]. Infected individuals can develop HTLV-1 associated pathologies such as adult T-cell leukemia/lymphoma (ATLL), HTLV-1-associated myelopathy/tropical spastic paraparesis (HAM/TSP) and HTLV-1-associated infectious dermatitis (IDH), among other inflammatory diseases, or can be classified as asymptomatic carriers (AC) [3–5]. The factors involved in the development of a particular clinical manifestation have not yet been elucidated and HTLV-1-infected individuals remain without specific treatment [6].

The HTLV-1 genome structure is composed by two flanking regions, known as long terminal repeat (5' and 3' LTR), and structural genes as gag, pol and env. Also, there is a non-structural region, pX, adjacent to 3' LTR, that codifies regulatory and accessory proteins, as Tax, Rex and HBZ [7]. Molecular characterization of the viral genome, based on sequencing combined with bioinformatics analysis, provides information on its genomic regions, such as viral integration sites, identification of mutations and epigenetic changes [8]. This information is important for the development of HTLV-1 specific vaccines and therapies.

Although HTLV-1 was the first human retrovirus described, there is a considerable difference in the amount of sequences generated when compared to other important retroviruses, such as Human Immunodeficiency Virus 1 (HIV-1). In March of 2021, there was 1,048,465HIV-1 published sequences, while for HTLV-1 there was only 9,980 sequences available in GenBank. This scenario demonstrates the importance of sequencing new HTLV-1 genomes in order to raise information and contribute to the understanding of pathogenesis and development of vaccines and therapeutics for HTLV-1.

In 1975, Sanger presented the first DNA sequencing technique, which composes the sequencing first-generation and was widely adopted, being used until today. This technique is based on the use of modified chain terminators, the dideoxynucleotides (ddNTP's) [9]. The sequence techniques evolution allowed the emergence of the Next-Generation Sequencing (NGS), starting with the second-generation technology. This technology bringing new methodologies for determining nucleotide sequences with greater efficiency and speed, like 454 from Roche Applied Science, Solexa by Illumina and Ion Torrent, that expanded the ways of sequencing the genetic material [10, 11]. The main examples of the second-generation

technology are pyrosequencing and sequencing by synthesis (SBS). In this generation the DNA polymerase acts in conjunction with a chemiluminescent enzyme, which when complementing a template of a DNA strand, emits chemiluminescent signals, allowing the determination of the sequence [12].

Recently, the third generation emerged, composed by nanopore sequencing (Oxford Nanopore Technologies) and Pacific Biosciences (PacBio) methodologies [11]. Unlike other sequencing technologies, these methodologies arise with the function of sequencing unique DNA molecules and longer read lengths in a shorter time, characteristic that cannot be found in the previous generations [13]. The nanopore methodology stand out not only for generating long nucleotide chains through larger devices such as GridION and PromethION, but also through small portable device such as MinION and Flongle. This technique is based on the passage of genetic material through a nanopore membrane that detects the electrical signals emitted in the passage of each nucleotide [14].

It should be noted that, in recent years, there has been a significant genome sequencing technological diversification, with more efficient, cheaper and faster devices. Investigating which sequencing technology is most used to generate HTLV-1 genomes allows us to understand the limitations and possibilities of research carried out on the viral genome. This may help to correctly fill the gaps of this virus knowledge, such as the factors involved in the development or not of the HTLV-1 associated diseases. In this sense, and considering the importance of technological choice for sequencing, this study aims to analyze the different technologies used to generate HTLV-1 sequences and the contributions of these techniques in new investigations on this retrovirus.

2. Material and methods

This study consists of a literature systematic review carried out in accordance with Preferred Reporting Items for Systematic Reviews and Meta-Analyzes (PRISMA®) guidelines. A systematic search was conducted for studies that performed HTLV-1 partial or total genome sequencing. The articles were searched at PubMed, Lilacs and SciELO databases in November 2020. The search algorithm used was composed by subjects from DeCS/MeSH database and additional keywords: *("Human T lymphotropic virus 1" OR "HTLV-1") AND "sequence*" AND*

("molecular sequence data" OR "sequencing"). Through the search algorithm, all titles were cross-checked to identify possible duplicate studies.

For the selection of articles, the following inclusion criteria were applied to select studies: (i) only articles in Portuguese, English or Spanish; (ii) original studies and (iii) that presented total or partial sequencing of HTLV-1. Articles published since 2000 were included. The exclusion criteria were: (i) studies not specifying the sequencing method, (ii) studies that did not generate HTLV-1 sequences or not specify number of generated sequences, (iii) animal studies, and (iv) studies that performed genome sequencing through cell line.

The articles found on the platforms were initially filtered and selected from reading the title and abstract. Subsequently, a new selection was made by reading the full text. After reading and analyzing the selected articles, the data were collected and included in this review. The search for published studies was independently performed by two authors (F.O.A. and M.S.C.) and disagreements about all outcomes were resolved by consensus among all authors.

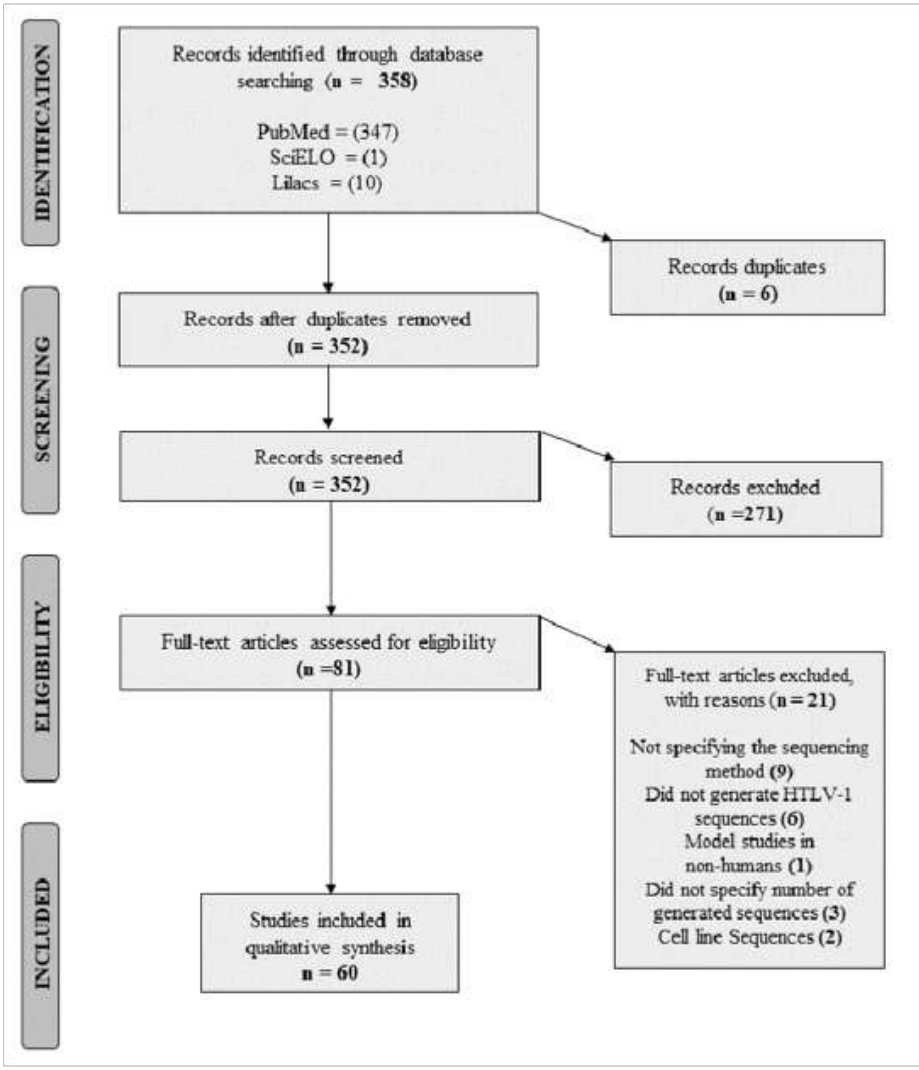
After reading the selected studies, the following contents were extracted from each one: (1) basic information (title, authors, year, objectives), (2) study design, (3) sequencing technology and method, (4) subjects (sample origin and HTLV-1 sequenced region), and (5) number of generated sequences. The data collected from the articles was tabulated using Microsoft Excel®□. The figures generated in this work were produced by the programs Adobe Photoshop®□ and Microsoft PowerPoint®□ 2019 versions.

This study was registered with the International Prospective Register of Systematic Reviews (PROSPERO) under number CRD42020218387.

3. Results

The search for studies identified a total of 358 articles, of which 347 were available in PubMed, one in SciELO and ten in Lilacs. Of these, six were excluded due to duplication, 271 were excluded after selection by title and abstract and 21 after reading the full text. Therefore, 60 articles were included in the systematic review (Figure 1).

Fig. 1 Flow diagram of the systematic review studies selection.



The articles indicated the use of three HTLV-1 sequencing methodologies: Sanger, Illumina and Ion Torrent. Sanger sequencing, which is the first generation of sequencing, was the most used technique. Even after the emergence of NGS methodologies in 2004, it is observed that the most HTLV-1 studies published continued to use Sanger as a preferential

methodology. Among the 60 articles used in this review, 56 used Sanger and, of these, 37 were carried out after 2004 (Table 1).

Table 1: Summary of information collected from the 60 selected studies, including samples origin, sequencing methodology, equipment and number of sequences generated.

References	Sample Origin	Sequencing Methodology	Equipment	Sequences Generated
[15]	NA[b]	Sanger	ABI[a] 377A DNA sequencer	218
[16]	NA	Sanger	Hitachi Fluorescent DNA sequencer SQ-5500	39
[17]	NA	Sanger	Hitachi Fluorescent DNA sequencer SQ-5500	64
[18]	France and UK	Sanger	ABI 373 automatic DNA sequencer	17
[19]	Chile	Sanger	Automated DNA sequencer	37
[20]	Brazil	Sanger	Fmol DNA sequencing system (Promega)	2
[21]	France	Sanger	ABI 377A DNA sequencer	208
[22]	Japan	Sanger	ABI 373 automatic DNA sequencer	178
[23]	NA	Sanger	SQ5500 automated sequencer	138
[24]	Colombia	Sanger	NA	12
[25]	Italy	Sanger	ABI PRISM automatic sequencer	6
[26]	Chile	Sanger	NA	50
[27]	Chile	Sanger	NA	128
[28]	Spain	Sanger	ABI 310 genetic analyzer	4
[29]	Colombia	Sanger	NA	12
[30]	France	Sanger	Applied Bioshstems 377 DNA sequencer	1
[31]	Japan	Sanger	ABI 377 DNA sequencer	231
[32]	Colombia	Sanger	ABI Prism serie 3700	11

[33]	Brasil	Sanger	ABI 373 DNA Sequencer	3
[34]	Russia	Sanger	ABI 377 automatic DNA sequencer	8
[35]	India	Sanger	ABI automated DNA sequencer	7
[36]	Brazil	Sanger	ABI 377 Automated DNA Sequencer	26
[37]	France, Gabon and Iran	Sanger	ABI Prism 377 and Ceq2000 sequencer	65
[38]	Argentina	Sanger	ABI model 377 automated DNA sequencer	12
[39]	Brazil	Sanger	ABI 377 Sequencer	134
[40]	Israel	Sanger	ABI automated sequencer	1
[41]	Brazil	Sanger	ABI Prism 377 DNA Sequencer	2
[42]	Brazil	Sanger	ABI 373 DNA Sequencer	5
[43]	NA	Sanger	ABI 310 sequencer	334
[44]	NA	Sanger	ABI 310 autosequencer	316
[45]	Japan	Sanger	ABI 377 DNA Sequence	445
[46]	Brazil	Sanger	ABI 3100 genetic analyzer	46
[47]	Argentina	Sanger	ABI Prism 3100 Genetic Analyzer	44
[48]	Gabon	Sanger	Automatic sequencing system (Euro Sequence Gene Services)	34

[49]	Brazil	Sanger	ABI 3100 genetic analyzer	8
[50]	Brazil	Sanger	ABI 3100 genetic analyzer	5
[51]	Argentina	Sanger	ABI PRISM 377 Automated DNA sequencer	114
[52]	Japan	Sanger	ABI 3730 Sequencer	19
[53]	Brazil	Sanger	ABI PRISM 310 Genetic Analyzer	25
[54]	Brazil	Sanger	ABI 3100 genetic analyzer	8
[55]	Mozambique	Sanger	ABI 3730 Automated DNA Sequencer	25
[56]	Colombia	Sanger	ABI PRISM 310 sequencer	30
[57]	Portugal and Spain	Sanger	Automated DNA sequencing	47
[58]	Brazil	Sanger	NA	1
[59]	Brazil	Sanger	ABI 3130 genetic analyzer	13
[60]	Brazil	Sanger	ABI 3100 genetic analyzer	32
[61]	Brazil	Sanger	ABI 3100 genetic analyzer	146
[62]	Brazil	Sanger	ABI 3100 genetic analyzer	18
[63]	Cuba	Sanger	Genome Lab Dye Terminator Cycle Sequence	12
[71]	Brazil	Illumina	Illumina MiSeq System	90
[64]	Brazil	Sanger	ABI 1373 Automated DNA Sequencer	14
[72]	Brazil	Ion Torrent	ABI 3130xl Genetic Analyzer	22

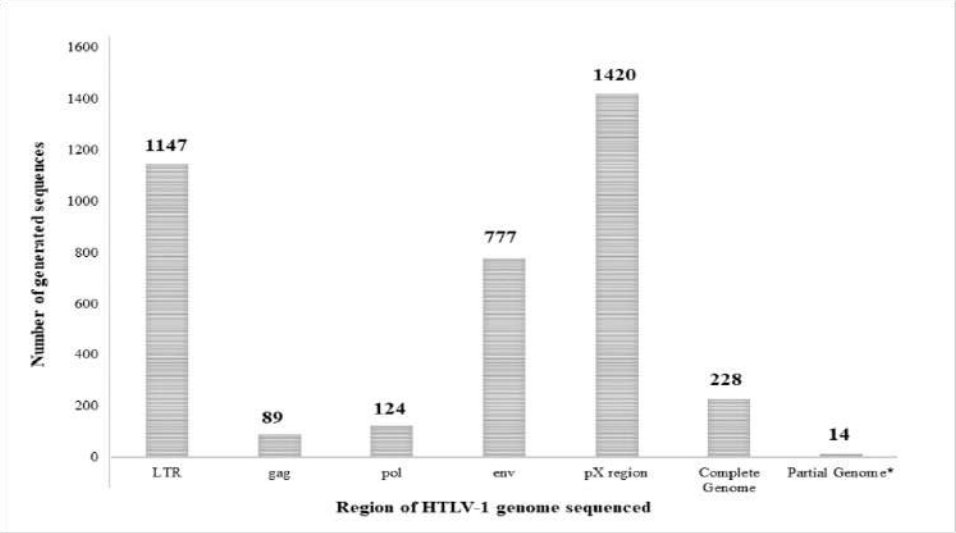
[65]	Iran	Sanger	ABI 3730 Sequencer	2
[66]	Brazil	Sanger	ABI PRISM 310 Genetic Analyzer	2
[67]	Japan and Brazil	Sanger	ABI PRISM 3740 Genetic Analyzer	14
[68]	Brazil	Sanger	ABI PRISM 3100 Genetic Analyzer	24
[8]	Japan	Illumina	Illumina MiSeq or NextSeq	98
[69]	Iran	Sanger	NA	5
[73]	Brazil	Ion Torrent	Ion 318™ Chip PGM	31
[70]	Brazil	Sanger	ABI 3130 Genetic Analyzer	132

[a]ABI: Applied Biosystems; [b]NA: Not available.

In most studies, partial HTLV-1 genome sequencing was performed. Of the 56 articles that used the Sanger methodology [15–70], 53 performed partial genome sequencing and three studies performed complete sequencing of the HTLV-1 genome. In the four articles that used NGS methodologies [8, 71–73], two partial sequencing studies were found and two studies performed complete genome sequencing.

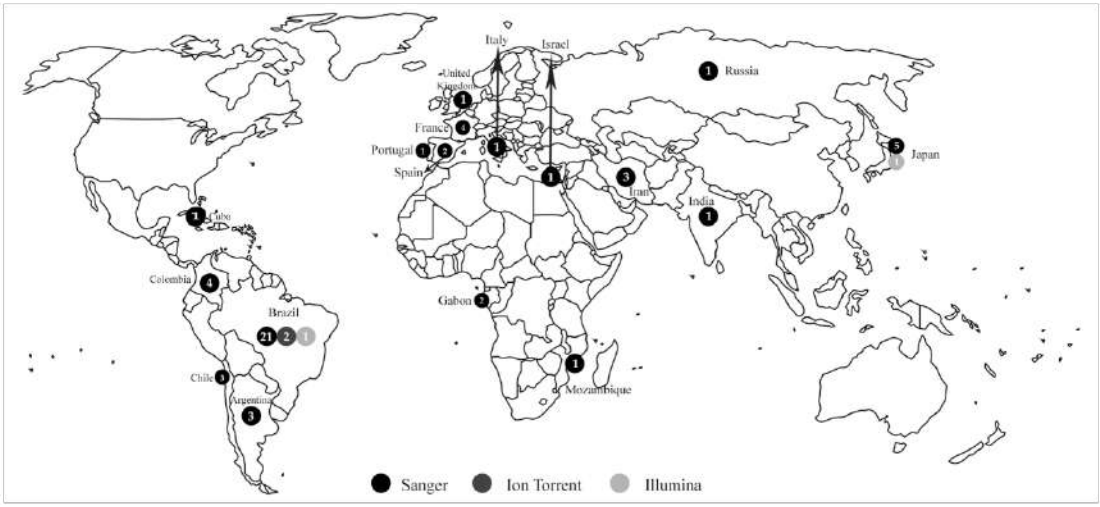
Another important aspect observed in these articles was the difference in the number of generated sequences of each HTLV-1 genome region: 1147 sequences of LTR, 89 sequences of gag, 124 of pol, 777 of env and 1420 of the pX region (Figure 2). From the pX region it is important to highlight that there are 4 different Overlapping Open Reading Frames (ORFs), that encodes the regulatory HTLV-1 proteins, and the HTLV-1 bzip domain gene (hbz) that is transcribed in the antisense direction by a promoter present in the 3'LTR. The quantity of each ORF and hbz sequences generated are as follows: ORF-I (311), ORF-II (54), ORF-III (54), ORF-IV (1153) and hbz (10). In addition, 14 partial genome sequences with the precise regions not described were found and 228 complete HTLV-1 genome sequences were reported.

Fig. 2 Description of the quantities of HTLV-1 sequences generated according to the regions of the genome.



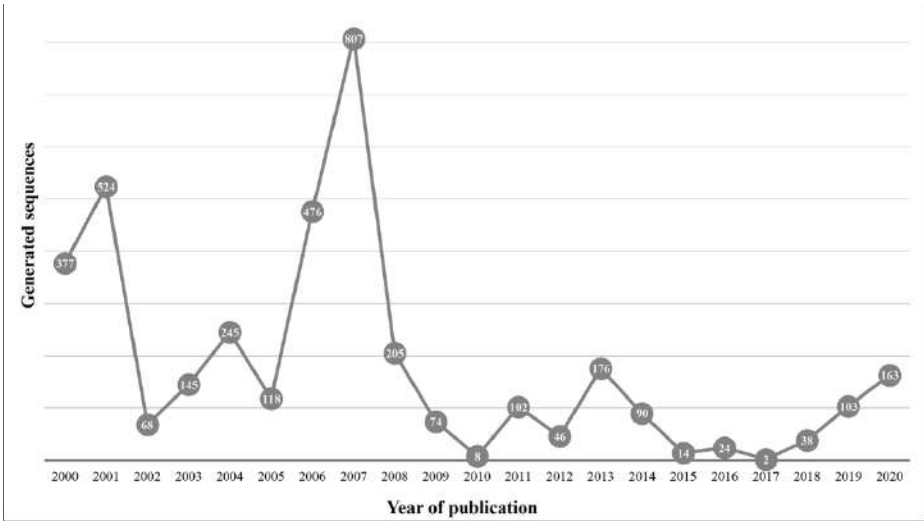
Brazil is the country with the highest number of sequences generated, distributed through 24 sequencing studies, followed by Japan with six. Colombia and France had four studies each, and Argentina and Chile had three each country. In Gabon, Africa, and in Spain, two studies were performed in each one. Other countries like Cuba, India, Israel, Italy, Mozambique, UK, Portugal, and Russia had only one study each. Finally, there was also six articles that did not inform the sequences origin (Figure 3).

Fig. 3 Geographic origin of HTLV-1 sequences and methodology used for sequence generation



It is important to note that, despite being a state-of-the-art methodology employed in genome sequencing of other retroviruses, like HIV [74], no published study using third generation sequencing for HTLV-1 was found. In addition, the most studies refer to information generated more than 10 years ago, in which more than 3,000 sequences were generated, while the most recent studies generated only approximately 800 sequences (Figure 4). Another interesting topic to be highlighted is the lack of clinical information about the patients included in the studies. Of the 60 articles included, 55% did not reported the clinical status of the studied population.

Fig. 4 Number of HTLV-1 sequences generated between 2000 and 2020.



4. Discussion

In the 41 years since the discovery of HTLV-1, no effective therapeutic methods and vaccines have been developed. In addition, it is still not clear what determines the different infection outcomes. During this period, diverse sequencing technologies have become available in order to assist the understanding of the genomes. The central aim of this systematic review was to summarize the different technologies used in HTLV-1 field, in order to guide the decision-making processes on the generation of new HTLV-1 genomes.

The Sanger methodology was the most used technology for generate HTLV-1 sequences, followed by Illumina and Ion Torrent. Among the different sequencing methodologies, all have advantages and disadvantages. The HTLV-1 characteristics, as well as the specific aspects of each methodology must be taken into consideration during technical decision-making processes in order to correctly respond the objective of the study.

One important aspect in HTLV-1 infection is that after infection, the virus integrates into the host cell DNA, and becomes known as a provirus. Therefore, different from HIV, in HTLV-1 infection the circulating viral RNA is not easily detected in the plasma or serum, usually needing additional techniques previously to sequencing, such as PBMC separations and nested-

PCRs [75, 76]. In this sense, the HTLV-1 sample extraction and preparation are an important point to consider during the choice of the sequencing platform.

Among the sequencing technology platforms, despite being the first generation, the Sanger sequencing is considered the gold standard, due to its low error rate. Regarding the second-generation technology, it is important to point out that even when the objective is to sequence larger regions and/or the complete proviral genome, the Ilumina and Ion Torrent technologies produces small sequences reads. This read size, as well as the polymerase chain reaction (PCR) step, can impair the understanding of an essential aspect involved in HTLV-1 infection: the clonality. While in patients with ATLL there is a monoclonal pattern, in patients with IDH, HAM/TSP and AC a polyclonal pattern is found [77, 78]. Therefore, the genome sequencing read size can make it difficult to identify viral quasispecies, and may demonstrate an unreal biological scenario.

It is important to highlight that the sequencing of viral genomes arises from the need to understand the aspects involved in the infection process [79]. Therefore, the use of few and old sequencing methodologies, despite the emergence of more innovative, faster and often less expensive technologies, makes the process of developing better alternatives for infection control and the understanding of viral pathogenesis increasingly distant.

The emergence of new sequencing protocols has contributed to the genome reading in a less time and to the reduction of production costs [80]. Despite that, no article included in this study used more recent technologies, such as the third generation. The MinION and PacBio could be an interestingly alternative, due to the shorter processing time, despite providing sequences with regular quality, when compared to other technologies such as older generations. These methodologies can be useful in HTLV-1 research, increasing the number of sequences available on the platforms, partial and/or complete, and contributing to a better clarification of the virus and host relationship.

In addition to the predominance of the older techniques used, the most studies focus on the sequencing of specific regions of the genome, with few studies generating complete genomes. The LTR and pX regions were the most sequenced. This could be because of the LTR important for the subtyping studies and the fact that pX encoded the HTLV-1 regulatory proteins. In this context, it is relevant to point out that the complete genome sequencing is essential and invaluable to the identification of gene functions and their involvement in disease.

More than that, the knowledge of the HTLV-1 complete genome is essential for the vaccine development.

This systematic review demonstrated a deficit in the amount of HTLV-1 sequences and in this sense, this study has an important limitation, since sequences can be deposited in databases, as GenBank, without being necessarily associated with a published article. However, our data corroborates with an ongoing study carried out by our group that highlight the deficit of HTLV-1 complete genomes available on GenBank. In this study, we verified that only 242 complete HTLV-1 genomes were available in GenBank and most of these sequences did not inform the patient's clinical and epidemiological information (Unpublished results).

On the other hand, the majority of studies provided geographic information about the samples sequenced. Most of them come from endemic regions such as Japan and Brazil. Another country that deserves attention is Colombia. The island of Tumaco, Colombian territory, has a high population density and a very high prevalence of HAM/TSP which is why this region is a focus of study of HTLV-1 [29]. Moreover, a few articles of Africa were found, despite being the largest endemic continent for HTLV-1 in the world [1]. The European continent also contributes to the generation of HTLV-1 sequences, although relatively few articles describe the sequencing, considering the extent and socioeconomic importance of the continent. An important highlight is that some studies did not provided the sequences origin, which makes it difficult to trace a more coherent epidemiological distribution of this scenario. The sum of studies from each country does not correspond to the number of articles included once some studies performed the sequencing of samples from different countries, as we can see in Bandeira et al., 2018.

Interestingly, only 19 articles included in this review were published in the last 10 years, which is equivalent to almost 30% of the total number of studies, revealing that there is still low investment in research of HTLV-1 field. The encouragement of more investments in HTLV-1 studies may contribute to the increased number of HTLV-1 sequences generated by geographic region and this can assist in the understanding of the global and regional distribution of this virus [1].

There are gaps to be filled in relation to information on HTLV-1 infection. Although it was the first human retrovirus described, and has been proven to be associated with the development of diseases, studies on the pathogenesis and treatment of this virus are not

encouraged, and worse, more and more investments in research are decreasing [81], demonstrating how HTLV-1 is still a neglected virus [82, 83]. Thus, more investments in HTLV-1 research, as well as the implementation of worldwide prevention strategies, will be the main motor for the eradication of these infection.

5. Conclusion

The analysis of the articles selected by this systematic review showed that the number of studies sequencing the HTLV-1 genome is much lower when compared to other retroviruses and the most of these studies still opt Sanger sequencing despite the emergence of new methodologies. This demonstrates the lack of investment in this field. It is important to note that Sanger has advantages in relation to others methodologies. However, the NGS methodologies also have technical characteristics that may be important to assist in the understanding of questions that remain unanswered about HTLV-1 infection. Finally, this study can corroborate with the fact that investments in HTLV-1 research are needed, mainly investments in the use of more current methodologies, since they are methodologies that have been developed through lessons learned and improved by previous generation.

Funding This research was funded by Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq), N. 421342/2018-8.

Compliance with ethical standards

Conflicts of Interest The authors declare no conflict of interest. The funder had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript, or in the decision to publish the results.

References

1. Gessain A, Cassar O (2012) Epidemiological Aspects and World Distribution of HTLV-1 Infection. Front Microbiol 3:388. https://doi.org/10.3389/fmicb.2012.00388

2. Poiesz BJ, Ruscetti FW, Gazdar AF, et al (1980) Detection and isolation of type C retrovirus particles from fresh and cultured lymphocytes of a patient with cutaneous T-cell lymphoma. Proc Natl Acad Sci U S A 77:7415–7419. https://doi.org/10.1073/pnas.77.12.7415

3. Gessain A, Barin F, Vernant JC, et al (1985) Antibodies to human T-lymphotropic virus type-I in patients with tropical spastic paraparesis. Lancet 2:407–410. https://doi.org/10.1016/s0140-6736(85)92734-5

4. La Grenade L, Manns A, Fletcher V, et al (1998) Clinical, pathologic, and immunologic features of human T-lymphotrophic virus type I-associated infective dermatitis in children. Arch Dermatol 134:439–444. https://doi.org/10.1001/archderm.134.4.439

5. Yoshida M, Miyoshi I, Hinuma Y (1982) Isolation and characterization of retrovirus from cell lines of human adult T-cell leukemia and its implication in the disease. Proc Natl Acad Sci U S A 79:2031–2035. https://doi.org/10.1073/pnas.79.6.2031

6. Futsch N, Mahieux R, Dutartre H (2017) HTLV-1, the Other Pathogenic Yet Neglected Human Retrovirus: From Transmission to Therapeutic Treatment. Viruses 10:. https://doi.org/10.3390/v10010001

7. Barreto FK, Araújo THA, Rego FF de A, Alcantara LCJ (2017) A Fully Annotated Genome Sequence of Human T-Cell Lymphotropic Virus Type 1 (HTLV-1)

8. Katsuya H, Islam S, Tan BJY, et al (2019) The Nature of the HTLV-1 Provirus in Naturally Infected Individuals Analyzed by the Viral DNA-Capture-Seq Approach. Cell Rep 29:724-735.e4. https://doi.org/10.1016/j.celrep.2019.09.016

9. Sanger F, Coulson AR (1975) A rapid method for determining sequences in DNA by primed synthesis with DNA polymerase. Journal of Molecular Biology 94:441–448. https://doi.org/10.1016/0022-2836(75)90213-2

10. Ansorge WJ (2009) Next-generation DNA sequencing techniques. N Biotechnol 25:195–203. https://doi.org/10.1016/j.nbt.2008.12.009

11. Kchouk M, Gibrat J-F, Elloumi M (2017) Generations of Sequencing Technologies: From First to Next Generation. Biology and Medicine 09: https://doi.org/10.4172/0974-8369.1000395

12. Yoshinaga Y, Daum C, He G, O'Malley R (2018) Genome Sequencing. Methods Mol Biol 1775:37–52. https://doi.org/10.1007/978-1-4939-7804-5_4

13. Rhoads A, Au KF (2015) PacBio Sequencing and Its Applications. Genomics Proteomics Bioinformatics 13:278–289. https://doi.org/10.1016/j.gpb.2015.08.002

14. Jain M, Olsen HE, Paten B, Akeson M (2016) The Oxford Nanopore MinION: delivery of nanopore sequencing to the genomics community. Genome Biol 17:239. https://doi.org/10.1186/s13059-016-1103-0

15. Leclercq I, Mortreux F, Cavrois M, et al (2000) Host sequences flanking the human T-cell leukemia virus type 1 provirus in vivo. J Virol 74:2305–2312. https://doi.org/10.1128/jvi.74.5.2305-2312.2000

16. Seki M, Higashiyama Y, Mizokami A, et al (2000) Up-regulation of human T lymphotropic virus type 1 (HTLV-1) tax/rex mRNA in infected lung tissues. Clin Exp Immunol 120:488–498. https://doi.org/10.1046/j.1365-2249.2000.01237.x

17. Nakane S, Shirabe S, Moriuchi R, et al (2000) Comparative molecular analysis of HTLV-I proviral DNA in HTLV-I infected members of a family with a discordant HTLV-I-associated myelopathy in monozygotic twins. J Neurovirol 6:275–283. https://doi.org/10.3109/13550280009030753

18. Morozov VA, Lagaye S, Taylor GP, et al (2000) Chimeric Matrix Proteins Encoded by Defective Proviruses with Large Internal Deletions in Human T-Cell Leukemia Virus Type 1-Infected Humans. J Virol 74:3933–3940

19. Sonoda S, Li HC, Cartier L, et al (2000) Ancient HTLV type 1 provirus DNA of Andean mummy. AIDS Res Hum Retroviruses 16:1753–1756. https://doi.org/10.1089/08892220050193263

20. Caterino-de-Araujo A, Favero A, de los Santos-Fortuna E, et al (2000) HTLV-I/HTLV-II coinfection in an AIDS patient from São Paulo, Brazil. AIDS Res Hum Retroviruses 16:715–719. https://doi.org/10.1089/088922200308710

21. Mortreux F, Leclercq I, Gabet AS, et al (2001) Somatic mutation in human T-cell leukemia virus type 1 provirus and flanking cellular sequences during clonal expansion in vivo. J Natl Cancer Inst 93:367–377. https://doi.org/10.1093/jnci/93.5.367

22. Furukawa Y, Kubota R, Tara M, et al (2001) Existence of escape mutant in HTLV-I tax during the development of adult T-cell leukemia. Blood 97:987–993. https://doi.org/10.1182/blood.v97.4.987

23. Okazaki S, Moriuchi R, Yosizuka N, et al (2001) HTLV-1 proviruses encoding non-functional TAX in adult T-cell leukemia. Virus Genes 23:123–135. https://doi.org/10.1023/a:1011840918149

24. Dominguez MC, Castillo A, Cabrera J, et al (2002) Envelope sequence variation and phylogenetic relations of human T cell lymphotropic virus type 1 from endemic areas of Colombia. AIDS Res Hum Retroviruses 18:887–890. https://doi.org/10.1089/08892220260190371

25. Manca N, Perandin F, De Simone N, et al (2002) Detection of HTLV-I tax-rex and pol gene sequences of thymus gland in a large group of patients with myasthenia gravis. J Acquir Immune Defic Syndr 29:300–306. https://doi.org/10.1097/00126334-200203010-00012

26. Ramirez E, Cartier L, Villota C, Fernandez J (2002) Genetic characterization and phylogeny of human T-cell lymphotropic virus type I from Chile. Virus Res 84:135–149. https://doi.org/10.1016/s0168-1702(02)00005-9

27. Ramirez E, Fernandez J, Cartier L, et al (2003) Defective human T-cell lymphotropic virus type I (HTLV-I) provirus in seronegative tropical spastic paraparesis/HTLV-I-associated myelopathy (TSP/HAM) patients. Virus Res 91:231–239. https://doi.org/10.1016/s0168-1702(02)00276-9

28. Toro C, Rodés B, Poveda E, Soriano V (2003) Rapid development of subacute myelopathy in three organ transplant recipients after transmission of human T-cell lymphotropic virus type I from a single donor. Transplantation 75:102–104. https://doi.org/10.1097/00007890-200301150-00019

29. Balcázar N, Sánchez GI, Garcia-Vallejo F (2003) Sequence and phylogenetic analysis of human T cell lymphotropic virus type 1 from Tumaco, Colombia. Memórias do Instituto Oswaldo Cruz 98:641–648. https://doi.org/10.1590/S0074-02762003000500010

30. Leclercq I, Mortreux F, Rabaaoui S, et al (2003) Naturally occurring substitutions of the human T-cell leukemia virus type 1 3' LTR influence strand-transfer reaction. J Virol Methods 109:105–117. https://doi.org/10.1016/s0166-0934(03)00052-1

31. Furukawa Y, Usuku K, Izumo S, Osame M (2004) Human T cell lymphotropic virus type I (HTLV-I) p12I is dispensable for HTLV-I transmission and maintenance of infection in vivo. AIDS Res Hum Retroviruses 20:1092–1099. https://doi.org/10.1089/aid.2004.20.1092

32. Chávez M, Domínguez MC, Blank A, et al (2004) Reconstrucción de la evolución molecular de la infección actual por el virus linfotrópico humano tipo I en Colombia. Biomédica (Bogotá) 20–32

33. Vallinoto ACR, Muto NA, Pontes GS, et al (2004) Serological and molecular evidence of HTLV-I infection among Japanese immigrants living in the Amazon region of Brazil. Jpn J Infect Dis 57:156–159

34. Morozov VA, Syrtsev AV, Ellerbrok H, et al (2005) Mycosis fungoides in European Russia: no antibodies to human T cell leukemia virus type I structural proteins, but virus-like sequences in blood and saliva. Intervirology 48:362–371. https://doi.org/10.1159/000086063

35. Ohkura S, Yamashita M, Ishida T, et al (2005) Phylogenetic heterogeneity of new HTLV type 1 isolates from southern India in subgroup A. AIDS Res Hum Retroviruses 21:325–330. https://doi.org/10.1089/aid.2005.21.325

36. Iñiguez AM, Otsuki K, Magalhães GP, et al (2005) Genetic markers on the HTLV-1 p12I protein sequences from Brazilian HAM/TSP patients and asymptomatic HTLV-1 carrier isolates. AIDS Res Hum Retroviruses 21:580–582. https://doi.org/10.1089/aid.2005.21.580

37. Capdepont S, Londos-Gagliardi D, Joubert M, et al (2005) New insights in HTLV-I phylogeny by sequencing and analyzing the entire envelope gene. AIDS Res Hum Retroviruses 21:28–42. https://doi.org/10.1089/aid.2005.21.28

38. Gastaldello R, Otsuki K, Barbas MG, et al (2005) Molecular evidence of HTLV-1 intrafamilial transmission in a non-endemic area in Argentina. J Med Virol 76:386–390. https://doi.org/10.1002/jmv.20370

39. Kashima S, Alcantara LC, Takayanagui OM, et al (2006) Distribution of human T cell lymphotropic virus type 1 (HTLV-1) subtypes in Brazil: genetic characterization of LTR and tax region. AIDS Res Hum Retroviruses 22:953–959. https://doi.org/10.1089/aid.2006.22.953

40. Shohat M, Shohat B, Mimouni D, et al (2006) Human T-cell lymphotropic virus type 1 provirus and phylogenetic analysis in patients with mycosis fungoides and their family relatives. Br J Dermatol 155:372–378. https://doi.org/10.1111/j.1365-2133.2006.07312.x

41. Vallinoto ACR, Pontes GS, Muto NA, et al (2006) Identification of human T-cell lymphotropic virus infection in a semi-isolated Afro-Brazilian quilombo located in the Marajó Island (Pará, Brazil). Memórias do Instituto Oswaldo Cruz 101:103–105. https://doi.org/10.1590/S0074-02762006000100020

42. Souza LA, Lopes IGL, Maia EL, et al (2006) Molecular characterization of HTLV-1 among patients with tropical spastic paraparesis/HTLV-1 associated myelopathy in Belém,

Pará. Revista da Sociedade Brasileira de Medicina Tropical 39:504–506. https://doi.org/10.1590/S0037-86822006000500017

43. Kim FJ, Lavanya M, Gessain A, et al (2006) Intrahost variations in the envelope receptor-binding domain (RBD) of HTLV-1 and STLV-1 primary isolates. Retrovirology 3:29. https://doi.org/10.1186/1742-4690-3-29

44. Miyazaki M, Yasunaga J-I, Taniguchi Y, et al (2007) Preferential selection of human T-cell leukemia virus type 1 provirus lacking the 5' long terminal repeat during oncogenesis. J Virol 81:5714–5723. https://doi.org/10.1128/JVI.02511-06

45. Kubota R, Hanada K, Furukawa Y, et al (2007) Genetic stability of human T lymphotropic virus type I despite antiviral pressures by CTLs. J Immunol 178:5966–5972. https://doi.org/10.4049/jimmunol.178.9.5966

46. Mota AC de A, Van Dooren S, Fernandes FM de C, et al (2007) The close relationship between South African and Latin American HTLV type 1 strains corroborated in a molecular epidemiological study of the HTLV type 1 isolates from a blood donor cohort. https://doi.org/10.1089/aid.2006.0203

47. Eirin ME, Dilernia DA, Berini CA, et al (2008) Divergent strains of human T-lymphotropic virus type 1 (HTLV-1) within the Cosmopolitan subtype in Argentina. AIDS Res Hum Retroviruses 24:1237–1244. https://doi.org/10.1089/aid.2008.0024

48. Etenna SL-D, Caron M, Besson G, et al (2008) New insights into prevalence, genetic diversity, and proviral load of human T-cell leukemia virus types 1 and 2 in pregnant women in Gabon in equatorial central Africa. J Clin Microbiol 46:3607–3614. https://doi.org/10.1128/JCM.01249-08

49. Magalhães TS de, Miranda ACAM, Alcantara LCJ, et al (2008) Phylogenetic and molecular analysis of HTLV-1 isolates from a medium sized town in northern of Brazil: tracing a common origin of the virus from the most endemic city in the country

50. Rego FF de A, Alcantara LCJ, Moura Neto JP de, et al (2008) HTLV type 1 molecular study in Brazilian villages with African characteristics giving support to the post-Columbian introduction hypothesis.

51. Gastaldello R, Iñiguez AM, Otsuki K, et al (2008) HTLV type 1 genetic types among native descendants in Argentina. AIDS Res Hum Retroviruses 24:1139–1146. https://doi.org/10.1089/aid.2007.0299

52. Eguchi K, Fujii H, Oshima K, et al (2009) Human T-lymphotropic virus type 1 (HTLV-1) genetic typing in Kakeroma Island, an island at the crossroads of the ryukyuans and Wajin in Japan, providing further insights into the origin of the virus in Japan. J Med Virol 81:1450–1456. https://doi.org/10.1002/jmv.21540

53. Santos EL dos, Tamegão-Lopes B, Machado LFA, et al (2009) Molecular characterization of HTLV-1/2 among blood donors in Belém, State of Pará: first description of HTLV-2b subtype in the Amazon region. Revista da Sociedade Brasileira de Medicina Tropical 42:271–276. https://doi.org/10.1590/S0037-86822009000300007

54. Rego FF de A, Miranda AM, Santos E de S, et al (2010) Seroprevalence and molecular epidemiology of HTLV-1 isolates from HIV-1 co-infected women in Feira de Santana, Bahia, Brazil. https://doi.org/10.1089/aid.2009.0298

55. Vicente ACP, Gudo ES, Iñiguez AM, et al (2011) Genetic characterization of human T-cell lymphotropic virus type 1 in Mozambique: transcontinental lineages drive the HTLV-1 endemic. https://doi.org/10.1371/journal.pntd.0001038

56. Salcedo-Cifuentes M, Domínguez MC, García-Vallejo F (2011) [Genome epidemiology and tropical spastic paraparesis associated with human T-cell lymphotropic virus type 1]. Rev Panam Salud Publica 30:422–430

57. Pádua E, Rodés B, Pérez-Piñar T, et al (2011) Molecular characterization of human T cell leukemia virus type 1 subtypes in a group of infected individuals diagnosed in Portugal and Spain. AIDS Res Hum Retroviruses 27:317–322. https://doi.org/10.1089/aid.2010.0195

58. Zanella L, Otsuki K, Marin MA, et al (2012) Complete Genome Sequence of Central Africa Human T-Cell Lymphotropic Virus Subtype 1b. J Virol 86:12451. https://doi.org/10.1128/JVI.02258-12

59. Magri MC, Brigido LF de M, Rodrigues R, et al (2012) Tax Gene Characterization of Human T-Lymphotropic Virus Type 1 Strains from Brazilian HIV-Coinfected Patients. AIDS Res Hum Retroviruses 28:1775–1778. https://doi.org/10.1089/aid.2011.0389

60. Magri MC, de Macedo Brigido LF, Rodrigues R, et al (2012) Phylogenetic and Similarity Analysis of HTLV-1 Isolates from HIV-Coinfected Patients from the South and Southeast Regions of Brazil. AIDS Res Hum Retroviruses 28:110–114. https://doi.org/10.1089/aid.2011.0117

61. Miranda ACAM, Barreto FK, Amarante MF de C, et al (2013) Molecular characterization of HTLV-1 gp46 glycoprotein from health carriers and HAM/TSP infected individuals. https://doi.org/10.1186/1743-422X-10-75

62. Mota-Miranda ACA, Barreto FK, Baptista E, et al (2013) Molecular study of HBZ and gp21 human T cell leukemia virus type 1 proteins isolated from different clinical profile infected individuals. AIDS Res Hum Retroviruses 29:1370–1372. https://doi.org/10.1089/AID.2013.0015

63. Machado LY, Navea LM, Díaz HM, et al (2013) Phylogenetic analysis of human T cell lymphotropic virus type 1 isolated from Cuban individuals. AIDS Res Hum Retroviruses 29:1168–1172. https://doi.org/10.1089/AID.2012.0225

64. Bandeira LM, Uehara SNO, Asato MA, et al (2015) High prevalence of HTLV-1 infection among Japanese immigrants in non-endemic area of Brazil. PLoS Negl Trop Dis 9:e0003691. https://doi.org/10.1371/journal.pntd.0003691

65. Rafatpanah H, Torkamani M, Valizadeh N, et al (2016) Prevalence and phylogenetic analysis of HTLV-1 in a segregated population in Iran. J Med Virol 88:1247–1253. https://doi.org/10.1002/jmv.24448

66. de Aguiar SA, de Souza França SA, Santana BB, et al (2017) Human T-lymphotropic virus 1aA circulation and risk factors for sexually transmitted infections in an Amazon geographic area with lowest human development index (Marajó Island, Northern Brazil). BMC Infect Dis 17:. https://doi.org/10.1186/s12879-017-2859-x

67. Bandeira LM, Uehara SNO, Puga MAM, et al (2018) HTLV-1 intrafamilial transmission among Japanese immigrants in Brazil. J Med Virol 90:351–357. https://doi.org/10.1002/jmv.24938

68. Ribeiro IP, Kozlowski AG, Dias de Matos MA, et al (2018) HTLV-1 and -2 in a first-time blood donor population in Northeastern Brazil: Prevalence, molecular characterization, and evidence of intrafamilial transmission. J Med Virol 90:1651–1657. https://doi.org/10.1002/jmv.25231

69. Mirhosseini A, Mohareri M, Arab R, et al (2019) Complete sequence of human T cell leukemia virus type 1 in ATLL patients from Northeast Iran, Mashhad revealed a prematurely terminated protease and an elongated pX open reading frame III. Infect Genet Evol 73:460–469. https://doi.org/10.1016/j.meegid.2019.05.012

70. Campos KR, Caterino-de-Araujo A (2020) Provirus Mutations of Human T-Lymphotropic Virus 1 and 2 (HTLV-1 and HTLV-2) in HIV-1-Coinfected Individuals. mSphere 5:. https://doi.org/10.1128/mSphere.00923-20

71. Pessôa R, Watanabe JT, Nukui Y, et al (2014) Molecular Characterization of Human T-Cell Lymphotropic Virus Type 1 Full and Partial Genomes by Illumina Massively Parallel Sequencing Technology. PLoS One 9:. https://doi.org/10.1371/journal.pone.0093374

72. Rego FF de A, Oliveira T de, Giovanetti M, et al (2016) Deep Sequencing Analysis of Human T Cell Lymphotropic Virus Type 1 Long Terminal Repeat 5' Region from Patients with Tropical Spastic Paraparesis/Human T Cell Lymphotropic Virus Type 1-Associated Myelopathy and Asymptomatic Carriers. https://doi.org/10.1089/aid.2015.0273

73. Araújo THA, Barreto FK, Menezes ADL, et al (2020) Complete genome sequence of human T-cell lymphotropic type 1 from patients with different clinical profiles, including infective dermatitis. Infect Genet Evol 79:104166. https://doi.org/10.1016/j.meegid.2019.104166

74. Nguyen Quang N, Goudey S, Ségéral E, et al (2020) Dynamic nanopore long-read sequencing analysis of HIV-1 splicing events during the early steps of infection. Retrovirology 17:. https://doi.org/10.1186/s12977-020-00533-1

75. Cabral F, Arruda LB, de Araújo ML, et al (2012) Detection of human T-cell lymphotropic virus type 1 in plasma samples. Virus Res 163:87–90. https://doi.org/10.1016/j.virusres.2011.08.014

76. Demontis MA, Sadiq MT, Golz S, Taylor GP (2015) HTLV-1 viral RNA is detected rarely in plasma of HTLV-1 infected subjects. J Med Virol 87:2130–2134. https://doi.org/10.1002/jmv.24264

77. Bangham CRM, Cook LB, Melamed A (2014) HTLV-1 clonality in adult T-cell leukaemia and non-malignant HTLV-1 infection. Seminars in Cancer Biology 26:89–98. https://doi.org/10.1016/j.semcancer.2013.11.003

78. Wattel E, Vartanian JP, Pannetier C, Wain-Hobson S (1995) Clonal expansion of human T-cell leukemia virus type I-infected cells in asymptomatic and symptomatic carriers without malignancy. J Virol 69:2863–2868

79. Capobianchi MR, Giombini E, Rozera G (2013) Next-generation sequencing technology in clinical virology. Clin Microbiol Infect 19:15–22. https://doi.org/10.1111/1469-0691.12056

80. Heather JM, Chain B (2016) The sequence of sequencers: The history of sequencing DNA. Genomics 107:1–8. https://doi.org/10.1016/j.ygeno.2015.11.003

81. Martin F, Tagaya Y, Gallo R (2018) Time to eradicate HTLV-1: an open letter to WHO. The Lancet 391:1893–1894. https://doi.org/10.1016/S0140-6736(18)30974-7

82. Cao F, Ji Y, Huang R, et al (2000) Sequence Note: Nucleotide Sequence Analyses of Partial envgp46 Gene of Human T-Lymphotropic Virus Type I from Inhabitants of Fujian Province in Southeast China. AIDS Research and Human Retroviruses 16:921–923. https://doi.org/10.1089/08892220050042855

83. S.Chou K, Okayama A, Tachibana N, et al (1995) Nucleotide sequence analysis of a full-length human T-cell leukemia virus type I from adult T-cell leukemia cells: A prematurely terminated PX open reading frame II. International Journal of Cancer 60:701–706. https://doi.org/10.1002/ijc.2910600522

REFERÊNCIAS

AL-MAWSAWI, L. Q. *et al.* Discovery of a small-molecule HIV-1 integrase inhibitor-binding site. **Proceedings National Academy of Science.** v. 103, n. 26, p. 10080-5, jun. 2006.

BARRETO, F. K. *et al.* Analyses of HTLV-1 sequences suggest interaction between ORF-I mutations and HAM/TSP outcome. **Infection, Genetics and Evolution : journal of molecular epidemiology and evolutionary genetics in infectious diseases**, v. 45, p. 420–425, nov. 2016.

BARRETO, F.K. *et al.* A Fully Annotated Genome Sequence of Human T-Cell Lymphotropic Virus Type 1 (HTLV-1). **Journal of Bioinformatics, Computational and Systems Biology.** v. 1, n. 1, p. 3, 2017.

BATISTA, E. S. *et al.* HTLV-1 proviral load in infective dermatitis associated with HTLV-1 does not increase after the development of HTLV-1-associated myelopathy/tropical spastic paraparesis and does not decrease after IDH remission. **PLoS Neglected Tropical Diseases**, v. 13, n. 12, p. e0007705, 2019.

BINDHU, M.; NAIR, A.; LAIRMORE, M. D. Role of accessory proteins of HTLV-1 in viral replication, T cell activation, and cellular gene expression. **Frontiers in Bioscience: A Journal and Virtual Library**, v. 9, p. 2556–2576, 1 set. 2004.

BITTENCOURT, A. L. *et al.* Human T-cell lymphotropic virus type 1 infection among pregnant women in northeastern Brazil. **Journal of Acquired Immune Deficiency Syndromes (1999)**, v. 26, n. 5, p. 490–494, 15 abr. 2001.

BITTENCOURT, A. L. *et al.* Adult T-cell leukemia/lymphoma in Bahia, Brazil: analysis of prognostic factors in a group of 70 patients. **American Journal of Clinical Pathology**, v. 128, n. 5, p. 875–882, nov. 2007.

BITTENCOURT, A. L.; OLIVEIRA, M. DE F. Dermatite infecciosa associada ao HTLV-I (DIH) infanto-juvenil e do adulto. **Anais Brasileiros de Dermatologia**, v. 80, p. S364–S369, dez. 2005.

BURKE, W. *et al.*. Recommendations for follow-up care of individuals with an inherited predisposition to cancer. I. Hereditary nonpolyposis colon cancer. Cancer Genetics Studies Consortium. **JAMA**, v. 277, n. 11, p. 915–919, 19 mar. 1997.

CALATTINI, S. *et al.* Discovery of a new human T-cell lymphotropic virus (HTLV-3) in Central Africa. **Retrovirology**, v. 2, p. 30, 9 maio 2005.

CARNEIRO-PROIETTI, A. B. F. *et al.* Infection and disease caused by the human T cell lymphotropic viruses type I and II in Brazil. **Revista da Sociedade Brasileira de Medicina Tropical**, v. 35, n. 5, p. 499–508, out. 2002.

CARPENTIER, A. *et al.* Modes of Human T Cell Leukemia Virus Type 1 Transmission, Replication and Persistence. **Viruses**, v. 7, n. 7, p. 3603–3624, 7 jul. 2015a.

CARPENTIER, A. *et al.* Modes of Human T Cell Leukemia Virus Type 1 Transmission, Replication and Persistence. **Viruses**, v. 7, n. 7, p. 3603–3624, 7 jul. 2015b.

CATALAN-SOARES, B. C.; PROIETTI, F. A.; CARNEIRO-PROIETTI, A. B. DE F. Os vírus linfotrópicos de células T humanos (HTLV) na última década (1990-2000): aspectos epidemiológicos. **Revista Brasileira de Epidemiologia**, v. 4, n. 2, p. 81–95, ago. 2001.

DELAMARRE, L. *et al.* The HTLV-I envelope glycoproteins: structure and functions. **Journal of Acquired Immune Deficiency Syndromes and Human Retrovirology,** v. 13 Suppl 1, p. S85-91, 1996.

DOURADO, I. *et al.* HTLV-I in the General Population of Salvador, Brazil: A City With African Ethnic and Sociodemographic Characteristics. **JAIDS Journal of Acquired Immune Deficiency Syndromes**, v. 34, n. 5, p. 527–531, dez. 2003.

EDWARDS, D. *et al.* Orf-I and Orf-II-Encoded Proteins in HTLV-1 Infection and Persistence. **Viruses**, v. 3, n. 6, p. 861–885, 17 jun. 2011.

ETOH, K. *et al.* Persistent clonal proliferation of human tlymphotropic virus type i-infected cells in vivo. **Cancer Research**, 57: 4862-4867, 1997.

FAN, N. *et al.* Infection of peripheral blood mononuclear cells and cell lines by cell-free human T-cell lymphoma/leukemia virus type I. **Journal of Clinical Microbiology**, v. 30, n. 4, p. 905–910, abr. 1992.

GALLO, R. C.; SLISKI, A.; WONG-STAAL, F. Origin of human T-cell leukaemia-lymphoma virus. **Lancet**, v. 2, n. 8356, p. 962–963, 22 out. 1983.

GAUDRAY, G. *et al.* The complementary strand of the human T-cell leukemia virus type 1 RNA genome encodes a bZIP transcription factor that down-regulates viral transcription. **Journal of Virology**, v. 76, n. 24, p. 12813–12822, dez. 2002.

GESSAIN, A. *et al.* Antibodies to human T-lymphotropic virus type-I in patients with tropical spastic paraparesis. **Lancet**, v. 2, n. 8452, p. 407–410, 24 ago. 1985.

GESSAIN, A.; CASSAR, O. Epidemiological Aspects and World Distribution of HTLV-1 Infection. **Frontiers in Microbiology**, v. 3, 15 nov. 2012a.

GESSAIN, A.; CASSAR, O. Epidemiological Aspects and World Distribution of HTLV-1 Infection. **Frontiers in Microbiology**, v. 3, 15 nov. 2012b.

GESSAIN, A.; GALLO, R. C.; FRANCHINI, G. Low degree of human T-cell leukemia/lymphoma virus type I genetic drift in vivo as a means of monitoring viral transmission and movement of ancient human populations. **Journal of Virology**, v. 66, n. 4, p. 2288–2295, abr. 1992.

GILLET, N. A. *et al.* Strongyloidiasis and Infective Dermatitis Alter Human T Lymphotropic Virus-1 Clonality in vivo. **PLoS Pathogens**, v. 9, n. 4, 4 abr. 2013.

GONÇALVES, D. U. *et al.* Epidemiology, treatment, and prevention of human T-cell leukemia virus type 1-associated diseases. **Clinical Microbiology Reviews**, v. 23, n. 3, p. 577–589, jul. 2010.

HANCHARD, B. *et al.* Childhood infective dermatitis evolving into adult T-cell leukaemia after 17 years. **Lancet**, v. 338, n. 8782–8783, p. 1593–1594, 21 dez. 1991.

IGAKURA, T. *et al.* Spread of HTLV-I between lymphocytes by virus-induced polarization of the cytoskeleton. **Science**, v. 299, n. 5613, p. 1713–1716, 14 mar. 2003.

JIN, J.; SHERER, N.; MOTHES, W. Surface Transmission or Polarized Egress? Lessons Learned from HTLV Cell-to-Cell Transmission. **Viruses**, v. 2, n. 2, p. 601–605, 10 fev. 2010.

JONES, K. S. *et al.* Cell-free HTLV-1 infects dendritic cells leading to transmission and transformation of CD4(+) T cells. **Nature Medicine**, v. 14, n. 4, p. 429–436, abr. 2008.

KALYANARAMAN, V. S. *et al.* A new subtype of human T-cell leukemia virus (HTLV-II) associated with a T-cell variant of hairy cell leukemia. **Science**, v. 218, n. 4572, p. 571–573, 5 nov. 1982.

KORALNIK, I. J. *et al.* Protein isoforms encoded by the pX region of human T-cell

leukemia/lymphotropic virus type I. **Proceedings of the National Academy of Sciences of the United States of America**, v. 89, n. 18, p. 8813–8817, 15 set. 1992.

LA GRENADE, L. HTLV-I-associated infective dermatitis: past, present, and future. **Journal of Acquired Immune Deficiency Syndromes and Human Retrovirology,** v. 13 Suppl 1, p. S46-49, 1996.

LU, H.; GIORDANO, F.; NING, Z. Oxford Nanopore MinION Sequencing and Genome Assembly. **Genomics, Proteomics & Bioinformatics**, v. 14, n. 5, p. 265–279, out. 2016.

MATSUOKA, M.; JEANG, K.-T. Human T-cell leukaemia virus type 1 (HTLV-1) infectivity and cellular transformation. **Nature Reviews. Cancer**, v. 7, n. 4, p. 270–280, abr. 2007.

MESNARD, J.-M.; BARBEAU, B.; DEVAUX, C. HBZ, a new important player in the mystery of adult T-cell leukemia. **Blood**, v. 108, n. 13, p. 3979–3982, 15 dez. 2006.

MITCHELL, R. S. *et al.* Retroviral DNA Integration: ASLV, HIV, and MLV Show Distinct Target Site Preferences. **PLoS Biology**, v. 2, n. 8, ago. 2004.

NEJMEDDINE, M.; BANGHAM, C. R. M. The HTLV-1 Virological Synapse. **Viruses**, v. 2, n. 7, p. 1427–1447, 7 jul. 2010.

NETO, W. K. *et al.* Correlation between LTR point mutations and proviral load levels among Human T cell Lymphotropic Virus type 1 (HTLV-1) asymptomatic carriers. **Virology Journal**, v. 8, p. 535, 13 dez. 2011.

OSAME, M. *et al.* HTLV-I associated myelopathy, a new clinical entity. **Lancet**, v. 1, n. 8488, p. 1031–1032, 3 maio 1986.

PISE-MASISON, C. A. *et al.* Co-dependence of HTLV-1 p12 and p8 Functions in Virus Persistence. **PLoS Pathogens**, v. 10, n. 11, 6 nov. 2014.

POIESZ, B. J. *et al.* Detection and isolation of type C retrovirus particles from fresh and cultured lymphocytes of a patient with cutaneous T-cell lymphoma. **Proceedings of the National Academy of Sciences of the United States of America**, v. 77, n. 12, p. 7415–7419, dez. 1980.

PRIMO, J. *et al.* High HTLV-1 proviral load, a marker for HTLV-1 associated myelopathy/tropical spastic paraparesis, is also detected in patients with infective dermatitis associated with HTLV-1. **Brazilian Journal of Medical and Biological Research**, v. 42, n. 8, p. 761–764, ago. 2009.

PRIMO, J. R. L. vInfective dermatitis and human T cell lymphotropic virus type 1-associated myelopathy/tropical spastic paraparesis in childhood and adolescence. **Clinical Infectious Diseases**, v. 41, n. 4, p. 535–541, 15 ago. 2005.

PROIETTI, F. A. Global epidemiology of HTLV-I infection and associated diseases. **Oncogene**, v. 24, n. 39, p. 6058–6068, 5 set. 2005.

PROOYEN, N. V. *et al.* Human T-cell leukemia virus type 1 p8 protein increases cellular conduits and virus transmission. **Proceedings of the National Academy of Sciences**, v. 107, n. 48, p. 20738–20743, 30 nov. 2010.

SCHADT, E. E.; TURNER, S.; KASARSKIS, A. A window into third-generation sequencing. **Human Molecular Genetics**, v. 19, n. R2, p. R227-240, 15 out. 2010.

SEIKI, M.; HATTORI, S.; YOSHIDA, M. Human adult T-cell leukemia virus: molecular cloning of the provirus DNA and the unique terminal structure. **Proceedings of the National Academy of Sciences of the United States of America**, v. 79, n. 22, p.

6899–6902, nov. 1982.

SHIMOYAMA, M. Diagnostic criteria and classification of clinical subtypes of adult T-cell leukaemia-lymphoma. A report from the Lymphoma Study Group (1984-87). **British Journal of Haematology**, v. 79, n. 3, p. 428–437, nov. 1991.

SOCORRO DE ALMEIDA VIANA, M. DE N. DO *et al.*. Stability of the HTLV-1 glycoprotein 46 (gp46) gene in an endemic region of the Brazilian Amazon and the presence of a significant mutation (N93D) in symptomatic patients. **Virology Journal**, v. 15, 2 maio 2018.

TAKEMOTO, S. *et al.*A novel diagnostic method of adult T-cell leukemia: monoclonal integration of human T-cell lymphotropic virus type I provirus DNA detected by inverse polymerase chain reaction. **Blood**, 84: 3080-3085, 1994.

VAN DOOREN, S. *et al.* Evidence for a post-Columbian introduction of human T-cell lymphotropic virus [type I] [corrected] in Latin America. **The Journal of General Virology**, v. 79 (Pt 11), p. 2695–2708, nov. 1998.

VERDONCK, K. *et al.*. Human T-lymphotropic virus 1: recent knowledge about an ancient infection. **The Lancet. Infectious Diseases**, v. 7, n. 4, p. 266–281, abr. 2007.

WATTEL, E. *et al.* Clonal expansion of human T-cell leukemia virus type I-infected cells in asymptomatic and symptomatic carriers without malignancy. **Journal of Virology**, v. 69, n. 5, p. 2863–2868, maio 1995.

YOSHIDA, M.; MIYOSHI, I.; HINUMA, Y. Isolation and characterization of retrovirus from cell lines of human adult T-cell leukemia and its implication in the disease. **Proceedings of the National Academy of Sciences of the United States of America**, v. 79, n. 6, p. 2031–2035, mar. 1982.

YOUNIS, I.; GREEN, P. L. The human T-cell leukemia virus Rex protein. **Frontiers in Bioscience,** v. 10, p. 431–445, 1 jan. 2005.

ZALA, C. *et al.*Human T-cell lymphotropic virus type I disease in Argentine intravenous drug users with human immunodeficiency virus type 1 infection. **Journal of Acquired Immune Deficiency Syndromes.** v. 7, n. 8, p. 870-1, Aug 1994.