Establishment and cryptic transmission of Zika virus in Brazil and the Americas

1 **Establishment and cryptic transmission of Zika virus in Brazil and**
2 **the Americas**

3

4 Faria, N. R.*[1,2], Quick, J.[3]*, Morales, I.[4]*, Thézé, J.[1]*, Jesus, J.G.[5]*, Giovanetti, M.[5,6]*,
5 Kraemer, M. U. G. [1,7,8]*, Hill, S. C.[1]*, Black, A.[9,10]*, da Costa, A. C.[3], Franco, L.C.[2], Silva, S.
6 P.[2], Wu, C.-H.[11], Raghwani, J.[1], Cauchemez, S.[12,13], du Plessis, L.[1], Verotti, M. P.[14], de
7 Oliveira, W. K.[15,16], Carmo, E. H.[17], Coelho, G. E.[18,19], Santelli, A. C. F. S.[18,20], Vinhal, L.
8 C.[18], Henriques, C. M.[17], Simpson, J. T.[21], Loose, M.[22], Andersen, K. G.[23], Grubaugh, N. D.[23],
9 Somasekar, S.[24], Chiu, C. Y.[24], Muñoz-Medina, J. E.[25], Gonzalez-Bonilla, C. R.[25], Arias, C. F.
10 [26], Lewis-Ximenez, L. L.[27], Baylis, S.A.[28], Chieppe, A. O.[29], Aguiar, S. F.[29], Fernandes, C.
11 A.[29], Lemos, P. S.[2], Nascimento, B. L. S.[2], Monteiro, H. A. O.[2], Siqueira, I. C.[5], de Queiroz,
12 M. G.[30], de Souza, T. R.[30,31], Bezerra, J. F.[30,32], Lemos, M. R.[33], Pereira, G. F.[33], Loudal, D.[33],
13 Moura, L. C.[33], Dhalia, R.[34], França, R. F.[34], Magalhães, T.[34], Marques, E. T. Jr.[34,35], Jaenisch,
14 T.,[36], Wallau, G. L.[34], de Lima, M. C.[37], Nascimento, V.[37], de Cerqueira, E. M.[38], de Lima, M.
15 M.[38], Mascarenhas, D. L.[39], Moura Neto, J. P.[40], Levin, A. S.[4], Tozetto-Mendoza, T. R.[4],
16 Fonseca, S. N.[41], Mendes-Correa, M. C. [4], Milagres, F.P.[42], Segurado, A.[4], Holmes, E. C.[43],
17 Rambaut, A.[44,45], Bedford, T.[7], Nunes, M. R. T.*[2,46], Sabino, E. C.[4]¶*, Alcantara, L. C. J.[5]¶*,
18 Loman, N.[3]¶*, Pybus, O. G.[1,47]*¶

19

20

21 **Affiliations:**

22 1. Department of Zoology, University of Oxford, Oxford OX3 1PS, UK

23 2. Evandro Chagas Institute, Ministry of Health, Ananindeua, Brazil

24 3. Institute of Microbiology and Infection, University of Birmingham, UK

25 4. Department of Infectious Disease, School of Medicine & Institute of Tropical Medicine,
26 University of São Paulo, Brazil

27 5. Fundação Oswaldo Cruz (FIOCRUZ), Salvador, Bahia, Brazil

28 6. University of Rome Tor Vergata, Rome, Italy

29 7. Harvard Medical School, Boston, MA, USA

30 8. Boston Children's Hospital, Boston, MA, USA

31 9. Vaccine and Infectious Disease Division, Fred Hutchinson Cancer Research Center,
32 Seattle, WA, USA

33 10. Department of Epidemiology, University of Washington, Seattle, WA, USA

34 11. Department of Statistics, University of Oxford, Oxford OX3 1PS, UK

35 12. Mathematical Modelling of Infectious Diseases and Center of Bioinformatics,
36 Biostatistics and Integrative Biology, Institut Pasteur, Paris, France

13. Centre National de la Recherche Scientifique, URA3012, Paris, France

14. Coordenação dos Laboratórios de Saúde (CGLAB/DEVIT/SVS), Ministry of Health, Brasília, Brazil

15. Coordenação Geral de Vigilância e Resposta às Emergências em Saúde Pública (CGVR/DEVIT), Ministry of Health, Brasília, Brazil

16. Center of Data and Knowledge Integration for Health (CIDACS), Fundação Oswaldo Cruz (FIOCRUZ), Brazil

17. Departamento de Vigilância das Doenças Transmissíveis, Ministry of Health, Brasilia, Brazil

18. Coordenação Geral dos Programas de Controle e Prevenção da Malária e das Doenças Transmitidas pelo *Aedes*, Ministry of Health, Brasília, Brazil

19. Pan American Health Organization (PAHO), Buenos Aires, Argentina

20. Fundação Oswaldo Cruz (FIOCRUZ), Rio de Janeiro, Brazil

21. Ontario Institute for Cancer Research, Toronto, Canada

22. University of Nottingham, Nottingham, UK

23. Department of Immunology and Microbial Science, The Scripps Research Institute, La Jolla, CA 92037, USA

24. Departments of Laboratory Medicine and Medicine & Infectious Diseases, University of California, San Francisco, USA

25. División de Laboratorios de Vigilancia e Investigación Epidemiológica, Instituto Mexicano del Seguro Social, Ciudad de México, Mexico

26. Instituto de Biotecnología, Universidad Nacional Autónoma de México, Cuernavaca, Mexico

27. Instituto Oswaldo Cruz (FIOCRUZ), Rio de Janeiro, Brazil

28. Paul-Ehrlich-Institut, Langen, Germany

29. Laboratório Central de Saúde Pública Noel Nutels, Rio de Janeiro, Brazil

30. Laboratório Central de Saúde Pública do Estado do Rio Grande do Norte, Natal, Brazil

31. Universidade Potiguar do Rio Grande do Norte, Natal, Brazil

32. Faculdade Natalense de Ensino e Cultura, Rio Grande do Norte, Natal, Brazil

33. Laboratório Central de Saúde Pública do Estado da Paraíba, João Pessoa, Brazil

34. Fundação Oswaldo Cruz (FIOCRUZ), Recife, Pernambuco, Brazil

68 35. Center for Vaccine Research, Graduate School of Public Health, University of Pittsburgh, Pittsburgh, PA, USA

70 36. Section Clinical Tropical Medicine, Department for Infectious Diseases, Heidelberg University Hospital, Heidelberg, Germany

72 37. Laboratório Central de Saúde Pública do Estado de Alagoas, Maceió, Brazil

73 38. Universidade Estadual de Feira de Santana, Feira de Santana, Bahia, Brazil

74 39. Secretaria de Saúde de Feira de Santana, Feira de Santana, Bahia, Brazil

75 40. Universidade Federal do Amazonas, Manaus, Brazil

76 41. Hospital São Francisco, Ribeirão Preto, Brazil

77 42. Universidade Federal do Tocantins, Palmas, Brazil

78 43. University of Sydney, Sydney, Australia

79 44. Institute of Evolutionary Biology, University of Edinburgh, Edinburgh EH9 3FL, UK

80 45. Fogarty International Center, National Institutes of Health, Bethesda, MD 20892, USA

81 46. Department of Pathology, University of Texas Medical Branch, Galveston, TX 77555, USA

83 47. Metabiota, San Francisco, CA 94104, USA

85 *** Joint first or senior author**

90 **One Sentence Summary**: Virus genomes reveal the establishment of Zika virus in 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]) and the most cases associated with microcephaly and other birth defects (2,366 confirmed by 31 Dec 2016[2]). Following the initial detection of ZIKV in Brazil, >45 countries in the Americas have reported local ZIKV transmission, with 24 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 this information for interpreting observed trends in reported microcephaly. To address this we generated 54 complete or partial ZIKV genomes, mostly from Brazil, and report data generated by a mobile genomics lab that travelled across Northeast (NE) Brazil in 2016. One sequence represents the earliest confirmed ZIKV infection in Brazil. Analyses of viral genomes with ecological and epidemiological data estimate that ZIKV was present in NE Brazil by February 2014 and likely disseminated from there, nationally and internationally, before the first detection of ZIKV in the Americas. Estimated dates of the international spread of ZIKV from Brazil indicate the duration of pre-detection cryptic transmission in recipient 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 an 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 to be 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 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; http://www.zibraproject.org). 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 is so severely affected[8]. Further, NE Brazil is the most populous region of Brazil that also has potential for year-round ZIKV transmission[9]. With support from the Brazilian Ministry of Health and other institutions (**Acknowledgements**), the ZiBRA lab screened 1330 samples (almost exclusively serum or blood) from patients residing in 82 municipalities across five federal states (**Fig. 1**; **Extended Data Table 1a**). Samples provided by the 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 after ~6 days and can last 1-2 weeks[10]. In line with previous observations in Colombia[11], we found that RT-qPCR+ samples in NE Brazil were, on average, collected only two days after onset of symptoms. The median RT-qPCR cycle threshold (Ct) value of positive samples was

4

139   corresponding high, at 36 (**Extended Data Fig. 1a,b**). For NE Brazil, the time series of RT-
140   qPCR+ cases was positively correlated with the number of weekly-notified cases (Pearson's
141   ρ=0.62; **Fig. 1b**).

142   The ability of the mosquito vector *Aedes aegypti* to transmit ZIKV is determined by
143   ecological factors that affect adult survival, viral replication, and infective periods[12]. To
144   investigate the receptivity of Brazilian regions to ZIKV transmission we used a measure of
145   vector climatic suitability, derived from monthly temperature, relative humidity, and
146   precipitation data[9]. Using linear regression we find that, for each Brazilian region, there is a
147   strong association between estimated climatic suitability and weekly notified cases (**Figs.**
148   **1b,1c**; adjusted R^2>0.84, P<0.001; **Extended Data Table 1b**). Similar to previous findings
149   from dengue virus outbreaks[13,14], notified ZIKV cases lag climatic suitability by ~4 to 6
150   weeks in all regions, except NE Brazil, where no time lag is evident. Despite these
151   associations, numbers of notified cases should be interpreted cautiously because (i) co-
152   circulating dengue and Chikungunya viruses exhibit symptoms similar to ZIKV, and (ii) the
153   Brazilian case reporting system has evolved through time (**Methods**). We estimated basic
154   reproductive numbers (R_0) for ZIKV in each Brazilian region from the weekly notified case
155   data and found that R_0 is high in NE Brazil (R_0~3 for both epidemic seasons; **Extended Data**
156   **Table 1c**). Although our R_0 values are approximate, in part due to spatial variation in
157   transmission across the large regions analysed here, they are consistent with estimates from
158   other approaches[15,16].

159   Encouraged by the utility of portable genomic technologies during the West African Ebola
160   virus epidemic[17] we used our open protocol[18] to sequence ZIKV genomes directly from
161   clinical material using MinION DNA sequencers. We were able to generate virus sequences
162   within 48 hours of the mobile lab's arrival at each LACEN. In pilot experiments using a
163   cultured ZIKV reference strain[19] we recovered 98% of the virus genome (**Extended Data Fig.**
164   **1c**). However, due to low viral copy numbers in clinical samples (**Extended Data Fig. 1a**)
165   many sequences exhibited incomplete genome coverage and required additional sequencing
166   efforts in static labs once fieldwork was completed. Whilst average genome coverage was
167   typically high for samples with lower Ct-values (85% for Ct<33; **Fig. 2a, Extended Data**
168   **Table 2**), samples with higher Ct values had variable coverage (mean=72% for Ct$\geq$33 (**Fig.**
169   **2a**). Unsequenced genome regions were non-randomly distributed (**Fig. 2b**) suggesting that
170   the efficiency of PCR amplification varied among primer pair combinations. We generated 36
171   near-complete or partial genomes from the NE, SE and N regions of Brazil, supplemented by
172   9 sequences from samples from Rio de Janeiro municipality. To further elucidate Zika virus
173   transmission in the Americas, we include 5 new ZIKV complete genomes from Colombia and
174   4 from Mexico. Further, we append to our dataset 115 publicly available sequences and 85
175   additional genomes from a companion paper[20]. The final dataset comprised 254 ZIKV
176   sequences, 241 of which were sampled in the Americas (**Methods**).

177   The American ZIKV epidemic comprises a single founder lineage[4,21,22] (hereafter termed Am-
178   ZIKV) derived from Asian genotype viruses (hereafter termed PreAm-ZIKV) from Southeast
179   Asia and the Pacific[4]. A sliding window analysis of pairwise genetic diversity along the ZIKV
180   genome shows that the diversity of PreAm-ZIKV strains is on average ~2.1-fold greater than
181   Am-ZIKV viruses (**Fig. 2d**), reflecting a longer period of ZIKV circulation in Asia and the
182   Pacific than in the Americas. Genetic diversity of Am-ZIKV strains will increase in future
183   and updated diagnostic assays are recommended to guarantee RT-qPCR sensitivity[23].

184 It has been suggested that recent ZIKV epidemics may be linked causally to a higher apparent
185 evolutionary rate for the Asian genotype than the African genotype[24,25]. However, such
186 comparisons are confounded by an inverse relationship between the timescale of observation
187 and estimated evolutionary rates[26]. Regression of sequence sampling dates against root-to-tip
188 genetic distances indicates that molecular clock models can be applied reliably to the Asian-
189 ZIKV lineage (**Fig. 2c; Extended Data Figs. 2,3**). We estimate the whole genome
190 evolutionary rate of Asian ZIKV to be 0.97×10^{-3} substitutions per site per year (s/s/y; 95%
191 Bayesian credible interval, BCI=0.87-1.01×10^{-3}), consistent with other estimates for this
192 lineage[4,25]. We found no significant differences in evolutionary rates among ZIKV genome
193 regions (**Extended Data Table 3a**). The estimated d_N/d_S ratio of the Am-ZIKV lineage is low
194 (0.11, 95% CI= 0.10-0.13), as observed for other vector-borne flaviviruses[27] but is higher than
195 that of PreAm-ZIKV viruses (0.061, 0.047-0.077), likely due to the raised probability of
196 observing slightly deleterious changes in short-term datasets, as observed during previous
197 epidemics[28].

198 We used two phylogeographic approaches with different assumptions[29,30] to reconstruct the
199 origins and spread of ZIKV in Brazil and the Americas. We dated the common ancestor of
200 ZIKV in the Americas (node B, **Fig. 3**) to Jan 2014 (95% BCIs=Oct 2013-Apr 2014;
201 **Extended Data Tables 3b,c**), in line with previous estimates[4,25]. We find evidence that NE
202 Brazil played a central role in the establishment and dissemination of Am-ZIKV. Whilst NE
203 Brazil is the most probable location of node B (location posterior support=0.83, **Fig. 3**),
204 current data cannot exclude the hypothesis that node B was in the Caribbean (**Fig. 3** dashed
205 branches) due the presence of two sequences from Haiti in one of its descendant lineages.
206 More importantly, most Am-ZIKV sequences descend from a radiation of lineages (node C
207 and its immediate descendants; **Fig. 3**) dated to late Feb 2014 (95% BCIs of node C=Dec
208 2013-Jun 2014). Node C is more strongly inferred to have existed in NE Brazil (location
209 posterior support=0.99, **Fig. 3**). All 20 replicate analyses performed on sub-sampled data sets
210 place node C in Brazil, 14 of which place node C in NE Brazil (**Extended Data Fig. 4**).
211 Consequently, we conclude that node C reflects the crucial turning point in the emergence of
212 ZIKV in the Americas. If further data show that node B did indeed exist in Haiti, then it is
213 likely that Haiti acted as an intermediate 'stepping stone' for Am-ZIKV's arrival and
214 establishment in Brazil, from where the virus subsequently spread to other regions. This
215 perspective is consistent with the lower population size of Haiti compared to Brazil. We infer
216 that node C was present in NE Brazil several months before three notable events, each of
217 which also occurred in NE Brazil: (i) the retrospective identification of a cluster of suspected
218 but unconfirmed ZIKV cases in Dec 2014[1], (ii) the oldest ZIKV genome sequence from
219 Brazil, reported here, sampled in Feb 2015, and (iii) confirmed cases of ZIKV transmission in
220 NE Brazil in Mar 2015[31,32].

221 Our results further indicate that viruses from NE Brazil were important in the continental
222 spread of ZIKV. Within Brazil, we find instances of virus lineage movement from NE to SE
223 Brazil; most of these events are dated to the second half of 2014 and led to onwards
224 transmission in Rio de Janeiro (RJ1 to RJ4; **Fig. 3**) and São Paulo states (SP1; **Fig. 3**). We
225 infer that ZIKV lineages disseminated from NE Brazil to elsewhere in Central America, the
226 Caribbean, and South America. Most Am-ZIKV strains sampled outside Brazil fall into four
227 well-supported phylogenetic groups (**Fig 3**); three (SA1/CB1, CA1 and SA2) are inferred to
228 have been exported from NE Brazil between Jul 2014 and Apr 2015, while the Caribbean
229 clade CB2 appears to originate from SE Brazil around Mar 2015 (**Figs. 3, 4**). Each viral
230 lineage export occurred during a period of climatic suitability for vector transmission in the

recipient location (**Fig. 4**). For the earliest exports to Central America (CA1) and South America (SA1), there is a 7-9 month gap between the estimated date of exportation and the date of ZIKV detection in the recipient location, suggesting a complete or partial season of undetected transmission. These periods of cryptic transmission are relevant to studies of spatio-temporal trends in reported microcephaly, because they help to define the appropriate timeframe for baseline (pre-ZIKV) microcephaly in each region.

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 with overlapping symptoms (e.g. dengue, Chikungunya, Mayaro, and Oropouche viruses). However combining virus genomic and epidemiological data can generate insights into vector-borne virus transmission. A system of continuous and structured virus sequencing in Brazil, integrated with surveillance data, could provide timely information to inform effective responses against Zika and other viruses, including the recently re-emerged yellow fever virus[33].

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-Brarré 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).

270   9   Bogoch, II *et al.* Potential for Zika virus introduction and transmission in
271   resource-limited countries in Africa and the Asia-Pacific region: a modelling
272   study. *The Lancet Infectious Diseases* **16**, 1237-1245, doi:10.1016/S1473-
273   3099(16)30270-5 (2016).

274   10   Lessler, J. T., Ott, C.T., Carcelen, A.C., Konikoff, J.M., Williamson, J., Bi, Q., et al. .
275   Times to key events in the course of Zika infection and their implications: a
276   systematic review and pooled analysis *[Submitted]. Bull World Health Organ*
277   **DOI: 10.2471/BLT.16.174540** (2016).

278   11   Pacheco, O. *et al.* Zika Virus Disease in Colombia - Preliminary Report. *The New*
279   *England JKournal of Medicine*, doi:10.1056/NEJMoa1604037 (2016).

280   12   Liu-Helmersson, J., Stenlund, H., Wilder-Smith, A. & Rocklov, J. Vectorial capacity
281   of Aedes aegypti: effects of temperature and implications for global dengue
282   epidemic potential. *PloS One* **9**, e89783, doi:10.1371/journal.pone.0089783
283   (2014).

284   13   Cuong, H. Q. *et al.* Quantifying the emergence of dengue in Hanoi, Vietnam: 1998-
285   2009. *PLoS Negl Trop Dis* **5**, e1322, doi:10.1371/journal.pntd.0001322 (2011).

286   14   Gharbi, M. *et al.* Time series analysis of dengue incidence in Guadeloupe, French
287   West Indies: forecasting models using climate variables as predictors. *BMC*
288   *Infectious Diseases* **11**, 166, doi:10.1186/1471-2334-11-166 (2011).

289   15   Caminade, C. *et al.* Global risk model for vector-borne transmission of Zika virus
290   reveals the role of El Nino 2015. *PNAS* **114**, 119-124,
291   doi:10.1073/pnas.1614303114 (2017).

292   16   Rocklov, J. *et al.* Assessing Seasonal Risks for the Introduction and Mosquito-
293   borne Spread of Zika Virus in Europe. *EBioMedicine* **9**, 250-256,
294   doi:10.1016/j.ebiom.2016.06.009 (2016).

295   17   Quick, J. *et al.* Real-time, portable genome sequencing for Ebola surveillance.
296   *Nature* **530**, 228-232, doi:10.1038/nature16996 (2016).

297   18   Quick J., *et al.* Multiplex PCR method for MinION and Illumina sequencing of Zika
298   and other virus genomes directly from clinical samples. *Nature Protocols* in press
299   (2017).

300   19   Trosemeier, J. H. *et al.* Genome Sequence of a Candidate World Health
301   Organization Reference Strain of Zika Virus for Nucleic Acid Testing. *Genome*
302   *Announcements* **4**, doi:10.1128/genomeA.00917-16 (2016).

303   20   Metsky, H. C. *et al.* Genome sequencing reveals Zika virus diversity and spread in
304   the Americas. *bioRxiv* https://doi.org/10.1101/109348 (2017).

305   21   Giovanetti, M. *et al.* Zika virus complete genome from Salvador, Bahia, Brazil.
306   *Infection, Genetics and Evolution* **41**, 142-145, doi:10.1016/j.meegid.2016.03.030
307   (2016).

308 22 Naccache, S. N. *et al.* Distinct Zika Virus Lineage in Salvador, Bahia, Brazil.
309   *Emerging Infectious Diseases* **22**, doi:10.3201/eid2210.160663 (2016).

310 23 Corman, V. M. *et al.* Assay optimization for molecular detection of Zika virus.
311   *Bulletin of the World Health Organization* **94**, 880-892,
312   doi:10.2471/BLT.16.175950 (2016).

313 24 Liu, H. *et al.* From discovery to outbreak: the genetic evolution of the emerging
314   Zika virus. *Emerg Microbes Infect* **5**, e111, doi:10.1038/emi.2016.109 (2016).

315 25 Pettersson, J. H. O., Eldholm, V., Seligmna, S. J., Lundkvist, A., Falconar, A. K.,
316   Gaunt, M. W., Musso, D., Nougairede, A., Charrel, R., Gould, E. A., Lamballerie, X.
317   How Did Zika Virus Emerge in the Pacific Islands and Latin America? *mBio* **7**,
318   201239-201216 (2016).

319 26 Holmes, E. C., Dudas, G., Rambaut, A. & Andersen, K. G. The evolution of Ebola
320   virus: Insights from the 2013-2016 epidemic. *Nature* **538**, 193-200,
321   doi:10.1038/nature19790 (2016).

322 27 Holmes, E. C. Patterns of intra- and interhost nonsynonymous variation reveal
323   strong purifying selection in dengue virus. *Journal of Virology* **77**, 11296-11298
324   (2003).

325 28 Park, D. J. *et al.* Ebola Virus Epidemiology, Transmission, and Evolution during
326   Seven Months in Sierra Leone. *Cell* **161**, 1516-1526,
327   doi:10.1016/j.cell.2015.06.007 (2015).

328 29 De Maio, N., Wu, C. H., O'Reilly, K. M. & Wilson, D. New Routes to
329   Phylogeography: A Bayesian Structured Coalescent Approximation. *PLoS*
330   *Genetics* **11**, e1005421, doi:10.1371/journal.pgen.1005421 (2015).

331 30 Lemey, P., Rambaut, A., Drummond, A. J. & Suchard, M. A. Bayesian
332   phylogeography finds its roots. *PLoS Computational Biology* **5**, e1000520,
333   doi:10.1371/journal.pcbi.1000520 (2009).

334 31 Campos, G. S., Bandeira, A. C. & Sardi, S. I. Zika Virus Outbreak, Bahia, Brazil.
335   *Emerging Infectious Diseases* **21**, 1885-1886, doi:10.3201/eid2110.150847
336   (2015).

337 32 Zanluca, C. *et al.* First report of autochthonous transmission of Zika virus in
338   Brazil. *Memorias do Instituto Oswaldo Cruz* **110**, 569-572, doi:10.1590/0074-
339   02760150192 (2015).

340 33 Paules, C. I., Fauci, A. S. Yellow Fever — Once Again on the Radar Screen in the
341   Americas. *The New England Journal of Medicine* (2017).

342

343 **Supplementary Information** is available in the online version of the paper.

344

Author Contributions: NRF, LCJA, MRTN, ECS, NL and OGP designed the study. NRF, JQ, NL, IM, JGJ, MG, SCH, AB, ACdC, LCF, SPS, TB, PSL, BLN, HAOM, MRTN, and LCJA undertook fieldwork and experiments. NRF, JT, C-HW, OGP, JR and LdP performed genetic analyses. NRF, MUG, OGP and SC performed epidemiological analyses. NRF, JQ, MUGK, NL and OGP wrote the manuscript. ECH, AR, TB, MRTN, ECS and LCJA edited the manuscript. Other authors were critical for coordination, collection, processing,

391   sequencing and bioinformatics of samples. All authors read and approved the contents of the
392   manuscript.

393

394   **Author Information:** Reprints and permissions information is available at
395   www.nature.com/reprints. Correspondence and requests for materials should be addressed to
396   L.C.J.A. (lalcan@bahia.fiocruz.br), E.C.S. (sabinoec@usp.br), N.J.L.
397   (n.j.loman@bham.ac.uk), and O.G.P. (oliver.pybus@zoo.ox.ac.uk).

403 **Fig. 1**. **Geographic and temporal distribution of ZIKV in Brazil. a**. Sampling location of
404 genome sequences from Brazil and the Americas. Federal states in Brazil are coloured
405 according to 5 geographic regions (lower inset). A red line surrounds the states surveyed by
406 the ZiBRA mobile lab in 2016. State codes are PA=Pará, MA=Maranhão, CE=Ceará,
407 TO=Tocantins, RN=Rio Grande do Norte, PB=Paraíba, PE=Pernambuco, AL=Alagoas,
408 BA=Bahia, RJ=Rio de Janeiro, SP=São Paulo. Underlined states represent those from which
409 sequences in this study were generated (upper inset). Publicly available sequences were also
410 collated from non-underlined states. **b.** Confirmed and notified ZIKV cases in NE Brazil.
411 Upper panel shows the temporal distribution of RT-qPCR+ cases detected during ZiBRA
412 fieldwork. Only samples with known collection dates are included (n=138 out of 181
413 confirmed cases). Lower panel shows notified ZIKV cases in NE Brazil between 01 Jan 2015
414 and 19 Nov 2016 (n=122,779). The dashed line represents the average climatic vector
415 suitability score for NE Brazil (**Methods**). The vertical arrow indicates date of ZIKV
416 confirmation in NE Brazil/Americas[1]. **c.** Notified ZIKV cases in the Centre-West, Southeast,
417 North, and South regions of Brazil (clockwise from top left). The dashed lines represent the
418 average climatic vector suitability score for each region.

419

420 **Fig. 2**. **Zika virus genetic diversity and sequencing statistics. a**. The percentage of ZIKV
421 genome sequenced plotted against RT-qPCR Ct-value, for each sample. Each circle represents
422 a sequence recovered from an infected individual in Brazil and is coloured by sampling
423 location. **b**. Illustration of sequencing coverage across the ZIKV genome for the ZiBRA
424 sequences, including data generated by both mobile and static laboratories. **c.** Regression of
425 sequence sampling dates against root-to-tip genetic distances in a maximum likelihood
426 phylogeny of the Asian-ZIKV lineage. **Extended Data Fig. 2b** contains a comparable
427 analysis that also includes P6-740 (the oldest Asian-ZIKV strain collected in 1966). **d**.
428 Average pairwise genetic diversity of the PreAm-ZIKV strains (grey line) and of the Am-
429 ZIKV lineage (black line), calculated using a sliding window of 300 nucleotides with a step
430 size of 50 nucleotides.

431

432 **Fig. 3**. **Phylogeography of ZIKV in the Americas.** Maximum clade credibility phylogeny,
433 estimated from complete and partial Am-ZIKV genomes using a molecular clock
434 phylogeographic approach (**Methods**). Terminal branches with yellow circles indicate
435 sequences reported in this study. Terminal branches with no circles and reduced opacity are
436 those reported in a companion paper[20]. Thin vertical grey boxes indicate statistical uncertainty
437 of estimated dates of nodes A, B and C (**Extended Data Table 3c**). Branch colours indicate
438 the most probable ancestral lineage locations. Diamonds at internal nodes are sized in
439 proportion to clade posterior probabilities. For selected nodes, coloured numbers show the
440 posterior probabilities of ancestral locations and numbers in grey are clade posterior
441 probabilities. Asterisks indicate the three available genomes from microcephaly cases. A
442 black arrow indicates the oldest Brazilian ZIKV sequence. The grey arrow and dotted line
443 denotes when ZIKV was first confirmed in the Americas[1]. Nodes A and B are equivalent to
444 the nodes named identically in[4]. Text labels along the bottom of the figure denote clades of
445 sequences from regions outside of NE Brazil. RJ1 to RJ4 are clades from Rio de Janeiro state,
446 TO from Tocantins, and SP1 from São Paulo state. Clades from outside Brazil are denoted
447 CB1 and CB2 (Caribbean), SA1 and SA2 (South America excluding Brazil), and CA1

448   (Central America). Thin grey horizontal lines along the bottom of the figure denote sequences

449   from Brazil.

450

451   **Fig. 4. Establishment of Am-ZIKV in the Americas.** The earliest inferred dates of lineage

452   export to non-Brazilian regions, represented by box-and-whisker plots. Each plot corresponds

453   to the earliest movement between a pair of locations with well-supported virus lineage

454   migration. The first exports to South America outside Brazil (SA1 in **Fig. 3**), to Central

455   America (CA1) and to the Caribbean (CB1) are shown in panels **a-c,** respectively. Box and

456   whisker plots were generated in ggplot2, with boxes representing the median and interquartile

457   ranges of the estimated date of earliest movement. In each of **a-c**, dashed lines show the

458   estimated climatic vector suitability score for each recipient region, averaged across the

459   countries for which sequence data is available (see **Methods**). In each of **a-c**, the bar plots

460   show available notified ZIKV case data (plots adapted from PAHO) for the countries with the

461   earliest confirmed cases (Colombia[61] in panel **a**, Mexico[62] in **b**, and Puerto Rico[63] in **c**).

462   Coloured arrows indicate the earliest confirmation of ZIKV autochthonous cases in each non-

463   Brazilian region. The vertical dashed line represents the date of ZIKV confirmation in the

464   Americas.

465

13

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 been previously 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 supported by the Brazilian Ministry of Health (MoH) as part of the emergency public health response to Zika. Samples had been previously obtained for routine diagnostic purposes from persons visiting local clinics by the Brazilian National Health Surveillance network as part of Zika virus surveillance activities. In these cases, we used samples without informed consent with the approval of the Brazilian Ministry of Health. Specifically, residual anonymised clinical diagnostic samples, with no or minimal risk to patients, were provided for research and surveillance purposes within the terms of Resolution 510/2016 of CONEP (Comissão Nacional de Ética em Pesquisa, Ministério da Saúde; National Ethical Committee for Research, Ministry of Health). For samples obtained from patients engaged in longitudinal studies of Zika virus in São Paulo and Tocantins states, informed consent was obtained (IRB CAAE 53153916.7.0000.0065). Samples from patients followed in Salvador and Feira de Santana were analysed under institutional approval from CPqGM/FioCruz/BA (1.184.454). Urine and plasma samples from Rio de Janeiro were obtained from patients at the Fiocruz Viral Hepatitis Ambulatory (Oswaldo Cruz Institute, Rio de Janeiro, Brazil) with Institutional Review Board approval (IRB142/01) from the Oswaldo Cruz Institute. 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 (**Extended Data Table 1a**). Viral RNA was isolated from 200 µl 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[34]. 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

14

511 included in each batch; however, due to the large number of samples tested in a short time it
512 was possible only to run each sample without replication.

513 **Whole genome sequencing**

514 Sequencing was attempted on all positive samples obtained from NE Brazil regardless of Ct
515 value. All samples collected in Brazil that are reported in this study were sequenced with the
516 Oxford Nanopore MinION. Sequencing statistics can be found in **Extended Data Table 2**.
517 The protocol employed cDNA synthesis with random primers followed by gene specific
518 multiplex PCR and is presented in detail in Quick et al. [18]. In brief, extracted RNA was
519 converted to cDNA using the Protoscript II First Strand cDNA synthesis Kit (New England
520 Biolabs, Hitchin, UK) and random hexamer priming. ZIKV genome amplification by
521 multiplex PCR was attempted using the ZikaAsianV1 primer scheme and 40 cycles of PCR
522 using Q5 High-Fidelity DNA polymerase (NEB) as described in Quick et al.[18]. PCR products
523 were cleaned-up using AmpureXP purification beads (Beckman Coulter, High Wycombe,
524 UK) and quantified using fluorimetry with the Qubit dsDNA High Sensitivity assay on the
525 Qubit 3.0 instrument (Life Technologies). PCR products for samples yielding sufficient
526 material were barcoded and pooled in an equimolar fashion using the Native Barcoding Kit
527 (Oxford Nanopore Technologies, Oxford, UK). Sequencing libraries were generated from the
528 barcoded products using the Genomic DNA Sequencing Kit SQK-MAP007/SQK-LSK208
529 (Oxford Nanopore Technologies). Sequencing libraries were loaded onto a R9/R9.4 flowcell
530 and data was collected for up to 48 hours but generally less. As described[18], consensus
531 genome sequences were produced by alignment of two-direction reads to a Zika virus
532 reference genome (strain H/PF/2013, GenBank Accession number: KJ776791) followed by
533 nanopore signal-level detection of single nucleotide variants. Only positions with ≥20x
534 genome coverage were used to produce consensus alleles. Regions with lower coverage, and
535 those in primer-binding regions were masked with N characters. Validation of our sequencing
536 approach on the MinION platform was undertaken by using the MinION platform to sequence
537 a WHO reference strain of Zika virus that was also sequenced using the Illumina Miseq
538 platform[19]; identical consensus sequences were recovered regardless of the MinION
539 chemistry version employed (R7.3, R9 and R9.4) (**Extended Data Fig. 1c**).

540

541 **Collation of genome-wide data sets**

542 Our complete and partial genome sequences were appended to a global data set of all
543 available published ZIKV genome sequences (up until January 2017) using an in-house script
544 that retrieves updated GenBank sequences on a daily basis. In addition to the genomes
545 generated from samples collected in NE Brazil during ZiBRA fieldwork, samples were sent
546 directly to University of São Paulo and elsewhere for sequencing. Thirteen genomes from
547 Ribeirão Preto, São Paulo state (SP; SE-Brazil region) and seven genomes from Tocantins
548 (TO; N-Brazil region) were sequenced at University of São Paulo. Nine genomes from Rio de
549 Janeiro (RJ; SE-Brazil region) were sequenced in Birmingham, UK, and added to our dataset.
550 All these genomes were generated using the same primer scheme as the ZiBRA samples
551 collected in NE Brazil[18]. In addition to these 45 sequences from Brazil, we further included in
552 analysis 9 genomes from ZIKV strains sampled outside of Brazil in order to contextualise the
553 genetic diversity of Brazilian ZIKV, giving rise to a final data set of 54 sequences.
554 Specifically, we included 5 genomes from samples collected in Colombia and 4 new genomes

555   from Mexico, which were generated using the protocols described in refs. [35] and [22],
556   respectively.

557   GenBank sequences belonging to the African genotype of ZIKV were identified using the
558   Arboviral genotyping tool (http://bioafrica2.mrc.ac.za/rega-genotype/typingtool/aedesviruses)
559   and excluded from subsequent analyses, as our focus of study was the Asian genotype of
560   ZIKV, and the Am-ZIKV lineage in particular. To assess the robustness of molecular clock
561   dating estimates to the inclusion of older sequences, analyses were performed both with and
562   without the P6-740 strain, the oldest known strain of the ZIKV-Asian genotype (sampled in
563   1966 in Malaysia). Our final alignment comprised the sequences reported in this study (*n*=54)
564   plus publicly available ZIKV-Asian genotype sequences, as of 1st March 2017 (*n*=115). We
565   also included in our analysis 85 additional genomes from a companion paper[20]. The dataset
566   used for analysis therefore included sequences from 254 Zika virus isolates, 241 of which
567   were from the Americas. Unpublished but publicly available genomes were included in our
568   analysis only if we had written permission from those who generated the data (see
569   **Acknowledgments**).

570

571   **Maximum likelihood analysis and recombination screening**

572   Preliminary maximum likelihood (ML) trees were estimated with ExaMLv3[36] using a per-site
573   rate category model and a gamma distribution of among site rate variation. For the final
574   analyses, ML trees were estimated using PhyML[37] under a GTR nucleotide substitution
575   model[38], with a gamma distribution of among site rate variation, as selected by
576   jModeltest.v.2[39]. Branch support was inferred using 100 bootstrap replicates[37]. Final ML
577   trees were estimated with NNI and SPR heuristic tree search algorithms; equilibrium
578   nucleotide frequencies and substitution model parameters were estimated using ML[37] (see
579   **Extended Data Fig. 3**).

580   Recombination may impact evolutionary estimates[40] and has been shown to be present in the
581   ZIKV-African genotype[41]. In addition to restricting our analysis to the Asian genotype of
582   ZIKV, we employed the 12 recombination detection methods available in RDPv4[42] and the
583   Phi-test approach[43] available in SplitsTree[44] to further search for evidence of recombination in
584   the ZIKV-Asian lineage. No evidence of recombination was found.

585   Analysis of the temporal molecular evolutionary signal in our ZIKV alignments was
586   conducted using TempEst[45]. In brief, collection dates in the format yyyy-mm-dd (ISO 8601
587   standard) were regressed against root-to-tip genetic distances obtained from the ML
588   phylogeny. When precise sampling dates were not available, a precision of 1 month or 1 year
589   in the collection dates was taken into account.

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

595

596   **Molecular clock phylogenetics and gene-specific d_N/d_S estimation**

597   To estimate Bayesian molecular clock phylogenies, analyses were run in duplicate using
598   BEASTv.1.8.4[46] for 30 million MCMC steps, sampling parameters and trees every 3000
599   steps. We employed a model selection procedure using both path-sampling and stepping stone
600   models[47] to estimate the most appropriate combination of molecular clock and coalescent
601   models for Bayesian phylogenetic analysis. The best fitting combination was a Bayesian
602   skyline tree prior and a relaxed molecular clock model, with log-normally distributed
603   variation in rates among branches (**Extended Data Table 3b**). A non-informative continuous
604   time Markov chain reference prior[49] on the molecular clock rate was used. Convergence of
605   MCMC chains was checked with Tracer v.1.6. After removal of burn-in, posterior tree
606   distributions were combined and subsampled to generate an empirical distribution of 1,500
607   molecular clock trees.

608   To estimate rates of evolution per gene we partitioned the alignment into 10 genes (3
609   structural genes *C*, *prM*, *E*, and 7 non-structural genes *NS1*, *NS2A*, *NS2B*, *NS3*, *NS4A*, *NS4B*
610   and *NS5*) and employed a SDR06 substitution model[48] and a strict molecular clock model,
611   using an empirical distribution of molecular clock phylogenies. To estimate the ratio of
612   nonsynonymous to synonymous substitutions per site (d_N/d_S) for the PreAm-ZIKV and the
613   Am-ZIKV lineages, we used the single likelihood ancestor counting (SLAC) method[50]
614   implemented in HyPhy[51]. This method was applied to two distinct codon-based alignments
615   and their corresponding ML trees which comprised the PreAm-ZIKV and Am-ZIKV
616   sequences, respectively.

617

618   **Phylogeographic analysis**

619   We investigated virus lineage movements using our empirical distribution of phylogenetic
620   trees and the sampling location of each ZIKV sequence. The sampling location of sequences
621   collected from returning travellers was set to the travel destination in the Americas where
622   infection likely occurred. We discretised sequence sampling locations in Brazil into the
623   geographic regions defined in the main text. The number of sequences per region available for
624   analysis was 10 for N Brazil, 41 for NE Brazil and 54 for SE Brazil. No viral genetic data was
625   available for the Centre-West (CW) and the South (S) Brazilian regions. We similarly
626   discretised the locations of ZIKV sequences sampled outside of Brazil. These were grouped
627   according to the United Nations M49 coding classification of macro-geographical regions.
628   Our analysis included 53 sequences from the Caribbean, 38 from Central America, 17 from
629   Polynesia, 37 from South America (excluding Brazil), 3 from Southeast Asia and 1 from
630   Micronesia. To account for the possibility of sampling bias arising from a larger number of
631   sequences from particular locations, we repeated all phylogeographic analyses using (i) the
632   full dataset (*n*=254) and (ii) ten jackknife resampled datasets (*n*=74) in which taxa from each
633   location (except for Southeast Asia and Micronesia) were randomly sub-sampled to 10
634   sequences (the number of sequences available for N-Brazil).

635   Phylogeographic reconstructions were conducted using two approaches; (i) using the
636   asymmetric[52] discrete trait evolution models implemented in BEASTv1.8.4[46] and (ii) using
637   the Bayesian structured coalescent approximation (BASTA)[29] implemented in BEAST2v.2.
638   The latter has been suggested to be less sensitive to sampling biases[53]. For both approaches,
639   maximum clade credibility trees were summarized from the MCMC samples using
640   TreeAnnotator after discarding 10% as burn-in. The posterior estimates of the location of

641 nodes A, B and C (depicted in **Fig. 3**) from these two analytical approaches (applied to both
642 the complete and jackknifed data sets) can be found in **Extended Data Fig. 4**.

643 For the discrete trait evolution approach, we counted the expected number of transitions
644 among each pair of locations (net migration) using the robust counting approach[54,55] available
645 in BEASTv1.8.4[46]. We then used those inferred transitions to identify the earliest estimated
646 ZIKV introductions into new regions. These viral lineage movement events were statistically
647 supported (with Bayes factors > 3) using the BSSVS (Bayesian stochastic search variable
648 selection) approach implemented in BEASTv.1.8.4[30]. Box plots for node ages were generated
649 using the ggplot2[56] package in R software[57].

650

651 **Epidemiological analysis**

652 Weekly suspected ZIKV data per Brazilian region were obtained from the Brazilian Ministry
653 of Health (MoH). Cases were defined as suspected ZIKV infection when patients presented
654 maculopapular rash and at least two of the following symptoms: fever, conjunctivitis,
655 polyarthralgia or periarticular edema. Because notified suspected ZIKV cases are based on
656 symptoms and not molecular diagnosis, it is possible that some notified cases represent other
657 co-circulating viruses with related symptoms, such as dengue and Chikungunya viruses.
658 Further, case reporting may have varied among regions and through time. Data from 2015
659 came from the pre-existing MoH sentinel surveillance system that comprised 150 reporting
660 units throughout Brazil, which was eventually standardised in Feb 2016 in response to the
661 ZIKV epidemic. We suggest that these limitations should be borne in mind when interpreting
662 the ZIKV notified case data and we consider the R_0 values estimated here to be approximate.
663 That said, our time series of RT-qPCR+ ZIKV diagnoses from NE Brazil qualitatively match
664 the time series of notified ZIKV cases from the same region (**Fig. 1b**). To estimate the
665 exponential growth rate of the ZIKV outbreak in Brazil, we fit a simple exponential growth
666 rate model to each stage of the weekly number of suspected ZIKV cases from each region
667 separately:

668

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

670

671 where I_w is the number of cases in week w. As described in main text, the Brazilian regions
672 considered here were NE Brazil, N-Brazil, S-Brazil, SE-Brazil, and CW-Brazil. The time
673 period over which exponential growth occurs was determined by plotting the log of I_w and
674 selecting the period of linearity (**Extended Data Fig. 5**). A linear model was then fitted to
675 this period to estimate the weekly exponential growth rate r_W:

676

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

678

679 Let $g(.)$ be the probability density distribution of the epidemic generation time (i.e. the
680 duration between the time of infection of a case and the mean time of infection of its
681 secondary infections). The following formula can be used to derive the reproduction number
682 R from the exponential growth rate r and density $g(\cdot)$[58].

18

683

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

685 In our baseline analysis, following Ferguson et al.[59] we assume that the ZIKV generation time
686 is Gamma-distributed with a mean of 20.0 days and a standard deviation (SD) of 7.4 days. In
687 a sensitivity analysis, we also explored scenarios with shorter mean generation times (10.0
688 and 15.0 days) but unchanged coefficient of variation SD/mean=7.4/20=0.37 (**Extended Data**
689 **Table 1c**).

690

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

692 To account for seasonal variation in the geographical distribution of the ZIKV vector *Aedes*
693 *aegypti* in Brazil we fitted high-resolution maps[60] to monthly covariate data. Covariate data
694 included time-varying variables, such as temperature-persistence suitability, relative
695 humidity, and precipitation, as well as static covariates such as urban versus rural land use.
696 Maps were produced at a 5km x 5km resolution for each calendar month and then aggregated
697 to the level of the five Brazilian regions used in this study (**Extended Data Fig. 6**). For
698 consistency, we rescaled monthly suitability values so that the sum of all monthly maps
699 equalled the annual mean map[9].

700 We then assessed the correlation between monthly *Aedes aegypti* climatic suitability and the
701 number of weekly ZIKV notified cases in each Brazilian region, to test how well vector
702 suitability explains the variation in the number of ZIKV notified cases. To account for the
703 correlation in each Brazilian region we fit a linear regression model with a lag and two
704 breakpoints. As there may be a lag between trends in suitability and trends in notified cases,
705 we include a temporal term in the model to allow for a shift in the respective curves. Thus for
706 each region, different sets of the constant and linear terms are fitted to different time periods.
707 More formally,

708

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

710

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

714 We then find the values of T and l that provide the highest adjusted-R^2 by stepwise iterative
715 optimisation. For each value of T evaluated, the optimal value of l (i.e. that which gives the
716 highest adjusted-R^2 for the model above) is found by the optim function in R[57]. Climatic
717 suitability values were only calculated for each month, so to calculate suitability values for
718 any given point in time we interpolated between the monthly values using a linear function.
719 We found no significant effect of residual autocorrelation in our data (**Extended Data Fig.**
720 **7**).

721

722 **Data availability**

19

Details of the primers and probes used here have been available at
http://www.zibraproject.org since the beginning of the project. BEAST XML files, tree files,
and sequence datasets analysed in this study are archived at
https://github.com/zibraproject. New Brazilian sequences are available in GenBank under
accession numbers KY558989 to KY559032 and KY817930. New Colombian and Mexican
sequences are available under accession numbers KY317936-40 and KY606271-4,
respectively. See Extended Data Table 2 for further details.

34   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).

35   Grubaugh, N. D. *et al.* Multiple introductions of Zika virus into the United States
      revealed through genomic epidemiology. *bioRxiv*
      https://doi.org/10.1101/104794 (2017).

36   Kozlov, A. M., Aberer, A. J., Stamatakis, A. ExaML version 3: a tool for
      phylogenomic analyses on supercomputers. *Bioinformatics* **31**, 2577-2579
      (2015).

37   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).

38   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).

39   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).

40   Schierup, M. H. & Hein, J. Consequences of recombination on traditional
      phylogenetic analysis. *Genetics* **156**, 879-891 (2000).

41   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).

42   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).

43   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).

761 44 Huson, D. H. & Bryant, D. Application of phylogenetic networks in evolutionary
762 studies. *Molecular Biology and Evolution* **23**, 254-267,
763 doi:10.1093/molbev/msj030 (2006).

764 45 Rambaut, A., Lam, T. T., Fagundes de Carvalho, L., Pybus, O. G. Exploring the
765 temporal structure of heterochronous sequences using TempEst (formerly Path-
766 O-Gen). *Virus Evolution* **2** (2016).

767 46 Drummond, A. J., Suchard, M. A., Xie, D. & Rambaut, A. Bayesian phylogenetics
768 with BEAUti and the BEAST 1.7. *Molecular Biology and Evolution* **29**, 1969-1973,
769 doi:10.1093/molbev/mss075 (2012).

770 47 Baele, G., Li, W. L., Drummond, A. J., Suchard, M. A. & Lemey, P. Accurate model
771 selection of relaxed molecular clocks in bayesian phylogenetics. *Molecular*
772 *Biology and Evolution* **30**, 239-243, doi:10.1093/molbev/mss243 (2013).

773 48 Shapiro, B., Rambaut, A. & Drummond, A. J. Choosing appropriate substitution
774 models for the phylogenetic analysis of protein-coding sequences. *Molecular*
775 *Biology and Evolution* **23**, 7-9, doi:10.1093/molbev/msj021 (2006).

776 49 Ferreira, M. A. R. & Suchard, M. A. Bayesian analysis of elapsed times in
777 continuous-time Markov chains. *Can J Stat* **36**, 355-368 (2008).

778 50 Kosakovsky Pond, S. L., Frost, S. D. Not so different after all: a comparison of
779 methods for detecting amino acid sites under selection. *Molecular Biology and*
780 *Evolution* **22**, 1208-1222 (2005).

781 51 Pond, S. L., Frost, S. D. & Muse, S. V. HyPhy: hypothesis testing using phylogenies.
782 *Bioinformatics* **21**, 676-679, doi:10.1093/bioinformatics/bti079 (2005).

783 52 Edwards, C. J. *et al.* Ancient hybridization and an Irish origin for the modern
784 polar bear matriline. *Current Biology : CB* **21**, 1251-1258,
785 doi:10.1016/j.cub.2011.05.058 (2011).

786 53 Bouckaert, R. *et al.* BEAST 2: a software platform for Bayesian evolutionary
787 analysis. *PLoS Computational Biology* **10**, e1003537,
788 doi:10.1371/journal.pcbi.1003537 (2014).

789 54 Minin, V. N. & Suchard, M. A. Fast, accurate and simulation-free stochastic
790 mapping. *Philos Trans R Soc Lond B Biol Sci* **363**, 3985-3995,
791 doi:10.1098/rstb.2008.0176 (2008).

792 55 O'Brien, J. D., Minin, V. N. & Suchard, M. A. Learning to count: robust estimates
793 for labeled distances between molecular sequences. *Molecular Biology and*
794 *Evolution* **26**, 801-814, doi:10.1093/molbev/msp003 (2009).

795 56 Wickham, H. *ggplot2: elegant graphics for data analysis*. (Springer New York,
796 2009).

797 57 R: A Language and Environment for Computing (R Foundation for Statistical
798 Computing, Vienna, Austria, 2014).

799   58   Cori, A., Ferguson, N. M., Fraser, C. & Cauchemez, S. A new framework and
800   software to estimate time-varying reproduction numbers during epidemics.
801   *American Journal of Epidemiology* **178**, 1505-1512, doi:10.1093/aje/kwt133
802   (2013).

803   59   Ferguson, N. M. *et al.* EPIDEMIOLOGY. Countering the Zika epidemic in Latin
804   America. *Science* **353**, 353-354, doi:10.1126/science.aag0219 (2016).

805   60   Kraemer, M. U. *et al.* The global distribution of the arbovirus vectors Aedes
806   aegypti and Ae. albopictus. *eLife* **4**, e08347, doi:10.7554/eLife.08347 (2015).

807   61   PAHO/WHO. Zika Epidemiological Update - Colombia (21 Dec 2016).
808   (Washington, D. C., 2016).

809   62   PAHO/WHO. Zika Epidemiological Update - Mexico (20 Dec 2016). (Washington,
810   D. C., 2016).

811   63   PAHO/WHO. Zika Epidemiological Update - Puerto Rico (20 Dec 2016).
812   (Washington, D. C., 2016).

813

814

Extended Data Figure Legends

Extended Data Fig. 1. **a.** The distribution of CT-values for the RT-qPCR+ samples tested during the ZiBRA journey in Brazil (*n*=181 samples; median CT = 35.96). **b.** shows the distribution of the temporal lag between the date of onset of clinical symptoms and the date of sample collection of RT-qPCR+ samples (median lag = 2 days). Red dashed lines represent the median of the distributions. **(c)** Validation of sequencing approaches. A phylogeny of the ZIKV Asian genotype estimated using PhyML[37] is shown. The expanded clade highlighted in blue contains the WHO reference ZIKV sequence[19] (accession number KX369547), which was generated using Illumina MiSeq. Sequences generated using MinION chemistries R9.4 2D, R9.4 1D, R9 1D, R9 2D and R7.3 2D contain no nucleotide differences and hence were also placed in this clade. Scale bars represent expected nucleotide substitutions per site (s/s). Am-ZIKV=American Zika virus lineage.

Extended Data 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 using PhyML[37], was explored using TempEst[45]. **a.** Estimates for the dataset used in the phylogenetic analysis presented in **Fig. 3c**, and **b.** estimates for the same dataset with the addition of the P6-740 strain sampled in 1966 (accession number HQ234499).

Extended Data Fig. 3. A non-clock maximum likelihood phylogeny of our ZIKV data set. Bootstrap branch support values are shown at each node. The phylogeny was estimated using PhyML[37]. Sequences generated in this study are highlighted in red. Scale bar represents expected nucleotide substitutions per site.

Extended Data Fig. 4. Ancestral node location posterior probabilities (ANLPP), for nodes A, B and C, estimated using the complete dataset (top row) and ten replicate subsampled data sets (other rows). See **Methods** for details. ANLPPs were calculated using two approaches: DTA=discrete trait analysis method[30] (left side columns) and BASTA=Bayesian structured coalescent approximation method[29] (right side columns). For each method, we employed an asymmetric model of location exchange to estimate ancestral node locations and to infer patterns of virus spread among regions.

Extended Data Fig. 5. 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.

Extended Data Fig. 6. Seasonal suitability for ZIKV transmission in the Americas. These maps were estimated by collating data on *Aedes* mosquitoes, temperature, relative humidity

854 and precipitation, and are the basis of the trends in suitability for different regions shown in
855 main text **Figs. 1 and 4**. For method details, see [9,60].

856

857 **Extended Data Fig. 7.** Partial autocorrelation functions for the linear model associating
858 climatic suitability and ZIKV notified cases in each geographic region in Brazil. The residuals
859 for the North, Northeast, Centre-West and Southeast regions show no autocorrelation, while a
860 small amount of autocorrelation cannot be excluded for the South region.

Extended Data Table Legends

Extended Data Table 1. a. Summary of the clinical samples tested (n=1330, of which 181 were RT-qPCR+) by the ZiBRA mobile lab in June 2016, NE Brazil. 84% of samples with known collection dates (n=698 of 826) were from 2016. ZIKV notified cases were confirmed using RT-qPCR (see **Methods**). Collection lag represents the median time interval (in days) between the date of onset of clinical symptoms and date of sample collection (both dates available for n=219) for all samples (including those that subsequently tested RT-qPCR negative). Federal states are RN: Rio Grande do Norte, PB: Paraíba, PE: Pernambuco, AL: Alagoas, BA: Bahia. Sample numbers in the FioCruz, PE row include RT-PCR+ cases from Pernambuco generated at Fiocruz Pernambuco. **b.** Parameters of the model measuring the link between climatic vector suitability and notified ZIKV cases in different Brazilian regions (CW: Centre-West, N: North, NE: Northeast, SE: Southeast, S: South). 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). **c.** For each region, estimates of the basic reproductive number (R) of ZIKV are shown for several values of generation time (g) parameter, together with the corresponding estimates of exponential growth rate (r) (per day) obtained from notified ZIKV case counts (see Extended Data Fig. 7). 1st: epidemic wave in 2015; 2nd: epidemic wave in 2016.

Extended Data Table 2. Sequencing statistics. Accession numbers, sample IDs, sequencing coverage, RT-qPCR values and epidemiological information for the samples from Brazil generated in this study. For the sequences from RJ state, alignments were performed against version 2 (KJ776791.2) of the genome reference; all other sequences used version 1 (KJ776791.1).

Extended Data Table 3. a. Estimated per-gene rates of evolution (mean and 95% Bayesian credible intervals=BCIs) are shown in units of 10^{-3} substitutions per site per year. **b.** Log-marginal likelihood estimates using the path-sampling (PS) and Stepping-Stone (SS) model selection approaches[47]. The overall ranking of the models is shown in parentheses for each estimator and the best-fitting combination is underscored. Two molecular clock models were tested here. SC: Strict clock model, UCLN: uncorrelated relaxed clock with lognormal distribution[46]. **c.** Estimated 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.

a.

Am-ZIKV genomes

■ This study
□ Other studies

Jan Apr Jul Oct Jan Apr Jul Oct
2015 2016

Brazil regions:
■ North
■ Northeast
■ Southeast
■ South
■ Centre West
— ZiBRA area

Other:
■ Caribbean
■ Central America
■ South America
○ Sequence data

MA CE RN PB PE AL BA RJ SP
PA TO

b.

■ RT-qPCR+ cases (ZiBRA)   ■ Notified cases   ···· Aedes suitablity per region

No. ZIKV weekly cases in NE-Brazil

Aedes suitability

Jan15 Apr15 Jul15 Oct15 Jan16 Apr16 Jul16 Oct16

c.

No. ZIKV weekly notified cases (x1,0000)

Aedes suitability

Jan15 Jul15 Jan16 Jul15
Jan15 Jul15 Jan16 Jul15

Month

a.

b.

c.

d.