Epidemic establishment and cryptic transmission of Zika virus in Brazil and the Americas

Faria, N. R.*[1,2], Quick, J.[3]*, Morales, I.[4]*, Thézé, J.[1]*, Jesus, J.G.[5]*, Giovanetti, M.[5,6]*, Kraemer, M. U. G.[1]*, Hill, S. C.[1]*, Black, A.[7,8]*, da Costa, A. C.[3], Franco, L.C.[2], Silva, S. P.[2], Wu, C.-H.[9], Ragwhani, J.[1], Cauchemez, S.[10,11], du Plessis, L.[1], Verotti, M. P.[12], de Oliveira, W. K.[13,14], Carmo, E. H.[15], Coelho, G. E.[16,17], Santelli, A. C. F. S.[16,18], Vinhal, L. C.[16], Henriques, C. M.[15], Simpson, J. T.[19], Loose, M.[20], Andersen, K. G.[21], Grubaugh, N. D.[21], Somasekar, S.[22], Chiu, C. Y.[22], Lewis-Ximenez, L. L.[23], Baylis, S.A.[24], Chieppe, A. O.[25], Aguiar, S. F.[25], Fernandes, C. A.[25], Lemos, P. S.[2], Nascimento, B. L. S.[2], Monteiro, H. A. O.[2], Siqueira, I. C.[5], de Queiroz, M. G.[26], de Souza, T. R.[26,27], Bezerra, J. F.[26,28], Lemos, M. R.[29], Pereira, G. F.[29], Loudal, D.[29], Moura, L. C.[29], Dhalia, R.[30], França, R. F.[30], Magalhães, T.[30,31], Marques, E. T. Jr.[30,32], Jaenish, T.[33], Wallau, G. L.[30], de Lima, M. C.[34], Nascimento, V.[34], de Cerqueira, E. M.[35], de Lima, M. M.[36], Mascarenhas, D. L.[36], Moura Neto, J. P.[37], Levin, A. S.[4], Tozetto-Mendoza, T. R.[4], Fonseca, S. N.[38], Mendes-Correa, M. C. [4], Milagres, F.P.[39], Segurado, A.[4], Holmes, E. C.[40], Rambaut, A.[41,42], Bedford, T.[7], Nunes, M. R. T.*[2,43], Sabino, E. C.[4¶]*, Alcantara, L. C. J.[5¶]*, Loman, N.[3¶]*, Pybus, O. G.[1]*¶

Affiliations:

1. Department of Zoology, University of Oxford, Oxford OX1 3PS, UK
2. Evandro Chagas Institute, Ministry of Health, Ananindeua, Brazil
3. Institute of Microbiology and Infection, School of Biosciences, University of Birmingham, UK
4. Department of Infectious Disease, School of Medicine & Institute of Tropical Medicine, University of São Paulo, Brazil
5. Fundação Oswaldo Cruz (FIOCRUZ), Salvador, Bahia, Brazil
6. University of Rome Tor Vergata, Rome, Italy
7. Vaccine and Infectious Disease Division, Fred Hutchinson Cancer Research Center, Seattle, WA, USA
8. Department of Epidemiology, University of Washington, Seattle, WA, USA
9. Department of Statistics, University of Oxford, Oxford OX1 3LB, UK
10. Mathematical Modelling of Infectious Diseases and Center of Bioinformatics, Biostatistics and Integrative Biology, Institut Pasteur, Paris, France
11. Centre National de la Recherche Scientifique, URA3012, Paris, France
12. Coordenação dos Laboratórios de Saúde (CGLAB/DEVIT/SVS), Ministry of Health, Brasília, Brazil
13. Coordenação Geral de Vigilância e Resposta às Emergências em Saúde Pública (CGVR/DEVIT), Ministry of Health, Brasília, Brazil
14. Center of Data and Knowledge Integration for Health (CIDACS), Fundação Oswaldo Cruz (FIOCRUZ), Brazil
15. Departamento de Vigilância das Doenças Transmissíveis, Ministry of Health, Brasilia, Brazil
16. Coordenação Geral dos Programas de Controle e Prevenção da Malária e das Doenças Transmitidas pelo *Aedes*, Ministry of Health, Brasília, Brazil
17. Pan American Health Organization (PAHO), Buenos Aires, Argentina
18. Fundação Oswaldo Cruz (FIOCRUZ), Rio de Janeiro, Brazil

19. Ontario Institute for Cancer Research, Toronto, Canada
20. University of Nottingham, Nottingham, UK
21. Department of Immunology and Microbial Science, The Scripps Research Institute, La Jolla, CA 92037, USA
22. Departments of Laboratory Medicine and Medicine & Infectious Diseases, University of California, San Francisco, USA
23. Instituto Oswaldo Cruz (FIOCRUZ), Rio de Janeiro, Brazil
24. Paul-Ehrlich-Institut, Langen, Germany
25. Laboratório Central de Saúde Pública Noel Nutels, Rio de Janeiro, Brazil
26. Laboratório Central de Saúde Pública do Estado do Rio Grande do Norte, Natal, Brazil
27. Universidade Potiguar do Rio Grande do Norte, Natal, Brazil
28. Faculdade Natalense de Ensino e Cultura, Rio Grande do Norte, Natal, Brazil
29. Laboratório Central de Saúde Pública do Estado da Paraíba, João Pessoa, Brazil
30. Instituto Aggeu Magalhães, Fundação Oswaldo Cruz (FIOCRUZ), Recife, Pernambuco, Brazil
31. Arthropod-borne and Infectious Diseases Laboratory (AIDL), Department of Microbiology, Immunology & Pathology, Colorado State University, USA
32. Center for Vaccine Research, Graduate School of Public Health, University of Pittsburgh, Pittsburgh, PA, USA
33. Section Clinical Tropical Medicine, Department for Infectious Diseases, Heidelberg University Hospital, Heidelberg, Germany
34. Laboratório Central de Saúde Pública do Estado de Alagoas, Maceió, Brazil
35. Universidade Estadual de Feira de Santana, Feira de Santana, Bahia, Brazil
36. Secretaria de Saúde de Feira de Santana, Feira de Santana, Bahia, Brazil
37. Universidade Federal do Amazonas, Manaus, Brazil
38. Hospital São Francisco, Ribeirão Preto, Brazil
39. Universidade Federal do Tocantins, Palmas, Brazil
40. University of Sydney, Sydney, Australia
41. Institute of Evolutionary Biology, University of Edinburgh, Edinburgh EH9 3FL, UK
42. Fogarty International Center, National Institutes of Health, Bethesda, MD 20892, USA
43. Department of Pathology, University of Texas Medical Branch, Galveston, TX 77555, USA

*** Joint first or senior author**

¶**Correspondence to:**

Prof. Luiz Carlos Junior Alcantara
Gonçalo Moniz Institute/FIOCRUZ
Rua Waldemar Falcão 121
Candeal, Salvador, Bahia, 40295-001
Tel: +55 (71)3351-9525
Email: lalcan@bahia.fiocruz.br

Prof. Ester C. Sabino
Institute of Tropical Medicine, University of São Paulo
Av Dr Eneas de Carvalho Aguiar 470.
São Paulo 05403-000 Brazil.
Tel: +551130618702
Email: sabinoec@usp.br

Dr. Nicholas Loman
Institution of Microbiology and Infection, School of Biosciences, University of Birmingham
Edgbaston, Birmingham B15 2TT UK
Tel: +44(0)121 4145564
Email: n.j.loman@bham.ac.uk

Prof. Oliver G. Pybus
Department of Zoology, University of Oxford
South Parks Road, OX1 3PS
Tel: +44(0)1865271274
Email: oliver.pybus@zoo.ox.ac.uk

01 February 2017

One Sentence Summary: Virus genomes reveal the establishment of Zika virus in Northeast Brazil and the Americas, and provide an appropriate timeframe for baseline (pre-Zika) microcephaly in different regions.

Zika virus (ZIKV) transmission in the Americas was first confirmed in May 2015 in Northeast Brazil[1] . Brazil has the highest number of reported ZIKV cases worldwide (>200,000 by 24 Dec 2016[2]) as well as the greatest number of cases associated with microcephaly and other birth defects (2,366 confirmed cases by 31 Dec 2016[2]). Following the initial detection of ZIKV in Brazil, 47 countries and territories in the Americas have reported local ZIKV transmission, with 22 of these reporting ZIKV-associated severe disease[3]. Yet the origin and epidemic history of ZIKV in Brazil and the Americas remain poorly understood, despite the value of such information for interpreting past trends in reported microcephaly. To address this we generated 53 complete or partial ZIKV genomes, mostly from Brazil, including data generated by the ZiBRA project – a mobile genomics lab that travelled across Northeast Brazil in 2016. One sequence represents the earliest confirmed ZIKV infection in Brazil. Joint analyses of viral genomes with ecological and epidemiological data estimate that the ZIKV epidemic first became established in NE Brazil by March 2014 and likely disseminated from there, both nationally and internationally, before the first detection of ZIKV in the Americas. Estimated dates of the international spread of ZIKV from Brazil coincide with periods of high vector suitability in recipient regions and indicate the duration of pre-detection cryptic transmission in those regions. NE Brazil's role in the establishment of ZIKV in the Americas is further supported by geographic analysis of ZIKV transmission potential and by estimates of the virus' basic reproduction number.

Previous phylogenetic analyses indicated that the ZIKV epidemic was caused by the introduction of a single Asian genotype lineage into the Americas around late 2013, at least one year before its detection there[4]. An estimated 100 million people in the Americas are predicted at risk of acquiring ZIKV once the epidemic has reached its full extent[5]. However, little is known about the genetic diversity and transmission history of the virus in different regions in Brazil[6]. Reconstructing ZIKV spread from case reports alone is challenging because symptoms (typically fever, headache, joint pain, rashes, and conjunctivitis) overlap with those caused by co-circulating arthropod-borne viruses[7] and due to a lack of nationwide ZIKV-specific surveillance in Brazil before 2016.

To address this we undertook a collaborative investigation of ZIKV molecular epidemiology in Brazil, including results from a mobile genomics laboratory that travelled through NE Brazil during June 2016 (the ZiBRA project; zibraproject.github.io). Of five regions of Brazil (**Fig. 1a**), the Northeast region (NE Brazil) has the most notified ZIKV cases (40% of Brazilian cases) and the most confirmed microcephaly cases (76% of Brazilian cases, to 31 Dec 2016[2]), raising questions about why the region has been so severely affected[8]. Further, NE Brazil is the most populous region of Brazil that exhibits year-round ZIKV transmission potential[9]. With the support of the Brazilian Ministry of Health and other institutions (**Acknowledgements**), the ZiBRA lab screened 1330 serum, blood and urine samples from patients residing in 82 municipalities across five federal states in NE Brazil (**Fig. 1**; **SI Table 1**). Samples provided by the central public health laboratory of each state (LACEN) and FIOCRUZ were screened for the presence of ZIKV by real time quantitative PCR (RT-qPCR).

On average, ZIKV viremia persists for 10 days after infection; symptoms develop ~6 days after infection and can last 1-2 weeks[10]. In support of the hypothesis that ZIKV

viremia is low by the time most patients seek medical care[11], we found that RT-qPCR+ samples were, on average, collected only two days after onset of symptoms. The median RT-qPCR cycle threshold (Ct) value of positive samples was correspondingly high, at 36 (**SI Fig. 1**). For NE Brazil, the time series of RT-qPCR+ cases was positively correlated with the number of weekly-notified cases (Pearson's ρ=0.62; **Fig. 1b**).

The ability of the mosquito vector *Aedes aegypti* to transmit ZIKV is determined by ecological factors that impact adult survival, viral replication, and infective periods. To investigate the receptivity of each Brazilian region to ZIKV transmission, we used a measure of vector climatic suitability derived from monthly temperature, relative humidity, and precipitation data[9]. Using a linear regression we find that, for each Brazilian region, there is a strong association between estimated climatic suitability and weekly notified cases (**Figs. 1b,1c**; adjusted R^2>0.84, P<0.001; **SI Table 2**). Notified ZIKV cases lag vector climatic suitability by ~4 to 6 weeks in all regions except NE Brazil, where no time lag is evident (awareness of ZIKV may have been higher in NE Brazil because that is where transmission was first confirmed). Despite these associations, numbers of notified cases should be interpreted cautiously because (i) co-circulating dengue and chikungunya viruses exhibit similar symptoms to ZIKV, and (ii) the Brazilian case reporting system has evolved through time (see Methods). We estimated the basic reproductive numbers (R_0) for ZIKV in each Brazilian region from the weekly notified case data and found that R_0 is high in NE Brazil (R_0~3 for both epidemic seasons; **SI Table 3**). Although our R_0 values are approximate, in part due to spatial variation in transmission across the large regions analysed here, they are consistent with previous estimates from a variety of approaches[12].

Encouraged by the utility of portable genomic technologies during the West African Ebola virus epidemic[13] we used our openly-developed protocol[14] to sequence ZIKV genomes directly from clinical material using MinION DNA sequencers. We were able to generate virus sequences within 48 hours of the mobile lab's arrival at each LACEN. In pilot experiments using a cultured ZIKV reference strain[15] our protocol recovered 98% of the virus genome[14]. However, due to low viral copy numbers in clinical samples (**SI Fig. 1**), many sequences exhibited incomplete genome coverage and were subjected to additional sequencing efforts in static labs once fieldwork was completed. Whilst average genome coverage was higher for samples with lower Ct-values (84% for Ct<33; **Fig. 2a**), samples with higher Ct values had highly variable coverage (mean=63% for Ct≥33 (**Fig. 2a**). Unsequenced genome regions were non-randomly distributed (**Fig. 2b**), suggesting that the efficiency of PCR amplification varied among primer pair combinations. We generated 36 near-complete or partial genomes from the NE, SE and N regions of Brazil, supplemented by 8 from serum/urine samples from Rio de Janeiro municipality.

The American ZIKV epidemic comprises a single founder lineage[4,16,17] (hereafter termed Am-ZIKV) derived from Asian genotype viruses (hereafter termed PreAm-ZIKV) from Southeast Asia and the Pacific[4]. A sliding window analysis of pairwise genetic diversity along the ZIKV genome shows that the diversity of PreAm-ZIKV strains is on average ~2.6-fold greater than Am-ZIKV viruses (**Fig. 2d**), reflecting a longer period of ZIKV circulation in Asia and the Pacific than in the Americas. Genetic diversity of the Am-ZIKV lineage will increase in future and diagnostic assays are recommended to consider two genome regions to increase sensitivity[18].

It has been suggested that recent ZIKV epidemics may be causally linked to a higher apparent evolutionary rate for the Asian genotype than the African genotype[19,20]. However, such comparisons are confounded by an inverse relationship between the timescale of observation and estimated viral evolutionary rates[21]. Regression of sequence sampling dates against root-to-tip genetic distances indicates that molecular clocks can be applied reliably to the Asian-ZIKV lineage (**Fig. 2c; SI Fig. 2**). We estimate the whole genome evolutionary rate of Asian ZIKV to be 0.97×10^{-3} substitutions per site per year (s/s/y; 95% Bayesian credible interval, BCI=0.79-1.16×10^{-3}), consistent with other estimates for the Asian genotype[4,20]. We found no significant differences in evolutionary rates among ZIKV genome regions (**Fig. 2d**). The estimated d_N/d_S ratio of the PreAm-ZIKV sequences is 0.061 (95% CI= 0.047-0.077) consistent with strong purifying selection, as observed for other vector-borne flaviviruses[22]. The d_N/d_S ratio of the Am-ZIKV lineage is higher (0.12, 95% CI= 0.10-0.14) likely due to the raised probability of observing slightly deleterious changes in short-term datasets, as observed in previous emerging epidemics[23].

We used two phylogeographic approaches with different assumptions[24,25] to reconstruct the spatial origins and spread of ZIKV in Brazil and the Americas from an alignment comprising the data generated here plus 61 available sequences. We dated the common ancestor of ZIKV in the Americas (node B, **Fig. 3**) to Feb 2014 (95% BCIs = Oct 2013-May 2014; **Ext Data Table 5**), in line with previous estimates[4,20]. We find evidence that NE Brazil played a central role in the establishment and dissemination of the Am-ZIKV lineage. Our results suggest that NE Brazil is the most probable location of node B (location posterior support =0.88, **Fig. 3**). However current data cannot exclude the hypothesis that node B was located in the Caribbean due the presence of two sequences from Haiti in one of its descendant lineages (**Fig. 3** dashed branches). More importantly, most Am-ZIKV sequences descend from a rapid radiation of lineages (denoted node C, **Fig. 3**) that is dated to around Mar 2014 (95% BCIs =Dec 2013-Jun 2014) and is more strongly inferred to have existed in NE Brazil (location posterior support =1.00, **Fig. 3**). This placement is also seen in 5 of 6 analyses on sub-sampled datasets (**SI Figs. 3 and 4**). Consequently, we conclude that node C represents the establishment of ZIKV in the Americas. If further data show that node B did indeed exist in Haiti, then it is mostly likely that Haiti acted as an intermediate 'stepping stone' for Am-ZIKV, and was not the place where the continental-wide epidemic took hold. This perspective is consistent with the lower population size of Haiti compared to Brazil (which receives 6 million annual visitors). We infer that node C was present in NE Brazil several months before three events that also all occurred in NE Brazil: (i) the retrospective identification of a cluster of suspected but unconfirmed ZIKV cases in Dec 2014 [1], (ii) the oldest ZIKV genome sequence from Brazil, reported here, sampled in Feb 2015, and (iii) the initial identification of ZIKV in the Americas in Mar 2015[26,27].

Our results further suggest that node C viruses from NE Brazil were important in the continental spread of the epidemic. Within Brazil, we find several instances of virus lineage movement from NE to SE Brazil; most of such events can be dated back to the second half of 2014 and led to onwards transmission in Rio de Janeiro (RJ1 and RJ2 in **Fig. 3**) and São Paulo states (SP1 and SP2 in **Fig. 3**). We also infer that ZIKV lineages disseminated from NE Brazil to elsewhere in Central America, the Caribbean, and South America. Most Am-ZIKV strains sampled outside Brazil fall into four well-supported monophyletic groups in Fig 3; three (SA1, SA2, CA1) are inferred to have been exported from NE Brazil between Jul 2014 and Feb 2015, and

one (CB1) from SE Brazil between Mar and Sep 2015 (**Figs. 3, 4)**. Notably, each independent viral lineage export occurred during a period of high climatic suitability for vector transmission in the recipient location (**Fig. 4**). For the three earliest exports, there is a 6-12 month gap between the estimated date of exportation and the date of ZIKV detection in the recipient location, suggesting an complete or partial season of undetected transmission. These periods of cryptic transmission are relevant to studies of spatio-temporal trends in reported microcephaly in the Americas, because they help define the appropriate timeframe for baseline (pre-ZIKV) microcephaly in each region.

Combing both virus genomic and epidemiological data can generate insights into the patterns and drivers of vector-borne virus transmission. Large-scale surveillance of ZIKV is challenging because (i) many cases may be asymptomatic and (ii) ZIKV co-circulates in some regions with other arthropod-borne viruses that exhibit overlapping symptoms (e.g. dengue, chikungunya, Mayaro, and Oropouche viruses). A system of continuous and proportional virus sequencing, integrated with surveillance data, could provide timely information on the distribution of Zika and other viruses and thereby inform effective response and control measures.

Methods

Sample collection

Between the 1st and 18th June 2016, 1330 samples from cases notified as ZIKV infected were tested for ZIKV infection in the Northeast region of Brazil (NE Brazil). During this period, 4 of the 5 laboratories in the region visited by the ZiBRA project were in the process of implementing molecular diagnostics for ZIKV. The ZiBRA team spent 2-3 days in each state central public health laboratory (LACEN). The samples analysed had previously been collected from patients who had attended a municipal or state public health facility, presenting maculopapular rash and at least two of the following symptoms: fever, conjunctivitis, polyarthralgia or periarticular edema. The majority of samples were linked to a digital record that collated epidemiological and clinical data: date of sample collection, location of residence, demographic characteristics, and date of onset of clinical symptoms (when available). The ZiBRA project was conducted under the auspices of the *Coordenação Geral de Laboratórios de Saúde Pública* in Brazil (CGLAB), part of the Brazilian Ministry of Health (MoH) in support of the ongoing emergency public health response to Zika. Urine and plasma samples from Rio were obtained from patients followed at the Viral Hepatitis Ambulatory/FIOCRUZ/Rio de Janeiro following Institutional Review Board approval. RNA was extracted at the Paul-Ehrlich-Institut and sequenced at the University of Birmingham, UK.

Nucleic acid isolation and RT-qPCR

Serum, blood and urine samples were obtained from patients 0 to 228 days after first symptoms (**SI Table 1**). Viral RNA was isolated from 200 ul Zika-suspected samples using either the NucliSENS easyMag system (BioMerieux, Basingstoke, UK) (Ribeirão Preto samples), the ExiPrep Dx Viral RNA Kit (BIONEER, Republic of Korea) (Rio de Janeiro samples) or the QIAamp Viral RNA Mini kit (QIAGEN, Hilden, Germany) (all other samples) according to the manufacturer's instructions. Ct values were determined for all samples by probe-based RT-qPCR against the prM target (using 5'FAM as the probe reporter dye) as previously described[28]. RT-qPCR assays were performed using the QuantiNova Probe RT-qPCR Kit (20 ul reaction volume; QIAGEN) with amplification in the Rotor-Gene Q (QIAGEN) following the manufacturer's protocol. Primers/probe were synthesised by Integrated DNA Technologies (Leuven, Belgium). The following reaction conditions were used: reverse transcription (50°C, 10 min), reverse transcriptase inactivation and DNA polymerase activation (95°C, 20 sec), followed by 40 cycles of DNA denaturation (95°C, 10 secs) and annealing-extension (60°C, 40 sec). Positive and negative controls were included in each batch; however, due to the large number of samples tested in a short time it was possible only to run each sample without replication.

Whole genome sequencing

Sequencing was attempted on all positive samples regardless of Ct value. For these samples, extracted RNA was converted to cDNA using the Protoscript II First Strand cDNA synthesis Kit (New England Biolabs, Hitchin, UK) and random hexamer priming. Zika genome amplification by multiplex PCR was attempted using the ZikaAsianV1 primer scheme and 40 cycles of PCR using Q5 High-Fidelity DNA

polymerase (NEB) as described in Quick et al.[29]. PCR products were cleaned-up using AmpureXP purification beads (Beckman Coulter, High Wycombe, UK) and quantified using fluorimetry with the Qubit dsDNA High Sensitivity assay on the Qubit 3.0 instrument (Life Technologies). PCR products for samples yielding sufficient material were barcoded and pooled in an equimolar fashion using the Native Barcoding Kit (Oxford Nanopore Technologies, Oxford, UK). Sequencing libraries were generated from the barcoded products using the Genomic DNA Sequencing Kit SQK-MAP007/SQK-LSK208 (Oxford Nanopore Technologies). Sequencing libraries were loaded onto a R9/R9.4 flowcell and data was collected for up to 48 hours but generally less. As described[29], consensus genome sequences were produced by alignment to a Zika virus reference genome (strain H/PF/2013, GenBank accession: KJ776791.1) followed by nanopore signal-level detection of single nucleotide variants. Only positions with ≥20x genome coverage were used to produce consensus alleles. Regions with lower coverage, and those in primer-binding regions were masked with N characters.

Collation of genome-wide data sets

Our complete and partial genome sequences were appended to a global data set of all available published ZIKV genome sequences (up until January 2017) using an in-house script that retrieves updated GenBank sequences on a daily basis. In addition to the genomes generated from samples collected in NE Brazil during ZiBRA fieldwork, samples were sent directly to University of São Paulo and elsewhere for sequencing. Thirteen genomes from Ribeirão Preto, São Paulo state (SP; SE-Brazil region) and seven genomes from Tocantins (TO; N-Brazil region) were sequenced at University of São Paulo. Eight genomes from Rio de Janeiro (RJ; SE-Brazil region) were sequenced in Birmingham, UK, and added to our dataset. All these genomes were generated using the same primer scheme as the ZiBRA samples collected in NE Brazil[29]. In addition to these 44 sequences from Brazil, we further included in analysis 9 genomes from ZIKV strains sampled outside of Brazil in order to contextualise the genetic diversity of Brazilian ZIKV, giving rise to a final data set of n=53 sequences. Specifically, we included 5 genomes from samples collected in Colombia (accession numbers KY317936- KY317940) and 4 from samples collected in Mexico (accession numbers XXX-XXX).

GenBank sequences belonging to the ZIKV-African genotype were identified using the Arboviral genotyping tool (http://bioafrica2.mrc.ac.za/rega-genotype/typingtool/aedesviruses) and excluded from subsequent analyses, as our focus of study was the ZIKV-Asian genotype and the Am-ZIKV lineage in particular. To assess the robustness of dating estimates to the inclusion of older sequences, all analyses were performed with and without the P6-740 strain, the oldest known strain of the ZIKV-Asian genotype (sampled during 1966 in Malaysia). Our final alignment comprised the data reported in this study (n=53) plus publicly available ZIKV-Asian genotype sequences (n=61). Unpublished but publicly available genomes were included in our analysis only if we had permission to do so from those who generated the data.

Maximum likelihood analysis and recombination screening

Preliminary maximum likelihood (ML) trees were estimated with ExaMLv3 [30] using a per-site rate category model and a gamma distribution of among site rate variation. For the final analyses, ML trees were estimated using PhyML[31] under a GTR nucleotide substitution model[32], with a gamma distribution of among site rate variation, as selected by jModeltest.v.2 [33]. An approximate likelihood ratio test was used to estimate branch support[31]. Final ML trees were estimated with NNI and SPR tree search algorithms; equilibrium nucleotide frequencies and substitution model parameters were estimated using ML[31] (see **SI Fig. 5**).

Recombination may impact evolutionary estimates[34] and has been shown to be present in the ZIKV-African genotype[35]. In addition to restricting our analysis to the Asian genotype of ZIKV, we used the 12 recombination detection methods available in RDPv4 [36] and the Phi-test approach [37] available in SplitsTree [38] to further search for evidence of recombination in the ZIKV-Asian lineage. No evidence of recombination was found.

Analysis of the temporal molecular evolutionary signal in the alignments was conducted using TempEst[39]. In brief, collection dates in the format yyyy-mm-dd (ISO 8601 standard) were regressed against root-to-tip genetic distances obtained from the ML phylogeny. When precise dates were not available, a precision of 1 month or 1 year in the collection dates was taken into account.

To compare the pairwise genetic diversity of PreAm-ZIKV from Asia and the Pacific with Am-ZIKV from the Americas we used a sliding window approach with 300 nt wide windows and a step size of 50 nt. Sequence gaps were ignored; hence the average pairwise difference per window was obtained by dividing the total pairwise nucleotide differences by the total number of pairwise comparisons.

Molecular clock phylogenetics and gene-specific d_N/d_S estimation

To estimate Bayesian molecular clock phylogenies, analyses were run in duplicate using BEASTv.1.8.4 [40] for 30 million MCMC steps, sampling parameters and trees every 3000th step. We employed a stringent model selection analysis using both path-sampling and stepping stone models[41] to estimate the most appropriate model combination for Bayesian phylogenetic analysis (**SI Tables 4 and 5**). The best fitting model was a HKY codon position-structured SDR06 nucleotide substitution model[42] with a Bayesian skygrid tree prior[43] (with 49 grid points and a cut off = 10) and a relaxed molecular clock model[44]. A non-informative continuous time Markov chain reference prior[45] on the molecular clock rate was used. Convergence of MCMC chains was checked with Tracer v.1.6. After removal of 10% as burn-in, posterior tree distributions were combined and subsampled to generate an empirical distribution of 1,500 molecular clock trees.

To estimate rates of evolution per gene we partitioned the alignment into 10 genes (3 structural genes C, prM, E, and 7 non-structural genes NS1, NS2A, NS2B, NS3, NS4A2k, NS4B and NS5) and employed a SDR06 substitution model[42] and a strict molecular clock model, using a set of empirical molecular clock trees. To estimate the ratio of nonsynonymous to synonymous substitutions per site (d_N/d_S) for the PreAm-ZIKV and the Am-ZIKV lineages, we used the single likelihood ancestor counting (SLAC) method[46] implemented in HyPhy[47]. This method was applied to two distinct

codon-based alignments and their corresponding ML trees, that comprised the PreAm-ZIKV and Am-ZIKV sequences, respectively.

Phylogeographic analysis

We investigated virus lineage movements using our empirical distribution of phylogenetic trees and the sampling location of each ZIKV sequence. The sampling location of sequences from returning travellers was set to the travel destination in the Americas where infection likely occurred. We discretised sequence sampling locations in Brazil into the geographic regions defined in main text. The number of sequences per region available for analysis was 10 for N-Brazil, 35 for NE Brazil and 24 for SE-Brazil. No viral genetic data was available for the Centre-West (CW) and the South (S) Brazilian regions. We similarly discretised the locations of ZIKV sequences sampled outside of Brazil. These were grouped according to the United Nations M49 coding classification of macro-geographical regions. Our analysis included 7 sequences from the Caribbean, 8 from Central America, 17 from Polynesia, 11 from South America (excluding Brazil) and 3 from Southeast Asia. To account for the possibility of sampling bias arising from a larger number of sequences from particular locations, we repeated all phylogeographic analyses using (i) the full dataset (n=113) and (ii) three jackknife resampled datasets (n=52) in which taxa from the five most frequently sampled locations (NE Brazil, SE-Brazil, Polynesia, South America, N-Brazil) were randomly sub-sampled to seven sequences (the number of sequences available for the Caribbean).

Phylogeographic reconstructions were conducted using two approaches; (i) using the symmetric[25] and asymmetric[48] discrete trait evolution models implemented in BEASTv1.8.4[40] and (ii) using the Bayesian structured coalescent approximation (BASTA)[24] implemented available in BEAST2v.2. The latter has been suggested to be less sensitive to sampling biases[49]. For both approaches, maximum clade credibility trees were summarized from the MCMC samples using TreeAnnotator after discarding 10% as burn-in. The posterior estimates of the location of the Am-ZIKV clade root node from these two analytical approaches (applied to both the complete and jackknifed data sets) can be found in **SI Figs. 3** and **4**.

For the discrete trait evolution approach, we counted the expected number of transitions among each pair of locations (net migration) using the robust counting approach[50,51] available in BEASTv1.8.4[40]. We then used those inferred transitions to identify the earliest estimated ZIKV introductions into new regions. These viral lineage movement events were statistically supported (with Bayes factors > 10) using the BSSVS (Bayesian stochastic search variable selection) approach implemented in BEAST[25]. Box plots for node ages were generated using the ggplot2[52] package in R software[53].

Epidemiological analysis

Weekly suspected ZIKV data per Brazilian region were obtained from the Brazilian Ministry of Health (MoH). Cases were defined as suspected ZIKV infection when patients presented maculopapular rash and at least two of the following symptoms: fever, conjunctivitis, polyarthralgia or periarticular edema. Because notified suspected ZIKV cases are based on symptoms and not molecular diagnosis, it is possible that some cases represent other co-circulating viruses with related symptoms, such as

dengue and chikungunya viruses. Further, case reporting may have varied among regions and through time. Data from 2015 came from the pre-existing MoH sentinel surveillance system that comprised 150 reporting units throughout Brazil, which was eventually standardised in Feb 2016 in response to the ZIKV epidemic. We suggest that these limitations should be borne in mind when interpreting the ZIKV notified case data, and we consider the R_0 values estimated here to be approximate. That said, our time series of RT-qPCR+ ZIKV diagnoses from NE Brazil qualitatively match the time series of notified ZIKV cases from the same region (**Fig. 1b**). To estimate the exponential growth rate of the ZIKV outbreak in Brazil, we fit a simple exponential growth rate model to each stage of the weekly number of suspected ZIKV cases from each region separately:

$$I_w = I_0 exp(r_W . w) \tag{1}$$

where I_w is the number of cases in week w. As described in main text, the Brazilian regions considered here were the NE Brazil, N-Brazil, S-Brazil, SE-Brazil, and CW-Brazil. The time period for which exponential growth occurs is determined by plotting the log of I_w and selecting the period of linearity (**SI Fig. 6**). A simple linear model is then fitted to this period to estimate the weekly exponential growth rate r_W:

$$ln(I_w) = ln(I_0) + r_W . w \tag{2}$$

Let $g(.)$ be the probability density distribution of the epidemic generation time (i.e. the duration between the time of infection of a case and the mean time of infection of its secondary infections). The following formula can be used to derive the reproduction number R from the exponential growth rate r and density $g(.)$[54].

$$R = \frac{1}{\int_0^\infty exp(-r.t)g(t)dt} \tag{3}$$

In our baseline analysis, following Ferguson et al.[55] we assume that the ZIKV generation time is Gamma-distributed with a mean of 20.0 days and a standard deviation (SD) of 7.4 days. In a sensitivity analysis, we also explored scenarios with shorter mean generation times (10.0 and 15.0 days) but unchanged coefficient of variation SD/mean=7.4/20=0.37 (**SI Table 3**).

Association between *Aedes aegypti* climatic suitability and ZIKV notified cases

To account for seasonal variation in the geographical distribution of the ZIKV vector *Aedes aegypti* in Brazil we fitted high-resolution maps[56] to monthly covariate data. Covariate data included time-varying variables, such as temperature-persistence suitability, relative humidity, and precipitation, as well as static covariates such as urban versus rural land use. Maps were produced at a 5km x 5km resolution for each calendar month and then aggregated to the level of the five Brazilian regions used in

this study (**SI Fig. 7**). For consistency, we rescaled monthly suitability values so that the sum of all monthly maps equalled the annual mean map[9].

We then assessed the correlation between monthly *Aedes aegypti* climatic suitability and the number of weekly ZIKV notified cases in each Brazilian region, to test how well vector suitability explains the variation in the number of ZIKV notified cases. To account for the correlation in each Brazilian region (n=5) we fit a linear regression model with a lag and two breakpoints. As there may be a lag between trends in suitability and trends in notified cases, we fit a flexible temporal term in the model to allow for a shift in the respective curves. Thus for each region, different sets of the constant and linear terms are fitted to different time period. More formally,

$$\log(y_i + 1) = \alpha + \mathbb{I}(i \notin T)\alpha' + [b + \mathbb{I}(i \notin T)b']x_{i-l} \qquad (4)$$

where y_i represents notified cases in a particular region in month i, x_i is the climatic suitability in that region in month i, l is the time lag that yields the highest correlation between y_i and x_i and T is the set of time indexes in the correlated region.

We then find the values of T and l that provide the highest adjusted-R^2 by stepwise iterative optimisation. For each value of T evaluated, the optimal value of l (i,e. that which gives the highest adjusted-R^2 for the model above) is found by the optim function in R[53]. Climatic suitability values were only calculated for each month, so to calculate suitability values for continuous time points we interpolated them using a linear function between the successive monthly data points. We found no significant effect of residual autocorrelation in our data (**SI Fig. 8**).

Data availability

Sequences of the primers and probes have been available at zibra.io since the beginning of the project. Genome sequences were made publicly available once generated and confirmed at http://www.zibraproject.org. New Brazilian sequences are deposited in GenBank under accession numbers KY558989 to KY559032.

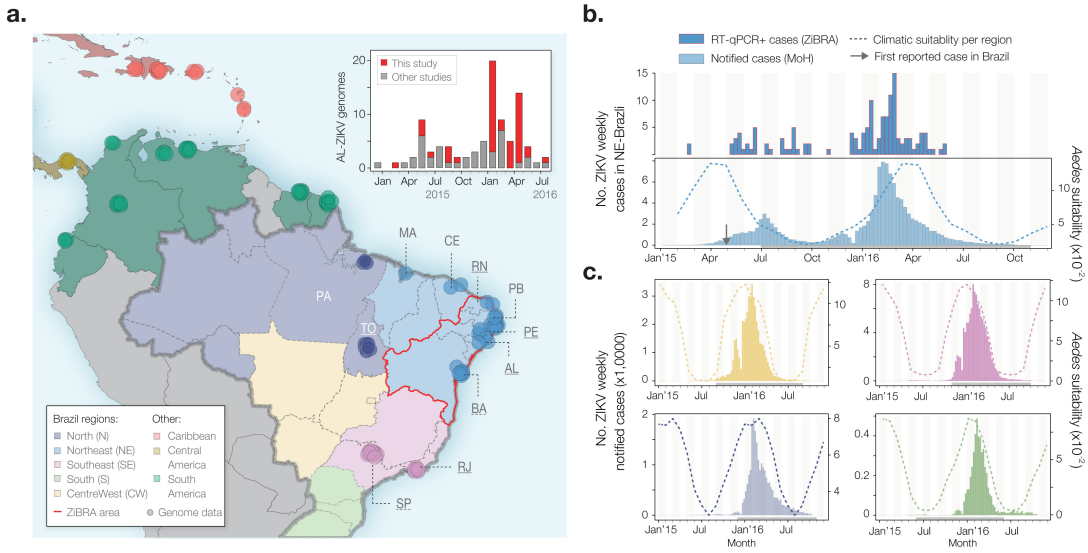


Fig. 1. **Geographic and temporal distribution of ZIKV in Brazil. a**. Location of sampling of genome sequences in Brazil and in the Americas. Federal states in Brazil have been coloured according to 5 geographic regions. A red contour line indicates the federal states surveyed during June 2016 by the ZiBRA mobile lab. Two letter state codes are as follows: PA: Pará, MA: Maranhão, CE: Ceará, TO: Tocantins, RN: Rio Grande do Norte, PB: Paraíba, PE: Pernambuco, AL: Alagoas, BA: Bahia, RJ: Rio de Janeiro, SP: São Paulo. Underlined states represent states from which sequences were generated in this study. Non-underlined states represent states from which states from which sequences were publicly available. **b.** ZIKV confirmed and notified cases in NE Brazil. The upper panel shows the temporal distribution of RT-qPCR+ cases (n=181) detected during the ZiBRA journey. Only confirmed cases for which the exact collection date was known (138 out of 181) were included. The lower panel shows notified ZIKV cases in NE Brazil between 01 Jan 2015 and 19 Nov 2016 (n = 122,779). **c.** Notified ZIKV cases in the Centre-West, Southeast, North and South regions of Brazil (clockwise). The dashed lines represent the average climatic vector suitability score across each region (see Methods). The R^2, P-value and estimated lag (T) values of the model used to compare these two trends are provided in **SI Table 2**. Grey horizontal bars below each time series indicate the time period for which correlation between suitability and ZIKV notified cases was highest.

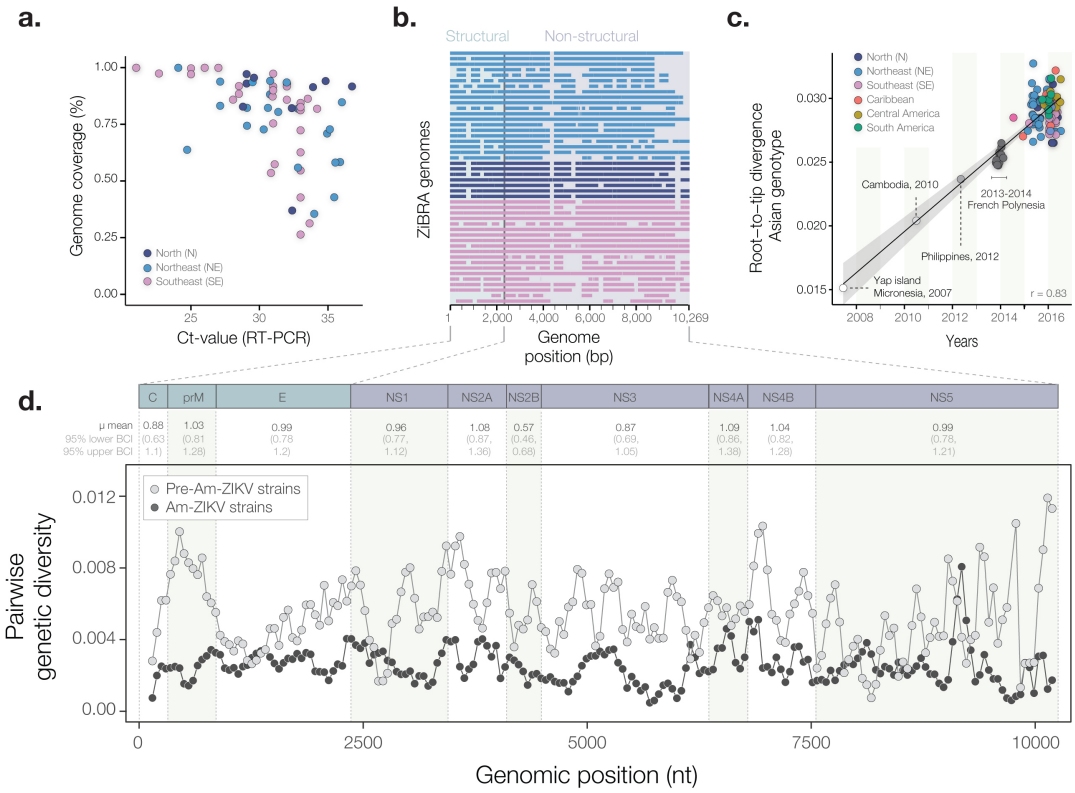


Fig. 2. **Zika virus genetic diversity and sequencing statistics. a**. Plot showing the percentage of the genome sequenced as a function of the corresponding RT-qPCR Ct-value. Each circle represents a sequence sample recovered from an infected individual in Brazil and has been colour coded according to the location of sampling. **b**. Plot showing sequencing coverage across the entire genome for the ZiBRA sequences. **c.** Regression of sequence sampling date and root-to-tip genetic in ML phylogenetic tree of the Asian-ZIKV lineage. This plot excludes P6-740 (the oldest Asian-ZIKV strain, isolated from Malaysia in 1966). A comparable analysis that include P6-740 is shown in **SI Fig. 2**. **d**. Average pairwise genetic diversity for the PreAm-ZIKV strains (grey line) and the Am-ZIKV lineage (black line), calculated using a sliding window of 300 nucleotides with a step size of 50 nucleotides.

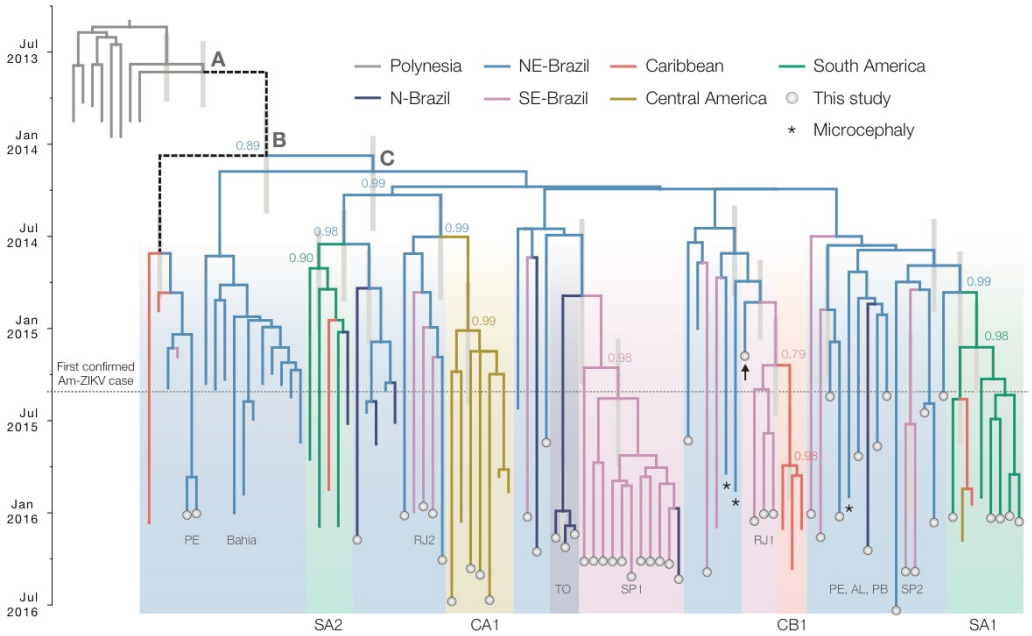


Fig. 3. **Phylogeography of ZIKV in the Americas.** Maximum clade credibility phylogeographic tree estimated from publicly available published genomes and the genomes reported in this study (highlighted by grey circles). Dashed vertical grey boxes illustrate the dating uncertainty of key internal nodes (see also **SI Table 5**). Branch colours indicate most probable ancestral locations. Coloured numbers show location state posterior probabilities. Asterisks indicate the three available genomes from microcephaly cases (accession numbers: KU497555, KU729217 and KU527068). An arrow indicates the oldest ZIKV sequence from Brazil. The horizontal line denotes when ZIKV was first confirmed in the Americas, in early May 2015. The nodes denoted A and B are equivalent to the nodes named identically in [4]. Acronyms along the bottom of the figure denote clades comprising three or more taxa from regions outside NE Brazil. Those from Brazil are denoted PE (Pernambuco) RJ1 and RJ2 (Rio de Janeiro), TO (Tocantins), SP1 and SP2 (São Paulo), BA (Bahia), AL (Alagoas), PB (Paraíba). Clades from outside Brazil are denoted CB (Caribbean), SA1 and SA2 (South America excluding Brazil), CA (Central America).

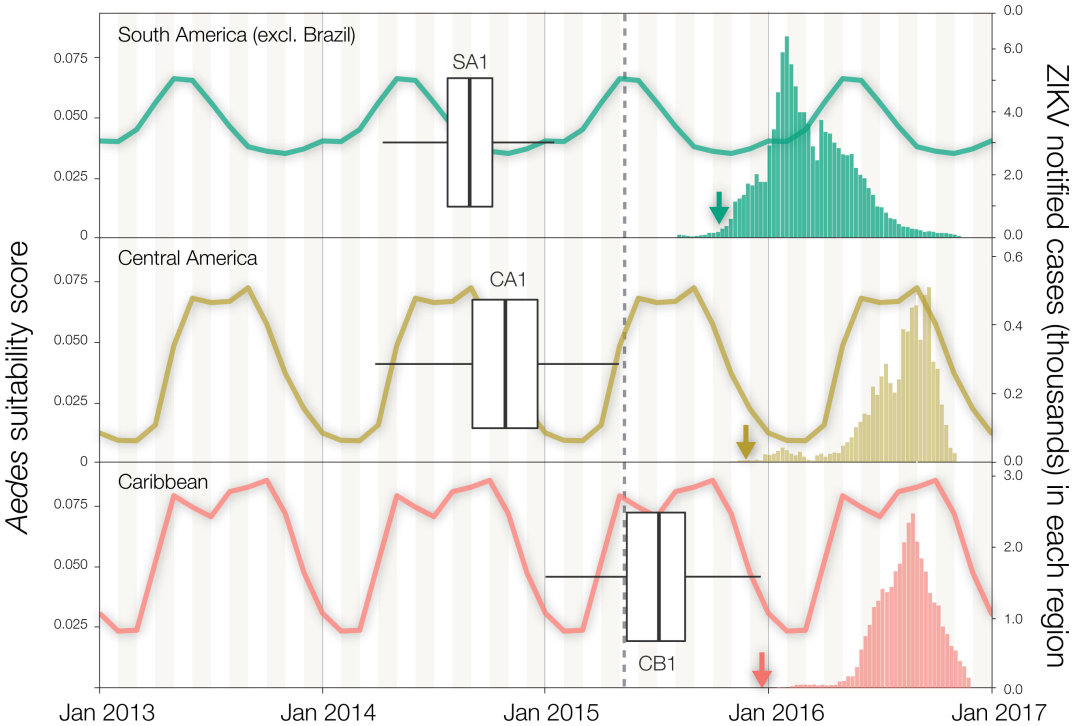


Fig. 4. Establishment of Am-ZIKV in the Americas. The earliest inferred dates of lineage export to each region are represent by box-and-whisker plots. Each refers to the branch immediately ancestral to the node denoted beneath the box-and-whisker plot. The vertical line within the box indicates the midpoint of that branch. The box edges and whiskers, from left to right, represent the 2.5%, 25%, 75% and 97.5% percentiles of the branch midpoint posterior distribution. Panel **a** shows the earliest export to South America outside Brazil (SA1 in **Fig. 3**), **b** shows an export to Central America (CA1), and **c** shows an export to the Caribbean (CB1). In each of **a-c**, dashed lines show an estimate of climatic vector suitability for each recipient region, averaged across the countries for which sequence data is available (see Methods). In each of **a-c**, the bar plots show available notified ZIKV case data for each recipient region. Coloured arrows indicate the earliest date of confirmation of ZIKV transmission in each recipient region. The vertical grey dashed line represents the date of ZIKV confirmation in the Americas and the colour scheme is identical to **Fig. 3**.

References

1 Kindhauser, M. K., Allen, T., Frank, V., Santhana, R. S. & Dye, C. Zika: the origin and spread of a mosquito-borne virus. *Bulletin of the World Health Organization* **94**, 675-686C, doi:10.2471/BLT.16.171082 (2016).

2 Ministério da Saúde, Informe Epidemiológico No. 57 - Semana epidemiológica 52/2016 - Monitoramento dos casos de microcefalia no Brasil. http://www.combateaedes.saude.gov.br/images/pdf/Informe-Epidemiologico-n57-SE-52_2016-09jan2017.pdf, 1-3 (2017).

3 WHO. Situation Report - Zika virus, microcephaly, Guillain-Barré syndrome (18 Jan 2017). (http://apps.who.int/iris/bitstream/10665/253604/1/zikasitrep20Jan17-eng.pdf?ua=1, 2017).

4 Faria, N. R. *et al.* Zika virus in the Americas: Early epidemiological and genetic findings. *Science* **352**, 345-349, doi:10.1126/science.aaf5036 (2016).

5 Alex Perkins, T., Siraj, A. S., Ruktanonchai, C. W., Kraemer, M. U. & Tatem, A. J. Model-based projections of Zika virus infections in childbearing women in the Americas. *Nat Microbiol* **1**, 16126, doi:10.1038/nmicrobiol.2016.126 (2016).

6 Lessler, J. *et al.* Assessing the global threat from Zika virus. *Science* **353**, aaf8160, doi:10.1126/science.aaf8160 (2016).

7 Vasconcelos, P. F. & Calisher, C. H. Emergence of Human Arboviral Diseases in the Americas, 2000-2016. *Vector Borne and Zoonotic Diseases* **16**, 295-301, doi:10.1089/vbz.2016.1952 (2016).

8 Vogel, G. One year later, Zika scientists prepare for a long war. *Science* **354**, 1088-1089 (2016).

9 Bogoch, II *et al.* Potential for Zika virus introduction and transmission in resource-limited countries in Africa and the Asia-Pacific region: a modelling study. *The Lancet Infectious Diseases* **16**, 1237-1245, doi:10.1016/S1473-3099(16)30270-5 (2016).

10 Lessler, J. T., Ott, C.T., Carcelen, A.C., Konikoff, J.M., Williamson, J., Bi, Q., et al. . Times to key events in the course of Zika infection and their implications: a systematic review and pooled analysis *[Submitted]. Bull World Health Organ* **DOI: 10.2471/BLT.16.174540** (2016).

11 Pacheco, O. *et al.* Zika Virus Disease in Colombia - Preliminary Report. *The New England Journal of Medicine*, doi:10.1056/NEJMoa1604037 (2016).

12 Caminade, C. *et al.* Global risk model for vector-borne transmission of Zika virus reveals the role of El Nino 2015. *Proceedings of the National Academy of Sciences of the United States of America* **114**, 119-124, doi:10.1073/pnas.1614303114 (2017).

13 Quick, J. *et al.* Real-time, portable genome sequencing for Ebola surveillance. *Nature* **530**, 228-232, doi:10.1038/nature16996 (2016).

14 Quick J., G., N. D., Pullan, S. T., Claro, I. M., Smith, A. D., Gangavarapu, k., Oliveira, G., Robles-Sikisaka, R., Rogers, T. F., Beutler, N. A., Burton, D. R., Lewis-Ximenez, L. L., de Jesus, J. G., Giovanetti, M., Hill, S., Black, A., Bedford, T., Carroll, M. W., Nunes, M., Alcantara, L. C., Sabino, E. C., Baylis, S. A., Faria, N. R., Loose, M., Simpson, J. T., Pybus, O. G, Andersen, K. G., Loman, N. J. Multiplex PCR method for MinION and Illumina sequencing of Zika and other virus genomes directly from clinical samples. *BioRxiv* https://doi.org/10.1101/098913 (2017).

15 Trosemeier, J. H. *et al.* Genome Sequence of a Candidate World Health Organization Reference Strain of Zika Virus for Nucleic Acid Testing. *Genome Announcements* **4**, doi:10.1128/genomeA.00917-16 (2016).
16 Giovanetti, M. *et al.* Zika virus complete genome from Salvador, Bahia, Brazil. *Infection, Genetics and Evolution* **41**, 142-145, doi:10.1016/j.meegid.2016.03.030 (2016).
17 Naccache, S. N. *et al.* Distinct Zika Virus Lineage in Salvador, Bahia, Brazil. *Emerging infectious diseases* **22**, doi:10.3201/eid2210.160663 (2016).
18 Corman, V. M. *et al.* Assay optimization for molecular detection of Zika virus. *Bulletin of the World Health Organization* **94**, 880-892, doi:10.2471/BLT.16.175950 (2016).
19 Liu, H. *et al.* From discovery to outbreak: the genetic evolution of the emerging Zika virus. *Emerg Microbes Infect* **5**, e111, doi:10.1038/emi.2016.109 (2016).
20 Pettersson, J. H. O., Eldholm, V., Seligmna, S. J., Lundkvist, A., Falconar, A. K., Gaunt, M. W., Musso, D., Nougairede, A., Charrel, R., Gould, E. A., Lamballerie, X. How Did Zika Virus Emerge in the Pacific Islands and Latin America? *mBio* **7**, 201239-201216 (2016).
21 Holmes, E. C., Dudas, G., Rambaut, A. & Andersen, K. G. The evolution of Ebola virus: Insights from the 2013-2016 epidemic. *Nature* **538**, 193-200, doi:10.1038/nature19790 (2016).
22 Holmes, E. C. Patterns of intra- and interhost nonsynonymous variation reveal strong purifying selection in dengue virus. *Journal of Virology* **77**, 11296-11298 (2003).
23 Park, D. J. *et al.* Ebola Virus Epidemiology, Transmission, and Evolution during Seven Months in Sierra Leone. *Cell* **161**, 1516-1526, doi:10.1016/j.cell.2015.06.007 (2015).
24 De Maio, N., Wu, C. H., O'Reilly, K. M. & Wilson, D. New Routes to Phylogeography: A Bayesian Structured Coalescent Approximation. *PLoS Genetics* **11**, e1005421, doi:10.1371/journal.pgen.1005421 (2015).
25 Lemey, P., Rambaut, A., Drummond, A. J. & Suchard, M. A. Bayesian phylogeography finds its roots. *PLoS Computational Biology* **5**, e1000520, doi:10.1371/journal.pcbi.1000520 (2009).
26 Campos, G. S., Bandeira, A. C. & Sardi, S. I. Zika Virus Outbreak, Bahia, Brazil. *Emerging Infectious Diseases* **21**, 1885-1886, doi:10.3201/eid2110.150847 (2015).
27 Zanluca, C. *et al.* First report of autochthonous transmission of Zika virus in Brazil. *Memorias do Instituto Oswaldo Cruz* **110**, 569-572, doi:10.1590/0074-02760150192 (2015).
28 Lanciotti, R. S. *et al.* Genetic and serologic properties of Zika virus associated with an epidemic, Yap State, Micronesia, 2007. *Emerging Infectious Diseases* **14**, 1232-1239, doi:10.3201/eid1408.080287 (2008).
29 Quick, J., Grubaugh, N. D., Pullan, S. T., Claro, I. M., Smith, K. G., Oliveira, G., Robles-Sikisaka, R., Rogers, T. F., Beutler, N. A., Burton, D. R., Lewis-Ximenez, L. L., de Jesus, J. G., Giovanetti, M., Hill, S., Black, A., Bedford, T., Carroll, M. W., Nunes, M. R. T., Alcantara, L. C. J., Sabino, E. C., Baylis, S. A., Faria, N. R., Loose, M., Simpson, J. T., Pybus, O. G., Andersen, K. G., Loman, N. J. Multiplex PCR method for MinION and Illumina sequencing of Zika and other virus genomes directly from clinical samples. *bioRxiv* **pre-print**, https://doi.org/10.1101/098913 (2017).

30 Kozlov, A. M., Aberer, A. J., Stamatakis, A. ExaML version 3: a tool for phylogenomic analyses on supercomputers. *Bioinformatics* **31**, 2577-2579 (2015).
31 Guindon, S. *et al.* New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. *Systematic Biology* **59**, 307-321, doi:10.1093/sysbio/syq010 (2010).
32 Hasegawa, M., Kishino, H. & Yano, T. Dating of the human-ape splitting by a molecular clock of mitochondrial DNA. *Journal of Molecular Evolution* **22**, 160-174 (1985).
33 Darriba, D., Taboada, G. L., Doallo, R. & Posada, D. jModelTest 2: more models, new heuristics and parallel computing. *Nature Methods* **9**, 772, doi:10.1038/nmeth.2109 (2012).
34 Schierup, M. H. & Hein, J. Consequences of recombination on traditional phylogenetic analysis. *Genetics* **156**, 879-891 (2000).
35 Faye, O. *et al.* Molecular evolution of Zika virus during its emergence in the 20(th) century. *PLoS Negl Trop Dis* **8**, e2636, doi:10.1371/journal.pntd.0002636 (2014).
36 Martin, D. P., Murrell, B., Golden, M., Khoosal, A. & Muhire, B. RDP4: Detection and analysis of recombination patterns in virus genomes. *Virus Evol* **1**, vev003, doi:10.1093/ve/vev003 (2015).
37 Bruen, T. C., Philippe, H. & Bryant, D. A simple and robust statistical test for detecting the presence of recombination. *Genetics* **172**, 2665-2681, doi:10.1534/genetics.105.048975 (2006).
38 Huson, D. H. & Bryant, D. Application of phylogenetic networks in evolutionary studies. *Molecular Biology and Evolution* **23**, 254-267, doi:10.1093/molbev/msj030 (2006).
39 Rambaut, A., Lam, T. T., Fagundes de Carvalho, L., Pybus, O. G. Exploring the temporal structure of heterochronous sequences using TempEst (formerly Path-O-Gen). *Virus Evolution* **2** (2016).
40 Drummond, A. J., Suchard, M. A., Xie, D. & Rambaut, A. Bayesian phylogenetics with BEAUti and the BEAST 1.7. *Molecular Biology and Evolution* **29**, 1969-1973, doi:10.1093/molbev/mss075 (2012).
41 Baele, G., Li, W. L., Drummond, A. J., Suchard, M. A. & Lemey, P. Accurate model selection of relaxed molecular clocks in bayesian phylogenetics. *Molecular Biology and Evolution* **30**, 239-243, doi:10.1093/molbev/mss243 (2013).
42 Shapiro, B., Rambaut, A. & Drummond, A. J. Choosing appropriate substitution models for the phylogenetic analysis of protein-coding sequences. *Molecular Biology and Evolution* **23**, 7-9, doi:10.1093/molbev/msj021 (2006).
43 Gill, M. S. *et al.* Improving Bayesian population dynamics inference: a coalescent-based model for multiple loci. *Molecular Biology and Evolution* **30**, 713-724, doi:10.1093/molbev/mss265 (2013).
44 Drummond, A. J., Ho, S. Y., Phillips, M. J. & Rambaut, A. Relaxed phylogenetics and dating with confidence. *PLoS Biology* **4**, e88, doi:10.1371/journal.pbio.0040088 (2006).
45 Ferreira, M. A. R. & Suchard, M. A. Bayesian analysis of elapsed times in continuous-time Markov chains. *Can J Stat* **36**, 355-368 (2008).
46 Kosakovsky Pond, S. L., Frost, S. D. Not so different after all: a comparison of methods for detecting amino acid sites under selection. *Molecular Biology and Evolution* **22**, 1208-1222 (2005).

47 Pond, S. L., Frost, S. D. & Muse, S. V. HyPhy: hypothesis testing using phylogenies. *Bioinformatics* **21**, 676-679, doi:10.1093/bioinformatics/bti079 (2005).
48 Edwards, C. J. *et al.* Ancient hybridization and an Irish origin for the modern polar bear matriline. *Current Biology : CB* **21**, 1251-1258, doi:10.1016/j.cub.2011.05.058 (2011).
49 Bouckaert, R. *et al.* BEAST 2: a software platform for Bayesian evolutionary analysis. *PLoS Computational Biology* **10**, e1003537, doi:10.1371/journal.pcbi.1003537 (2014).
50 Minin, V. N. & Suchard, M. A. Fast, accurate and simulation-free stochastic mapping. *Philos Trans R Soc Lond B Biol Sci* **363**, 3985-3995, doi:10.1098/rstb.2008.0176 (2008).
51 O'Brien, J. D., Minin, V. N. & Suchard, M. A. Learning to count: robust estimates for labeled distances between molecular sequences. *Molecular biology and evolution* **26**, 801-814, doi:10.1093/molbev/msp003 (2009).
52 Wickham, H. *ggplot2: elegant graphics for data analysis*. (Springer New York, 2009).
53 R: A Language and Environment for Computing (R Foundation for Statistical Computing, Vienna, Austria, 2014).
54 Cori, A., Ferguson, N. M., Fraser, C. & Cauchemez, S. A new framework and software to estimate time-varying reproduction numbers during epidemics. *American Journal of Epidemiology* **178**, 1505-1512, doi:10.1093/aje/kwt133 (2013).
55 Ferguson, N. M. *et al.* EPIDEMIOLOGY. Countering the Zika epidemic in Latin America. *Science* **353**, 353-354, doi:10.1126/science.aag0219 (2016).
56 Kraemer, M. U. *et al.* The global distribution of the arbovirus vectors Aedes aegypti and Ae. albopictus. *eLife* **4**, e08347, doi:10.7554/eLife.08347 (2015).

Acknowledgments

We are grateful for the essential contributions to this project of Fundação Oswaldo Cruz in Bahia and Pernambuco states, University of São Paulo, and Instituto Evandro Chagas. We thank the Brazilian Zika virus surveillance network for their assistance to the ZiBRA project. We thank FIOCRUZ Bahia for providing the vehicle for the mobile lab. Robert Lanciotti (US CDC) kindly gave permission to use his unpublished genomes in analyses. We thank Pedro Fernando da Costa Vasconcelos, Sueli Guerreiro Rodrigues and João Vianez Junior (Instituto Evandro Chagas, Brazil), Juliana Gil Melgaço (Fiocruz Rio de Janeiro, Brazil), Johannes Blumel (Paul-Ehrlich-Institut, Langen, Germany), Marcia Cristina Brito Lobato, Liliana Nunes Fava (Tocantins State Department of Health, Brazil) and Constancia Ayres (Instituto Aggeu Magalhães, Fundação Oswaldo Cruz (FIOCRUZ), Recife, Pernambuco, Brazil). CYC thanks José E. Muñoz-Medina and Cesar R. González-Bonilla (Instituto Mexicano del Seguro Social) and Carlos F. Arias and Susana López (Universidad Nacional Autónoma de México). LCJA thanks QIAGEN for reagents and equipment for the ZiBRA project. MRTN thanks FERPEL for providing consumables. We are grateful to the staff of Oxford Nanopore for technical support to the project, with particular thanks to Rosemary Dokos, Gordon Sanghera and Oliver Hartwell.

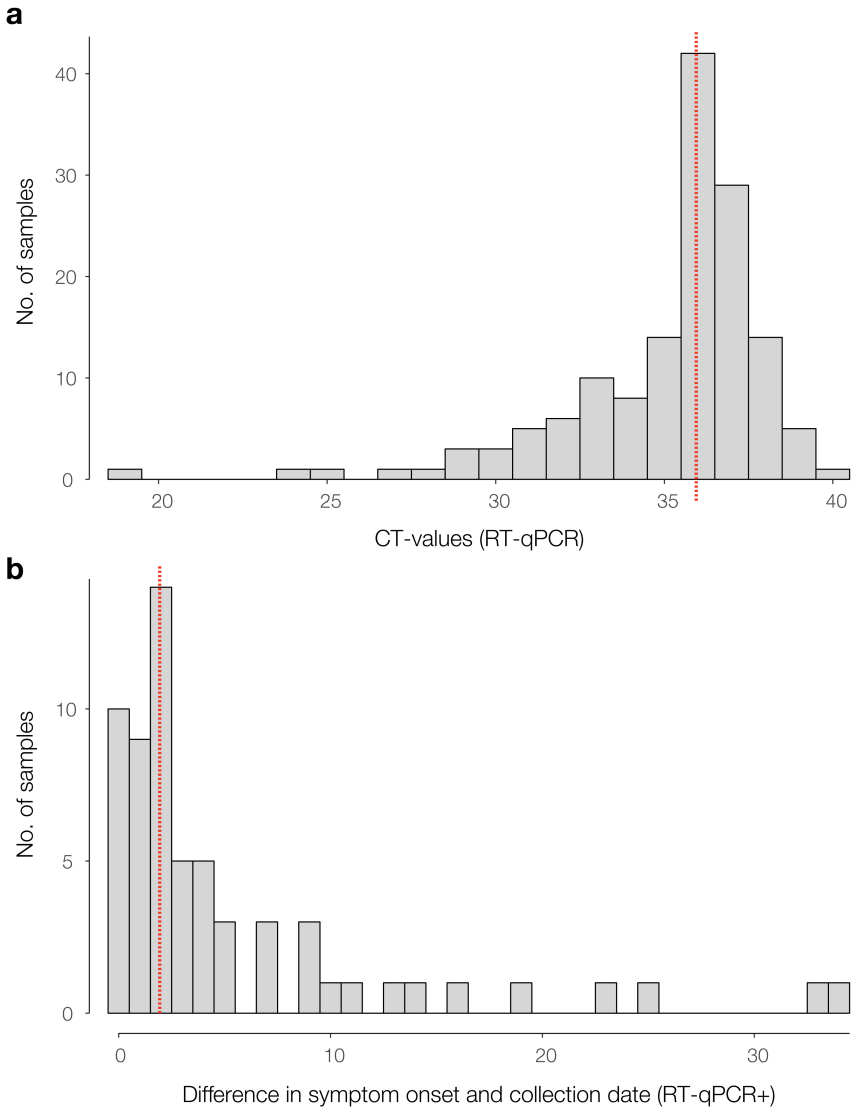
Funding

This work was supported by the Medical Research Council/Wellcome Trust/Newton Fund Zika Rapid Response Initiative (grant number MC_PC_15100/ ZK/16-078) which also supports JQ's salary (Grant) and by the generous support of the American people through the United States Agency for International Development Emerging Pandemic Threats Program-2 PREDICT-2 (Cooperative Agreement No. AID-OAA-A-14-00102), which also supports MUGK's salary. NJL is supported by a Medical Research Council Bioinformatics Fellowship as part of the Cloud Infrastructure for Microbial Bioinformatics (CLIMB) project. NRF is funded by a Sir Henry Dale Fellowship (Wellcome Trust / Royal Society Grant 204311/Z/16/Z). CNPq contributed to the trip expenses (grant no. 457480/2014-9). ACC was supported by FAPESP #2012/03417-7. MRTN is supported by the Brazilian National Council of Scientific and Technological Development (CNPq) grant no. 302584/2015-3. AB and TB were supported by NIH award R35 GM119774. AB is supported by the National Science Foundation Graduate Research Fellowship Program under Grant No. DGE-1256082. TB is a Pew Biomedical Scholar. CYC is partially supported by NIH grant R01 HL105704 and a pathogen discovery award from Abbott Laboratories, Inc. EH is supported by a National Health and Medical Research Council Australia Fellowship (GNT1037231). SCH is supported by Wellcome Trust Grant 102427. This research received funding from the ERC under the European Union's Seventh Framework Programme (FP7/2007-2013)/ERC, grant agreement numbers 614725-PATHPHYLODYN and 278433-PREDEMICS, and from European Union Horizon 2020 under grant agreements 643476-COMPARE and 734548-ZIKAlliance (to SC). TJ and ETJM and acknowledge funding from IDAMS, DENFREE, DengueTools, and from PPSUS-FACEPE (project Number APQ-0302-4.01/13). RFF received funding from FACEPE, grant number: APQ-0044.2.11/16 and APQ-0055.2.11/16 and from CNPq 439975/2016-6. SAB was supported by the Sicherheit von Blut und Geweben hinsichtlich der Abwesenheit von Zikaviren from the German Ministry of Health.

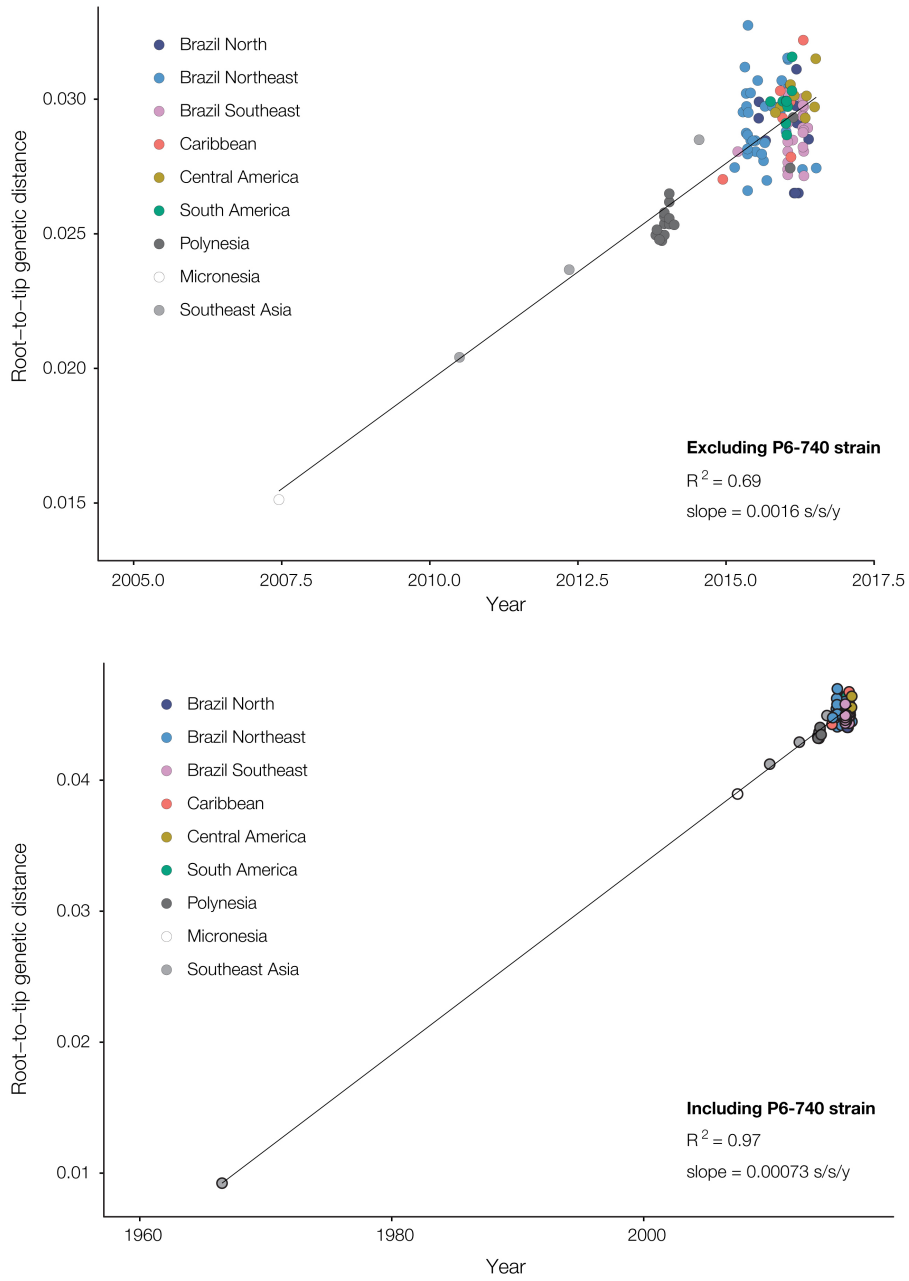
Conflicts of Interest

NJL received an honorarium for speaking at an Oxford Nanopore meeting. NJL and NRF received travel and accommodation to attend London Calling meetings. NJL has ongoing research collaborations with Oxford Nanopore Technologies and has received free-of-charge reagents in support of the ZiBRA project. OGP receives consultancy income from Metabiota Inc, CA, USA. CYC is the director of the UCSF-Abbott Viral Diagnostics and Discovery Center and receives research support from Abbott Laboratories, Inc.

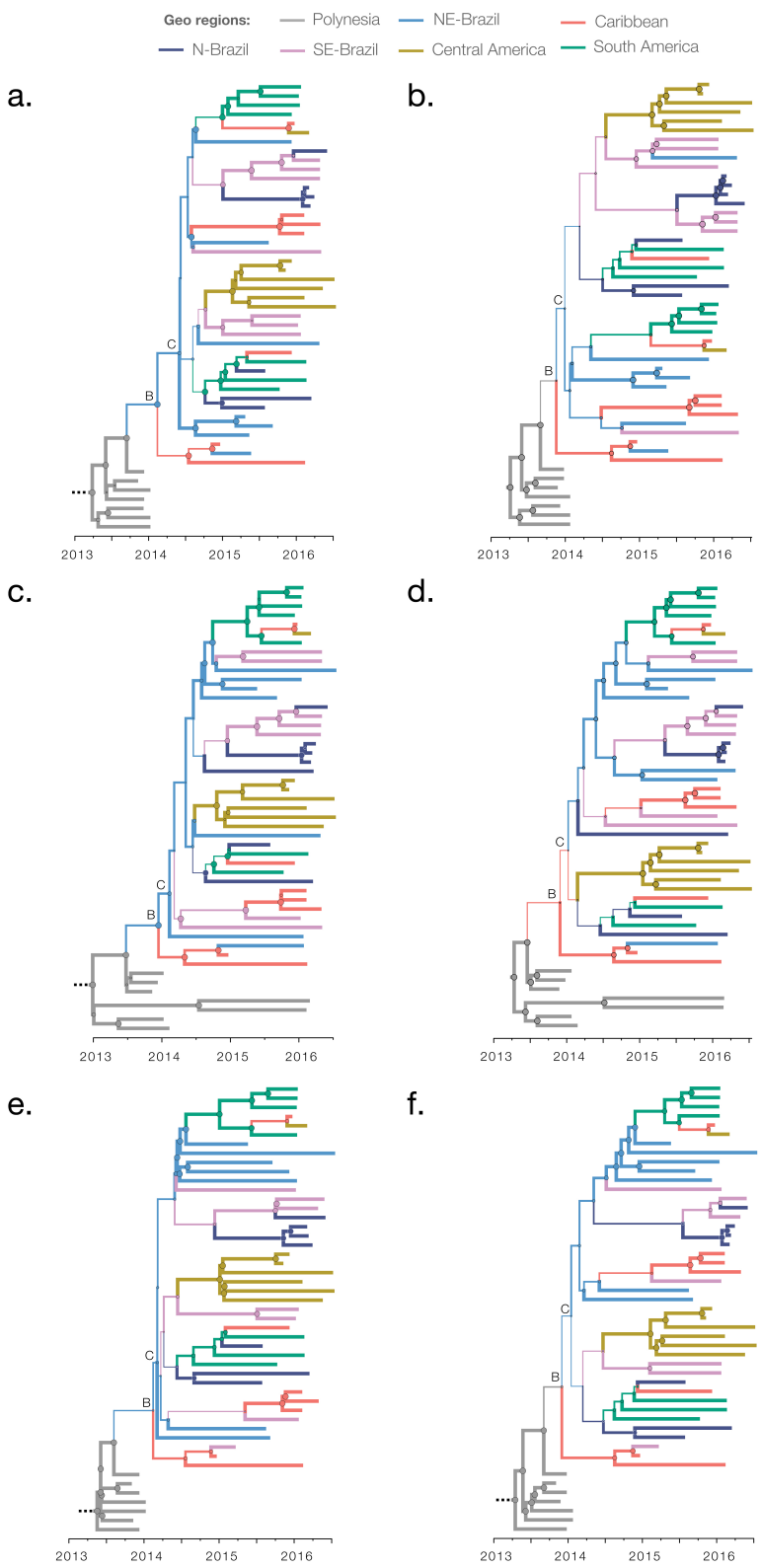
Supplementary Information



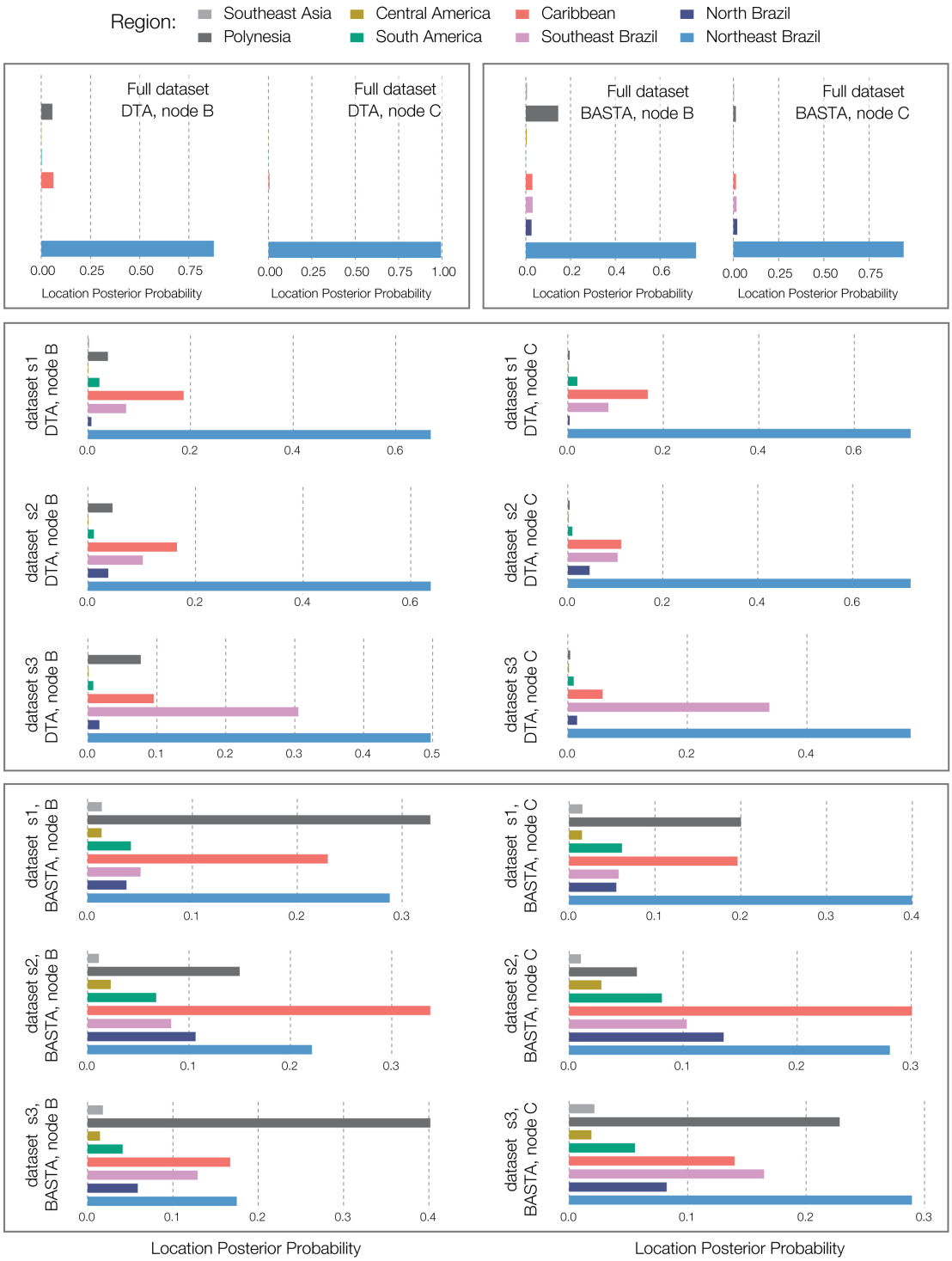
SI Fig. 1. Panel **a** shows the distribution of RT-qPCR+ samples tested during the ZiBRA journey in Brazil (*n*=181; median = 35.96). Panel **b** shows the temporal distribution of the difference between the date of onset of clinical symptoms and the date of sample collection of RT-qPCR+ samples (median = 2 days). Vertical dashed lines represent the median of the distributions.



SI Fig. 2. Temporal signal of the ZIKV Asian genotype. The correlation between sampling dates and genetic distances from the tips to the root of a maximum likelihood (ML) tree estimated PhyML[1] was explored using TempEst[2]. (**a**) Estimates for the dataset used for the phylogenetic analysis in **Fig. 3**, and (**b**) estimates for the same dataset including the P6-740 strain sampled in 1966 (accession number HQ234499).

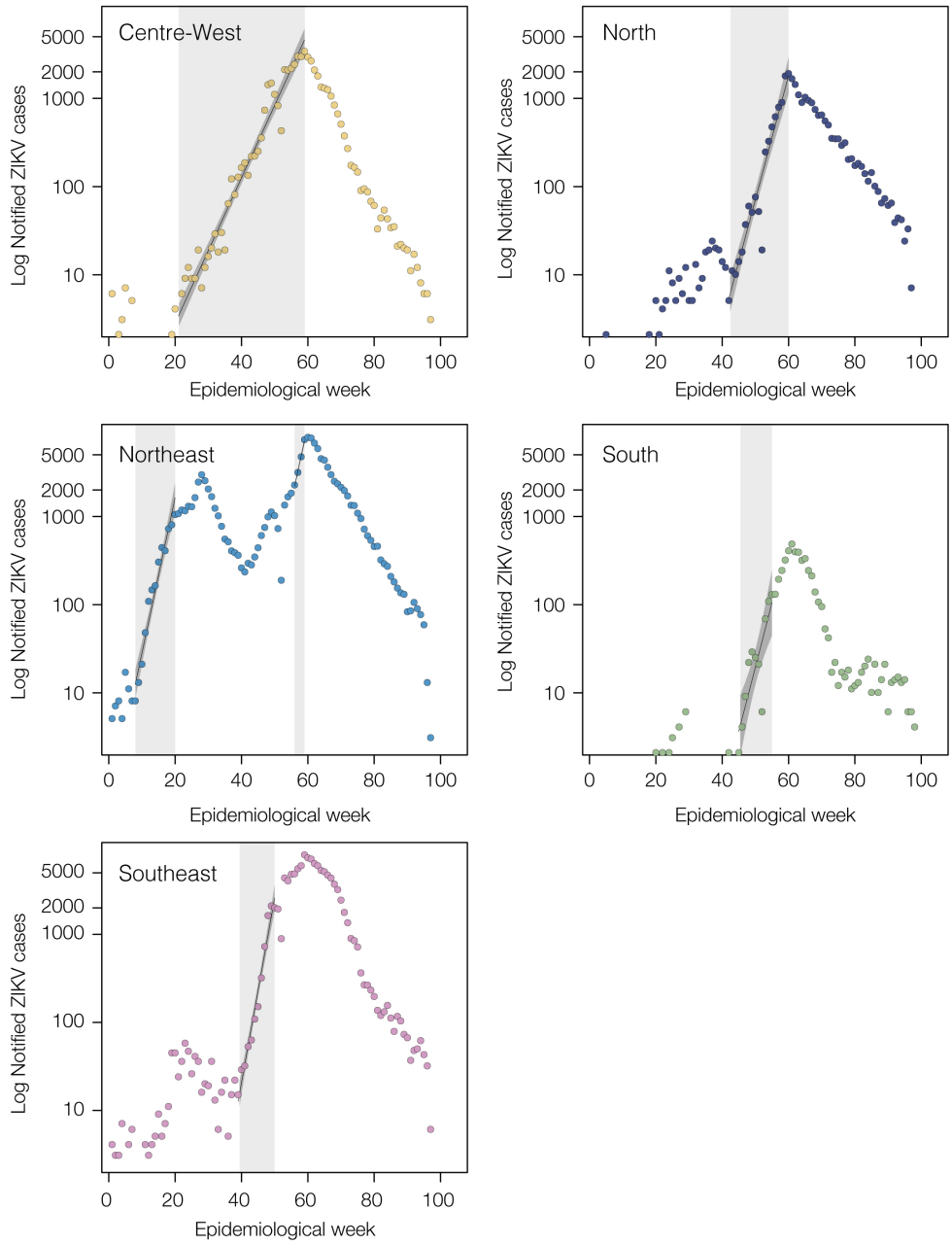


SI Fig. 3. Maximum clade credibility (MCC) trees for three subsampled datasets (dataset s1: panels **a** and **b**; dataset s2: panels **c** and **d**; dataset s3: panels **e** and **f**). On the left (panels **a**, **c** and **e**), the MCC trees generated using an asymmetric discrete trait model[3,4]; while on the right (panels **b**, **d** and **f**) MCC reconstructions were generated using a structured coalescent approximation approach[5]. Branch width represents the posterior modal location.



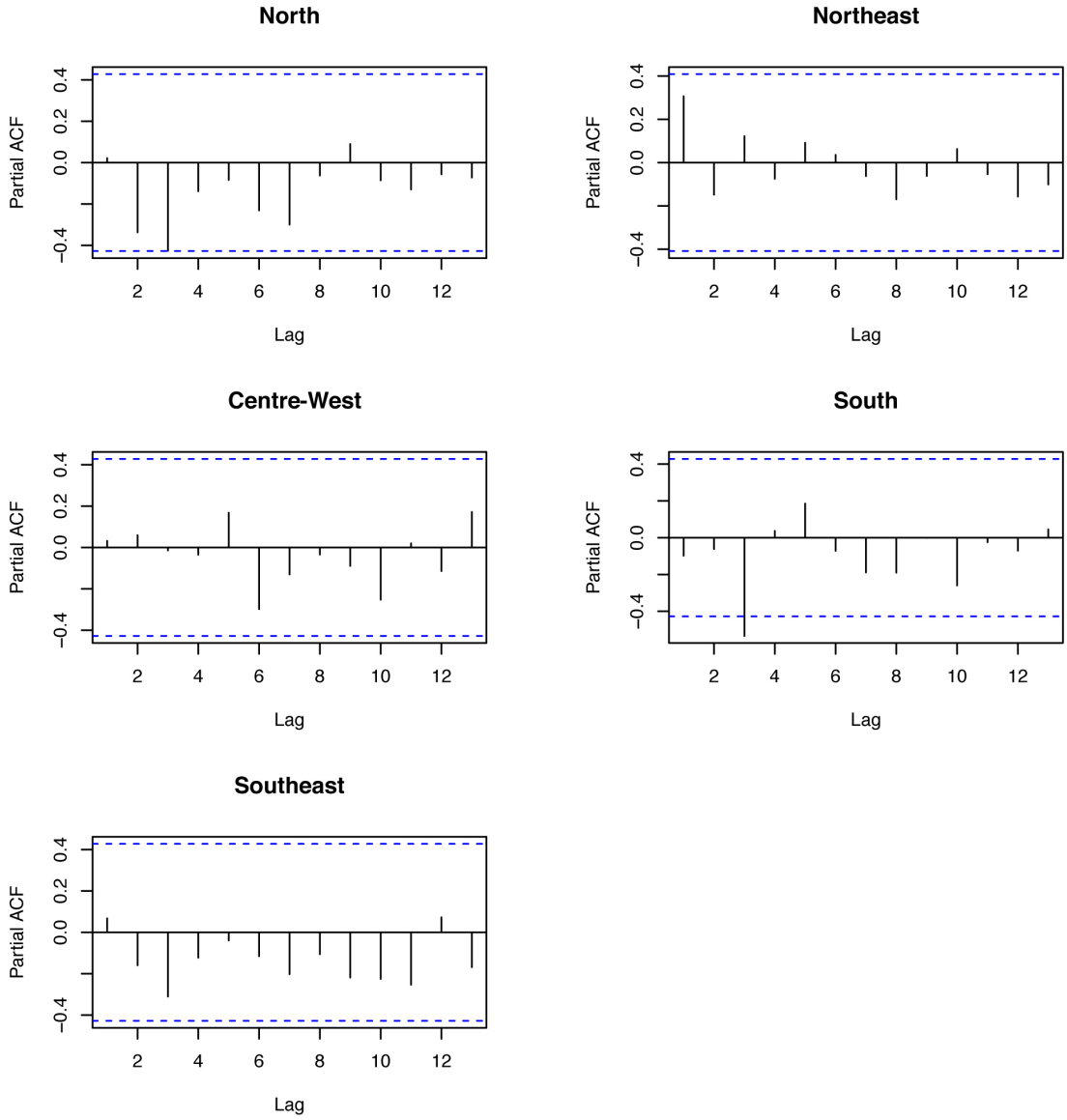
SI Fig. 4. Posterior probabilities of the locations of nodes A, B and C, estimated using the full dataset (upper panel) and three subsampled datasets (s1, s2 and s3; see Methods and SI Fig. 3). DTA=discrete trait analysis method[4]. BASTA: Bayesian structured coalescent approximation method[5]. For each method, we used an asymmetric model of location exchange to estimate ancestral node locations and to infer patterns of virus spread among regions.

SI Fig. 5. ZIKV non-clock maximum likelihood tree. Branch support is shown at each node. Tree was estimated using PhyML[1]. Sequences generated in this study are highlighted in red. Scale bar is shown in units of nucleotide substitutions per site.



SI Fig. 6. Epidemic growth rates estimated from weekly ZIKV notified cases in Brazil. Time series show the number of ZIKV notified cases in each region of Brazil. Periods from which exponential growth were estimated are highlighted in grey.

SI Fig. 7. Seasonal suitability for ZIKV transmission in the Americas. These maps were estimated by collating data on *Aedes* mosquitoes, temperature, relative humidity and precipitation, and are the basis of the trends in suitability for different regions shown in main text **Figs. 1 and 4**. For details, see[6,7].



SI Fig. 8. Partial autocorrelation functions for the linear model associating climatic suitability and ZIKV notified cases in each geographic region in Brazil. The residuals for the North, Northeast, Centre-West and Southeast regions show no autocorrelation, while a small amount of autocorrelation cannot be excluded for the South region.

SI Tables

SI Table 1. Summary of the clinical samples tested (*n*=1330, of which 181 were RT-qPCR positive) by the ZiBRA mobile lab in June 2016, NE Brazil. ZIKV notified cases were confirmed using RT-qPCR (see Methods). The collection lag represents the median time interval (in days) between the date of onset of clinical symptoms and the date of sample collection (both dates available for *n*=219) for all samples (including those that subsequently tested RT-qPCR negative). Northeast Brazilian states where samples were tested were RN: Rio Grande do Norte, PB: Paraíba, PE: Pernambuco, AL: Alagoas, BA: Bahia.

Laboratory, Federal state	**No. Positives / Tested (%)**	**Ct value (mean, min-max)**	**Collection lag (median, min-max)**
LACEN, RN	27/335 (8.1%)	35.9 (18.6-39.1)	5 (4-16)
LACEN, PB	26/276 (9.4%)	35.7 (30.7-37.0)	6 (0-88)
FioCruz, PE	95/315 (30%)[1]	34.6 (24.1-38.3)	2.5 (0-33)
LACEN, AL	16/140 (11%)	34.1 (27.1-40.2)	2 (0-3)
FioCruz, BA	17/264 (6.4%)	35.8 (24.7-39.2)	4 (0-228)

[1] Includes RT-PCR+ cases from Pernambuco that were generated at Fiocruz Pernambuco.

SI Table 2. Parameters of the model linking climatic vector suitability and notified ZIKV cases in different Brazilian regions. For each region the table provides the estimated correlated time period (T), P-value of the linear term of suitability in T, adjusted-R^2 of the model, and time lag (l).

	North	**Northeast**	**Centre-West**	**South**	**Southeast**
Correlated time period	12/2015 to 10/2016	7/2015 to 10/2016	9/2015 to 8/2016	6/2015 to 05/2016	11/2015 to 9/2016
***P*-value**	<0.0001	0.00013	<0.0001	<0.0001	<0.0001
Adjusted-R^2	0.929	0.8448	0.987	0.9543	0.953
Time lag (months)	1.27	0	1.12	1.19	1.33

SI Table 3. For each region, estimates of the basic reproductive number (*R*) of ZIKV, for several values of generation time (*g*) parameter are shown together with the corresponding estimates of exponential growth rate (*r*) (per day) obtained from notified ZIKV case counts (see also **SI Fig. 7**). CW: Centre-West, N: North, NE: Northeast (1st = epidemic wave in 2015; 2nd = epidemic wave in 2016), SE: Southeast, S: South. CI: 95% confidence interval.

Region	***R* (mean, CI), *g* =20 days**	***R* (mean, CI), *g* =15 days**	***R* (mean, CI), *g*=10 days**	**Growth rate (*r*, CI)**
CW	1.71 (1.65-1.78)	1.46 (1.20-1.77)	1.29 (1.13-1.46)	0.027 (0.02-0.03)
N	2.48 (2.19-2.81)	1.98 (1.80-2.18)	1.58 (1.48-1.69)	0.046 (0.04-0.05)
NE, 1st	3.12 (2.69-3.60)	2.36 (2.11-2.63)	1.78 (1.65-1.91)	0.06 (0.05-0.07)
NE, 2nd	3.03 (2.74-3.36)	2.31 (2.14-2.49)	1.75 (1.66-1.84)	0.06 (0.05-0.06)
SE	3.85 (3.35-4.42)	2.77 (2.49-3.07)	1.98 (1.84-2.12)	0.07 (0.06-0.076)
S	2.57 (1.72-3.82)	2.04 (1.50-2.75)	1.61 (1.31-1.97)	0.05 (0.04-0.07)

SI Table 4. Differences in log-marginal likelihood estimates using the path-sampling (PS) and Stepping-Stone (SS) model selection approaches[8,9]. The molecular clock and coalescent model combinations are ordered by log Bayes factor (BF) estimates; the best-fitting combination is underscored. BF was calculated as the ratio between two models, H0 and H1, where H0 is the simpler model combination. Two molecular clock models were tested here. SC: Strict clock model, and the UCLD: uncorrelated relaxed clock with lognormal distribution[10].

Clock	Coalescent	PS	BF (PS)	SS	BF (SS)
UCLD	Skygrid	-24147.6	61.49	-24148.73	60.73
UCLD	Skyline	-24150.26	58.83	-24151.14	58.32
SC	Skyline	-24162.58	46.51	-24163.84	45.62
UCLD	Logistic	-24169.53	39.56	-24170.30	39.16
SC	Skygrid	-24172.99	36.1	-24174.08	25.38
UCLD	Exponential	-24182.5	26.59	-24183.09	26.37
UCLD	Constant	-24185.9	23.19	-24186.49	22.97
SC	Logistic	-24192.97	16.12	-24193.37	16.09
SC	Exponential	-24206.01	3.08	-24206.57	2.89
SC	Constant	-24209.09	-	-24209.46	-

SI Table 5. Estimates dates of nodes A, B and C (Fig 3) under various different molecular clock and coalescent model combinations. TMRCA: time of the most recent common ancestor, BCI: Bayesian credible interval, SC: strict molecular clock model, UCLN: uncorrelated clock with lognormal distribution.

Clock model	**Coalescent prior**	***Node A*** **TMRCA (95% BCIs)**	***Node B*** **TMRCA (95% BCIs)**	***Node C*** **TMRCA (95% BCIs)**
SC	Constant	2013.5 (2013.2, 2013.7)	2013.6 (2013.4, 2013.9)	2013.7 (2013.5, 2014)
SC	Exponential	2013.4 (2013.1, 2013.7)	2013.6 (2013.7, 2013.9)	2013.9 (2016.5, 2013.4)
SC	Logistic	2013.4 (2013.2, 2013.7)	2013.7 (2013.4, 2013.9)	2013.73 (2013.5, 2014.0)
SC	Skygrid	2013.7 (2013.5, 2013.8)	2013.87 (2013.6, 2014.1)	2013.97 (2013.7, 2014.2)
SC	Skyline	2013.6 (2013.5, 2013.8)	2013.83 (2013.6, 2014.0)	2013.91 (2013.7, 2014.1)
UCLN	Constant	2013.5 (2013.2, 2013.8)	2013.71 (2013.3, 2014.1)	2013.86 (2013.5, 2014.2)
UCLN	Exponential	2013.6 (2013.3, 2013.9)	2013.77 (2013.4, 2014.1)	2013.9 (2013.6, 2014.2)
UCLN	Logistic	2013.6 (2013.4, 2013.9)	2013.8 (2013.5, 2014.1)	2013.93 (2013.6, 2014.2)
UCLN	Skygrid	2013.7 (2013.5, 2013.9)	2013.95 (2013.7, 2014.2)	2014.08 (2013.8, 2014.3)
UCLN	Skyline	2013.7 (2013.6, 2013.9)	2013.98 (2013.7, 2014.3)	2014.12 (2013.9, 2014.4)

References

1 Guindon, S., Delsuc, F., Dufayard, J. F. & Gascuel, O. Estimating maximum likelihood phylogenies with PhyML. *Methods in Molecular Biology* **537**, 113-137, doi:10.1007/978-1-59745-251-9_6 (2009).

2 Rambaut, A., Lam, T. T., Fagundes de Carvalho, L., Pybus, O. G. Exploring the temporal structure of heterochronous sequences using TempEst (formerly Path-O-Gen). *Virus Evolution* **2** (2016).

3 Edwards, C. J. *et al.* Ancient hybridization and an Irish origin for the modern polar bear matriline. *Current Biology : CB* **21**, 1251-1258, doi:10.1016/j.cub.2011.05.058 (2011).

4 Lemey, P., Rambaut, A., Drummond, A. J. & Suchard, M. A. Bayesian phylogeography finds its roots. *PLoS Computational Biology* **5**, e1000520, doi:10.1371/journal.pcbi.1000520 (2009).

5 De Maio, N., Wu, C. H., O'Reilly, K. M. & Wilson, D. New Routes to Phylogeography: A Bayesian Structured Coalescent Approximation. *PLoS Genetics* **11**, e1005421, doi:10.1371/journal.pgen.1005421 (2015).

6 Bogoch, II *et al.* Potential for Zika virus introduction and transmission in resource-limited countries in Africa and the Asia-Pacific region: a modelling study. *The Lancet Infectious Diseases* **16**, 1237-1245, doi:10.1016/S1473-3099(16)30270-5 (2016).

7 Kraemer, M. U. *et al.* The global distribution of the arbovirus vectors Aedes aegypti and Ae. albopictus. *eLife* **4**, e08347, doi:10.7554/eLife.08347 (2015).

8 Baele, G. *et al.* Improving the accuracy of demographic and molecular clock model comparison while accommodating phylogenetic uncertainty. *Molecular Biology and Evolution* **29**, 2157-2167, doi:10.1093/molbev/mss084 (2012).

9 Baele, G., Li, W. L., Drummond, A. J., Suchard, M. A. & Lemey, P. Accurate model selection of relaxed molecular clocks in bayesian phylogenetics. *Molecular Biology and Evolution* **30**, 239-243, doi:10.1093/molbev/mss243 (2013).

10 Drummond, A. J., Ho, S. Y., Phillips, M. J. & Rambaut, A. Relaxed phylogenetics and dating with confidence. *PLoS Biology* **4**, e88, doi:10.1371/journal.pbio.0040088 (2006).