Genome sequencing reveals Zika virus diversity and spread in the Americas

Metsky, H.C.*[1,2], Matranga, C.B.*[1], Wohl, S.*[1,3], Schaffner, S.F.[1,3,4], Freije, C.A.[1,3], Winnicki, S.M.[1], West, K.[1], Qu, J.[1], Baniecki, M.L.[1], Gladden-Young, A.[1], Lin, A.E.[1,3], Tomkins-Tinch, C.H.[1], Park, D.J.[1], Luo, C.Y.[1,3], Barnes, K.G.[1,3], Chak, B.[1,3], Barbosa-Lima, G.[5], Delatorre, E.[6], Vieira, Y.R.[5], Paul, L.M.[7], Tan, A.L.[7], Porcelli, M.C.[8], Vasquez, C.[8], Cannons, A.C.[9], Cone, M.R.[9], Hogan, K.N.[9], Kopp, E.W. IV[9], Anzinger, J.J.[10], Garcia, K.F.[11], Parham, L.A.[11], Gélvez Ramírez, R.M.[12], Miranda Montoya, M.C.[12], Rojas, D.P.[13], Brown, C.M.[14], Hennigan, S.[14], Sabina, B.[14], Scotland, S.[14], Gangavarapu, K.[15], Grubaugh, N.D.[15], Oliveira, G.[16], Robles-Sikisaka, R.[15], Rambaut, A.[17,18], Gehrke, L.[19,20], Smole, S.[14], Halloran, M.E.[21,22], Villar Centeno, L.A.[12], Mattar, S.[23], Lorenzana, I.[11], Cerbino-Neto, J.[5], Degrave, W.[24], Bozza, P.T.[25], Gnirke, A.[1], Andersen, K.G.†[15,16,26], Isern, S.†[7], Michael, S.†[7], Bozza, F.†[5,27], Souza, T.M.L.¶†[28,29], Bosch, I.†[19], Yozwiak, N.L.†[1,3], MacInnis, B.L.¶†[1,4], Sabeti, P.C.†[1,3,4]

Affiliations:

1. Broad Institute of MIT and Harvard, Cambridge MA, USA
2. Department of Electrical Engineering and Computer Science, Massachusetts Institute of Technology, Cambridge MA, USA
3. Center for Systems Biology, Department of Organismic and Evolutionary Biology, Harvard University, Cambridge MA, USA
4. Harvard T.H. Chan School of Public Health, Harvard University, Boston MA, USA
5. National Institute of Infectious Diseases Evandro Chagas, Fundação Oswaldo Cruz (FIOCRUZ), Rio de Janeiro RJ, Brazil
6. Laboratório de AIDS e Imunologia Molecular, Instituto Oswaldo Cruz, FIOCRUZ, Rio de Janeiro RJ, Brazil
7. Department of Biological Sciences, College of Arts and Sciences, Florida Gulf Coast University, Fort Myers FL, USA
8. Miami-Dade County Mosquito Control, Miami FL, USA
9. Bureau of Public Health Laboratories, Division of Disease Control and Health Protection, Florida Department of Health, Tampa FL, USA
10. Department of Microbiology, The University of the West Indies, Mona, Kingston, Jamaica
11. Instituto de Investigacion en Microbiologia (IIM) - Universidad Nacional Autónoma de Honduras, Honduras
12. Grupo de Epidemiología Clínica, Universidad Industrial de Santander, Bucaramanga, Colombia
13. Department of Epidemiology, College of Public Health and Health Professions, University of Florida, Gainesville FL, USA
14. Massachusetts Department of Public Health, Jamaica Plain MA, USA
15. Department of Immunology and Microbial Science, The Scripps Research Institute, La Jolla CA, USA
16. Scripps Translational Science Institute, La Jolla CA, USA
17. University of Edinburgh, Edinburgh EH9 3FL, UK
18. Fogarty International Center, National Institutes of Health, Bethesda MD, USA
19. Institute for Medical Engineering and Science, Massachusetts Institute of Technology, Cambridge MA, USA
20. Department of Microbiology and Immunobiology, Harvard Medical School, Boston MA, USA
21. Vaccine and Infectious Disease Division, Fred Hutchinson Cancer Research Center, Seattle WA, USA
22. Department of Biostatistics, University of Washington, Seattle WA, USA
23. Institute for Tropical Biology Research, Universidad de Córdoba, Colombia
24. Fiocruz, Instituto Oswaldo Cruz, Laboratório de Genômica Funcional e Bioinformática, Rio de Janeiro RJ, Brazil
25. Laboratório de Imunofarmacologia, Instituto Oswaldo Cruz (IOC), Fundação Oswaldo Cruz (FIOCRUZ), Rio de Janeiro RJ, Brazil
26. Department of Integrative Structural and Computational Biology, The Scripps Research Institute, La Jolla CA, USA
27. D'Or Institute for Research and Education (IDOR), Brazil
28. National Institute for Science and Technology on Innovation on Neglected Diseases (INCT/IDN), Fiocruz, Rio de Janeiro RJ, Brazil
29. Center for Technological Development in Health (CDTS), Fiocruz, Rio de Janeiro RJ, Brazil

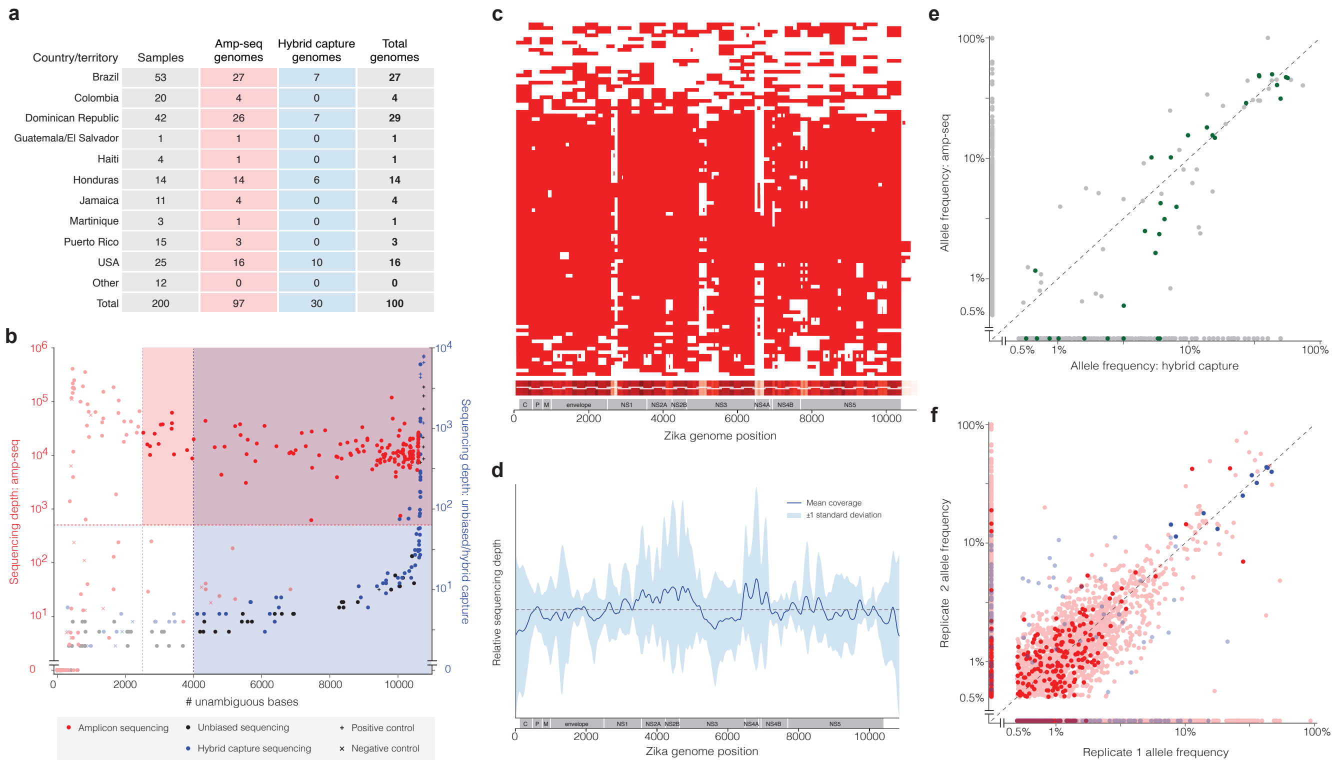
* co-first author
† co-senior author
¶ co-corresponding author: B.L.M. (bronwyn@broadinstitute.org); T.M.L.S. (tmoreno@cdts.fiocruz.br)

Despite great attention given to the recent Zika virus (ZIKV) epidemic in the Americas, much remains unknown about its epidemiology and evolution, in part due to a lack of genomic data. We applied multiple sequencing approaches to generate 100 ZIKV genomes from clinical and mosquito samples from 10 countries and territories, greatly expanding the observed viral genetic diversity from this outbreak. We analyzed the timing and patterns of introductions into distinct geographic regions, confirming phylogenetic evidence for the origin and rapid expansion of the outbreak in Brazil[1], and for multiple introductions from Brazil into Honduras, Colombia, Puerto Rico, other Caribbean islands, and the continental US. We find that ZIKV circulated undetected in many regions of the Americas for up to a year before the first locally transmitted cases were confirmed, highlighting the challenge of effective surveillance for this virus. We further characterize genetic variation across the outbreak to identify mutations with possible functional implications for ZIKV biology and pathogenesis.

Since its introduction into Brazil in 2013[1], mosquito-borne ZIKV (Family: *Flaviviridae*) has spread rapidly throughout the Americas, causing hundreds of thousands of cases of ZIKV disease, as well as ZIKV congenital syndrome and likely other neurological complications[2–4]. Comparative phylogenomic analysis of ZIKV can reveal the trajectory of the outbreak and detect mutations that may be associated with new disease phenotypes or affect molecular diagnostics. Despite the nearly 60 years since its discovery and the scale of the recent outbreak, however, fewer than 100 ZIKV genomes have been sequenced directly from clinical samples. This is due in part to technical challenges posed by low viral loads (often orders of magnitude lower than in Ebola or dengue virus infection[5–7]), as well as issues of RNA preservation in samples collected without the unique requirements of sequencing in mind. Culturing the virus can greatly increase the material available for sequencing, but it can introduce artefacts and is time-consuming and difficult.

We sought to gain a deeper understanding of the viral populations underpinning the ZIKV epidemic by extensive genome sequencing of the virus directly from samples collected as part of ongoing surveillance. We initially pursued metagenomic RNA sequencing in order to capture both ZIKV and other potential co-infections in an unbiased way. In most of the 37 samples examined by this approach, however, the amount of ZIKV material was not sufficient for genome assembly. Unbiased RNA sequencing still proved valuable because it provided ZIKV data to verify results from other methods. We also observed 3 other viruses in 7 samples (Extended Data Table 1) in our metagenomic data; notably, these did not include chikungunya or dengue, viruses known to be co-circulating with ZIKV in many affected regions[8,9].

In order to capture sufficient ZIKV content for genome assembly, we turned to two targeted enrichment approaches: hybrid capture[10] and PCR amplicon-based sequencing[11] (amp-seq). We attempted to sequence ZIKV directly from 200 samples (192 clinical samples from confirmed and possible ZIKV disease cases and 8 mosquito pools) from across the epidemic in the Americas (Fig. 1a and Supplementary Table 1). Because these approaches also enrich any contaminant ZIKV content, it was critical that we take steps to avoid creating artefactual sequence. For this purpose, we relied heavily on negative control samples, and established robust, method-specific thresholds on coverage and completeness (Fig. 1b); these allowed us to identify high confidence ZIKV assemblies while discarding contamination. We assembled complete or partial genomes from 100 samples, which we used for further analysis. This dataset includes 97 genomes enriched by amp-seq (out of 197 attempted) and a partially overlapping set of 30 genomes enriched by hybrid capture (out of 37 attempted). Patient sample type (urine, serum or plasma) made no significant difference in sequencing success (Extended Data Fig. 1). Completeness and coverage for these genomes are shown in Fig. 1c and d; the median fraction of the genome



Country/territory	Samples	Amp-seq genomes	Hybrid capture genomes	Total genomes
Brazil	53	27	7	**27**
Colombia	20	4	0	**4**
Dominican Republic	42	26	7	**29**
Guatemala/El Salvador	1	1	0	**1**
Haiti	4	1	0	**1**
Honduras	14	14	6	**14**
Jamaica	11	4	0	**4**
Martinique	3	1	0	**1**
Puerto Rico	15	3	0	**3**
USA	25	16	10	**16**
Other	12	0	0	**0**
Total	200	97	30	**100**

Figure 1 | Sequence data generated directly from clinical and mosquito samples. 200 clinical and mosquito pool samples were sequenced using amp-seq and/or hybrid capture, generating 100 ZIKV genomes. **(a)** For each country or territory, the number of genomes generated by each sequencing method. Each genome counted is from a sample that has at least one "positive" assembly, i.e. a replicate passes thresholds in (b). The "Other" category includes all samples from countries that did not produce a positive assembly. In the final column, genomes are counted only once if both methods produced a positive assembly. **(b)** Thresholds used to select samples for downstream analysis. Each point is a replicate. Red and blue shading: regions of accepted amp-seq and hybrid capture genome assemblies, respectively; purple: positive assemblies by either method. Not shown: hybrid capture positive controls with depth >10,000x. **(c)** Amp-seq coverage by sample across the ZIKV genome. Red indicates sequencing depth ≥500x, and the heat map (bottom) sums coverage across all samples; white horizontal lines indicate amplicon locations. **(d)** Relative sequencing depth across hybrid capture genomes. **(e)** Within-sample variant frequencies across methods. Each point is a variant in an individual sample and points are plotted on a log-log scale. Green points represent "verified" variants detected by hybrid capture sequencing that pass strand bias and single-library frequency filters. **(f)** Within-sample variant frequencies across replicate libraries per method. Red points are variants identified using amp-seq; blue points are variants identified using hybrid capture. Light colored points do not pass a strand bias filter; dark points do. In (e-f), frequencies <0.5% are shown at 0%.

with unambiguous base calls was 94%. Per-base discordance between genomes produced by the two methods was 0.0082% across the genome, 0.080% at polymorphic positions, and 1.1% for minor allele base calls. Concordance of within-sample variants is shown in more detail in Fig. 1e and f.

To investigate the spread of ZIKV in the Americas we performed a phylogenetic molecular clock analysis on the 100 genomes from our dataset, together with 64 published and available genomes from NCBI GenBank and our companion papers[12,13] (Fig. 2a). The resulting phylogenetic reconstruction (Fig. 2b) is consistent with the outbreak having started in Brazil: Brazil ZIKV genomes appear on all deep branches of the tree, and their most recent common ancestor occurs at the root of the entire tree. Estimated time to the most recent common ancestor (tMRCA) dates it in late 2013 (95% credible interval, CI, in decimal years [2013.39, 2014.30]). Low posterior probabilities of ancestral nodes near the root provide little support for stratification early in the outbreak, particularly in Brazil; this suggests rapid early spread, consistent with the introduction of a new virus in an immunologically naive population. ZIKV genomes from Colombia (n=10), Honduras (n=14), and Puerto Rico

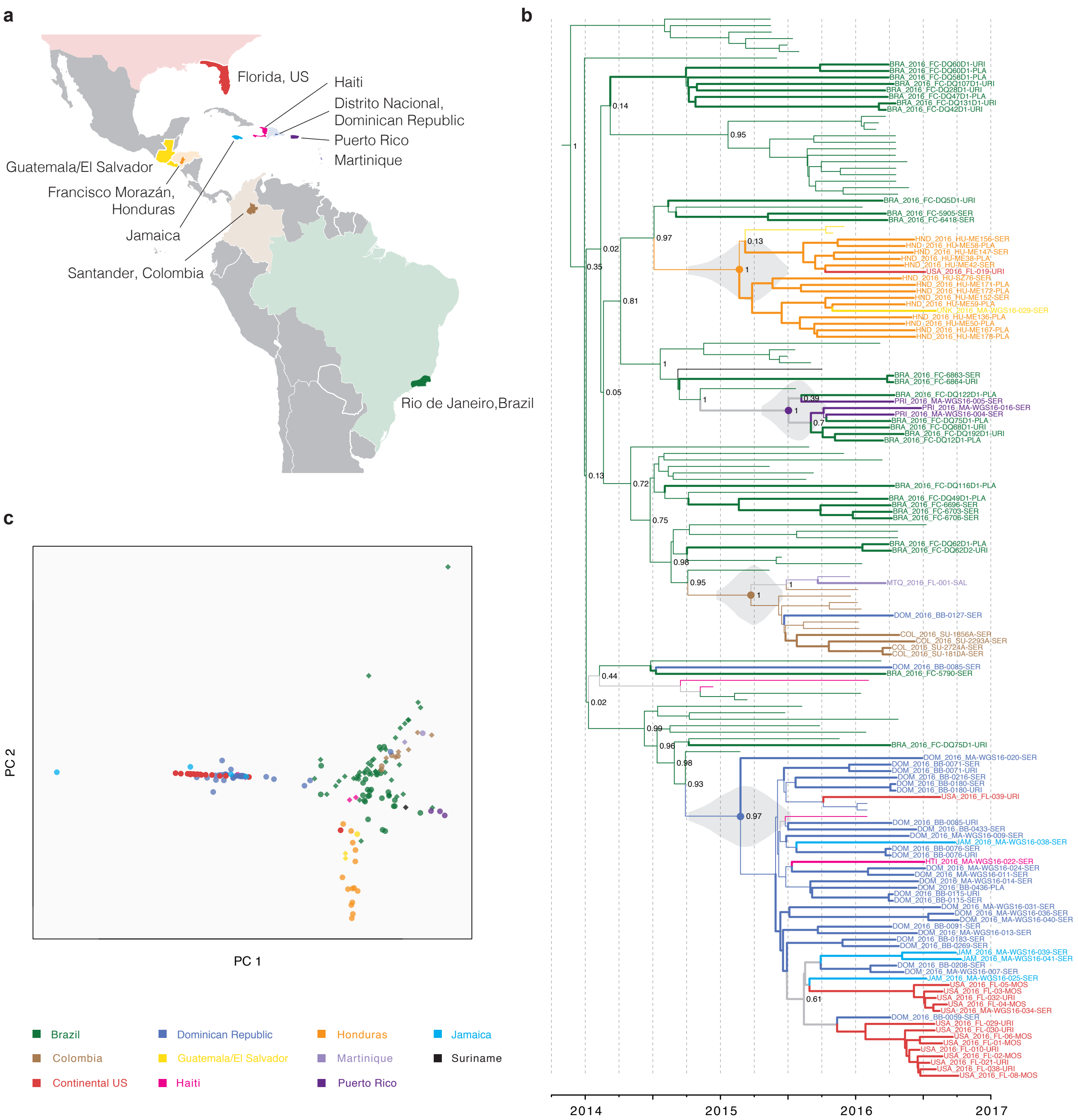


Figure 2 | Zika virus spread throughout the Americas. (a) Samples were collected in each of the colored countries or territories. Darker regions indicate the specific state, department, or province of sample origin for samples from this study, if known. **(b)** Maximum clade credibility tree generated using BEAST shows ZIKV introductions into various countries and territories in the Americas. Tips with bolded branches and labels represent genomes generated in this study. Grey violin plots denote probability distributions for the tMRCA of four major clades. **(c)** Principal component analysis of variants between samples shows geographic clustering. Circular points represent data generated in this study; diamond points represent other publicly available genomes from this outbreak that are used in this analysis.

(*n*=3) each cluster into distinct clades. We also observed a clade consisting entirely of samples from 3 Caribbean countries (the Dominican Republic, Jamaica, and Haiti) and the continental US. This clade contains most of the genomes from the Dominican Republic (29 of 31) and the continental US (15 of 16). We estimated the within-outbreak substitution rate to be 1.0×10^{-3} substitutions/site/year (95% CI [8.5×10^{-4}, 1.2×10^{-3}]), consistent with prior estimates[1]. Root-to-tip divergence based on genomes from Southeast Asia, the Pacific, and the Americas supports the use of a molecular clock analysis (Extended Data Fig. 2).

Principal component analysis (PCA) of the same genomes is consistent with the phylogenetic observations (Fig. 2c and Extended Data Fig. 3). We observed tight clustering among ZIKV from the continental US, the Dominican Republic, and Jamaica. ZIKV genomes from Brazil, Colombia, and Puerto Rico are similar and distinct from genomes sampled in other countries. ZIKV from Honduras forms a third cluster that contains genomes from Guatemala or El Salvador. The PCA results also show no clear stratification of genomes within Brazil.

Determining when ZIKV arrived in specific regions is important for understanding the epidemiology of the virus; more importantly, it tells researchers when and where to look for rising incidence of possible complications of ZIKV infection. We estimated the tMRCA for well-supported nodes within our phylogeny, including four highly supported clades (posterior probability >0.95), formed mostly by strains from Colombia, Honduras, Puerto Rico, the Caribbean, and the continental US. We found that these four clades originated in early to mid 2015, several months before the first ZIKV cases were confirmed in these regions, indicating ongoing local circulation of ZIKV before its detection by surveillance systems. We estimated tMRCA of ZIKV from Colombia to be in March 2015 (95% CI [2014.97, 2015.46]), 7 months before the first confirmed locally transmitted cases[14]; the tMRCA from Honduras to be in March 2015 (95% CI [2014.76, 2015.50]), 9 months before the first confirmed locally transmitted cases[15]; and the tMRCA from Puerto Rico to be in July 2015 (95% CI [2015.30, 2015.78]), 5 months before the first confirmed locally transmitted cases[16]. We estimated tMRCA of the Caribbean clade, consisting of genomes from three Caribbean countries and the continental US, to be in February 2015 (95% CI [2014.76, 2015.52]), 11 months before the first confirmed case in the Dominican Republic[17] and more than a year before the first confirmed case in Florida, USA[18]. We observed several introductions of ZIKV into the continental US and found that sequences from mosquito and human samples collected in Florida cluster together; this suggests local ZIKV transmission in Florida and is consistent with findings detailed in a companion paper[13]. Similar temporal gaps between the tMRCA of local transmission chains and the detection of early cases were observed in the emergence of chikungunya virus in the Americas[19].

Analysis of genetic variation can provide important clues to understanding ZIKV biology and pathogenesis, and may reveal potentially functional changes in the virus. Using the same dataset from 164 outbreak samples, we observed 1007 single nucleotide polymorphisms (SNPs), well distributed across the genome (Fig. 3a). These included 198 nonsynonymous SNPs (Supplementary Table 2) and 31 variants in the 5' and 3' untranslated regions (UTRs). We observed 4 positions with nonsynonymous mutations (at >5% frequency) that occur on two or more branches of the tree (Fig. 3b); two of these (at 4287 and 8991) occur together and might represent incorrect placement of a Brazil branch in the tree. The remaining two are more likely to represent multiple nonsynonymous mutations; one (at 9240) appears to involve nonsynonymous mutations to 2 different alleles.

To assess possible biological significance of these mutations, we looked for evidence of selection on the ZIKV genome. In particular, we searched for an excess of nonsynonymous mutations in the envelope (E) protein coding region; positive selection due to host antibody binding to this ZIKV surface protein could drive nonsynonymous changes there, as seen in other viral glycoproteins[20–22]. However, the nonsynonymous substitution rate in E proved to be similar to that in the rest of the coding region (Fig. 3c, left); moreover, amino acid changes were significantly more conservative in the region than elsewhere (Fig. 3c, middle and right). Any diversifying selection that is occurring in the surface protein appears to be operating under substantial selective constraint. In the ZIKV 3' UTR, we also found evidence for purifying selection (Fig. 3d and Supplementary Table 3); this confirms the functional importance of the region, which has previously been reported to play a role in viral replication[23].

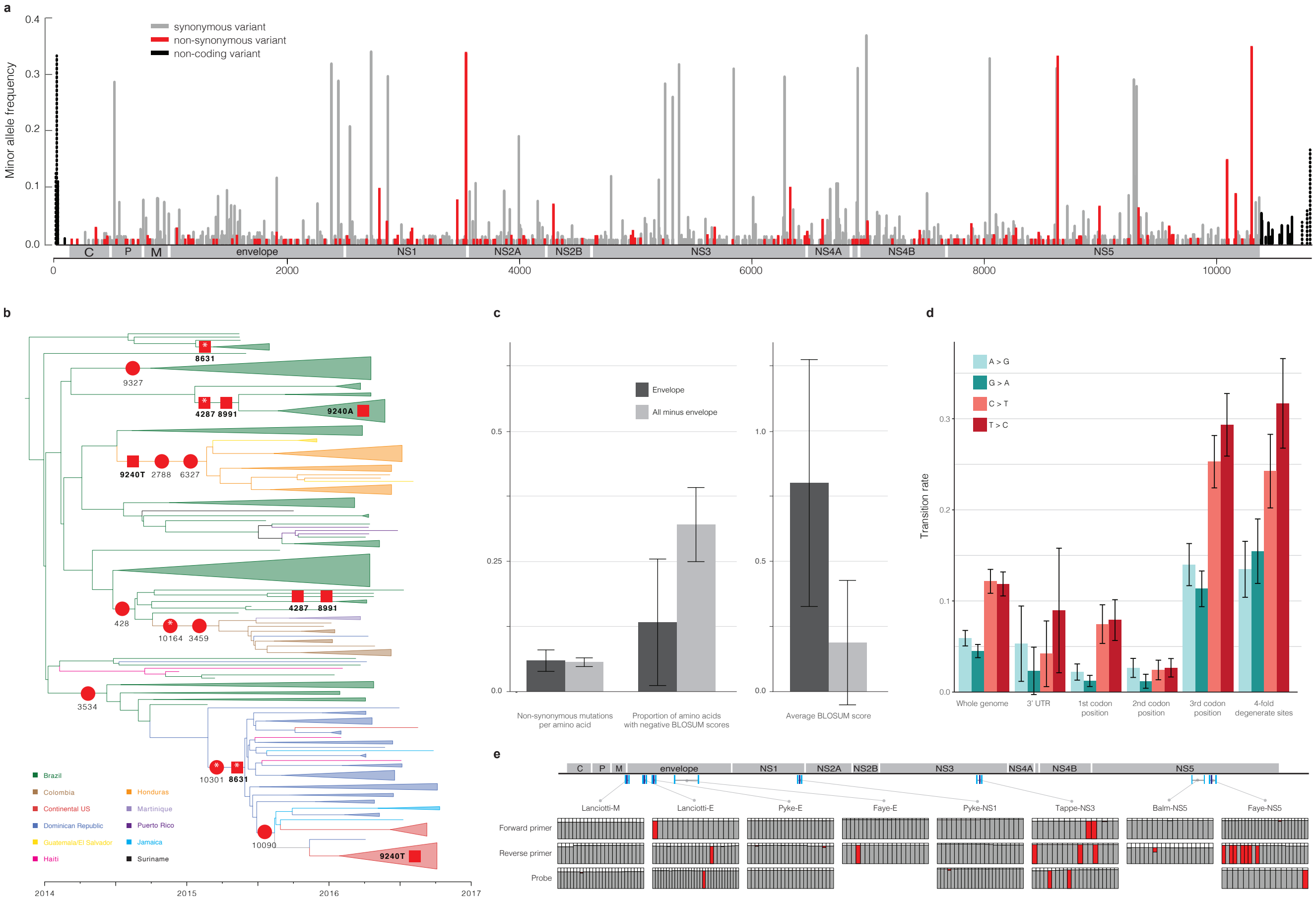


Figure 3 | Geographic and genomic distribution of Zika virus variation. (a) Location of variants in the ZIKV genome. The minor allele frequency is the proportion of genomes sharing a variant out of 164 genomes from the outbreak. Dotted bars (positions 1, 8-11, 13, 19, 10736, 10779, 10805) indicate <25% of samples had a base call at that position. **(b)** Phylogenetic distribution of nonsynonymous variants that have a derived frequency ≥5% (of the 164 genomes), shown on the branch where the mutation most likely occurred. White asterisks indicate the variant might be on the next-most ancestral branch (in one case, 2 branches upstream), but the exact location is unclear because of missing data. Squares denote a variant occurring at more than one location in the tree. **(c)** Conservation of the ZIKV envelope (E) region. Left: nonsynonymous variants per genome length for the E region (dark grey) and the rest of the coding region (light grey). Middle: proportion of nonsynonymous variants resulting in negative BLOSUM62 scores, which indicate unlikely or extreme substitutions ($p < 0.038$, $\chi 2$ test). Right: average of BLOSUM62 scores for nonsynonymous variants ($p < 0.029$, 2-sample t-test). **(d)** Constraint in the ZIKV 3' UTR and transition rates over the ZIKV genome. **(e)** ZIKV diversity in diagnostic primer and probe regions. Top: locations of published probes (dark blue) and primers (cyan)[53–58] on the ZIKV genome. Bottom: each column represents a nucleotide position in the probe or primer and each row one of the 164 ZIKV genomes. Cell color indicates that a sample's allele matches the probe/primer sequence (grey), differs from it (red), or has no data for that position (white).

While the transition-to-transversion ratio, which had a value of 7.0, was within the range seen in other viruses[24], we observed a significantly higher frequency of C-to-T and T-to-C substitutions than other transitions (Fig. 3d, Extended Data Fig. 4, Supplementary Table 3). This enrichment is apparent both in the genome as a whole and at 4-fold degenerate sites, where selection pressure is minimal. Many processes are possible contributors to this conspicuous mutation pattern, including mutational bias of the ZIKV RNA-dependent RNA polymerase, host RNA editing enzymes (e.g., APOBECs, ADARs) acting upon viral RNA, and chemical deamination, but further work is required to determine the actual cause of this phenomenon.

One potential concern for ZIKV surveillance is that genetic changes to the virus could reduce the sensitivity of diagnostic tests, which are based on specific viral RNA sequences. Accordingly, we evaluated 8 published qRT-PCR-based primer/probe sets *in silico* using all ZIKV genomes from this outbreak (Fig. 3e). We did find mutations in diagnostic regions that occurred during the outbreak, but they were all present at low frequencies. The bulk of mismatches between assays and ZIKV sequence appears to be the result of nucleotides that are fixed in the Americas. These observations suggest that diagnostic assays used in the Americas should be chosen with care, and should be systematically reevaluated and updated regularly[25].

Accurately calling within-host virus genetic diversity is important for understanding virus-host interactions and viral transmission. Therefore, in addition to the between-sample variants described above, we examined genetic variation within samples. We observed high concordance between the methods we used — amp-seq and hybrid capture — for base calls in our genome assemblies, including calls at polymorphic sites. By contrast, we found that most within-sample variants did not replicate between methods or between technical replicates (Fig. 1e-f and Extended Data Table 2). Even so, there is good reason to think that useful information about within-sample variation is accessible: when alleles do replicate, estimates of their frequency show good concordance (RMSE <5% in both methods). Imposing a previously described filter[26] eliminated all hybrid capture alleles not validated by a second replicate, but naively using these filters on amp-seq data had little effect (49% of alleles remain unvalidated). These results suggest that while these methods identify real within-sample variants, they also detect many spurious ones; current filtering can separate true from spurious hybrid capture variants, but technical improvements are needed to either reduce amp-seq noise or computationally filter it.

Sequencing low titer viruses like ZIKV directly from clinical samples presents several challenges, which have likely contributed to the paucity of genomes available from the current outbreak. While development of technical and analytical methods will surely continue, we note that factors upstream in the process, including collection site and cohort, were strong predictors of sequencing success (Extended Data Fig. 1). This suggests that study design and sample handling processes are key for this application of sequencing, and further that developing and implementing best practices for sample handling, without disrupting standard clinical workflows, will be vital to wider adoption of genome surveillance during outbreaks, especially for low titer viruses. Additional sequencing, however challenging, remains critical to ongoing investigation of ZIKV biology and pathogenesis. Together with two companion studies[12,13], this effort advances both technological and collaborative strategies for genomic surveillance in the face of unexpected outbreak challenges.

Acknowledgments

We are deeply grateful for the vision and support of Marc and Lynne Benioff, and for the continued support and guidance of Liliana Brown, Eun Mi Lee, and Maria Giovanni (NIAID), and Justine Levin-Allerhand and Eric S. Lander (Broad Institute). We thank Molly Schleicher, Emily Lipscomb, August Felix, Andrea Saltzman, and Stacey Donnelly (Broad Institute) for timely assistance with IRB and Ethics processes; Eva Mair and Liisa Nogelo (Broad Institute), and Erika Carmean (HHMI) for legal counsel; Tamara Mason and the Broad Institute Genomics Platform for sequencing support; Ashley Matthews, Sinéad Chapman, Daniel E. Neafsey, Bruce W. Birren (Broad Institute) for management and guidance; Oliver Pybus (Oxford University) and ZiBRA Project colleagues for sharing data prior to publication; Daniel Olson and Edwin Asturias (Children's Hospital Colorado) for sharing samples for preliminary studies; Marc Salit (National Institute of Standards and Technology) and Etienne Simon-Loriere (Pasteur Institute) for sharing reagents; and Gonzalo Bello (Fiocruz), Ryan Tewhey, Anne Piantadosi, Simon Ye, Chris Edwards and the Sabeti Lab (Broad Institute) for many helpful discussions and critical reading of the manuscript. We are indebted to Zika patients and clinical teams for making this work possible.

Funding

Funding for this work was generously provided by: Marc and Lynne Benioff (to P.C.S.), the National Institute of Allergy and Infectious Diseases, National Institutes of Health, Department of Health and Human Services, under Grant Number U19AI110818 (to the Broad Institute); the Howard Hughes Medical Institute (to P.C.S.); the Broad Institute "BroadNext10" program (to P.C.S. and A.G.); Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq, Brazil, grant number 440909/2016-3) and Fundação de Amparo a Pesquisa do Estado do Rio de Janeiro (FAPERJ, Brazil; grant numbers E-26/201.320/2016; E-26/201.332/2016; E-26/010.000194/2015) (to P.T.B.); National Institute of Allergy and Infectious Diseases 1R01AI099210-01A1 (to S.I. and S.M); MIDAS-National Institute of General Medical Sciences U54GM111274 (to M.E.H.); NIH NIAID AI100190 (to I.B. and L.G.); AEDES Network for support on sample collection and diagnostic (to I.B. and S.P.C.) and Colombian Science, Technology and Innovation Fund of Sistema General de Regalías-BPIN 2013000100011(to L.A.V, R.M.G., M.C.M. and I.B.); ASTMH Shope Fellowship (to K.G.B.); NSF DGE 1144152 (to A.E.L.); PNPD/CAPES Postdoctoral Fellowship (to E.D.); and Fulbright-Colciencias Doctoral Scholarship to D.P.R. K.G.A is Pew Biomedical Scholar, his work is supported by an NIH National Center for Advancing Translational Studies Clinical and Translational Science Award UL1TR001114 and NIAID contract HHSN272201400048C.

Methods

Ethics statement

The clinical studies from which samples were obtained were evaluated and approved by relevant by Institutional Review Boards/Ethics Review Committees at Hospital General de la Plaza de la Salud (Santo Domingo, Dominican Republic), University of the West Indies (Kingston, Jamaica), Universidad Nacional Autónoma de Honduras (Tegucigalpa, Honduras), Oswaldo Cruz Foundation (Rio de Janeiro, Brazil), Centro de Investigaciones Epidemiologicas - Universidad Industrial de Santander (Bucaramanga, Colombia), Massachusetts Department of Public Health (Jamaica Plain, Massachusetts), and Florida Department of Health (Tallahassee, Florida). Harvard University and Massachusetts Institute of Technology (MIT) Institutional Review Boards/Ethics Review Committees provided approval for sequencing and secondary analysis of samples collected by the aforementioned institutions.

Sample collections and study subjects

Suspected ZIKV cases (including high-risk travelers) were enrolled through study protocols at multiple aforementioned collection sites. Clinical samples (including blood, urine, cerebrospinal fluid, and saliva) were obtained from suspected or confirmed ZIKV cases and from high-risk travelers. De-identified information about study participants and other sample metadata are reported in Supplementary Table 1.

Viral RNA isolation

Viral samples were inactivated in AVL buffer (Qiagen) following standard operating protocol (0.2 mL sample) or large volume method (1 mL sample, Massachusetts Department of Health[27]). RNA (from AVL) was isolated using the QIAamp Viral RNA Minikit (Qiagen) according to the manufacturer's protocol, except that in some cases 0.1 M final concentration of β-mercaptoethanol was added to AVL buffer prior to inactivation, and in some cases 40 μg/mL final concentration of linear acrylamide (Ambion) was added. Extracted RNA was resuspended in water.

Carrier RNA and host rRNA depletion

In a subset of human samples, carrier poly(rA) RNA and host rRNA were depleted from RNA samples using RNase H selective depletion[10,28]. Briefly, oligo d(T) (40 nt long) and/or DNA probes complementary to human rRNA were hybridized to the sample RNA. The sample was then treated with 15 units of Hybridase Thermostable RNase H (Epicentre) for 30 minutes at 45°C. The complementary DNA probes were removed by treating each reaction with RNase-free DNase kit (Qiagen) according to the manufacturer's protocol. Following depletion, samples were purified using 1.8x volume AMPure RNAclean beads (Beckman Coulter Genomics) and eluted into 10 μl water for cDNA synthesis.

Illumina library construction and sequencing

cDNA synthesis was performed as described in previously published RNA-seq methods[10]. To track potential cross-contamination, 50 fg of synthetic RNA (ERCC, gift from M. Salit, NIST) was spiked into samples, using unique RNA for each individual ZIKV sample. Positive and negative control cDNA libraries were prepared from water, human K-562 total RNA (Ambion), ZIKV Senegal (HD78788) seed stock, Pernambuco (KX197192.1) seed stock, and EBOV Makona (KY425633.1) seed stock. The dual index Accel-NGS® 2S Plus DNA Library Kit (Swift Biosciences) was used for library preparation. Approximately half of the cDNA product was used for the library construction, and indexed libraries were generated using 18 cycles of PCR. Each individual sample was indexed with a unique barcode. Libraries were pooled equally based on molar concentration and sequenced on the Illumina HiSeq 2500 or MiSeq (paired-end reads) platforms.

Amplicon-based cDNA synthesis and library construction

ZIKV amplicons were prepared as described[11,13] with slight modifications. After PCR amplification, each amplicon pool was quantified on a 2200 Tapestation (Agilent Technologies) using High Sensitivity D1000 ScreenTape (Agilent Technologies). 2 μL of a 1:10 dilution of the amplicon cDNA was loaded and the concentration of the 350-550 bp fragments was calculated. The cDNA concentration, as reported by the Tapestation, was highly predictive of ZIKV positivity as determined by sequencing (Extended Data Fig. 5). cDNA from each of the two amplicon pools were mixed equally (10-25 ng each) and libraries were prepared using the dual index Accel-NGS® 2S Plus DNA Library Kit (Swift Biosciences) according to protocol. Libraries were indexed with a unique barcode using 7 cycles of PCR, pooled equally and sequenced on the Illumina MiSeq (250 bp paired-end reads) platform. Primer sequences were removed by hard trimming the first 30 bases for each insert read prior to analysis.

Zika virus hybrid capture

Viral hybrid capture was performed as previously described[10]. Probes were created to target ZIKV and chikungunya virus (CHIKV). Candidate probes were created by tiling across publicly available sequences for ZIKV and CHIKV (on GenBank). Probes were selected from among these candidate probes to minimize the number used while maintaining coverage of the observed diversity of the viruses. Alternating universal adapters were added to allow two separate PCR amplifications, each consisting of non-overlapping probes. (To download probe sequences, see Supplementary Information.)

The probes were synthesized on a 12k array (CustomArray). The synthesized oligos were amplified by two separate emulsion PCR reactions with primers containing T7 RNA polymerase promoters. Biotinylated baits were in vitro transcribed (MEGAshortscript, Ambion) and added to prepared ZIKV libraries. The baits and libraries were hybridized overnight (~16 hrs), captured on streptavidin beads, washed, and re-amplified by PCR using the Illumina adapter sequences. Capture libraries were then pooled and sequenced. Since samples often contained only a few ZIKV reads, a second round of hybrid capture was occasionally attempted (Extended Data Fig. 6). In the main text, “hybrid capture” refers to a combination of hybrid capture sequencing data and data from the same libraries without capture (unbiased), unless explicitly distinguished.

Genome assembly

We assembled reads from all sequencing methods into genomes using viral-ngs v1.13.3[29,30]. We taxonomically filtered reads from amp-seq against KU321639.1; we filtered reads from other approaches against the list of accessions provided in Supplementary Information. To compute results on individual replicates, we *de novo* assembled these and scaffolded against the KU321639.1 reference genome. To obtain final genomes for analysis, we pooled data from multiple replicates of a sample, *de novo* assembled, and scaffolded against KX197192.1. For all assemblies, we set the viral-ngs ‘assembly_min_length_fraction_of_reference’ and ‘assembly_min_unambig’ parameters to 0.01. For amp-seq, unambiguous base calls required at least 90% of reads to agree in order to call that allele (‘major_cutoff’ = 0.9); for hybrid capture data, we used the default threshold of 50%. We modified viral-ngs so that calls to GATK’s UnifiedGenotyper set ‘min_indel_count_for_genotyping’ to 2.

At 3 sites with insertions or deletions (indels) in the consensus genome CDS, we corrected the genome with results from Sanger sequencing (namely, at 3447 in DOM_2016_BB-0085-SER; at 5469 in BRA_2016_FC-DQ12D1-PLA; and at 6516-6564 in BRA_2016_FC-DQ107D1-URI, with coordinates in KX197192.1). At other indels, we replaced the indel with ambiguity.

Depth of coverage values from amp-seq include PCR and optical duplicates. In all other cases, we removed duplicates with viral-ngs.

Identification of non-ZIKV viruses in samples by unbiased sequencing

Using Kraken v0.10.6[31] in viral-ngs, we built a database that includes its default “full” database (which incorporates all bacterial and viral whole genomes from RefSeq[32] as of October 2015). Additionally, we included the whole human genome (hg38), genomes from PlasmoDB[33], sequences covering mosquito genomes (*Aedes aegypti*, *Aedes albopictus*, *Anopheles albimanus*, *Anopheles quadrimaculatus*, *Culex quinquefasciatus*, and the outgroup *Drosophila melanogaster*) from GenBank[34], protozoa and fungi whole genomes from RefSeq, SILVA LTP 16s rRNA sequences[35], and all sequences from NCBI’s viral accession list (as of October 2015) for viral taxa that have human as a host. (To download database, see Supplementary Information.)

For each replicate of unbiased sequencing data per sample (not including hybrid capture data), we ran Kraken and searched its output reports for viral taxa with more than 100 reported reads. We manually filtered the results, removing ZIKV, bacteriophages, and lab contaminants. For each sample and its associated taxa, we assembled genomes using viral-ngs as described above. We used the following genomes for taxonomically filtering reads and as the reference for assembly: KJ741267.1 (cell fusing agent virus), AY292384.1 (deformed wing virus), LC164349.1 (JC polyomavirus). When reporting sequence identity of an assembly with its taxon, we used the identity determined by BLASTN[36] when the assembly is compared against the reference genome used for its assembly.

Relationship between metadata and sequencing outcome

To determine if available sample metadata are predictive of sequencing outcome, we tested the following variables: sample collection site, patient gender, patient age, sample type, and the number of days between symptom onset and sample collection ("collection interval"). To describe sequencing outcome of a sample S, we used the following response variable Y_S:

mean({ I(R) * (number of unambiguous bases in R) for all amp-seq replicates R of S }),

where I(R)=1 if median depth of coverage of $R \geq 500$ and I(R)=0 otherwise

This value is listed in Supplementary Table 1 under "dependent variable used in regression on metadata". We excluded the saliva and cerebrospinal fluid sample types due to sample number (n=1), the samples from mosquito pools, and rows with missing values. We treated samples with type "Plasma EDTA" as having type "Plasma". We treated the "collection interval" variable as categorical (0-1, 2-3, 4-6, and 7+ days).

With a single model we underfit the zero counts, possibly because many zeros (no positive ZIKV assembly) are truly ZIKV-negative. We thus view the data as coming from two processes: one determining whether a sample is ZIKV-positive or ZIKV-negative, and another that determines, among the observed positive samples, how much of a ZIKV genome we are able to sequence. We modeled the first process with logistic regression (in R using GLM[37] with binomial family and logit link); the positive observed samples are the samples S for which $Y_S \geq 2500$. For the second, we performed a beta regression, using only the positive observed samples, of Y_S divided by ZIKV genome length on the predictor variables. We implemented this in R using the betareg package[38] and transformed fractions from the closed unit interval to the open unit interval as the authors suggest.

To test the significance of predictor variables, we used a likelihood ratio test. For variable X_i we compared a full model (with all predictors) against a model that uses all predictors except X_i. Results are shown in Extended Data Fig. 1.

Visualization of coverage depth across genomes

For amp-seq data, we plotted coverage across 97 samples that yielded a positive assembly by either method and for which we have amp-seq data (Fig. 1c). With viral-ngs, we aligned depleted reads to the reference sequence KX197192.1 using the novoalign aligner with options '-r Random -l 40 -g 40 -x 20 -t 100 –k'. There was no duplicate removal. We binarized depth at each nucleotide position, showing red if depth of coverage is at least 500x. Rows (samples) are hierarchically clustered to ease visualization.

For hybrid capture sequencing data, we plotted depth of coverage across the 32 samples that yielded a positive assembly by either method and for which we have hybrid capture data (Fig. 1d). (This is 2 more than the 30 samples that were positive based on their hybrid capture data.) We aligned reads as described above for amp-seq data, except we removed optical and PCR duplicates. For each sample, we calculated depth of coverage at each nucleotide position. We then shifted the values for each sample so that its mean depth is 0. At each nucleotide position in KX197192.1, we calculated the mean and standard deviation of the resulting values. We plotted the mean of these metrics within a 80 nt sliding window.

Criteria for pooling across replicates

We attempted to sequence one or more replicates of each sample and attempted to assemble a genome from each replicate. We discarded data from any replicates whose assembly showed high sequence similarity, in any part of the genome, to our assembly of a sample consisting of an African (Senegal) lineage (strain HD78788). We used this sample as a positive control throughout this study, and we considered its presence in the assembly of a clinical or mosquito pool sample to be evidence of contamination. We also discarded data from replicates that show evidence of contamination, at the RNA stage, by the baits we used for hybrid capture; we detected these by looking for adapters that were added to these probes for amplification.

For amp-seq, we consider an assembly positive if it contains at least 2500 unambiguous base calls and has a median depth of coverage of at least 500x over its unambiguous bases (depth is calculated including duplicate reads). For the unbiased and hybrid capture approaches, we consider an assembly of a replicate positive if it contains at least 4000 unambiguous base calls at any coverage depth. For each approach, we selected the unambiguous base threshold based on an observed density of negative controls below the threshold (Fig. 1b). For amp-seq assemblies, we added a threshold on depth of coverage because coverage depth was roughly binary across replicates, with negative controls falling in the lower class. Based on these thresholds, we found that 0 of 87 negative controls used throughout our sequencing runs yield positive assemblies and that 29 of 29 positive controls yield positive assemblies.

We consider a sample to have a positive assembly if any of its replicates, by either method, yields an assembly that passes the above thresholds. For each sample with at least one positive assembly, we pooled read data across replicates for each sample, including replicates with assemblies that do not pass the positivity thresholds. When data was available from both amp-seq and unbiased/hybrid capture approaches, we pooled amp-seq data separately from data produced by the unbiased and hybrid capture approaches, the latter two of which were pooled together (henceforth, the “hybrid capture” pool). We then assembled a genome from each set of pooled data. When assemblies on pooled data were available from both approaches, we selected the assembly from the hybrid capture approach if it had more than 10267 unambiguous base calls (95% of the reference, GenBank accession KX197192.1); when both assemblies had fewer than this number of unambiguous base calls, we selected the one that had more unambiguous base calls.

The number of ZIKV genomes publicly available prior to this study is the result of a GenBank[34] search for ZIKV in February 2017. We filtered any sequences with length <4000 nt, excluded sequences that are being published as part of this study or a co-submission, and excluded sequences labeled as having been passaged. We counted <100 sequences.

Multiple sequence alignments

We aligned ZIKV consensus genomes using MAFFT v7.221[39] with the following parameters: ‘--maxiterate 1000 --ep 0.123 --localpair’.

In Supplementary Data, we provide all sequences and alignments used in analyses.

Analysis of within- and between-sample variants

To measure overall per-base discordance between consensus genomes produced by amp-seq and hybrid capture, we considered all sites where base calls were made in both the amp-seq and hybrid capture consensus genomes of a sample, and we calculated the fraction in which the bases were not in agreement. To measure discordance at polymorphic sites, we took all of the consensus genomes generated in this study that we selected for downstream analysis and searched for positions with polymorphism (see **Criteria for pooling across replicates** for choosing among the amp-seq and hybrid capture genome when both are available). We then looked at these positions in genomes that were available from both methods, and we calculated the fraction in which the alleles were not in agreement. To measure discordance at minor alleles, we took all of the consensus genomes generated in this study that we selected for downstream analysis and searched for minor alleles. We then looked at all sites at which there was a minor allele and for which genomes from both methods were available, and we calculated the fraction in which the alleles were not in agreement. For both calculations, we tolerated partial ambiguity (e.g., ‘Y’ is concordant with ‘T’). If one genome had full ambiguity (‘N’) at a position and the other genome had an indel, we counted the site as discordant; otherwise, if one genome had full ambiguity, we did not count the site.

After assembling genomes, we determined within-sample allele frequencies for each sample by running V-Phaser 2.0 (via viral-ngs) on all pooled reads mapping to the sample assembly. When comparing allele frequencies between replicate libraries, we included only samples with two or more replicates. Similarly, the comparison of alleles across methods included only samples that have a positive assembly by either method and for which we have data from both methods. For these comparisons, we only included positions with a minor variant; i.e. positions for which both libraries/methods had an allele at 100% were removed, even if the single allele differed between the two libraries/methods. Additionally, we considered any allele with frequency <0.5% as not found (0%).

When comparing allele frequencies across methods: let f_a and f_{hc} be frequencies in amp-seq and hybrid capture, respectively. If both are nonzero, we only included an allele if the sequencing depth at its position was $\geq 1/\min(f_a, f_{hc})$ in both methods. If f_a=0, we required a sequencing depth of $1/f_{hc}$ at the position in the amp-seq method; similarly, if f_{hc}=0 we required a sequencing depth of $1/f_a$ at the position in the hybrid capture method. This was to eliminate lack of coverage as a reason for discrepancy between two methods. When comparing allele frequencies across sequencing replicates within a method, we eliminated effects of low coverage by imposing a minimum coverage depth in both libraries at the position in question: 500x for amp-seq; 100x for hybrid capture. In samples with more than two replicates, we only considered the two replicates with the highest coverage at a particular site for each allele.

We considered allele frequencies from hybrid capture sequencing “verified” if they passed the strand bias and frequency filters described in Gire et al. 2014[26], with the exception that we allowed a variant identified in only one library if its frequency was ≥5%. In Fig. 1f we applied the same strand bias filter but not the minimum frequency filter. In Fig. 1e and 1f, we considered variants “validated” if they were present at above 0.5% frequency in both libraries or methods. When comparing two libraries for a given method *M* (amp-seq or hybrid capture): the proportion unvalidated is the fraction, among all variants in *M* at ≥0.5% frequency in at least one library, of the variants that are at ≥0.5% frequency in exactly one of the two libraries. Similarly, when comparing methods: the proportion unvalidated for a method *M* is the fraction, among all variants at ≥0.5% frequency in *M*, of the variants that are at ≥0.5% frequency in *M* and <0.5% frequency in the other method. The root mean squared error (RMSE) includes only variants found in both methods or replicates (i.e. does not include unvalidated variants). Restricting the sample set used for comparison of variants across libraries to only samples with a positive assembly in both methods had no significant impact on the results.

We initially called SNPs on the aligned consensus genomes using Geneious version 9.1.7[40]. We converted all fully or partially ambiguous calls, which are treated by Geneious as variants, into missing data. We then removed all sites that were no longer polymorphic from the SNP set and re-calculated allele frequencies. A nonsynonymous SNP is shown on the tree (Fig. 3b) if it includes an allele that is nonsynonymous relative to the ancestral state (see **Molecular clock phylogenetics and ancestral state reconstruction** section below) and has an allele frequency of >5%; all occurrences of nonsynonymous alleles are shown. We placed mutations at a node such that the node leads only to samples with the mutation or with no call at that site. Uncertainty in placement occurs when a sample lacks a base call for the corresponding SNP; in this case, we placed the SNP on the most recent branch for which we have available data. We also used this ancestral ZIKV state to count the frequency of each type of substitution over various regions of the ZIKV genome, per number of available bases in each region (Fig. 3d and Supplementary Table 3).

We quantified the effect of nonsynonymous SNPs using the original BLOSUM62 scoring matrix for amino acids[41], in which positive scores indicate conservative amino acid changes and negative scores unlikely or extreme substitutions. We assessed statistical significance for equality of proportions by χ^2 test (Fig. 3c, middle), and for difference of means by 2-sample t-test with Welch-Satterthwaite approximation of df (Fig. 3c, right). Error bars are 95% confidence intervals derived from binomial distributions (Fig. 3c, left and middle) or Student's t-distributions (Fig. 3c, right). In Fig. 3d, error bars are 95% confidence intervals derived from binomial distributions.

Maximum likelihood estimation and root-to-tip regression

We generated a maximum likelihood tree using a multiple sequence alignment that includes sequences generated in this study, as well as a selection of other available sequences from the Americas, Southeast Asia, and the Pacific. (See the sequences listed in Supplementary Information.) We ran IQ-TREE[42] with options '-m HKY+G4 -bb 1000'[43]. In FigTree v1.4.2[44], we rooted the tree on the oldest sequence used as input (GenBank accession EU545988.1).

We used TempEst v1.5[45], which selects the best-fitting root with a residual mean squared function (also EU545988.1), to estimate root-to-tip distances. We performed regression in R with the lm function[37] of distances on dates.

In Supplementary Data, we provide the output of IQ-TREE, as well as the dates and distances used for root-to-tip regression.

Molecular clock phylogenetics and ancestral state reconstruction

For molecular clock phylogenetics, we made a multiple sequence alignment from the sequences generated in this study combined with a selection of other available sequences from the Americas. (See the sequences listed in Supplementary Information.)

We used BEAST v1.8.2 to perform molecular clock analyses[46]. We used the SDR06 substitution model on the CDS, which uses HKY with gamma site heterogeneity and partitions codons into two partitions (positions (1+2) and 3)[47]. For the noncoding regions, we used a HKY substitution model with gamma site heterogeneity and no codon partitioning. We used a strict clock model with a Bayesian Skyline tree prior (10 groups, piecewise-constant model)[48]. We set the molecular clock rate to use a continuous time Markov chain rate reference prior[49]. We ran BEAST twice, each with 75 million MCMC steps and sampling set to every 10,000th step, to verify convergence. We then removed the first 7.5 million states (10%) from each run as burn-in and combined the samples from the two runs. We extracted clock rate and tMRCA estimates, and their distributions, with Tracer v1.6.0 and identified the maximum clade credibility (MCC) tree using TreeAnnotator v1.8.2. The reported credible intervals around estimates are 95% highest posterior density (HPD) intervals.

We used BEAST v1.8.2 to estimate transition and transversion rates. The model was the same as above except that we used the Yang96 substitution model on the CDS, which uses GTR with gamma site heterogeneity and partitions codons into three partitions[50]; for the noncoding regions, we used a GTR substitution model with gamma site heterogeneity and no codon partitioning. There were four partitions in total: one for each codon position and another for the noncoding region. The number of steps, sampling, and burn-in was the same as above. At each sampled step of the MCMC, we calculated substitution rates for each partition using the overall substitution rate, the relative substitution rate of the partition, the relative rates of substitutions in the partition, and base frequencies. In Extended Data Fig. 4, we plot the means of these rates over the steps; the error bars shown are 95% HPD intervals of the rates over the steps.

We used BEAST v1.8.2 to reconstruct ancestral state at the root of the tree. The model was the same as the first described in this section except that, on the CDS, we used the HKY substitution model with gamma site heterogeneity and codons partitioned into three partitions (one per codon position). We ran BEAST twice, each with 40 million MCMC steps and sampling every 10,000th step, to verify convergence. We then removed the first 4 million states (10%) from each run as burn-in. On each run, we used TreeAnnotator v1.8.2 to find the state with the MCC tree. We selected the run with the higher log clade credibility for its MCC tree and used the ancestral state from this run.

In Supplementary Data, we provide BEAST input (XML) and output files. We also provide the sequence of the reconstructed ancestral state.

Principal component analysis

We carried out principal component analysis using the R package FactoMineR[51]. We imputed missing data with the package missMDA[52]. Removing the two most extreme outlier samples from the plot clarified population structure, and we show the results in Fig. 2c; we show the PC distribution for all samples in Extended Data Fig. 3. Removing additional outliers had little effect.

Diagnostic assay assessment

We extracted primer and probe sequences from eight published qRT-PCR assays[53–58] and aligned to our ZIKV genomes using Geneious version 9.1.7[40]. We then tabulated matches and mismatches to the diagnostic sequence for all outbreak genomes, allowing multiple bases to match where the diagnostic primer and/or probe sequence contained nucleotide ambiguity codes (Fig. 3e).

Data availability

Sequence data that support findings of this study will be deposited in GenBank prior to publication, and accession numbers will be listed here (BioProject accession PRJNA344504).

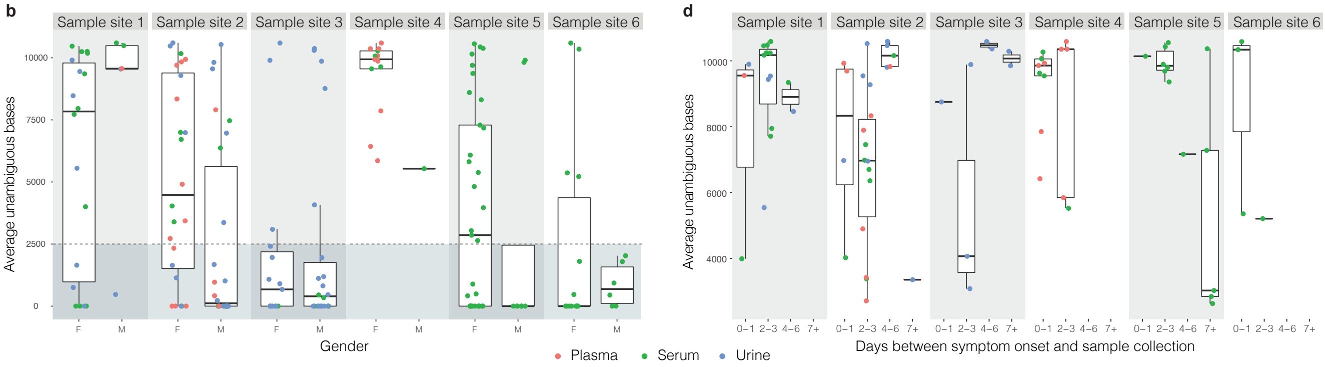
Extended Data

a

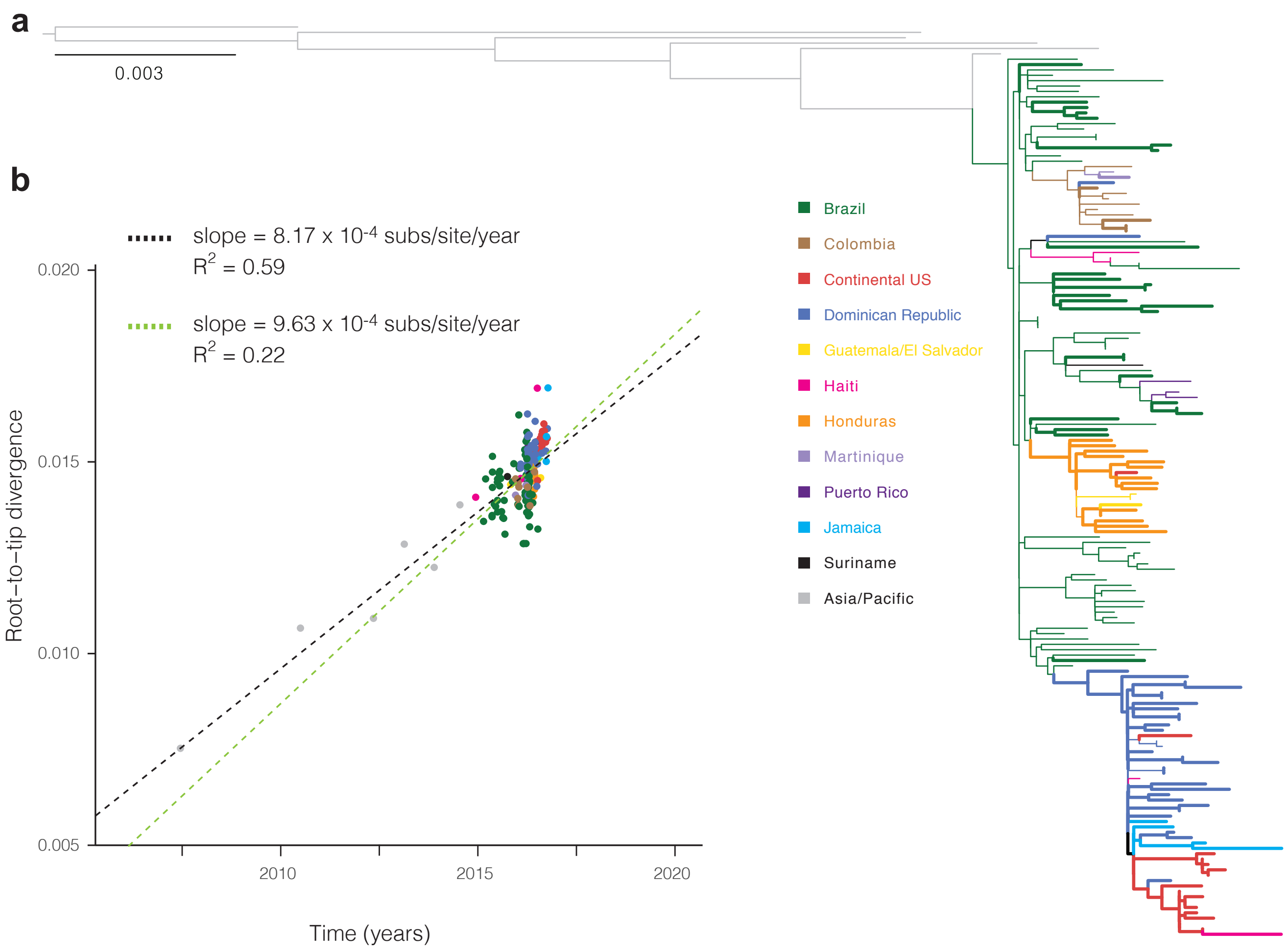
Model comparison all predictors *vs* all predictors except:	Df	$2\Delta \ln L$	Pr(>Chi-Squared)
sample site	5	27.801	<0.001***
patient gender	1	4.9961	0.0254*
patient age	1	0.9623	0.3266
sample type	2	0.0715	0.9649
collection interval	3	5.4396	0.1423

c

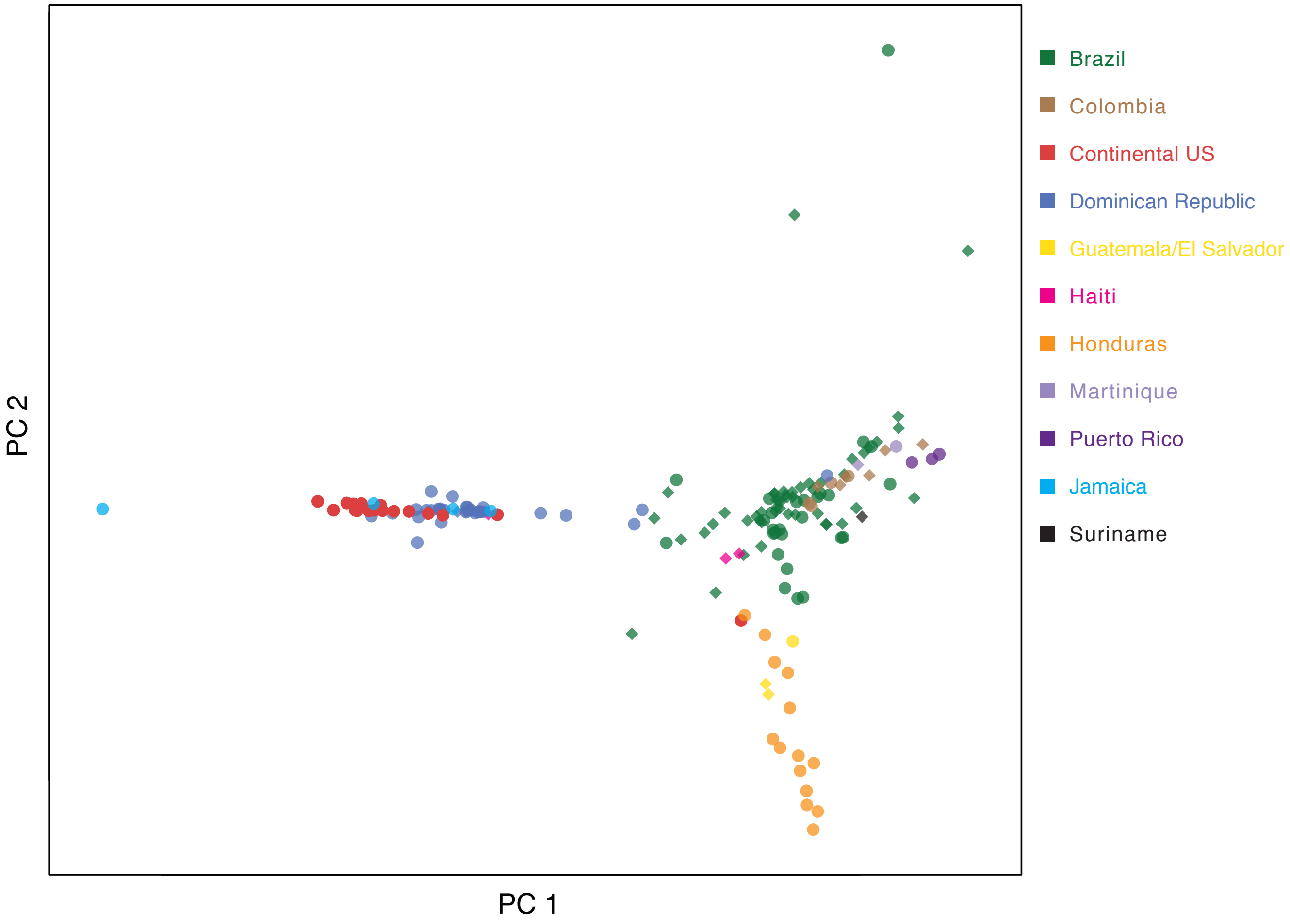
Model comparison all predictors *vs* all predictors except:	Df	$2\Delta \ln L$	Pr(>Chi-Squared)
sample site	5	7.762	0.1698
patient gender	1	1.5949	0.2066
patient age	1	0.3351	0.5627
sample type	2	0.5496	0.7597
collection interval	3	10.769	0.01304*



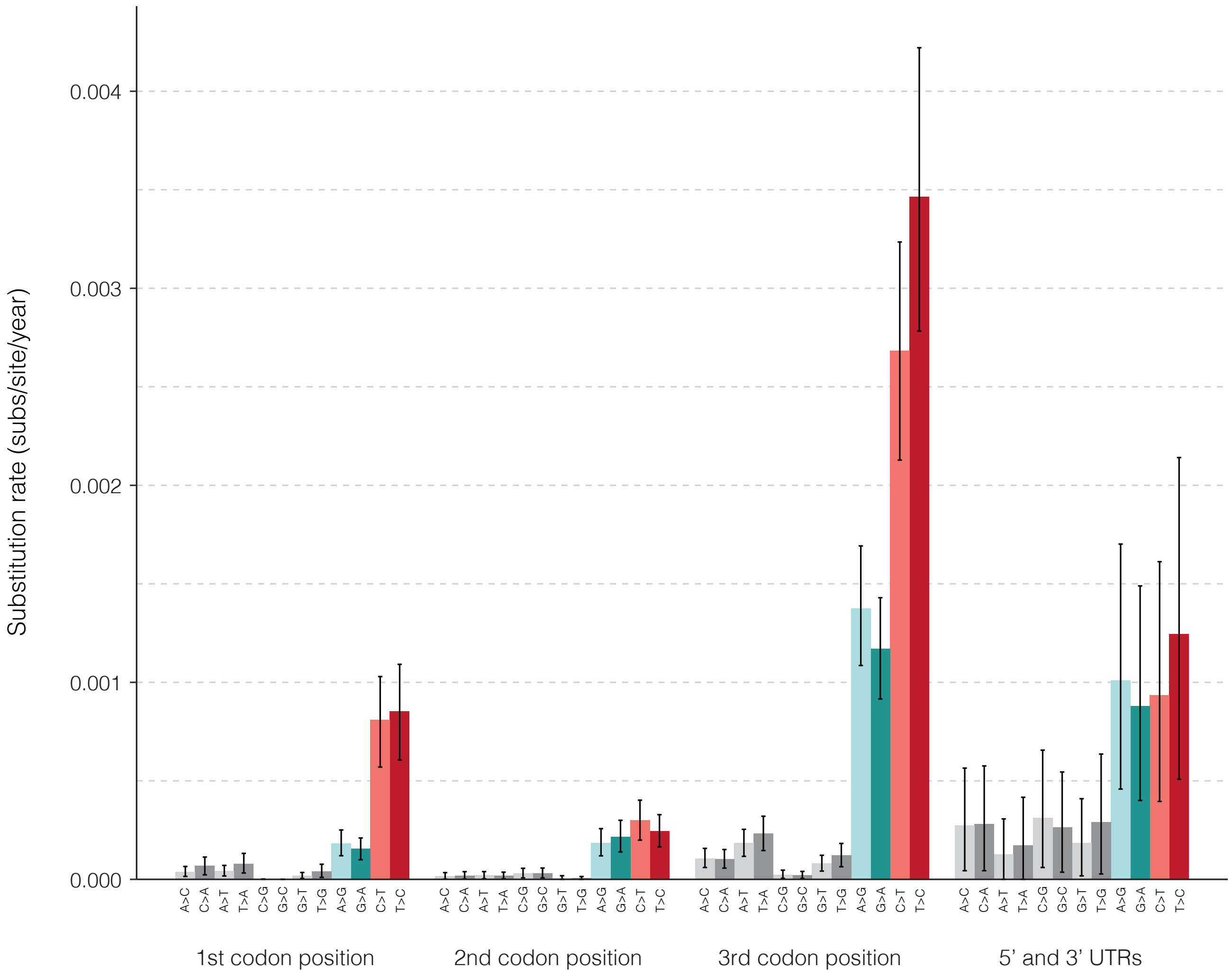
Extended Data Figure 1 | Relationship between metadata and sequencing outcome. Analysis of possible predictors of sequencing outcome: the site where a sample was collected, patient gender, patient age, sample type, and days between symptom onset and sample collection ("collection interval"). **(a)** Prediction of whether a sample is positive by sequencing. Rows show results of likelihood ratio tests on each predictor by omitting the variable from a full model that contains all predictors. Sample site and patient gender improve the model. **(b)** Sequencing outcome for each sample, divided by gender, across six samples sites. Shaded region below dotted line shows sequencing-negative values used in this model; region above is positive. The discrepancy in positivity between females and males is driven largely by Sample sites 2, 5, and 6. **(c)** Prediction of the percent genome identified, using sequencing-positive samples. Rows show results of likelihood ratio tests, as in (a). Collection interval improves the model. **(d)** Sequencing outcome for each sample, divided by collection interval, across six sample sites. Samples collected 7+ days after symptom onset produced, on average, the fewest unambiguous bases, though these observations are based on a limited number of data points. While the sample site variable accounts for differences in the composition of cohorts, the observed effects of gender and collection interval might be due to confounders in cohort composition that span multiple cohorts.



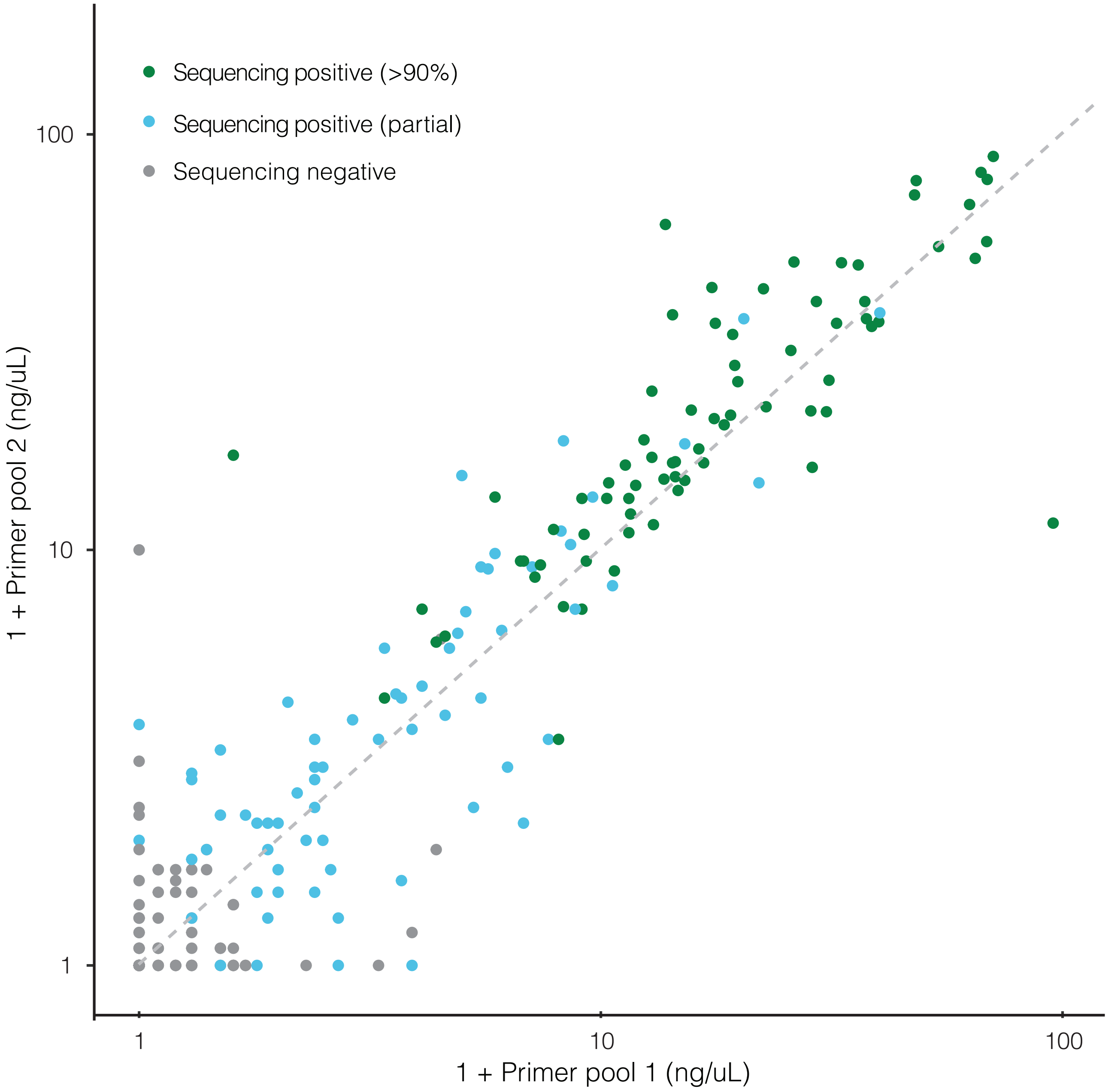
Extended Data Figure 2 | Maximum likelihood tree and root-to-tip regression. (a) Tips are colored by sample collection location. Bolded tips indicate those generated in this study; all other colored tips are other publicly available genomes from the outbreak in the Americas. Grey tips are samples from ZIKV cases in Southeast Asia and the Pacific. **(b)** Linear regression of root-to-tip divergence on dates. The substitution rate for the full tree, indicated by the slope of the black regression line, is consistent with rates of Asian lineage ZIKV estimated by molecular clock analyses[1]. The substitution rate for sequences within the Americas outbreak only, indicated by the slope of the green regression line, is consistent with rates estimated by BEAST ($1.04x10^{-3}$; 95% CI [$8.54x10^{-4}$, $1.21x10^{-3}$]) for this data set.



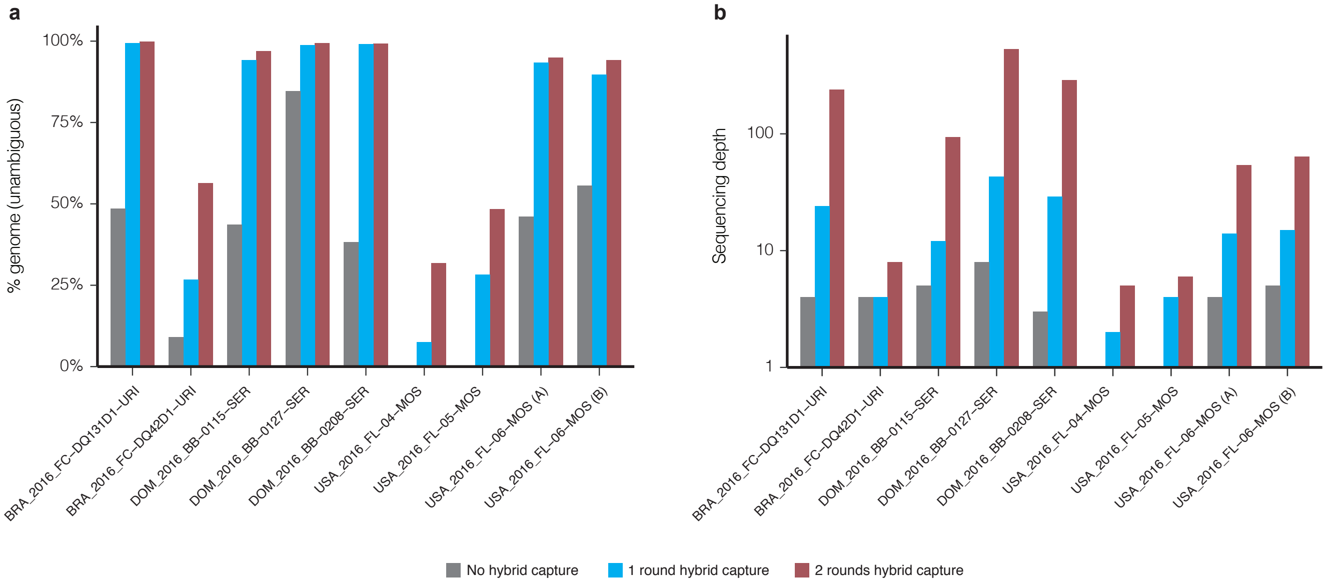
Extended Data Figure 3 | Principal component analysis with outliers. Principal component analysis of variants between samples in the outbreak. Circular points represent data generated in this study; diamond points represent other publicly available genomes from this outbreak that are used in this analysis. Outliers removed in Fig. 2c are included here; the relationship between countries is largely unchanged with inclusion of outliers.



Extended Data Figure 4 | Substitution rates estimated with BEAST. Substitution rates estimated in three codon positions and noncoding regions (5' and 3' UTRs). Transversions are shown in grey and transitions are colored by transition type. Plotted values show the mean of rates calculated at each sampled Markov chain Monte Carlo (MCMC) step of a BEAST run. Error bars indicate the 95% credible interval of the estimate. These calculated rates provide additional evidence for the observed high C-to-T and T-to-C transition rates shown in Fig. 3d.



Extended Data Figure 5 | cDNA concentration of amplicon primer pools predicts sequencing outcome. cDNA concentration of amplicon pools (as measured by Agilent 2200 Tapestation) is highly predictive of amp-seq outcome. On each axis, 1+primer pool concentration is plotted on a log scale. Each point is a technical replicate of a sample and colors denote observed sequencing outcome of the replicate. If a replicate is considered positive when at least one primer pool concentration is ≥0.8 ng/μL, then sensitivity=98.58% and specificity=91.47%.



Extended Data Figure 6 | Evaluating multiple rounds of Zika virus hybrid capture. Genome assembly statistics of samples prior to hybrid capture (grey), and after one (blue) or two (red) rounds of hybrid capture. 9 individual libraries (8 unique samples) were sequenced all three ways, had >1 million raw reads in each method, and generated at least one positive assembly. Raw reads from each method were downsampled to the same number of raw reads (8.5 million) before genomes were assembled. **(a)** Percent of the genome identified, as measured by number of unambiguous bases. **(b)** Median sequencing depth of Zika genomes, taken over the assembled regions.

Species	Sample	# reads from species (% of total)	% genome unambiguous
Cell fusing agent virus	USA_2016_FL-01-MOS	5662 (0.02%)	99.1%
	USA_2016_FL-04-MOS	1588 (0.003%)	91.1%
	USA_2016_FL-05-MOS	9614 (0.02%)	99.9%
	USA_2016_FL-06-MOS	2646 (0.007%)	82.2%
	USA_2016_FL-08-MOS	13608 (0.008%)	99.4%
Deformed wing virus	USA_2016_FL-06-MOS	6580 (0.02%)	8.34%
JC polyomavirus	BRA_2016_FC-DQ75D1-URI	8050 (0.20%)	99.2%
	USA_2016_FL-032-URI	316 (0.001%)	7.71%

Extended Data Table 1 | Viruses other than Zika uncovered by unbiased sequencing. Three viral species other than Zika were found by unbiased sequencing of 37 samples. Column 3: the number of reads in a sample that belong to the species as a raw count and a percent of total reads. Column 4: the percent genome assembled based on the number of unambiguous bases called. Cell fusing agent virus (a Flavivirus) and deformed wing virus were in mosquito pools, and JC polyomavirus was in clinical samples. All assemblies had ≥95% sequence identity to a reference sequence for the listed species, except cell fusing agent virus in USA_2016_FL-06-MOS (91%).

Extended Data Table 2a | Unvalidated variants across methods.

Method	% unvalidated by other method
Amp-seq	97.3%
Hybrid capture	82.8%
Hybrid capture, verified	29.0%

Extended Data Table 2b | Unvalidated variants within methods.

Method	% unvalidated in replicate	
	all variants	variants passing strand bias filter
Amp-seq	61.4%	49.3%
Hybrid capture	76.9%	0.00%

Extended Data Table 2 | Within-sample variant validation between and within sequencing methods. (a) For each method (amp-seq or hybrid capture), fraction of identified variants (≥0.5%) not identified at ≥0.5% by the other method. Verified hybrid capture variants are those passing strand bias and frequency filters, as described in Methods. **(b)** For each method, fraction of identified variants not validated by a second library. To pass the strand bias filter, a variant must meet filter criteria in both replicates.

Supplementary Files

Supplementary Information: Supplementary Methods. In particular, links to publicly available data used in analyses and listings of accession numbers of sequences used in analyses.

Supplementary Table 1: Table of information on 200 samples that we attempted to sequence in this study, including the 100 whose genomes we analyze. This provides sequencing outcome and metadata on the samples.

Supplementary Table 2: Table listing observed non-synonymous SNPs across the data used for SNP analysis. Includes frequency and count of ancestral and derived alleles at each position, as well as amino acid changes caused by each SNP.

Supplementary Table 3: Table giving substitution rates across the 164 genomes analyzed (100 of which were sequenced as part of this study). Includes observed mutations per available base (used in Fig. 3d), as well as substitution rates estimated by BEAST (used in Extended Data Fig. 4).

Supplementary Data: Sequences, alignments, BEAST input and output files, and root-to-tip data used in analyses. See README.txt for details.

References

1. Faria, N. R. *et al.* Zika virus in the Americas: Early epidemiological and genetic findings. *Science* **352,** 345–349 (2016).
2. *Zika situation report: Zika virus, Microcephaly and Guillian-Barré syndrome*. (World Health Organization, 2017).
3. de Vigilância em Saúde, S. *Protocolo de vigilância e resposta à ocorrência de microcefalia*. (Ministério da Saúde Brasília, 2016).
4. Dos Santos, T. *et al.* Zika Virus and the Guillain-Barré Syndrome - Case Series from Seven Countries. *N. Engl. J. Med.* **375,** 1598–1601 (2016).
5. Schieffelin, J. S. *et al.* Clinical illness and outcomes in patients with Ebola in Sierra Leone. *N. Engl. J. Med.* **371,** 2092–2100 (2014).
6. Sardi, S. I. *et al.* Coinfections of Zika and Chikungunya Viruses in Bahia, Brazil, Identified by Metagenomic Next-Generation Sequencing. *J. Clin. Microbiol.* **54,** 2348–2353 (2016).
7. Martina, B. E. E., Koraka, P. & Osterhaus, A. D. M. E. Dengue virus pathogenesis: an integrated view. *Clin. Microbiol. Rev.* **22,** 564–581 (2009).
8. Fauci, A. S. & Morens, D. M. Zika Virus in the Americas — Yet Another Arbovirus Threat. *N. Engl. J. Med.* **374,** 601–604 (2016).
9. Villamil-Gómez, W. E., González-Camargo, O., Rodriguez-Ayubi, J., Zapata-Serpa, D. & Rodriguez-Morales, A. J. Dengue, chikungunya and Zika co-infection in a patient from Colombia. *J. Infect. Public Health* **9,** 684–686 (2016).
10. Matranga, C. B. *et al.* Enhanced methods for unbiased deep sequencing of Lassa and Ebola RNA viruses from clinical and biological samples. *Genome Biol.* **15,** 519 (2014).
11. Quick, J. *et al.* Multiplex PCR method for MinION and Illumina sequencing of Zika and other virus genomes directly from clinical samples. *bioRxiv* 098913 (2017). doi:10.1101/098913
12. Faria, N. R. *et al.* Epidemic establishment and cryptic transmission of Zika virus in Brazil and the Americas. Preprint at https://doi.org/10.1101/105171 (2017).
13. Grubaugh, N. D. *et al.* Multiple introductions of Zika virus into the United States revealed through genomic epidemiology. Preprint at https://doi.org/10.1101/104794 (2017).
14. Pan American Health Organization. *Epidemiological Update: Zika virus infection*. (World Health Organization, 2015).
15. Pan American Health Organization. *Zika: Epidemiological Report Honduras*. (World Health Organization, 2016).
16. First case of Zika virus reported in Puerto Rico. *Centers for Disease Control and Prevention* (2015).
17. Pan American Health Organization. *Zika: Epidemiological Report Dominican Republic*. (World Health Organization, 2016).
18. Florida investigation links four recent Zika cases to local mosquito-borne virus transmission. *Centers for Disease Control and Prevention* (2016).
19. Nunes, M. R. T. *et al.* Emergence and potential for spread of Chikungunya virus in Brazil. *BMC Med.* **13,** 102 (2015).
20. Piantadosi, A. *et al.* HIV-1 evolution in gag and env is highly correlated but exhibits different relationships with viral load and the immune response. *AIDS* **23,** 579–587 (2009).
21. Ray, S. C. *et al.* Acute hepatitis C virus structural gene sequences as predictors of persistent viremia: hypervariable region 1 as a decoy. *J. Virol.* **73,** 2938–2946 (1999).
22. Villabona-Arenas, C. J. *et al.* Dengue Virus Type 3 Adaptive Changes during Epidemics in São Jose de Rio Preto,

Brazil, 2006–2007. *PLoS One* **8,** e63496 (2013).

23. Brinton, M. A. & Basu, M. Functions of the 3′ and 5′ genome RNA regions of members of the genus Flavivirus. *Virus Res.* **206,** 108–119 (2015).
24. Duchêne, S., Ho, S. Y. W. & Holmes, E. C. Declining transition/transversion ratios through time reveal limitations to the accuracy of nucleotide substitution models. *BMC Evol. Biol.* **15,** 36 (2015).
25. Charrel, R. N. *et al.* Background review for diagnostic test development for Zika virus infection. *Bull. World Health Organ.* **94,** 574–584D (2016).
26. Gire, S. K. *et al.* Genomic surveillance elucidates Ebola virus origin and transmission during the 2014 outbreak. *Science* **345,** 1369–1372 (2014).
27. *Zika Virus Response Updates from FDA*. (U.S. Food and Drug Administration, 2017).
28. Morlan, J. D., Qu, K. & Sinicropi, D. V. Selective depletion of rRNA enables whole transcriptome profiling of archival fixed tissue. *PLoS One* **7,** e42882 (2012).
29. Park, D. J. *et al.* Ebola Virus Epidemiology, Transmission, and Evolution during Seven Months in Sierra Leone. *Cell* **161,** 1516–1526 (2015).
30. Tomkins-Tinch, C. *et al. Broadinstitute/Viral-Ngs: V1.13.3*. (Zenodo, 2016). doi:10.5281/zenodo.200428
31. Wood, D. E. & Salzberg, S. L. Kraken: ultrafast metagenomic sequence classification using exact alignments. *Genome Biol.* **15,** R46 (2014).
32. O’Leary, N. A. *et al.* Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. *Nucleic Acids Res.* **44,** D733–45 (2016).
33. Aurrecoechea, C. *et al.* PlasmoDB: a functional genomic database for malaria parasites. *Nucleic Acids Res.* **37,** D539–43 (2009).
34. Clark, K., Karsch-Mizrachi, I., Lipman, D. J., Ostell, J. & Sayers, E. W. GenBank. *Nucleic Acids Res.* **44,** D67–72 (2016).
35. Yarza, P. *et al.* The All-Species Living Tree project: a 16S rRNA-based phylogenetic tree of all sequenced type strains. *Syst. Appl. Microbiol.* **31,** 241–250 (2008).
36. Altschul, S. F. *et al.* Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* **25,** 3389–3402 (1997).
37. R Core Team. R: A Language and Environment for Statistical Computing. *R Foundation for Statistical Computing* (2016). Available at: https://www.R-project.org/.
38. Cribari-Neto, F. & Zeileis, A. Beta Regression in R. *J. Stat. Softw.* **34,** 1–24 (2010).
39. Katoh, K. & Standley, D. M. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol. Biol. Evol.* **30,** 772–780 (2013).
40. Kearse, M. *et al.* Geneious Basic: an integrated and extendable desktop software platform for the organization and analysis of sequence data. *Bioinformatics* **28,** 1647–1649 (2012).
41. Henikoff, S. & Henikoff, J. G. Amino acid substitution matrices from protein blocks. *Proc. Natl. Acad. Sci. U. S. A.* **89,** 10915–10919 (1992).
42. Nguyen, L.-T., Schmidt, H. A., von Haeseler, A. & Minh, B. Q. IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. *Mol. Biol. Evol.* **32,** 268–274 (2015).
43. Minh, B. Q., Nguyen, M. A. T. & von Haeseler, A. Ultrafast approximation for phylogenetic bootstrap. *Mol. Biol. Evol.*

30, 1188–1195 (2013).

44. Rambaut, A. FigTree. Version 1.4.2. *Edinburgh, UK: Inst. Evol. Biol., Univ. Edinburgh.* (2014). Available at: http://tree.bio.ed.ac.uk/software/figtree/.

45. Rambaut, A., Lam, T. T., Max Carvalho, L. & Pybus, O. G. Exploring the temporal structure of heterochronous sequences using TempEst (formerly Path-O-Gen). *Virus Evol* **2,** vew007 (2016).

46. Drummond, A. J., Suchard, M. A., Xie, D. & Rambaut, A. Bayesian phylogenetics with BEAUti and the BEAST 1.7. *Mol. Biol. Evol.* **29,** 1969–1973 (2012).

47. Shapiro, B., Rambaut, A. & Drummond, A. J. Choosing appropriate substitution models for the phylogenetic analysis of protein-coding sequences. *Mol. Biol. Evol.* **23,** 7–9 (2006).

48. Drummond, A. J., Rambaut, A., Shapiro, B. & Pybus, O. G. Bayesian coalescent inference of past population dynamics from molecular sequences. *Mol. Biol. Evol.* **22,** 1185–1192 (2005).

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

50. Yang, Z. Maximum-Likelihood Models for Combined Analyses of Multiple Sequence Data. *J. Mol. Evol.* **42,** 587–596 (1996).

51. Lê, S., Josse, J. & Husson, F. FactoMineR: an R package for multivariate analysis. *J. Stat. Softw.* (2008).

52. Josse, J. & Husson, F. missMDA: A Package for Handling Missing Values in Multivariate Data Analysis. *J. Stat. Softw.* **70,** 1–31 (2016).

53. Pyke, A. T. *et al.* Imported zika virus infection from the cook islands into australia, 2014. *PLoS Curr.* **6,** (2014).

54. Lanciotti, R. S. *et al.* Genetic and serologic properties of Zika virus associated with an epidemic, Yap State, Micronesia, 2007. *Emerg. Infect. Dis.* **14,** 1232–1239 (2008).

55. Faye, O. *et al.* Quantitative real-time PCR detection of Zika virus and evaluation with field-caught mosquitoes. *Virol. J.* **10,** 311 (2013).

56. Faye, O. *et al.* One-step RT-PCR for detection of Zika virus. *J. Clin. Virol.* **43,** 96–101 (2008).

57. Balm, M. N. D. *et al.* A diagnostic polymerase chain reaction assay for Zika virus. *J. Med. Virol.* **84,** 1501–1505 (2012).

58. Tappe, D. *et al.* First case of laboratory-confirmed Zika virus infection imported into Europe, November 2013. *Euro Surveill.* **19,** (2014).