# METHODS

# Study Population

Between May 2015 and June 2016, we actively recruited patients presenting clinical symptoms of arbovirus-like infection in Bahia, Brazil (Campo Formoso = 230; Salvador = 948).

# Nucleic acid isolation and RT-qPCR

Whole blood, saliva, and urine samples from 1 to 9 days after the onset of symptoms were sent to Instituto Gonçalo Moniz, FIOCRUZ-BA, Bahia, Brazil, for molecular diagnostics. The viral RNA extraction from plasma, urine and/or saliva (oral swab was incubated in 200μL of nuclease-free H2O) samples from Campo Formoso was performed with QIAmp Viral RNA Mini Kit (Qiagen, Germany) following the manufacturer’s recommendations and tested by quantitative RT-PCR (RT-qPCR) using ZDC Molecular Kit (IBMP, Brazil) or Trioplex Real-time RT-PCR Assay (CDC, USA). The viral RNA extraction from sera samples from Salvador was extracted using Maxwell 16 Viral Total Nucleic Acid Purification Kit (Promega, USA), tested by conventional RT-PCR according to Balm et. al (2012) [1] and confirmed the results using Trioplex Real-time RT-PCR Assay (CDC, USA). Samples were selected for sequencing based on Ct value < 30 and availability of epidemiological metadata, such as date of sample collection and municipality of residence. A total of 14 samples from Salvador and seven samples from Campo Formoso, Bahia were included.

# Complete genome MinION nanopore sequencing

A protocol developed by Quick and collaborators [2] were used. The cDNA synthesis was generated using the ProtoScript II First Strand cDNA Synthesis Kit (New England Biolabs, United Kingdom) and random hexamer priming. The cDNA generated was subjected to multiplex PCR using Q5 High-Fidelity DNA polymerase (New England Biolabs, United Kingdom) and a set of specific primers (ZikaAsian V2) designed by the ZiBRA project (<https://github.com/zibraproject/zika-pipeline>).Amplicons were purified using 1x AMPure XP beads (Beckman Coulter, United Kingdom) and quantified on a Qubit 3.0 fluorimeter (Life Technologies, United States) using Qubit dsDNA BR assay. The library preparation was performed using the Ligation Sequencing Kit SQK-LSK108 (Oxford Nanopore Technologies, United Kingdom) and Native Barcoding Kit EXP-NBD103 (Oxford Nanopore Technologies, United Kingdom). Sequencing libraries were loaded into an R9.4/R9.4.1 flow cell (Oxford Nanopore Technologies, United Kingdom), and sequencing data were collected for up to 48hr on a MinION platform (Oxford Nanopore Technologies, United Kingdom).

# Generation of consensus sequences from nanopore

The fast5 files generated during sequencing were submitted to the pipeline defined by Black and colleagues [3] with minor modifications. In brief, the sequencing data were basecalled on the high-accuracy model performed by Guppy v.3.4.4 (Oxford Nanopore Technologies, United Kingdom). The basecalled fastq files with a minimum Q score of 7 were selected for subsequent demultiplex process using Guppy v.3.4.4 (Oxford Nanopore Technologies, United Kingdom). A re-demultiplex process, trimming adapters and chimaeras were performed by Porechop v.0.2.4 (<https://github.com/rrwick/Porechop>). Furthermore, the assembly was performed by Burrows-Wheeler Aligner (BWA) v.0.7.17-r1188 [4] using NCBI Genbank accession number KJ776791.1 as genome reference. The primer sequences were trimmed with align\_trim.py. The assembly was then polished, and the variant calling were performed by nanopolish v.0.11.3 (<https://github.com/jts/nanopolish>). The consensus sequences were then masked with “N” at regions with coverage depth <20, and the variant candidates were incorporated into the consensus genome by using VCFtools v.0.1.16 [5]. The assembly statistics were calculated with SAMtools v.1.10 (using HTSlib 1.10.2) [6] and Seqtk v.1.3-r106 (<https://github.com/lh3/seqtk>).

# Collation of sequence dataset

A dataset of ZIKV full or near full-length genomes annotated with the date of sample collection and location was compiled. All ZIKV sequence data were retrieved from GenBank on January 28, 2020, and gbmunge (<https://github.com/sdwfrost/gbmunge>) was used for extracting FASTA format sequences and associated metadata. Genotyping was conducted using the Genome Detective Virus Tool [7]. Full-length and near full-length (>7,000 pb) ZIKV genomes of Asian genotype were retained for further analyses. Sequences identified as duplicates were excluded, and only South American isolates were kept for the final dataset. The alignment of the sequence dataset was performed using MAFFT v7.455 [8,9] and manually edited using AliView v.1.26 [10].

# Maximum likelihood analysis and temporal signal estimation

Maximum Likelihood (ML) phylogenetic analyses were performed using IQ-TREE v.1.6.12 [11] under GTR+F+I+G4 with 1,000-replicate ultrafast bootstrap [12]. The best fitting model was inferred in ModelFinder [13] implemented in the IQ-TREE. The ML trees were visualised and edited using FigTree v.1.4.4 (<http://tree.bio.ed.ac.uk/software/figtree>). To investigate the evolutionary temporal signal, we apply a regression of root-to-tip genetic distances against the date of sample collection using TempEst v1.5.3 [14], which revealed sufficient temporal signal (R2 = 0.3958) to justify a molecular clock approach. The NCBI GenBank accession numbers for the sequences are as follows: KR872956, KU312312, KU321639, KU365777-KU365780, KU497555, KU501215, KU509998, KU527068, KU647676, KU707826, KU729217, KU729218, KU740184, KU744693, KU758877, KU761564, KU820897, KU820898, KU922923, KU922960, KU926309, KU926310, KU937936, KU940224, KU940227, KU940228, KU955590, KU991811, KX051563, KX056898, KX101060-KX101067, KX156774-KX156776, KX197205, KX198135, KX247646, KX280026, KX447510, KX447518, KX520666, KX548902, KX702400, KX766028, KX811222, KX879603, KX879604, KX893855, KY003153, KY003154, KY014296, KY014297, KY014301, KY014303, KY014307-KY014309, KY014313, KY014317, KY014320, KY075932-KY075934, KY120352, KY241788, KY272991, KY317936-KY317940, KY348640, KY379148, KY441401-KY441403, KY558989-KY558993, KY558995-KY559028, KY559030-KY559032, KY631492, KY693678, KY693679, KY693680, KY785409, KY785410, KY785417, KY785426, KY785427, KY785429, KY785433, KY785436, KY785437, KY785439, KY785446, KY785450, KY785451, KY785455, KY785456, KY785460, KY785462, KY785464, KY785466, KY785469, KY785477, KY785479-KY785481, KY785483, KY817930, KY989971, MF073357-MF073359, MF098767, MF167360, MF352141, MF574552, MF783073, MF794971, MG770183, MG770184-MG770186, MH179341, MH513598, MH513599, MH544701, MH675619-MH675628, MH882542, MH882549, MK028858, MK028859, MK028861, MK049245-MK049252, MK216687, MK216688, MK216690-MK216693, MK216695-MK216699, MK216703, MK216709-MK216723, MK216726-MK216730, MK216732, MK216733, MK216738, MK216741-MK216745, MK216747, MK216748, MK241415-MK241417, MK269360, MK269361, MK829152-MK829154, MN171421, MN185324, MN473453, MN473454, MN577543, MN577544; and the ENA accession nos.: SRR5309450-SRR5309457.

# Molecular clock phylogenetic analysis

To explore the evolutionary temporal signal using a molecular clock approach, a discrete phylogeographic analysis was performed using a Bayesian Markov Chain Monte Carlo (MCMC) method as implemented in the BEAST package v1.10.5pre [15,16] with BEAGLE v4.0.0 to improve the computational performance [17]. The substitution process was modelled with a codon-partitioned HKY+G4 model for the coding region of the genome [18–20]. An asymmetric discrete trait model integrated with Bayesian Stochastic Search Variable Selection (BSSVS) [21] was specified to reconstruct the spread history and the number of transitions between locations was estimated with a Markov Jumps approach [17,22,23]. The substitution rate was estimated with an Uncorrelated Lognormal Distribution (UCLD) relaxed-clock model [24]. We employed sampled tip dates to address inexact dates [25]. The Skyride model served as a flexible tree prior [26,27]. We used Tracer v1.7.1 [28] to diagnose mixing and convergence properties. Post burn-in trees were summarised as a Maximum Clade Credibility (MCC) tree using TreeAnnotator v.1.10.5 pre-implemented in BEAST. Trees were visualised and edited using FigTree v.1.4.4.

# Data availability statement

The new sequences have been deposited in NCBI GenBank under accession numbers OQ727565-OQ727578; and the XML files and datasets analysed in this study are available in the GitHub repository (<https://github.com/khourious/Early-ZIKV-genomes-NE-BR-to-the-Americas>).

# REFERENCES

1. Balm, M.N.D.; Lee, C.K.; Lee, H.K.; Chiu, L.; Koay, E.S.C.; Tang, J.W. A Diagnostic Polymerase Chain Reaction Assay for Zika Virus. *J. Med. Virol.* **2012**, *84*, 1501–1505, doi:10.1002/jmv.23241.

2. Quick, J.; Grubaugh, N.D.; Pullan, S.T.; Claro, I.M.; Smith, A.D.; Gangavarapu, K.; Oliveira, G.; Robles-Sikisaka, R.; Rogers, T.F.; Beutler, N.A.; et al. Multiplex PCR Method for MinION and Illumina Sequencing of Zika and Other Virus Genomes Directly from Clinical Samples. *Nat. Protoc.* **2017**, *12*, 1261–1276, doi:10.1038/nprot.2017.066.

3. Black, A.; Moncla, L.H.; Laiton-Donato, K.; Potter, B.; Pardo, L.; Rico, A.; Tovar, C.; Rojas, D.P.; Longini, I.M.; Halloran, M.E.; et al. Genomic Epidemiology Supports Multiple Introductions and Cryptic Transmission of Zika Virus in Colombia. *BMC Infect. Dis.* **2019**, *19*, 963, doi:10.1186/s12879-019-4566-2.

4. Li, H.; Durbin, R. Fast and Accurate Long-Read Alignment with Burrows–Wheeler Transform. *Bioinformatics* **2010**, *26*, 589–595, doi:10.1093/bioinformatics/btp698.

5. Danecek, P.; Auton, A.; Abecasis, G.; Albers, C.A.; Banks, E.; DePristo, M.A.; Handsaker, R.E.; Lunter, G.; Marth, G.T.; Sherry, S.T.; et al. The Variant Call Format and VCFtools. *Bioinformatics* **2011**, *27*, 2156–2158, doi:10.1093/bioinformatics/btr330.

6. Danecek, P.; Bonfield, J.K.; Liddle, J.; Marshall, J.; Ohan, V.; Pollard, M.O.; Whitwham, A.; Keane, T.; McCarthy, S.A.; Davies, R.M.; et al. Twelve Years of SAMtools and BCFtools. *Gigascience* **2021**, *10*, 1–4, doi:10.1093/gigascience/giab008.

7. Vilsker, M.; Moosa, Y.; Nooij, S.; Fonseca, V.; Ghysens, Y.; Dumon, K.; Pauwels, R.; Alcantara, L.C.; Vanden Eynden, E.; Vandamme, A.-M.; et al. Genome Detective: An Automated System for Virus Identification from High-Throughput Sequencing Data. *Bioinformatics* **2019**, *35*, 871–873, doi:10.1093/bioinformatics/bty695.

8. Katoh, K.; Standley, D.M. MAFFT Multiple Sequence Alignment Software Version 7: Improvements in Performance and Usability. *Mol. Biol. Evol.* **2013**, *30*, 772–780, doi:10.1093/molbev/mst010.

9. Katoh, K. MAFFT: A Novel Method for Rapid Multiple Sequence Alignment Based on Fast Fourier Transform. *Nucleic Acids Res.* **2002**, *30*, 3059–3066, doi:10.1093/nar/gkf436.

10. Larsson, A. AliView: A Fast and Lightweight Alignment Viewer and Editor for Large Datasets. *Bioinformatics* **2014**, *30*, 3276–3278, doi:10.1093/bioinformatics/btu531.

11. Minh, B.Q.; Schmidt, H.A.; Chernomor, O.; Schrempf, D.; Woodhams, M.D.; von Haeseler, A.; Lanfear, R. IQ-TREE 2: New Models and Efficient Methods for Phylogenetic Inference in the Genomic Era. *Mol. Biol. Evol.* **2020**, *37*, 1530–1534, doi:10.1093/molbev/msaa015.

12. Hoang, D.T.; Chernomor, O.; von Haeseler, A.; Minh, B.Q.; Vinh, L.S. UFBoot2: Improving the Ultrafast Bootstrap Approximation. *Mol. Biol. Evol.* **2018**, *35*, 518–522, doi:10.1093/molbev/msx281.

13. Kalyaanamoorthy, S.; Minh, B.Q.; Wong, T.K.F.; von Haeseler, A.; Jermiin, L.S. ModelFinder: Fast Model Selection for Accurate Phylogenetic Estimates. *Nat. Methods* **2017**, *14*, 587–589, doi:10.1038/nmeth.4285.

14. Rambaut, A.; Lam, T.T.; Max Carvalho, L.; Pybus, O.G. Exploring the Temporal Structure of Heterochronous Sequences Using TempEst (Formerly Path-O-Gen). *Virus Evol.* **2016**, *2*, vew007, doi:10.1093/ve/vew007.

15. Suchard, M.A.; Lemey, P.; Baele, G.; Ayres, D.L.; Drummond, A.J.; Rambaut, A. Bayesian Phylogenetic and Phylodynamic Data Integration Using BEAST 1.10. *Virus Evol.* **2018**, *4*, 1–5, doi:10.1093/ve/vey016.

16. Baele, G.; Ayres, D.L.; Rambaut, A.; Suchard, M.A.; Lemey, P. High-Performance Computing in Bayesian Phylogenetics and Phylodynamics Using BEAGLE. In *Evolutionary Genomics*; 2019; Vol. 1910, pp. 691–722.

17. Ayres, D.L.; Cummings, M.P.; Baele, G.; Darling, A.E.; Lewis, P.O.; Swofford, D.L.; Huelsenbeck, J.P.; Lemey, P.; Rambaut, A.; Suchard, M.A. BEAGLE 3: Improved Performance, Scaling, and Usability for a High-Performance Computing Library for Statistical Phylogenetics. *Syst. Biol.* **2019**, *68*, 1052–1061, doi:10.1093/sysbio/syz020.

18. Hasegawa, M.; Kishino, H.; Yano, T. Dating of the Human-Ape Splitting by a Molecular Clock of Mitochondrial DNA. *J. Mol. Evol.* **1985**, *22*, 160–174, doi:10.1007/BF02101694.

19. Yang, Z. Maximum Likelihood Phylogenetic Estimation from DNA Sequences with Variable Rates over Sites: Approximate Methods. *J. Mol. Evol.* **1994**, *39*, 306–314, doi:10.1007/BF00160154.

20. Edwards, C.J.; Suchard, M.A.; Lemey, P.; Welch, J.J.; Barnes, I.; Fulton, T.L.; Barnett, R.; O’Connell, T.C.; Coxon, P.; Monaghan, N.; et al. Ancient Hybridization and an Irish Origin for the Modern Polar Bear Matriline. *Curr. Biol.* **2011**, *21*, 1251–1258, doi:10.1016/j.cub.2011.05.058.

21. Drummond, A.J.; Suchard, M.A. Bayesian Random Local Clocks, or One Rate to Rule Them All. *BMC Biol.* **2010**, *8*, 114, doi:10.1186/1741-7007-8-114.

22. Minin, V.N.; Suchard, M.A. Fast, Accurate and Simulation-Free Stochastic Mapping. *Philos. Trans. R. Soc. B Biol. Sci.* **2008**, *363*, 3985–3995, doi:10.1098/rstb.2008.0176.

23. Minin, V.N.; Suchard, M.A. Counting Labeled Transitions in Continuous-Time Markov Models of Evolution. *J. Math. Biol.* **2007**, *56*, 391–412, doi:10.1007/s00285-007-0120-8.

24. Drummond, A.J.; Ho, S.Y.W.; Phillips, M.J.; Rambaut, A. Relaxed Phylogenetics and Dating with Confidence. *PLoS Biol.* **2006**, *4*, e88, doi:10.1371/journal.pbio.0040088.

25. Shapiro, B.; Ho, S.Y.W.; Drummond, A.J.; Suchard, M.A.; Pybus, O.G.; Rambaut, A. A Bayesian Phylogenetic Method to Estimate Unknown Sequence Ages. *Mol. Biol. Evol.* **2011**, *28*, 879–887, doi:10.1093/molbev/msq262.

26. Gill, M.S.; Lemey, P.; Faria, N.R.; Rambaut, A.; Shapiro, B.; Suchard, M.A. Improving Bayesian Population Dynamics Inference: A Coalescent-Based Model for Multiple Loci. *Mol. Biol. Evol.* **2013**, *30*, 713–724, doi:10.1093/molbev/mss265.

27. Ferreira, M.A.R.; Suchard, M.A. Bayesian Analysis of Elapsed Times in Continuous-Time Markov Chains. *Can. J. Stat.* **2008**, *36*, 355–368, doi:10.1002/cjs.5550360302.

28. Rambaut, A.; Drummond, A.J.; Xie, D.; Baele, G.; Suchard, M.A. Posterior Summarization in Bayesian Phylogenetics Using Tracer 1.7. *Syst. Biol.* **2018**, *67*, 901–904, doi:10.1093/sysbio/syy032.