NEXRES 3 (2026) 101110

Contents lists available at ScienceDirect

Next Research

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

Cross-platform protocol for HTLV-1 whole-genome sequencing using nanopore and Illumina technologies

Ana Carolina Marinho Monteiro Lima [a,b], Laise de Moraes [a], Rebeca Leão Amorim [c], Marina Silveira Cucco [d], Roberta Muniz Luz Silva [c], Marta Giovanetti [e,f], Filipe Ferreira de Almeida Rego [g], Thessika Hialla Almeida Araújo [g], Vagner Fonseca [h,i], Ricardo Khouri [a,b,d], Luiz Carlos Junior Alcantara [f,†], Luciane Amorim Santos [a,b,g,†], Fernanda Khouri Barreto [c,†,*]

[a] Laboratório de Medicina e Saúde Pública de Precisão, Instituto Gonçalo Moniz, Fundação Oswaldo Cruz, 121 Waldemar Falcão Street Candeal, Salvador, BA 40296-710, Brazil
[b] Faculdade de Medicina da Bahia, Universidade Federal da Bahia, Largo Terreiro de Jesus, Pelourinho, Salvador, BA 40026-010, Brazil
[c] Instituto Multidisciplinar em Saúde, Universidade Federal da Bahia, 58 Hormindo Barros Street Candeias, Vitória da Conquista, BA 45029-094, Brazil
[d] Plataforma de Vigilância Molecular, Instituto Gonçalo Moniz, Fundação Oswaldo Cruz, 121 Waldemar Falcão Street Candeal, Salvador, BA 40296-710, Brazil
[e] Sciences and Technologies for Sustainable Development and One Health, Università Campus Bio-Medico di Roma, Via Álvaro del Portillo 21, 00128 Rome, RM, Italy
[f] Instituto René Rachou, Fundação Oswaldo Cruz, Av. Augusto de Lima 1715 - Barro Preto, Belo Horizonte, MG, 30190-002, Brazil
[g] Escola Bahiana de Medicina e Saúde Pública, Av. Dom João VI 275 - Brotas, Salvador, BA 40290-000, Brazil
[h] Departamento de Ciências Exatas e da Terra, Universidade do Estado da Bahia, Rua Silveira Martins, 2555 Cabula, Salvador, BA 41.150-000, Brazil
[i] Centre for Epidemic Response and Innovation (CERI), School of Data Science and Computational Thinking, Stellenbosch 7600, South Africa

ARTICLE INFO

Keywords:
Human T-lymphotropic virus type 1
Whole Genome Sequencing
Illumina Sequencing
Nanopore Sequencing

ABSTRACT

Human T-lymphotropic virus type 1 (HTLV-1) is associated with several diseases, such as HTLV-1-associated myelopathy/tropical spastic paraparesis (HAM/TSP), adult T-cell leukemia/lymphoma, and infectious dermatitis. However, the mechanisms behind its diverse clinical manifestations remain poorly understood. Despite its global health impact, complete HTLV-1 genome sequences are limited, hindering progress in research and treatment development. In this study, we develop and validate a tiling amplicon-based sequencing protocol compatible with both Illumina and Oxford Nanopore platforms for generating HTLV-1 genomes. The protocol produced near-complete HTLV-1 genomes with up to 98.67 % coverage. BBMap (Illumina) and the Sup base-calling model (Nanopore) yielded the best results. Hac provided an optimal balance between processing time and accuracy. Our protocol is efficient, versatile, and suitable for broad implementation. Its dual-platform compatibility facilitates genomic surveillance and enhances the availability of high-quality HTLV-1 genomic data, supporting advances in viral pathogenesis, evolution, and transmission research.

1. Introduction

The Human T-lymphotropic virus type 1 (HTLV-1) is the etiological agent of several pathologies, including HTLV-1-associated myelopathy/tropical spastic paraparesis (HAM/TSP) [1], adult T-cell leukemia/-lymphoma (ATLL) [2], and HTLV-1-associated infectious dermatitis (IDH) [3], among others. These pathologies exhibit distinct pathogenic mechanisms, and the factors underlying the wide spectrum of clinical manifestations, including why many infected individuals remain asymptomatic, are still poorly understood [4].

It is estimated that 5 to 10 million people worldwide are infected with HTLV-1 [5]. Despite its global burden, there are currently no vaccines or specific antiviral therapies available. In addition, although HTLV-1 was the first human retrovirus to be discovered, the number of complete genome sequences available remains considerably lower compared to other retroviruses such as Human Immunodeficiency Virus 1 (HIV-1). In October of 2025, there were 1.421.999 published HIV-1 sequences, while for HTLV-1, there were only 10.918 sequences

* Corresponding author.
E-mail address: fernanda.khouri@hotmail.com (F.K. Barreto).
† These authors contributed equally to this work.

https://doi.org/10.1016/j.nexres.2025.101110
Received 18 August 2025; Received in revised form 2 November 2025; Accepted 17 November 2025
Available online 19 November 2025

available in the GenBank database, and among these, only 577 are complete genomes. A systematic review conducted by our group showed that, despite advances in sequencing technologies, the Sanger method remains the most widely used approach, primarily yielding partial HTLV-1 sequences [6]. Generating complete viral genomes is therefore crucial to advance our understanding of HTLV-1 biology and pathogenesis.

The HTLV-1 genome, approximately 9032 base pairs in length, shares structural features with other retroviruses but is highly conserved [7]. It comprises canonical retroviral genes including *gag* (antigenic group), *pro* (protease), *pol* (polymerase), and *env* (envelope). At the 3` end lies the pX region, which encodes key regulatory proteins such as *tax* and *rex,* as well as the HTLV-1 bZip factor (HBZ) transcribed from the antisense strand. Long terminal repeat sequences (LTR) flanking both ends of the genome are essential for viral integration in the host genome [8].

Here, we present a validated tiling amplicon-based protocol compatible with both Nanopore and Illumina sequencing for the generation of complete HTLV-1 genomes. This approach yielded sequences with an average coverage of up to 98,63 %, enabling efficient and high-throughput molecular characterization. The proposed method enhances genomic surveillance and may contribute to addressing key knowledge gaps in HTLV-1 research.

2. Material and methods

2.1. Primers design and evaluation

To design a complete set of primers covering the nearly entire HTLV-1 genome, we used the *Primal Scheme* tool v3.0.2 (https://primalscheme.com). A total of 29 primer pairs were designed (hereafter referred to as CCEM-HTLV1 primer scheme). The input consisted of a FASTA file containing 31 HTLV-1 genome sequences available in GenBank (accessions KY007244-KY007274; genome size: 8989 pb). An amplicon length of 400 base pairs (bp) with a 50 bp overlap between adjacent amplicons was specified (Fig. 1). The designed primer set was assessed for potential secondary structures and interactions using the Multiple Primer Analyzer tool (Thermo Fisher Scientific) and analyzed using BLAST, which showed no cross-reactivity and indicated that our primers are highly specific to the HTLV-1 genome.

To evaluate primer efficiency, we initially standardized PCR conditions to allow multiplex usage. Amplification was performed using DNA extracted from the MT-2 cell line (a T-cell line from normal human cord leukocytes of a male infant by co-culturing with leukemic T-cells from a female patient with adult T-cell leukemia) with the QIAamp DNA Mini Kit (Qiagen, Cat. No 56,304). Each of the 29 primer pairs was tested individually in a 25 µL reaction, containing 19.5 µL of Platinum PCR SuperMix High Fidelity (Thermo Fisher Scientific, Cat. No 12,532,016), 1.5 µL of DNA sample (100ng/µL), and 2 µL of each primer (10pmol/µL). For the negative control, the DNA sample volume was replaced by nuclease-free ultra-pure water. PCR cycling conditions were as follows: initial denaturation at 94 °C for 2 min (1 cycle); denaturation 94 °C for 25 s, annealing 65 °C for 30 s, extension 68 °C for 30 s (35 cycles); and final extension 72 °C for 8 min (1 cycle). The PCR products were visualized by electrophoresis in 1.5 % agarose gel. Some regions were not detected by gel electrophoresis and also showed low coverage in sequencing: specifically, those amplified with primers 4, 8, 9, and 27 (See Supplementary Material A).

2.2. Library preparation and DNA sequencing

The libraries for Illumina and Nanopore sequencing were prepared independently. For Illumina sequencing, the COVIDSeq Test (Illumina, Cat. No 20,043,675 and 20,043,137) was adapted using the CCEM-HTLV1 primer scheme. Fragment length distribution was assessed with the Agilent Bioanalyzer High Sensitivity DNA Kit (Agilent Technologies, Cat. No 5067–4626) using the Agilent 2100 Bioanalyzer (Agilent Technologies, Cat. No G2939BA) and DNA concentration was quantified with the Qubit dsDNA HS Assay Kit (Thermo Fisher Scientific, Cat. No Q32854). Libraries were denatured and diluted to a final concentration of 10pM before being loaded onto a MiSeq Reagent Micro Kit v2 (Illumina, Cat. No MS-103–1002). Paired-end sequencing (2 × 150 bp) was performed on the Illumina MiSeq platform (Illumina, Cat. No SY-410–1003). All steps followed the manufacturer's instructions.

For Nanopore sequencing, library preparation involved three main steps: end-repair, adapter ligation and clean-up. DNA was first amplified using the CCEM-HTLV1 primer scheme in a 25 µL PCR reaction containing 18.75 µL of Platinum PCR SuperMix High Fidelity (Thermo Fisher Scientific, Cat. No 12,532,016), 1 µL of DNA sample (100ng/µL), 2 µL of primers pooled (See Supplementary Material B) and 3.25 µL of nuclease free-water. Thermocycling conditions were: denaturation for 2 min 94 °C (1 cycle), denaturation 94 °C for 25 s, annealing 65 °C for 30 s, extension 68 °C for 30 s (35 cycles), and final extension 72 °C for 8 min (1 cycle). End-repair and A-tailing were performed using the NEBNext Ultra II End Repair/dA-Tailing Module (New England Biolabs, Cat. No E7546S). Library construction followed the Oxford Nanopore e Ligation Sequencing Kit V14 (Oxford Nanopore Technologies, Cat. No SQK-LSK114). Adapted ligation was carried out using the NEBNext Quick Ligation Module (New England Biolabs, Cat. No E6056L). Libraries were purified with AMPure XP beads (Beckman Coulter, Cat. No A63881), and quantified with the Qubit 1X dsDNA HS Assay Kit (Cat. No Q33230 on the *Qubit 4 Fluorometer*). The final library was loaded onto an R10.4.1

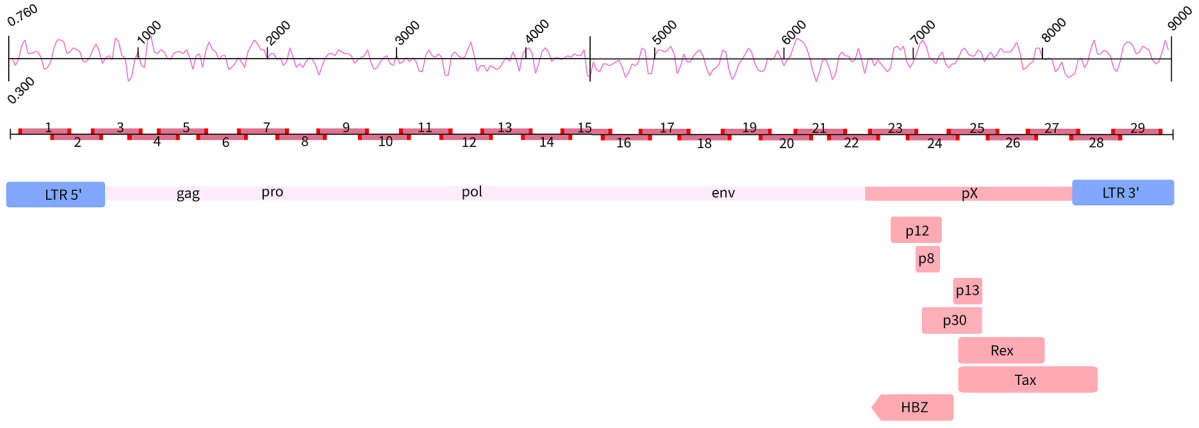

Fig. 1. Schematic representation of the 29 primer pairs designed to amplify the entire HTLV-1 genome using a tiling amplicon approach (See Supplementary Material A). Each red block represents an individual amplicon, with adjacent regions overlapping by 50 base pairs to ensure complete genome coverage. The purple waveform indicates GC content variation across the genome.

flow cell (Oxford Nanopore Technologies, Cat. No FLO-MIN114) and sequenced for up to 16 h using a MinION Mk1B (Oxford Nanopore Technologies, Cat. No Mk1B). All procedures followed the respective instructions.

We performed a comparative analysis of the HTLV-1 genome assembly from Illumina and Oxford Nanopore sequencing, including different assembly approaches for each platform. For Illumina reads, raw FASTQ files were quality-trimmed using fastp v0.24.1 [9] to remove low-quality base pairs. Reads were aligned to the HTLV-1 reference genome (NCBI GenBank Accession No J02029.1) using four assemblers: BBMap v38.84 [10], Bowtie2 v.2.4.5 [11], Burrows-Wheeler Aligner (BWA) v2.2.1 [12], and minimap2 v2.24 [13]. Primer sequences and variant calling were performed using iVAR v1.4.4 [14]. Consensus sequences were masked with "N" at positions with a coverage depth <10. Assembly metrics were calculated using SAMtools v1.21 (HTSlib v1.21) [15] and Seqtk v1.4-r122 [16].

For Nanopore data, raw POD5 files were basecalled using three accuracy models (fast, high-accuracy [Hac], and super-accuracy [Sup]) with Dorado v1.0.0 (Oxford Nanopore Technologies) executed on Google Colab with an NVIDIA Ada Lovelace L4 Tensor Core GPU. BAM files with a minimum quality score of 10 were retained for demultiplexing and adapter trimming using Dorado v1.0.0). Read quality was evaluated using the PyPNanoQC workflow (https://github.com/khourious/PyPNanoQC). Length filtering to remove chimeric reads and adapter dimers was carried out using BBDuk v39.01 [17] implemented in BBMap v39.01 [10]. Assembly was performed with minimap2 v2.28 [13] with the NCBI GenBank Accession No J02029.1 as genome reference. Trimming primer was performed using iVAR v1.4.2. The assembly was then polished, and variant calling was performed, both using Clair3 v1.1.0 [18] with model selection based on basecalling accuracy (fast: r1041_e82_400bps_fast_g632, hac: r1041_e82_400bps_hac_v520, sup: r1041_e82_400bps_sup_v520). The consensus sequences were masked with "N" at regions with coverage depth <20, and the variant candidates were incorporated into the consensus genome using BCFtools v1.17[15]. The assembly statistics were calculated with SAMtools v1.15.1 (using HTSlib v1.15.1) [15] and Seqtk v1.4 [16]. This entire workflow is available at https://github.com/khourious/vigeas.

3. Results

The highest-coverage HTLV-1 genome assemblies were achieved with BBMap for Illumina assembly (98.50 %) and sup accuracy level for Nanopore (98.67 %). Both platforms achieved high coverage, with Illumina showing only a single gap corresponding to the region targeted by primer pair 4 (Fig. 2).

Given the inherently higher error rates associated with Oxford Nanopore sequencing, selecting an optimal basecalling strategy was essential. The impact of basecalling accuracy on this sequencing performance was evaluated by comparing sequencing outputs generated with three basecalling models: Fast, Hac, and Sup. The Fast model, while computationally efficient with a processing time of 10 min, yielded suboptimal sequencing results, including low sequencing depth and incomplete genome recovery (58.08 % coverage). In contrast, the Hac model significantly improved sequencing performance, achieving 98.37 % coverage with only a modest increase in processing time (21 min; 2.1 × slower than fast), offering an optimal balance between accuracy and speed. The Sup model produced similar results (98.67 %) with the highest read depth, although with increased computational requirements (2 h and 9 min; 6.5 × slower than Fast and 3.1 × slower than Hac). (Table 1).

Therefore, among the tested models, Sup provided the best performance, yielding the highest sequencing depth, accuracy, and genome completeness (See Supplementary Material C). Although the Sup model requires greater computational resources and longer processing times, its improved accuracy significantly reduces sequencing errors and gaps, thereby enhancing downstream analyses. For applications where processing time is a constraint, the Hac model offers an effective balance between accuracy and efficiency. Additionally, the presence of chimeric

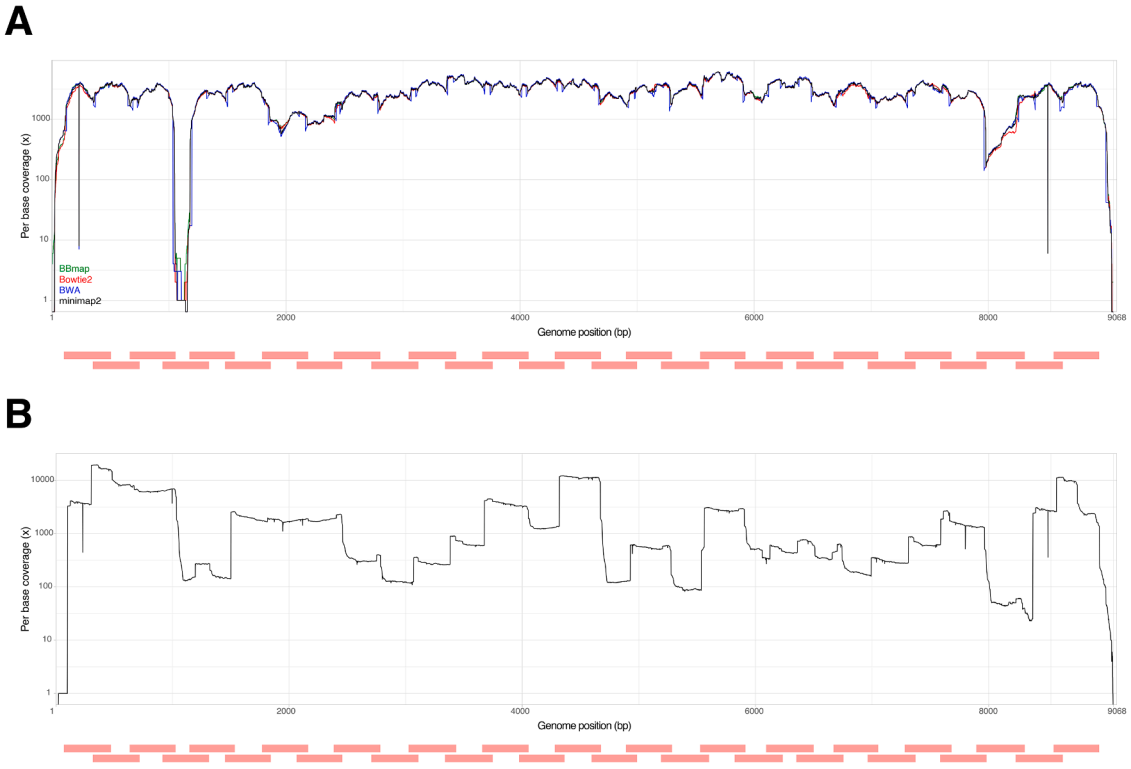

Fig. 2. HTLV-1 genome coverage using CCEM-HTLV1 primer scheme. (A) Illumina and (B) Oxford Nanopore MinION per-base coverage. Mapped assemblers are indicated by color: BBmap (green), Bowtie2 (blue), BWA (red), and minimap2 (black). Each red block represents an individual amplicon in the primer scheme.

Table 1
Comparison of HTLV-1 genome assembly obtained with the Illumina and Oxford Nanopore Technologies platforms.

Sequencing platform and Assembly strategy	Mean read length	Total reads	Total mapped bases (mb)	Mapped reads	Mean depth	Breadth of coverage %
Illumina + BWA	141	200,384	25.73	183,209	2714.10	98.15
Illumina + BBMap	141	200,384	25.85	183,493	2916.90	98.50
Illumina + Bowtie2	141	200,384	25.12	178,683	2724.50	98.30
Illumina + minimap2	141	200,384	25.87	185,710	2813.70	98.30
Nanopore + fast accuracy	443	636,019	1.48	66,061	169.58	58.08
Nanopore + hac accuracy	471	817,048	33.61	140,337	1525.58	98.37
Nanopore + sup accuracy	497	826,930	37.95	137,121	2367.43	98.67

reads was detected using nanopore sequencing. While the omission of barcodes is a common practice in single-sample workflows, it may have contributed to the formation of chimeric molecules and influenced the performance of the basecalling model, increasing the need for higher-accuracy basecalling to correctly interpret the sequencing data. These findings emphasize the importance of selecting an appropriate basecalling model to ensure high-quality data, particularly in applications requiring complete and reliable viral genome reconstruction.

4. Discussion

Several unresolved questions in the HTLV-1 field, particularly those related to the pathogenesis, transmission dynamics, and the development of effective treatments, are hindered by the limited availability of complete genomic data. The lack of standardized protocols for whole-genome sequencing using these platforms has contributed to the scarcity of complete HTLV-1 genomes, limiting progress in understanding the virus's genetic diversity, and its associations with distinct clinical outcomes — such as adult T-cell leukemia/lymphoma (ATLL), HTLV-1-associated myelopathy/tropical spastic paraparesis (HAM/TSP), and HTLV-1-associated infectious dermatitis (IDH). To address this limitation, we developed an amplicon-based sequencing protocol compatible with both Illumina and Nanopore platforms, and the successful implementation of this protocol on both platforms highlights its flexibility and broad applicability. We employed 29 primer pairs, designed to amplify ~400 bp fragments spanning the full HTLV-1 genome. The coverage of each amplicon appears satisfactory, except for primer pair 4. Redesigning this primer could be a viable option for future research.

It is important to highlight that among the Nanopore tested models, Sup provided the best performance, yielding the highest sequencing depth, accuracy, and genome completeness. Although the Sup model requires greater computational resources and longer processing times, its improved accuracy significantly reduces sequencing errors and gaps, thereby enhancing downstream analyses. For applications where processing time is a constraint, the Hac model offers an effective balance between accuracy and efficiency. Additionally, the presence of chimeric reads was detected using nanopore sequencing. While the omission of barcodes is a common practice in single-sample workflows, it may have contributed to the formation of chimeric molecules and influenced the performance of the basecalling model, increasing the need for higher-accuracy basecalling to correctly interpret the sequencing data. These findings emphasize the importance of selecting an appropriate basecalling model to ensure high-quality data, particularly in applications requiring complete and reliable viral genome reconstruction.

The dual-platform compatibility of the protocol validated here allows laboratories with diverse technological capabilities to adopt this approach, thereby facilitating expanded HTLV-1 genomic surveillance and research. The application of this method can substantially increase the availability of high-quality genomic data, thereby advancing our understanding of HTLV-1 pathogenesis, viral evolution, and transmission. Due to its portable characteristics, the use of MinION could be especially interesting in remote areas. Ultimately, this approach supports the development of improved therapeutic strategies and informed public health interventions. As such, despite here we performed with only one cell line (MT2), it represents a significant step forward in addressing key gaps in the field of HTLV-1 research.

Funding

This work was supported by the Fundação de Amparo à Pesquisa do Estado da Bahia (FAPESB), [N. APP.0048/2023. PROPCI - PROPG/UFBA 007/2022- JOVEMPESq] and the Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq) [N. 421342/2018-8]. L.d.M. was supported by Coordination for the Improvement of Higher Education Personnel (CAPES; financial code 001).

CRediT authorship contribution statement

Ana Carolina Marinho Monteiro Lima: Writing – original draft, Validation, Methodology, Investigation. **Laise de Moraes:** Writing – original draft, Methodology, Formal analysis, Data curation. **Rebeca Leão Amorim:** Writing – original draft, Methodology, Investigation. **Marina Silveira Cucco:** Writing – original draft, Methodology. **Roberta Muniz Luz Silva:** Validation, Investigation. **Marta Giovanetti:** Writing – review & editing, Conceptualization. **Filipe Ferreira de Almeida Rego:** Writing – review & editing, Methodology, Formal analysis, Data curation. **Thessika Hialla Almeida Araújo:** Writing – review & editing, Conceptualization. **Vagner Fonseca:** Writing – review & editing, Conceptualization. **Ricardo Khouri:** Writing – review & editing, Resources, Conceptualization. **Luiz Carlos Junior Alcantara:** Writing – review & editing, Supervision, Resources, Project administration, Methodology, Funding acquisition, Conceptualization. **Luciane Amorim Santos:** Writing – review & editing, Supervision, Resources, Project administration, Methodology, Funding acquisition, Conceptualization. **Fernanda Khouri Barreto:** Writing – review & editing, Supervision, Resources, Project administration, Methodology, Funding acquisition, Conceptualization.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

The authors would like to acknowledge the Structural Transformation to Attain Responsible BIOSciences (STARBIOS). We thank Dr. Bernardo Galvão-Castro for his valuable contribution to the study design and for his support as a consultant throughout the development of this article. We also thank the Rede de Plataformas Tecnológicas Fiocruz (RPT/FIOCRUZ)-RPT01Q Plataforma de Vigilância Molecular and Fiocruz Genomic Network Bahia for providing their molecular biology facilities and their assistance with DNA extraction and NGS sequencing.

Supplementary materials

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.nexres.2025.101110.

References

[1] M. Osame, K. Usuku, S. Izumo, N. Ijichi, H. Amitani, A. Igata, et al., HTLV-I associated myelopathy, a new clinical entity, Lancet 1 (1986) 1031–1032, https://doi.org/10.1016/S0140-6736(86)91298-5.

[2] T. Uchiyama, J. Yodoi, K. Sagawa, K. Takatsuki, H. Uchino, Adult T-cell leukemia: clinical and hematologic features of 16 cases, Blood 50 (1977) 481–492, https://doi.org/10.1182/BLOOD.V50.3.481.481.

[3] L. La Grenade, A. Manns, V. Fletcher, C. Carberry, B. Hanchard, E.M. Maloney, et al., Clinical, pathologic, and immunologic features of human T-lymphotrophic virus type I-associated infective dermatitis in children, Arch. Dermatol. 134 (1998) 439–444, https://doi.org/10.1001/ARCHDERM.134.4.439.

[4] Y. Yamano, T. Sato, Clinical pathophysiology of Human T-Lymphotropic virus-type 1-associated myelopathy/tropical spastic paraparesis, Front. Microbiol. 3 (2012) 389, https://doi.org/10.3389/FMICB.2012.00389.

[5] A. Gessain, O. Cassar, Epidemiological aspects and world distribution of HTLV-1 infection, Front. Microbiol. 3 (2012), https://doi.org/10.3389/FMICB.2012.00388.

[6] F. de Oliveira Andrade, M.S. Cucco, M.M.N. Borba, R.C. Neto, L.L. Gois, Almeida de, F.F. Rego, et al., An overview of sequencing technology platforms applied to HTLV-1 studies: a systematic review, Arch. Virol. 166 (2021) 3037–3048, https://doi.org/10.1007/S00705-021-05204-W.

[7] A. Gessain, R.C. Gallo, G. Franchini, 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 (1992) 2288–2295, https://doi.org/10.1128/jvi.66.4.2288-2295.1992.

[8] M. Seiki, S. Hattori, Y. Hirayama, M. Yoshida, Human adult T-cell leukemia virus: complete nucleotide sequence of the provirus genome integrated in leukemia cell DNA, Proc. Natl. Acad. Sci. U S A 80 (1983) 3618–3622, https://doi.org/10.1073/PNAS.80.12.3618.

[9] S. Chen, Y. Zhou, Y. Chen, J. Gu, fastp: an ultra-fast all-in-one FASTQ preprocessor, Bioinformatics 34 (2018) i884–i890, https://doi.org/10.1093/BIOINFORMATICS/BTY560.

[10] Bushnell B. BBMap: a fast, accurate, splice-aware aligner 2014.

[11] B. Langmead, S.L. Salzberg, Fast gapped-read alignment with Bowtie 2, Nat. Methods 9 (2012) 357–359, https://doi.org/10.1038/NMETH.1923. SUBJMETA=1647,208,212,48,514,631;KWRD=BIOINFORMATICS,GENOMICS, SEQUENCING.

[12] H. Li, R. Durbin, Fast and accurate short read alignment with Burrows-Wheeler transform, Bioinformatics 25 (2009) 1754–1760, https://doi.org/10.1093/BIOINFORMATICS/BTP324.

[13] H. Li, Minimap2: pairwise alignment for nucleotide sequences, Bioinformatics 34 (2018) 3094–3100, https://doi.org/10.1093/BIOINFORMATICS/BTY191.

[14] S. Castellano, F. Cestari, G. Faglioni, E. Tenedini, M. Marino, L. Artuso, et al., iVar, an interpretation-oriented tool to manage the update and revision of variant annotation and classification, Genes 12 (2021) 384, https://doi.org/10.3390/GENES12030384.

[15] P. Danecek, J.K. Bonfield, J. Liddle, J. Marshall, V. Ohan, M.O. Pollard, et al., Twelve years of SAMtools and BCFtools, Gigascience 10 (2021) 1–4, https://doi.org/10.1093/GIGASCIENCE/GIAB008.

[16] Li, Seqtk: Toolkit for Processing Sequences in FASTA/Q Formats, GitHub, 2016 [Internet], https://github.com/lh3/seqtk (accessed June 17, 2025).

[17] Whitham J. Impact of BBDuk metagenomic read trimming and decontamination 2021. https://doi.org/10.25982/77705.1341/1779218.

[18] Z. Zheng, S. Li, J. Su, A.W.S. Leung, T.W. Lam, R. Luo, Symphonizing pileup and full-alignment for deep learning-based long-read variant calling, Nat. Comput. Sci. 2 (2022) 797–803, https://doi.org/10.1038/S43588-022-00387-X.