JVI
Journals.ASM.org

Genome-Wide Patterns of Intrahuman Dengue Virus Diversity Reveal Associations with Viral Phylogenetic Clade and Interhost Diversity

Poornima Parameswaran,[a] Patrick Charlebois,[b] Yolanda Tellez,[c] Andrea Nunez,[c] Elizabeth M. Ryan,[b] Christine M. Malboeuf,[b] Joshua Z. Levin,[b] Niall J. Lennon,[b] Angel Balmaseda,[c] Eva Harris,[a] and Matthew R. Henn[b]*

Division of Infectious Diseases and Vaccinology, School of Public Health, University of California, Berkeley, Berkeley, California, USA[a]; Broad Institute of Massachusetts Institute of Technology and Harvard, Cambridge, Massachusetts, USA[b]; and Laboratorio Nacional de Virología, Centro Nacional de Diagnóstico y Referencia, Ministry of Health, Managua, Nicaragua[c]

Analogous to observations in RNA viruses such as human immunodeficiency virus, genetic variation associated with intrahost dengue virus (DENV) populations has been postulated to influence viral fitness and disease pathogenesis. Previous attempts to investigate intrahost genetic variation in DENV characterized only a few viral genes or a limited number of full-length genomes. We developed a whole-genome amplification approach coupled with deep sequencing to capture intrahost diversity across the entire coding region of DENV-2. Using this approach, we sequenced DENV-2 genomes from the serum of 22 Nicaraguan individuals with secondary DENV infection and captured ~75% of the DENV genome in each sample (range, 40 to 98%). We identified and quantified variants using a highly sensitive and specific method and determined that the extent of diversity was considerably lower than previous estimates. Significant differences in intrahost diversity were detected between genes and also between antigenically distinct domains of the Envelope gene. Interestingly, a strong association was discerned between the extent of intrahost diversity in a few genes and viral clade identity. Additionally, the abundance of viral variants within a host, as well as the impact of viral mutations on amino acid encoding and predicted protein function, determined whether intrahost variants were observed at the interhost level in circulating Nicaraguan DENV-2 populations, strongly suggestive of purifying selection across transmission events. Our data illustrate the value of high-coverage genome-wide analysis of intrahost diversity for high-resolution mapping of the relationship between intrahost diversity and clinical, epidemiological, and virological parameters of viral infection.

Nearly three billion people worldwide are at risk for infection with dengue virus (DENV), a *Flavivirus* transmitted by the mosquitoes *Aedes aegypti* and *A. albopictus* (for reviews, see references 24 and 25). DENV is an RNA virus with a single-stranded, nonsegmented RNA genome ~10.7 kb in length, which encodes three structural proteins (capsid [C], premembrane/membrane [prM/M], and envelope]E]) and seven nonstructural (NS) proteins (NS1, NS2A, NS2B, NS3, NS4A, NS4B, and NS5). There are four closely related serotypes of DENV (DENV-1 to DENV-4) that can cause asymptomatic infections or clinical illnesses ranging from the self-limiting but debilitating dengue fever to the life-threatening dengue hemorrhagic fever (DHF) and dengue shock syndrome (DSS), which are characterized by vascular leakage (76). Even though several factors, such as viral genetics, preexisting heterotypic immunity (i.e., immunity to a different serotype due to a prior infection), the sequence of infection with distinct serotypes, and host genetics appear to influence disease severity (23, 35, 53, 54, 58, 61, 69), the precise causes for progression to severe disease remain unknown.

The four DENV serotypes share limited identity, with 25 to 40% variability observed between serotypes at the amino acid level (30, 68). Considerable variation is also observed between viruses from the same serotype (~3 to 6% at the nucleotide level), which are phylogenetically divided into genotypes that are further subdivided into clades. This extensive genetic variability originates from the accumulation of genetically distinct genomes in individual hosts (referred to here as intrahost diversity) due to the error-prone nature of the enzyme responsible for viral RNA replication, the viral RNA-dependent RNA polymerase (RdRp) (16). The overall composition of intrahost variants determines the consensus viral genome, which is defined as the composite of all intrahost variants in one host (Fig. 1). As in other RNA viruses (18, 31, 32, 38, 51, 52, 66), these genetic intrahost variants are thought to serve as the templates on which evolutionary mechanisms, such as recombination, drift, bottlenecking, or positive/negative selective pressures, act to shape variation at the consensus level between hosts (i.e., interhost diversity) (Fig. 1).

Several investigations have uncovered associations between genetic determinants at the consensus level and viral fitness and pathogenicity (e.g., see references 57 and 72). In addition to consensus sequence, the genetic composition of intrahost viral populations has been shown to be essential for maintaining the fitness of poliovirus populations *in vivo* (70) and for influencing hepatitis C virus (HCV) and human immunodeficiency virus (HIV) pathogenesis and disease outcome (18, 32, 38, 51, 66). For instance, in the case of HCV, high viral diversity was observed in individuals who presented with mild chronic hepatitis rather than severe hepatic injury (66), while low viral diversity posttreatment was

Received 24 March 2012 **Accepted** 22 May 2012

Published ahead of print 30 May 2012

Address correspondence to Eva Harris, eharris@berkeley.edu, or Matthew R. Henn, mhenn@sereshealth.com.

* Present address: Matthew R. Henn, Seres Health, Cambridge, Massachusetts, USA.

Eva Harris and Matthew R. Henn contributed equally to this study.

Copyright © 2012, American Society for Microbiology. All Rights Reserved.

doi:10.1128/JVI.00736-12

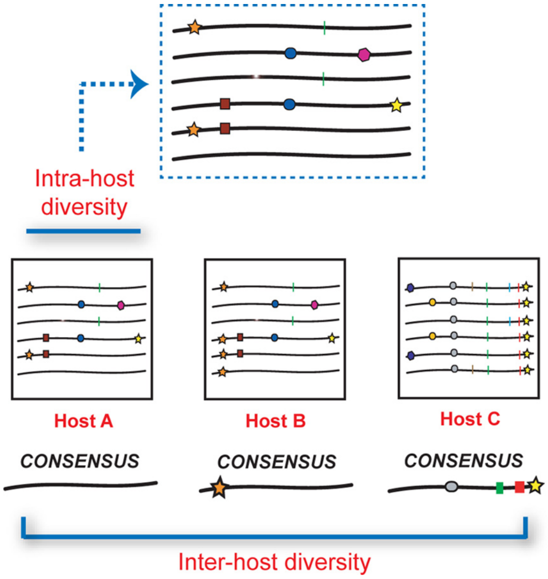

FIG 1 Schematic representation of the distinction between intrahost diversity and consensus-level (interhost) diversity. Viral genomes and genomic polymorphisms are represented by lines and symbols, respectively.

associated with sustained response to antiviral therapy (18, 51). The situation for HIV is more complex, with some studies reporting an association between higher viral diversity and slower disease progression (20), while others suggest that individuals harboring HIV populations with low diversity progress slower and mount more robust immune responses (32). In addition, recent work has shown that the level of diversity within an individual host during acute infection may be associated with viral set point, which refers to stabilized viral load after acute infection (28). Such studies illustrate the importance of assessing the composition of viral populations at the intrahost level during the course of an infection, which is distinct from assessing changes in consensus sequence at the interhost level between multiple infected individuals.

For DENV, most sequencing efforts have focused on assessing interhost diversity in viral consensus genomes from viruses isolated from serum (e.g., see references 7 and 47) or, as in a few studies, directly from viruses in human serum (e.g., see reference 39). Far fewer efforts have been directed at capturing DENV intrahost diversity and have considerable limitations, in that they either characterize a few sequences at the whole-genome level or sequence one or two genes of the viral genome, such as C, E, or NS2B (15, 43, 74, 75). In these studies, a wide range of intrahost diversity was observed in DENV populations in humans (15, 74, 75) and mosquitoes (43), with lower diversity observed in mosquitoes (43). Attempts have also been made to correlate intrahost viral diversity with disease severity. Descloux et al. (15) postulated that the level of viral diversity was lower in patients with severe dengue (DHF/DSS) than in patients with the less severe DF. However, in a recent study by Thai et al. (67), no correlation was observed between viral diversity in domain III of E and disease outcome or immune status. These authors also reported the detection of multiple viral lineages in some study subjects, indicating possible contributions from mixed infection to observed intrahost variation. Importantly, this study used rigorous algorithms to identify true variants from error introduced during PCR and documented far lower diversity than what had been previously reported.

We have developed a whole-genome segmented amplification approach which, when coupled with high-throughput sequencing, allows for capturing intrahost diversity across the entire coding region of the DENV genome with considerable depth of coverage. Using this approach, we captured diversity at 40 to 98% (average, ~75%) of the DENV-2 genome from 25 serum samples collected from 22 individuals, with an average coverage of 110 to 812 reads per nucleotide in each sample, and employed variant-calling algorithms (45) to identify true variants. The scale of this data set allowed us to critically compare gene-wise diversity within and between samples, detect rare mutation events, and correlate multiple measures of intrahost diversity with interhost viral diversity in a manner that was not possible before.

MATERIALS AND METHODS

Ethics statement. Written consent was obtained from the parents/legal guardians of all pediatric subjects, with subjects 6 years of age and older also providing assent. These studies were approved by the Institutional Review Boards (IRBs) of the University of California, Berkeley, and of the Nicaraguan Ministry of Health. The cohort study (see below) was also subject to review and approval by the IRB of the International Vaccine Institute in Seoul, South Korea.

Study population. Serum from subjects enrolled in one of two ongoing prospective studies of dengue in Nicaragua established by Eva Harris and her colleagues at the Nicaraguan Ministry of Health were used for this study. The first study (designated the hospital study; ongoing since 1998 [11, 26]) is a hospital-based study of pediatric dengue in patients who present with suspected dengue to the National Pediatric Reference Hospital in Nicaragua, Hospital Infantil Manuel de Jesús Rivera (HIMJR). Plasma and peripheral blood mononuclear cells (PBMCs) are collected from patients for three sequential days during the acute phase of illness, at convalescence (day 14), and longitudinally for 18 months. The second study is a prospective study referred to as the pediatric dengue cohort study, or PDCS (2004 to present) (2, 6, 33), which monitors ~3,800 children in District II of Managua. In this study, acute- and convalescent-phase serum samples collected from children who present with suspected dengue or undifferentiated febrile illness undergo rigorous laboratory diagnostic testing for DENV. Extensive epidemiological and clinical data are available for samples from both studies. Dengue cases are classified as DF, DHF, or DSS according to the 1997 WHO dengue guidelines (76).

Suspected dengue cases from both the hospital study and the PDCS are considered dengue positive if they meet at least one of the following criteria (27): (i) DENV genomes are detected using reverse transcription-PCR (RT-PCR) (27, 37); (ii) DENV is isolated from serum (5, 22); (iii) seroconversion of DENV-specific IgM antibodies is observed between paired acute- and convalescent-phase samples (3, 56); and/or (iv) there is a ≥4-fold increase in antibody titer between paired acute- and convalescence-phase samples, as measured by inhibition enzyme-linked immunosorbent assay (ELISA) (4, 19). Based on inhibition ELISA measurements, acute and convalescent antibody titers of <10 or <2,560, respectively, are indicative of primary DENV infection, whereas antibody titers of ≥10 (acute) or ≥2,560 (convalescent) are indicative of secondary DENV infection (27).

RNA extraction and DENV cDNA synthesis. Viral genomes were isolated from the sera of individuals using the QIAamp viral RNA isolation kit (Qiagen). For each sample, 5 μl of viral RNA (maximum of 5 μg), 50 ng random hexamers (Invitrogen), 0.5 μM DENV-2-specific primer (targeting the 3′ untranslated region; 5′-AGAACCTGTTGATTCAACAGCAC-3′), and 0.5 mM deoxynucleoside triphosphates (dNTPs) were incubated for 5 min at 65°C and then immediately placed on ice. In a final reaction volume of 20 μl, RT master mix (1× first-strand SuperScript III RT buffer, 5 mM $MgCl_2$, 10 mM dithiothreitol [DTT], 40 U RNase OUT, and 200 U SuperScript III RT; Invitrogen) was added to each sample and incubated for 10 min at 25°C and then for 50 min at 50°C. The reaction was terminated by heating to 85°C for 5 min. After cDNA synthesis, viral RNA was removed by digestion with RNase H (2 U/sample; Invitrogen) at 37°C for 20 min.

DENV cDNA amplification and sequencing. All of the DENV-2 genomes of Nicaraguan origin were sequenced at the Broad Institute as part of the Genome Resources in Dengue Consortium project. Genomic sequences were generated using either Sanger capillary sequencing or pyrosequencing strategies. One hundred sixty-one DENV-2 consensus genomes were generated using the Sanger sequencing protocol to determine codon and amino acid variability at the interhost level. These viruses were from sera collected between the years 2005 and 2009, which were also the years during which sera were collected for intrahost analysis.

For Sanger-based sequencing, cDNA was generated as described and amplified using degenerate primers (designed using PriSM [77]) to capture maximum diversity, and the high-fidelity DNA polymerase PfuUltra II Fusion HS DNA polymerase (Stratagene) was used to reduce PCR-related error. A total of 14 amplicons ranging in size from 1.5 to 2 kb were generated and were subsequently amplified using a panel of 96 primer pairs tailed with M13 sequence. The resulting 500- to 700-bp amplicons were sequenced in the forward and reverse directions, yielding a net sequence coverage of ~8-fold.

A subset of the 161 DENV-2 genomes was also sequenced using the Roche/454 pyrosequencing strategy (46) to obtain the higher coverage necessary for assessing intrahost diversity. Four PCRs were set up with PfuUltra II Fusion HS DNA polymerase (Stratagene) using the following primer pairs (designed using PriSM [77]): (i) 5′-CTACGTGGACCGACAAAGACAG-3′ (D2_F_11nt) and 5′-CCATRCCTATGTCATCYGTCAT-3′ (D2_R_3669nt); (ii) 5′-ATGGTGCARGCYGATAGTGGTT-3′ (D2_F_2411nt) and 5′-TCTGTCTGCRTAGTTGATRCCTTC-3′ (D2_R_6183nt); (iii) 5′-CTGGRACGTCAGGATCTCCAAT-3′ (D2_F_4919nt) and 5′-CGCTGTTGTCCRAAWGGAGT-3′ (D2_R_8618nt); and (iv) 5′-CTAGAWCCAATACCYTATGATCC-3′ (D2_F_7293nt) and 5′-CCATGCGTACAGCTTCCATGGT-3′ or 5′-GAACCTGTTGATTCAACAGCACC-3′ (D2_R_10477nt or D2_R_10712nt, respectively). Dimethylsulfoxide (DMSO) was added at a final concentration of 1% to all reaction mixtures. Thermocycling conditions were 95°C for 2 min (1 cycle); 94°C for 30 s, 62°C for 30 s, and 72°C for 3.5 min (40 cycles); 72°C for 10 min (1 cycle); and 10°C (hold). PCR products were quantified with the Quant-It DNA assay kit (Invitrogen), normalized, sheared, and barcoded by sample. Samples were pooled and sequenced on a 454/Roche high-throughput genome sequencer as previously described (28, 40), with a target average sequence coverage of 250-fold for each residue in the DENV genome per sample, allowing for the capture of variants that are present at >1% of the viral population.

Consensus genome assembly for intrahost and interhost analysis. For genomes sequenced using either Sanger- or pyrosequencing-based protocols, a consensus genome was generated using the AssembleViral454 *de novo* assembler (28). This assembler takes into consideration the inherent diversity in viral genomes. Briefly, reads were first clustered into small contigs using a lenient alignment algorithm and were subsequently assembled into larger contigs using an iterative algorithm that merges contigs, filters low-quality bases, and corrects assembly errors, as previously described (28). Resulting contigs for the assembly were ordered and oriented, corrected for frameshift errors, and annotated using in-house algorithms at the Broad Institute. Assemblies and annotations were then manually inspected, and any insertions/deletions (InDels) that were not resolved by frameshift correction algorithms and either were not supported by underlying reads or were in homopolymer regions were corrected. All consensus genome assemblies and annotations generated as part of this project were submitted to NCBI's GenBank database. 454 Read data for all samples is available in NCBI's Short Read Archive under BioProject 31235.

Calling variants at the interhost level. One hundred sixty-one full-length DENV-2 genomes of Nicaraguan origin sequenced at the Broad Institute (as described above) and 901 non-Nicaraguan full-length DENV-2 genomes obtained from NCBI (http://www.ncbi.nlm.nih.gov/) were used to determine interhost diversity in Nicaraguan and non-Nicaraguan DENV-2 populations, respectively. The Nicaraguan data set was composed of DENV-2 sequences from 27 primary, 2 probable primary, 99 secondary, 4 probable secondary, and 29 indeterminate immune status dengue cases. The non-Nicaraguan data set constituted genomes from 68 primary and 316 secondary dengue cases, with 517 of unknown immune status. Genomes were aligned using MUSCLE (17), and genes were manually annotated in the resulting alignments. Nucleotide, codon, and amino acid frequencies were calculated for every corresponding coordinate to yield percent interhost diversity values at every locus.

Variant-calling algorithms. 454/Roche pyrosequencing captures sequences of individual reads, thus providing information about the inherent diversity in nucleic acid repertoires and enabling assessment of polymorphisms at every nucleotide in the DENV genome (i.e., intrahost diversity). We utilized the ReadClean454 (RC454) (28) and V-Phaser (45) algorithms to call intrahost variants from the ultradeep 454 data sets. Briefly, RC454 was used to align reads to the *de novo* consensus assembly, and reads were subsequently corrected for sequencing-related artifacts (if present at <25% of the reads at a specific locus); errors evaluated included Carry Forward and Incomplete Extension errors (CAFIE) (9) and InDels resulting from overcalls and undercalls in homopolymeric regions. In addition, RC454 optimizes read alignments based on protein-coding information. The V-Phaser algorithm leverages sequence context information, such as base qualities and covariation of variants, and utilizes a composite Bernoulli model and expectation maximization to iteratively refine probabilities and to define the threshold at a given sequence coverage required to statistically define an observed variant as a true biological variant versus an amplification and/or sequencing artifact. The outputs from V-Phaser were further filtered to discard all regions spanning primer-binding sites, since it was difficult to ascertain whether mutations in these regions were true biological variants or were due to predefined heterogeneity in the degenerate primers. Further manual inspection revealed that tandem occurrence of true intrahost diversity with homopolymer miscalls caused alignment artifacts, resulting in the misalignment of multiple bases and creation of false mutations that occurred in phase with one another. Since such artifacts were found particularly at the ends of reads, we discarded variants that were only observed in the first 10 or last 10 nucleotides of alignments. Hence, there is a compelling need for manual inspection to identify and discard artifacts caused by complex factors, as they may otherwise not be detected by algorithms and filtering protocols.

Intrahost diversity hot-spot analysis. We utilized a permutation test to identify whether specific residues in the genome were hot spots for intrahost diversity, as previously described (28). First, residues in each genome were scored based on whether they harbored diversity. We then computed a total diversity score (i.e., total number of genomes with diversity at a given residue) for each residue. Scores associated with each residue were then permutated ($n = 1000$) across all residues in all genomes to generate a null distribution of diversity scores. Loci that demonstrated diversity scores in the 95th percentile or higher were considered to harbor statistically significant diversity and were defined as hot spots for intrahost diversity.

Nucleotide sequence accession numbers. All consensus genome assemblies and annotations generated as part of this project were submitted to NCBI's GenBank database under the following numbers (in parentheses): V3149 (HQ541788), V3241 (HQ541805), V3264 (JX079692), V4636 (HQ541793), V4637 (JX079694), V4639 (HQ541794), V4644 (HQ541801), V2595 (JX079686), V2596 (HQ541786), V2599 (HQ733861), V2605 (JF357905), V3136 (HQ541787), V3143 (HQ634199), V3152 (JX079688), V3223 (HQ541792), V3226 (JX079689), V3227 (JF357906), V3232 (JX079690), V3251 (JX079691), V3268 (JX079693), V4646 (JX079695), V4650 (JF357907), V4659 (JX079696), and V3148 (JX079687).

TABLE 1 Description of samples used in this study[a]

Sample	Day postinfection	Immune status	Disease severity	Yr	Sample ID	Clade[b]	Genes with ≥25-fold sequence coverage at every position	Fraction of the DENV-2 coding region covered	Avg read coverage across every nucleotide
V3149	5	Secondary	DHF	2007	314.1	NI-1	C to NS4B	0.74	812
V3241	4	Secondary	DF	2005	67.1	NI-1	NS1 to NS4B	0.51	162
V3264	4	Not available	DSS	2007	251.1	NI-1	prM to NS2B	0.41	122
V4636	3	Secondary	DF	2005	18.1	NI-1	prM to NS5	0.97	528
V4637	4	Secondary	DF	2005	18.2	NI-1	prM to NS5	0.97	164
V4639	3	Secondary	DF	2005	25.1	NI-1	C to NS5	0.98	493
V4642	3	Secondary	DF	2005	34.1	NI-1	prM to NS5	0.97	327
V4644	4	Secondary	DF	2005	80.1	NI-1	prM to NS4B	0.70	220
V2595	3	Secondary	DF	2006	113.1	NI-2B	prM to NS4B	0.70	214
V2596	2	Secondary	DF	2006	114.1	NI-2B	C to NS4B	0.74	304
V2599	3	Secondary	DHF	2006	128.1	NI-2B	C to NS5	0.98	363
V2605	3	Secondary	DF	2007	269.1	NI-2B	prM to NA4B	0.70	292
V3136	5	Secondary	DF	2007	266.1	NI-2B	prM to NS5	0.97	343
V3143	5	Secondary	DHF	2007	306.1	NI-2B	C to NS5	0.98	470
V3152	4	Secondary	DHF	2008	380.1	NI-2B	C to NS5	0.98	259
V3223	3	Secondary	DHF	2008	341.1	NI-2B	C to NS5	0.98	493
V3226	5	Secondary	DHF	2008	364.1	NI-2B	NS1 to NS4B	0.49	172
V3227	5	Secondary	DHF	2008	396.1	NI-2B	C to NS4B	0.74	291
V3232	5	Secondary	DSS	2007	275.1	NI-2B	prM to 2KPep	0.63	157
V3251	5	Secondary	DSS	2007	299.1	NI-2B	prM to NS2A, NS4A to NS4B	0.49	433
V3268	1	Secondary	DF	2008	410.1	NI-2B	NS1 to NS4B	0.50	174
V4646	4	Secondary	DF	2006	114.3	NI-2B	prM to NS4B	0.70	163
V4650	4	Secondary	DF	2007	269.2	NI-2B	prM to NS5	0.97	262
V4659	1	Secondary	DF	2007	2906.4.a.1	NI-2B	prM to NS4B	0.70	151
V3148	5	Secondary	DF	2007	313.1	NI-2B	prM to NS2B, NS4A to NS4B	0.53	110

[a] Serum samples were obtained from individuals who tested positive for DENV-2. More than 100 other clinical and laboratory parameters are available for these samples but are not shown due to space constraints.
[b] Nicaraguan clades NI-1 and NI-2B are 0.3% variant at the protein level.

RESULTS

General trends in DENV diversity. Clinical, virological, and epidemiological data for the samples used in this study are shown in Table 1. To minimize artifacts due to low coverage, we only considered genes in every sample that had sequence coverage of ≥25 reads at every nucleotide position (Table 1). We also discarded diversity observed at loci spanning primer-binding sites to minimize false contributions from degenerate bases. On average, 75% (range, 40 to 98%) of the DENV-coding region met these criteria in each sample (Table 1).

At the nucleotide level, approximately 0.027 to 0.66% of all sequenced nucleotide coordinates in the viral genome were associated with diversity (data not shown). Mutations were not biased toward a particular nucleotide (Fig. 2A). In fact, the mutation frequency of each nucleotide was proportional to its occurrence in the DENV genome (data not shown). As in other studies (10, 31, 63), transitions (A→G, G→A, C→T, or T→C) were the most common type of mutation (Fig. 2B). Also, mutations were more frequently observed in the third codon position, suggestive of selection against mutations in the first and second codon positions, which are more likely to affect amino acid encoding ($P < 0.0001$) (Fig. 2C).

We next computed multiple measures of intrahost diversity, based on percent divergence, calculated as the percentage of reads spanning each coordinate that was different from the consensus either in codon or amino acid space (referred to here as percent codon diversity and percent amino acid diversity, respectively). Codon mutations include variants that do or do not change amino acid encoding and may be variant in any of the three nucleotides comprising the codon. We classified intrahost variants according to abundance, with primary and secondary variants designating variants with higher and lower abundances, respectively. All diversity parameters were calculated per sample, either genome-wide or gene-wise, and either at the codon or amino acid level: (i) proportion (p) of coordinates (codons or amino acids) that exhibit diversity, computed as the number of coordinates with diversity divided by the total number of sequenced coordinates [p (divergence)]; (ii) sum percent divergence, computed as the sum of percent codon or percent amino acid diversity; (iii) average percent divergence, computed as the sum of percent codon or percent amino acid diversity divided by the number of codons or amino acids that show diversity; and (iv) variance in divergence, which is the relative variation in percent codon diversity or percent amino acid diversity and is computed as the relative standard errors of average percent divergence. These measures allow us to investigate various elements of diversity within an individual. For instance, a low p (divergence) is suggestive of a low frequency of mutations. High sum percent divergence and average percent divergence values indicate high levels of variants, which may either be confined to one coordinate or dispersed across multiple coordinates. High variance in divergence scores suggests contributions from both dominant and minor alleles, while low values suggest contributions from either dominant (if high average percent divergence) or minor (if low average percent divergence) alleles alone.

We observed significant variation in p (divergence), percent

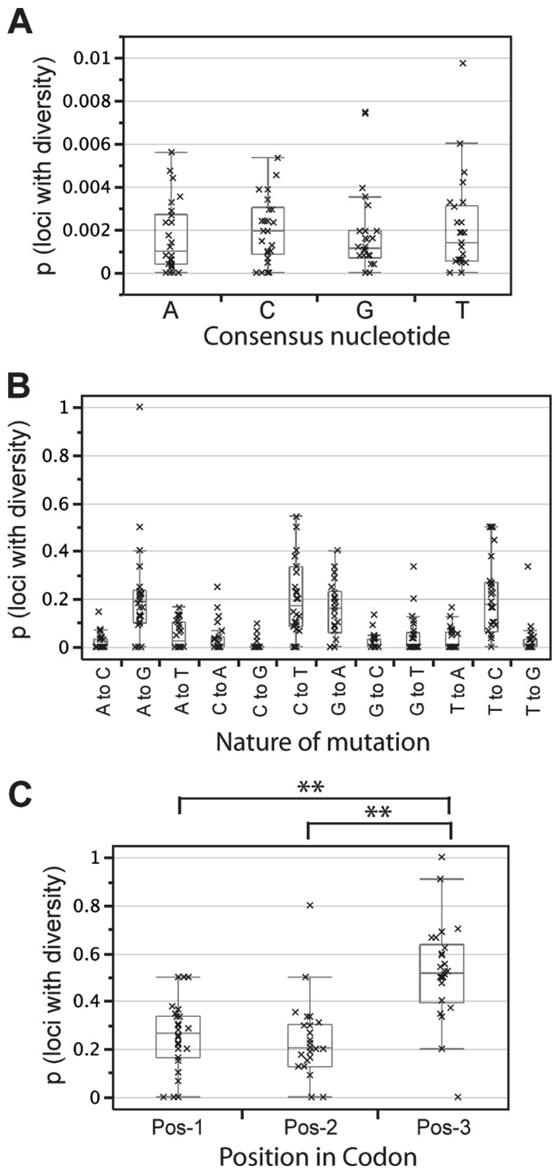

FIG 2 Nucleotide diversity in individuals infected with DENV-2. (A) Proportion (p) of loci that demonstrate diversity relative to the total number of sequenced nucleotides in each sample, categorized by the identity of the consensus nucleotide. (B) Proportion of specific nucleotide changes observed in each sample ($P < 0.0001$; Kruskal-Wallis). (C) Proportion of mutations that occur at codon position 1, 2, or 3 in each sample. $P < 0.0001$ (Kruskal-Wallis). **, $P < 0.0001$ for individual comparisons (one-way chi-square test).

codon diversity, and percent amino acid diversity between individuals (Fig. 3). On average, p (divergence) scores were 0.0047 (range, 0.00057 to 0.018) and 0.0025 (range, 0 to 0.009) for codons and amino acids, respectively, per sample, while average percentages of divergence for codons and amino acids were 5.69% (range, 1.83 to 16.26%) and 4.14% (range, 1.25 to 20.1%), respectively (Fig. 3). A wide range of variance in divergence values was observed at the codon (range, 0 to 73.98%; median, 25%) and amino acid (range, 0 to 100%; median, 45%) levels across samples, suggesting that there were no consistent biases toward minor or dominant alleles in DENV-2 intrahost populations. Also, 9.1 to 100% (median, 50%) of all detected mutations at the codon level in each sample resulted in changes to the amino acid sequence (i.e., nonsynonymous mutations) (Fig. 3). Even though this is a considerably wide range, it is interesting that in at least half the samples, the majority of detected mutations may affect protein function through changing amino acid encoding.

Across all samples, diversity was detected at 285 codon coordinates in the DENV-2 genome (data not shown). However, only six of these loci (0.02%) passed the threshold set by the permutation test to be considered a hot spot for diversity (Table 2). We further investigated whether the limited number of hot spots was due to the conservative filtering imposed by V-Phaser, which requires any mutation to be observed on at least two or more reads depending on sequence coverage. We included in our analysis variants that were observed on single reads and confirmed that very few mutations were observed in multiple samples (data not shown). The lack of mutations occurring in multiple samples supports a disseminated diversity model where mutations are distributed across distinct sets of loci, rather than a hot spot diversity model where diversity is observed in the same subset of loci across multiple hosts.

In addition to substitutions, our approach also allowed us to identify mutations that gave rise to stop codons in various genes (data not shown) and deletions that maintained or did not maintain frame (Table 2). The repeated detection of such defective genomes (Table 2) raises the possibility of their maintenance through genetic complementation with functional full-length genomes, as has been observed in other studies (1, 41). Indeed, some of these deletions are Nicaraguan (NI) clade specific, with most being propagated across epidemics and appearing in multiple consecutive years (Table 2). Alternatively, rapid DENV replication kinetics in human hosts could result in frequent production of such defective variants across multiple hosts.

Gene-specific trends in intrahost diversity. Surveys of DENV intrahost diversity have focused primarily on the Envelope (E) gene (15, 43, 67, 74), since intrahost diversity in this structural gene has direct implications for infectivity, antibody binding/neutralization, and major histocompatibility complex (MHC) presentation/T cell-mediated responses. Consequently, information about intrahost diversity at other regions of the viral genome is scarce, making it difficult to critically compare intrahost diversity between genes in the same host. Our genome-wide approach allowed us to capture intrahost diversity at several genes in the same host and to assess whether particular regions of the viral genome were subject to differential selection pressures.

For each sample, we calculated a gene-wise p (divergence) and a gene-wise average percent divergence score for codons and amino acids. No significant differences were observed in p (divergence) scores (data not shown). However, there were significant differences between genes in average percent divergence scores for codons ($P = 0.0005$) and amino acids ($P = 0.0055$) (Fig. 4A and B, box plots). Specifically, the most pronounced differences in codon average percent divergence were observed between E and NS2B ($P < 0.0001$) and between NS5 and NS2B ($P = 0.0005$) (Fig. 4A, box plots), while the most significant difference in amino acid average percent divergence was observed between NS5 and NS2B ($P = 0.0006$) (Fig. 4B, box plots). Thus, E and NS5 appear to have the highest evolutionary potential in the human host, while NS2B may be one of the slower-evolving DENV genes in the context of DENV-2 infections in humans. In contrast, very different trends in gene-specific distribution of diversity were found across con-

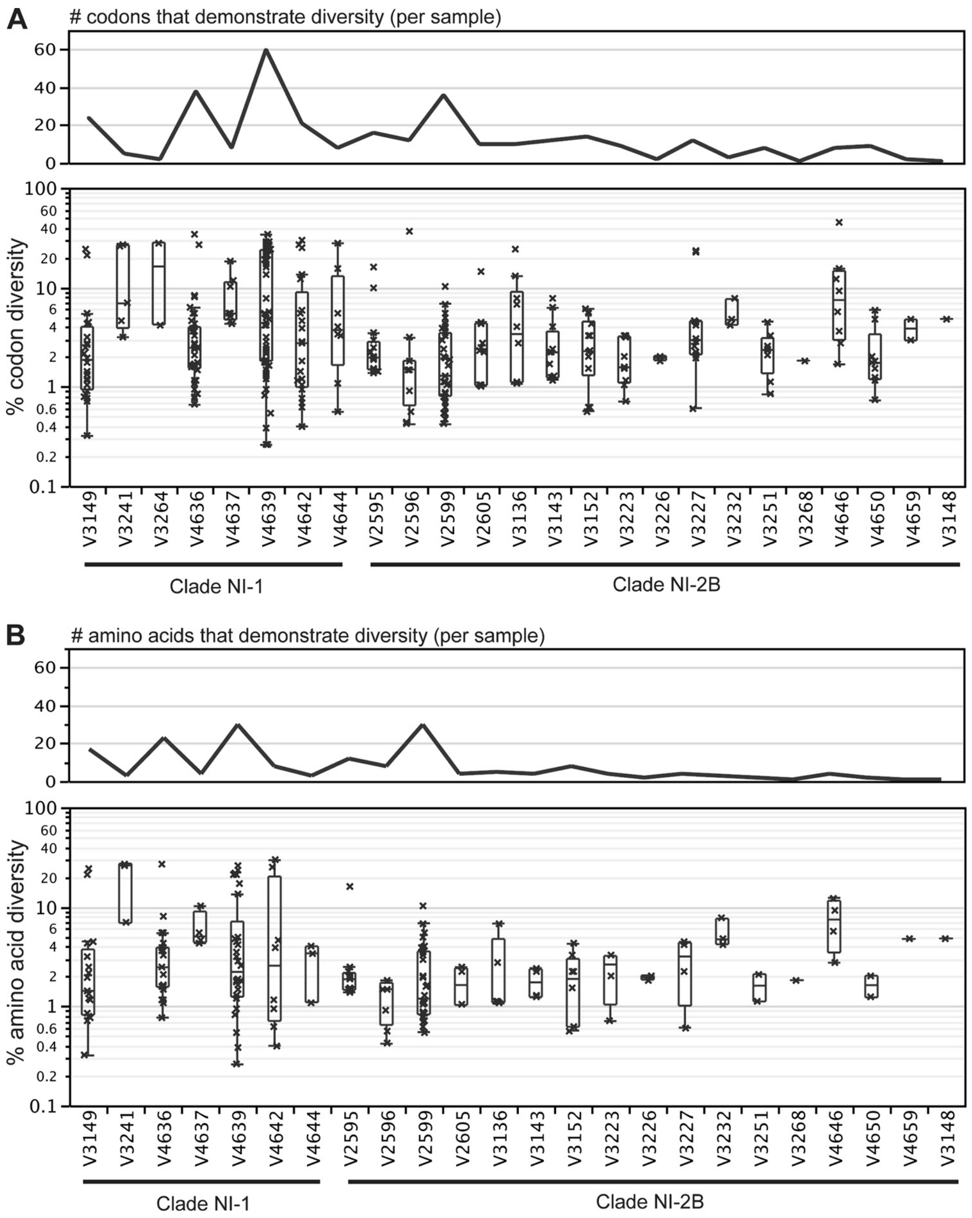

FIG 3 Wide range of DENV-2 intrahost diversity in secondary dengue cases in humans. (A and B) Total number of loci with variant alleles (line plots) and distribution of percent diversity (i.e., percentage of reads that are distinct from the consensus; box plots) at various loci in codon space (A) or amino acid space (B) for each sample. NI-1 and NI-2B refer to viruses from Nicaraguan clades 1 and 2B, respectively.

sensus genomes circulating in Nicaragua (i.e., interhost diversity) (Fig. 4A and B, line plots), and no significant differences were noted between NS5, E, and NS2B.

Intrahost diversity in the Envelope gene. Domains I, II, and III (which form the ectodomain) and the C-terminal hydrophobic domain (C-term) of the E protein not only have distinct structural functionalities in the folding and assembly of the protein on the surface of the virion but also elicit substantially different antibody responses in humans (49, 50, 59). Domain II (EDII) contains the fusion loop responsible for low-pH-mediated fusion with endosomal compartments, while EDIII is a putative receptor-binding domain with an immunoglobulin-like fold, as predicted from E structures in other flaviviruses (8, 60, 71, 78). Murine monoclonal antibody studies have determined that epitopes in EDI and EDII generally elicit serotype cross-reactive antibodies that are poorly to moderately neutralizing (12, 55), while antibodies targeting EDIII are, on the whole, more serotype specific and potently neutralizing (14, 21, 42, 44, 65), although some cross-reactive and nonneutralizing monoclonal antibodies binding to EDIII epitopes have been recently identified (62, 64, 65). Studies of the human polyvalent antibody repertoire have determined that the majority of the antibody response appears to be directed against EDI/II, and only 10 to 15% of the neutralizing response may be attributed to antibodies targeting EDIII (13, 36, 48, 73).

We observed intrahost diversity in all domains of E, with conspicuous differences in the distribution of loci with diversity among domains ($P = 0.0027$ and 0.0010 for codons and amino acids, respectively). The majority of codon mutations in E were

TABLE 2 Hot spots for intrahost diversity in the DENV-2 genome[a]

			Interhost diversity (%)					
Gene (affected codon position[s] in gene)	Type of mutation	Identical variants across samples	Nica (n = 161)	Non-Nica interhost diversity (n = 901)	Sample count	Sample IDs	Clade identity (no. of samples)	Epidemic season[b] (no. of samples)
C (7)	Substitution	No	No	Yes (0.11)	2	V3149, V4639	NI-1 (2), NI-2B (0)	2005–2006 (1), 2007–2008 (1)
C (8)	Substitution	No	No	Yes (0.22)	2	V3149, V4639	NI-1 (2), NI-2B (0)	2005–2006 (1), 2007–2008 (1)
C (66[c])	Substitution	Yes	No	Yes (69.37)	2	V3148, V4642	NI-1 (1), NI-2B (1)	2005–2006 (1), 2007–2008 (1)
NS2B (49[c])	Substitution	Yes	Yes (1.9)	Yes (8.99)	2	V2599, V3149	NI-1 (1), NI-2B (1)	2006–2007 (1), 2007–2008 (1)
NS5 (270[c])	Substitution	Yes	No	Yes (23.75)	2	V2599, V3152	NI-1 (0), NI-2B (2)	2006–2007 (1), 2008–2009 (1)
NS5 (424[c])	Substitution	Yes	No	Yes (4.66)	2	V2596, V2599	NI-1 (0), NI-2B (2)	2006–2007 (2)
NS5 (12)	Deletion, frameshift	Yes	No	No	4	V4636, V4639, V4642, V4637	NI-1 (4), NI-2B (0)	2005–2006 (4)
NS5 (53)	Deletion, no frameshift	Yes	No	No	3	V2596, V2599, V2605	NI-1 (0), NI-2B (3)	2006–2007 (2), 2007–2008 (1)
NS5 (238)	Deletion, frameshift	Yes	No	No	2	V2596, V2605	NI-1 (0), NI-2B (2)	2006–2007 (1), 2007–2008 (1)

[a] Only positions that demonstrate diversity in more samples than predicted by chance alone (i.e., with a score ≥95th percentile using the permutation test) are shown. Nica, Nicaraguan; Non-Nica, non-Nicaraguan.

[b] The epidemic season in Nicaragua lasts from the months of July/August until January/February of the following year.

[c] Loci where the proportions of observed variants in Nicaraguan and non-Nicaraguan virus populations are significantly different (Fisher's exact test).

located in EDII, followed by C-term and EDI, while EDIII had the least number of mutations (Fig. 4C and D, box plots; the pairwise comparison between EDII and EDIII was most significant, with $P = 0.0002$). The rank order for number of loci with amino acid diversity was EDII $>$ EDI $>$ EDIII $\approx$ C-term ($P = 0.0005$ for the most significant pairwise comparison between EDII and EDIII/C-term).

We then corrected for the lengths of the various domains by calculating the domain-specific p (divergence), which is the number of loci with diversity relative to the length of the domain. We did not detect any significant differences in codon p (divergence) scores (Fig. 4E, box plots). However, there were significant differences between domains in amino acid p (divergence) scores ($P = 0.0419$), with the most significant difference observed between EDII and EDIII (Fig. 4F, box plots; $P = 0.0094$). No significant differences were observed in codon p (divergence) or amino acid p (divergence) in consensus genomes (Fig. 4C to F, line plots). Overall, our observations suggest the presence of differential immune pressures on the various domains of E.

Phylogenetic clade influences the extent of intrahost diversity. We next investigated whether any of the diversity parameters [p (divergence), sum percent divergence, average percent divergence, or variance in divergence] correlated with various clinical, virological, and epidemiological factors. As in other studies (67), we observed no significant correlations between any of these measures and serum viremia (data not shown). Additionally, no association was discerned between any of these measures of diversity and severity of disease (i.e., DF versus DHF/DSS), although our power to detect such associations is limited given our sample size. Interestingly, the only factor that appeared to have a significant effect on all three parameters was viral clade identity (Fig. 5A to D). OhAinle et al. (54) had previously reported the cocirculation of distinct clades of DENV-2 in Nicaragua. At the consensus level, these clades are highly similar and only differ at nine positions in the coding region (54). At the intrahost level, we found that p (divergence) for genes NS4A and NS4B differed significantly between clades NI-1 and NI-2B ($P = 0.0012$ and 0.0155, respectively) (Fig. 5A), as did sum percent divergence scores for E, NS4A, and NS4B ($P = 0.0343$, 0.0003, and 0.0049, respectively) (Fig. 5B) and average percent divergence scores for NS4A and NS4B ($P = 0.0009$ and 0.0062, respectively) (Fig. 5C). In all cases, the measure of diversity was significantly higher for NI-1 viruses than NI-2B viruses. Apparent outliers were indistinguishable from the rest of the group with respect to various parameters, including position in the gene, nature of the variant, disease severity, and other clinical parameters associated with the sample. There were also no significant differences in variance in divergence scores, suggesting that there was no bias toward dominant or minor alleles in viruses from these clades. Intriguingly, the two clades also exhibited differential diversity within domains of E; a higher proportion of loci in EDII and EDIII exhibited diversity in NI-1 viruses compared to NI-2B viruses ($P = 0.0138$ and 0.0047 for EDII and EDIII, respectively) (Fig. 5D). These observations suggest that NI-2B viruses are less diverse at the intrahost level than NI-1 viruses.

We also examined whether distinct patterns of intrahost diversity contributed to the overall genetic variation between clades NI-1 and NI-2 at the consensus level. We evaluated intrahost diversity in samples at all critical clade-defining loci that distinguish viruses from either clade (54). Of the nine clade-defining loci, only two (encoding amino acids NS1 to 279 and NS3 to 245) harbored variants. At both of these loci, the observed dominant variant was the clade-defining allele and the minor variants were at low frequency, suggesting that the loci that distinguish the two primary

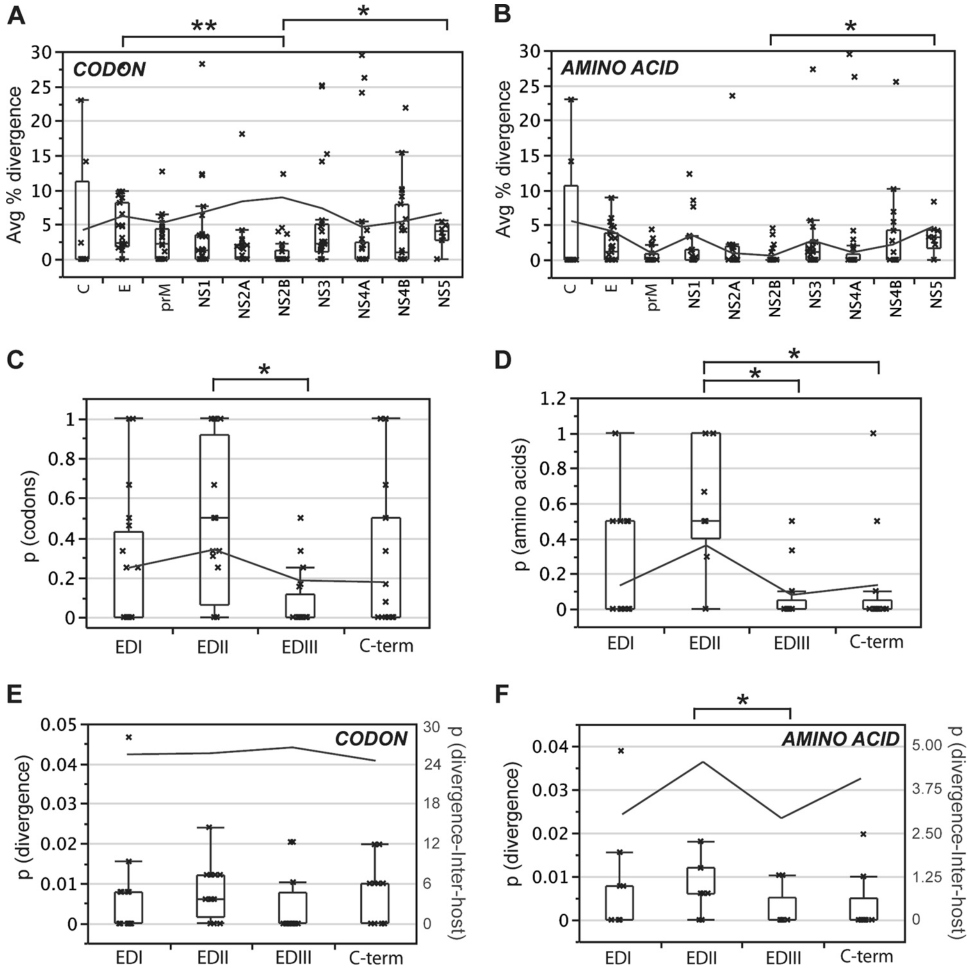

FIG 4 Gene-specific and domain-specific trends in DENV-2 intrahost diversity. (A and B) Average percent codon diversity (A) and percent amino acid diversity (B) (average percent divergence) per gene per sample. Averages are significantly different at $P = 0.0005$ (A) and 0.0055 (B) (Kruskal-Wallis), with individual comparisons significant at $P < 0.0001$ (**) for E and NS2B (A) by one-way chi-square test, $P = 0.0005$ (*) for NS5 and NS2B (A) by one-way chi-square test, and $P = 0.0006$ (*) for NS5 and NS2B (B) by one-way chi-square test. (C and D) Distribution of loci with intrahost (C) codon [p (codon)] or (D) amino acid [p (amino acid)] diversity between ectodomains EDI to EDIII and the C-terminal (C-term) region of the E gene in each sample. In panel C, $P = 0.0027$ (*) for comparisons between EDII and EDIII (one-way chi-square test). In panel D, $P = 0.0010$ (*) for comparisons between EDII and EDIII as well as EDII and C-term (one-way chi-square test). (E and F) Proportion of codons (E) and amino acids (F) with diversity [p (divergence)] in the three ectodomains (EDI to EDIII) and the C-terminal region of the Envelope gene calculated for each sample. For panel F, significant differences are observed in p (divergence) at the amino acid level across the domains ($P = 0.0419$ by Kruskal-Wallis), with comparison between EDII and EDIII significant at $P = 0.0094$ (*; one-way chi-square test test). The gray line (C to F) shows the corresponding value at the consensus level (values on the left y axis for A to D and on the right y axis for E and F [p (divergence-interhost)].

Nicaraguan clades are fixed and are not targeted for intrahost evolution.

Viral diversity at the intrahost-interhost interface. To evaluate potential bottlenecks that occur during events such as virus transmission, we compared patterns of diversity at the intrahost level to those observed at the interhost level. Interhost metrics were computed either across 161 viral isolates that were isolated from Managua during the same epidemic years as the samples used for intrahost analysis (i.e., between 2005 and 2009) or across 901 non-Nicaraguan DENV-2 genomes. Viruses isolated from primary and secondary cases of dengue were represented at similar ratios in both data sets. We classified mutations according to whether they appeared at the interhost level only, at the intrahost level only, or in both (intrahost and interhost). We observed that across all samples, the number of loci harboring variants (synonymous or nonsynonymous) that were shared between the intrahost and the Nicaraguan interhost data sets was significantly less than the number of loci in the intrahost data set alone (Fig. 6A). Therefore, the majority of synonymous and nonsynonymous mutations observed within hosts were not observed at the consensus level in the circulating Nicaraguan DENV-2 populations (Fig. 6A). In addition, sites that were identified as diversity hot spots in the intrahost data set by a permutation test were not necessarily hot spots of diversity in the Nicaraguan interhost population (Table 2); only one of the sites harbored low levels of mutation at the interhost level. Notably, all six hot spots for intrahost diversity in our samples harbored mutations in non-Nicaraguan DENV-2 genomes, with significant differences between Nicaraguan and non-Nicaraguan viruses observed at four loci (Table 2). This is suggestive of unique selection pressures on transmitted/

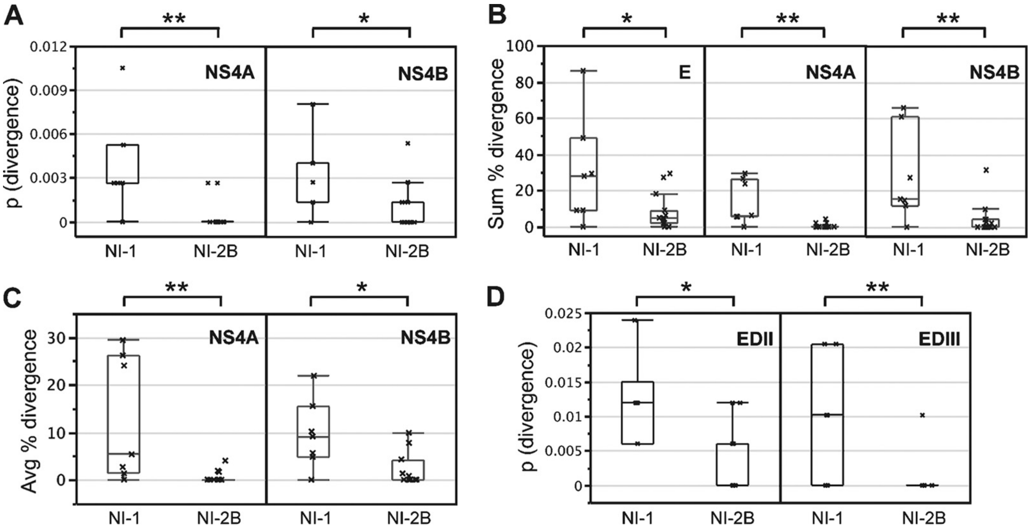

FIG 5 Nicaraguan clade NI-1 viruses demonstrate higher diversity than clade NI-2B viruses. All parameters were calculated per gene or per domain for each sample. Only significant differences between clades are shown. (A) Proportion of codons [p (divergence)] that exhibit diversity. Differences in p (divergence) values are significant at $P = 0.0012$ (**; NS4A) and 0.0155 (*; NS4B). (B) Percent codon diversity summed for all loci with diversity (sum percent divergence). Significant differences in sum percent divergence values are observed for E (*, $P = 0.0343$), NS4A (**, $P = 0.0003$), and NS4B (**, $P = 0.0049$). (C) Average percent codon diversity. Differences in average percent divergence values are significant at $P = 0.0009$ (**; NS4A) and $P = 0.0062$ (*; NS4B). (D) p (divergence) values for codon diversity in E ectodomains EDII and EDIII. $P = 0.0138$ (*; EDII) and 0.0047 (**; EDIII). All *P* values were calculated using one-way chi-square test tests. *, $P < 0.05$; **, $P < 0.005$ (highly significant because it passes the Bonferroni correction threshold).

infectious viral populations in Nicaragua compared to non-Nicaraguan DENV-2 populations.

We also tested whether the abundance of specific mutations in the intrahost data set correlated with the appearance of these mutations in consensus genomes at the interhost level in Nicaragua (or vice versa). We observed that loci associated with identical synonymous mutations (Fig. 6B; list of loci is available upon request) or identical nonsynonymous mutations (Fig. 6C; list of loci is available upon request) in both intrahost and interhost data sets had higher intrahost diversity than loci that exhibit distinct mutations in either data set only ($P < 0.0001$ and 0.0002 for synonymous and nonsynonymous comparisons, respectively). This suggests that a synonymous or nonsynonymous mutation has to be present at relatively high levels within a host (median, 5.2 and 8.49%, respectively) in order to also be observed at the interhost level.

In addition, homology-based functional predictions using SIFT analysis (34) revealed that only ~34% of all nonsynonymous substitutions that appeared only at the intrahost level were tolerated (i.e., were not predicted to affect protein function) (Fig. 6D, box plot). This was much lower than the 75% tolerance predicted for mutations that appeared at the interhost level only ($P < 0.0001$) (Fig. 6D, line plot). Very few nonsynonymous mutations were present at both intra- and interhost levels, but even the majority of these (mostly singleton) mutations were predicted to be tolerated ($P = 0.045$ for comparison with the intrahost-only data set) (data not shown). We surmise that most mutations that appear only at the intrahost level are deleterious and extinguished (i.e., not transmitted), whereas mutations that appear at both intrahost and interhost levels have a higher probability of being either neutral or favorable.

The combination of these results strongly suggests that purifying selection occurs during transmission, with selective pressures being more extreme for nonsynonymous mutations than synonymous mutations ($P = 0.03$) (Fig. 6A), for low-abundance intrahost variants compared to high-abundance tolerated variants (Fig. 6B and C), and for nonsynonymous substitutions that are predicted to affect protein function compared to ones predicted to be tolerated (Fig. 6D).

DISCUSSION

We used whole-genome amplification methodologies combined with deep sequencing to capture intrahost diversity across the coding region of the DENV genome. This approach allowed us to identify gene-specific and intragenic differences within samples and across samples and to investigate associations with disease severity and viral genetics. In addition, since the samples used for both intrahost and interhost analyses were from the same study population in Managua, Nicaragua, we were able to assess whether the spectra of variants observed within hosts were also present at the consensus level in actively circulating DENV-2 populations.

In our study, the numbers of variants within a host at the nucleotide or amino acid level were observed to be much lower than previous estimates (10, 15, 43, 74, 75). This discrepancy is most likely explained by our use of variant-calling algorithms that were designed specifically to minimize contributions from process errors to estimates of intrahost diversity. These algorithms also produced similar mutation frequencies when implemented by Thai et al. (67) to select for true variants. Such corrective measures were either not employed or were minimally employed in previous studies, suggesting that previous reports of DENV intrahost diversity represent overestimates of diversity. One limitation of our shotgun approach was that it did not allow us to determine whether detected variants (especially those separated by distances greater than read length) were present in *cis* (i.e., on the same genome) or in *trans* (i.e., on different genomes). Hence, it was difficult to estimate the exact proportion of variant genomes in

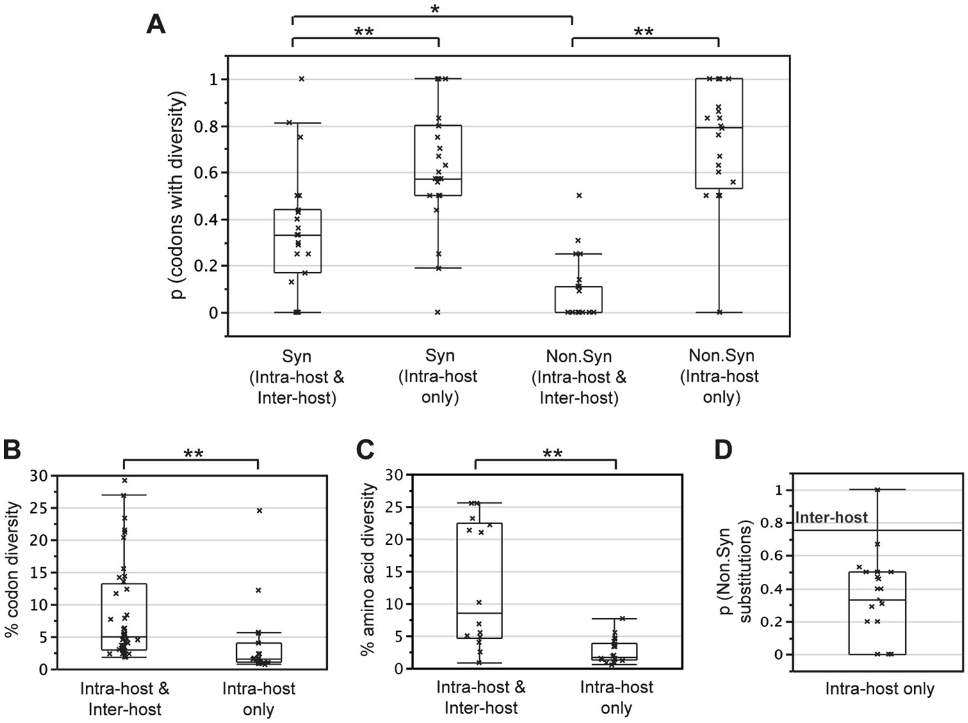

FIG 6 Purifying selection at the intrahost-interhost interface. All parameters were calculated per sample. (A) Proportion (p) of all synonymous (Syn) or nonsynonymous (Non.Syn) mutations that occur in both intrahost and interhost data sets or in the intrahost data set only. *P* values were calculated using the paired *t* test (*, $P < 0.05$; **, $P < 0.005$). (B and C) Relationship between the abundance of intrahost variants and the appearance of these variants in consensus genomes. (B) Percent intrahost codon diversity, categorized by whether the secondary variant is identical (Intrahost & Interhost) or nonidentical (Intrahost only) in intrahost and interhost data sets. Data are shown for loci that demonstrate synonymous mutations only. Differences are significant at $P < 0.0001$ (**; one-way chi-square test). (C) Percent intrahost amino acid diversity, categorized by whether the intrahost mutation is also found in interhost consensus genomes (Intrahost & Interhost) or is distinct from variation observed at the interhost level (Intrahost only). Data are shown for loci that can tolerate nonsynonymous mutations. Differences are significant at $P = 0.0002$ (**; one-way chi-square test). (D) Proportion (p) of predicted tolerated nonsynonymous substitutions relative to all nonsynonymous substitutions in intrahost data sets. The gray reference line indicates the proportion of predicted tolerated mutations at the consensus level in 161 Nicaraguan isolates.

our samples. As an alternative, we calculated the maximum proportion of variant genomes relative to consensus genomes by dividing the number of reads with variants by the total number of sequenced reads and estimated that in our samples, on average ≤2.4% (range, 0.25 to 13%) of the virus population was composed of variant genomes.

The genome-wide nature of our survey allowed us to identify rare variants, such as deletions and early termination STOP mutations, which were present at very low frequencies in human infections but were found in multiple individuals. The detection of such deletions in multiple samples and in multiple epidemics is suggestive either of cocirculation and/or maintenance with full-length genomes through epidemic seasons or of frequent genesis during viral replication in human and/or mosquito hosts (1, 41).

Complete coverage across the coding region also facilitated detection of nonuniform patterns of intrahost diversity across genes. NS5 and E appeared to exhibit the highest average diversity, suggesting either that these genes tolerated mutations well (indicative of neutral evolution) or that they are rapidly evolving. In contrast, the least amount of intrahost diversity was observed in NS2B, possibly as a consequence of strong selective pressures on this gene. Interestingly, differential diversity was also observed at the intragene level in E, with the highest diversity seen in EDII, which appears to be highly immunogenic in humans (13, 36, 48, 73), while the lowest diversity was observed in EDIII, which is poorly immunogenic in humans (13, 36, 48, 73). Notably, the extent of diversity in EDIII was similar to that observed in the C-terminal hydrophobic domain of E, which is known to be poorly immunogenic. The higher diversity in EDII observed in our samples thus could be reflective of diversity generated in response to immune-driven pressures. In contrast, the limited diversity in EDIII and the C-terminal domain may be due to lack of immune-driven pressure. This suggests an association between immunogenicity and viral intrahost evolution and implies the existence of immune-driven viral evolution even during the short infection periods associated with acute DENV infections. Notably, none of these gene- and domain-specific trends were observed in the consensus genome data sets, implying distinct patterns of selection at the intrahost and interhost levels.

Gene-specific differences in intrahost diversity were also observed between Nicaraguan clades, raising questions about possible relationships between intrahost diversity, inherent clade genetics, bottlenecks, and immune-driven selection. NI-1 and NI-2B viruses vary at nine amino acid-encoding loci (54), and these genetic differences may contribute to the higher diversity in NI-1 viruses than NI-2B viruses. For instance, differential consensus amino acid encoding in the RdRp region (amino acids NS5 to 200, NS5 to 290, and NS5 to 401 [54]) may influence error rates of the NS5 RdRp, causing distinct diversity profiles in NI-1 and NI-2B viruses. Alternatively, NI-2B viruses may be subject to more severe bottlenecks and/or selective pressures either within the host or during transmission, resulting in reduced diversity in these viruses. Yet another possibility is that the specificity of memory B and T cell responses dictated by prior infections in humans influence viral intrahost evolution. Since all study subjects had prior infections with a DENV serotype distinct from DENV-2, it is conceivable that preexisting immunity in these individuals exerted distinct selective pressures on NI-1 and NI-2B viruses and differentially influenced the intrahost evolution of these viruses. Indeed, a complex relationship between clade and preexisting immunity has already been suggested as a factor in determining disease severity (54). Further investigation is required to understand if the differences in intrahost diversity contribute to the differences in phenotypes between clades reported by OhAinle et al. (54).

Interestingly, we observed no correlation between disease severity and genome-wide diversity, suggesting that viral diversity sampled during disease manifestation does not correlate with disease outcome, although the limited number of samples in our data set warrants further investigation of associations between viral diversity and disease outcome. A similar lack of correlation between intrahost diversity in EDIII and disease severity was observed by Thai et al. (67) but contrasts with what was reported by Descloux et al. (15). A hypothesis that remains to be tested is whether disease severity is associated with population diversity inherent in the infecting viral population rather than with viral diversity at later time points after symptom onset (~1 to 2 weeks after infection). Indeed, in the three subjects in whom we analyzed DENV intrahost diversity at multiple time points, we observed substantial differences in the extent and spectrum of intrahost diversity even for small time scales (1 to 2 days; data not shown), indicating that substantial differences exist in intrahost diversity between infecting viral populations and viral populations at or after symptom onset.

We also identified fundamental relationships between intrahost diversity and interhost diversity. Extremely low proportions of intrahost variants were found in consensus genomes isolated from the circulating DENV-2 populations in Nicaragua (i.e., interhost diversity). This may be attributed to transmission bottlenecks, with contributions from purifying selection as suggested in previous studies (29, 43). There also appears to be a direct relationship between the level of intrahost diversity and the appearance of mutations in the interhost population (or vice versa), indicative of selection against nonabundant intrahost variants. In addition, a strong effect of purifying selection is evidenced by the discrepancy in the proportion of mutations that are predicted to affect protein function between intrahost and interhost data sets. We also discerned selective pressures that were possibly unique to viruses circulating in Nicaragua; certain loci with intrahost diversity show no diversity at the interhost level in Nicaraguan populations but appear to show considerable diversity in DENV-2 populations outside Nicaragua. These observations illustrate how diverse selection pressures, perhaps along with transmission bottlenecks, act together to shape evolutionary dynamics across the intrahost-interhost interface. Furthermore, our exploration of the relationship between clade-specific intrahost diversity and consensus changes that differentiate the two clades allowed us to determine that Nicaraguan clades were not evolving at loci that distinguished them from their nearest ancestor or other closely related clades. We also note that a paucity of intrahost diversity at Nicaraguan clade-distinguishing loci suggests the absence (or the complete resolution) of mixed infections in these individuals.

Our study demonstrates gene-wise and genome-wide trends in intrahost diversity and determines how these trends track with viral genetics and interhost diversity. These observations have direct implications for viral evolution in DENV as well as other RNA viruses and underscore the value of using a sequencing approach that captures large regions of viral genomes with considerable depth of coverage. Indeed, such methodological and analytical approaches will prime additional studies that are critical for further high-resolution mapping of the relationship between intrahost diversity and clinical, epidemiological, immunological, and virological parameters.

ACKNOWLEDGMENTS

We thank the Broad Institute's Sequencing and Biological Samples Repository Platforms and L. Green, R. Erlich, J. Levin, R. Newman, L. Larson, F. Viloria, and T. Poon for their assistance with sample handling and sequencing; B. Weiner for input on analysis; and M. OhAinle and K. Williams (Harris laboratory, UC Berkeley), S. Bennett (California Academy of Sciences), E. C. Holmes (The Pennsylvania State University), A. Fire (Stanford University), and B. W. Birren (Broad Institute) for helpful discussions.

This project was supported by federal funds from the National Institutes of Health, Department of Health and Human Services, from the National Institute of Medical Sciences under grant GM087405 (E.H.) and from the National Institute of Allergy and Infectious Diseases under grants AI65359 (A.B.), AI62100 (E.H. and A.B.), and U54AI065359 (PSWRCE Career Development subaward to P.P.) and contracts HHSN272200900018C (awarded to B. W. Birren) and HHSN272200900006C (awarded to B. W. Birren), as well as grant VE-1 (EH) from the Pediatric Dengue Vaccine Initiative.

M. R. Henn, E. Harris, P. Parameswaran, A. Balmaseda, and P. Charlebois conceived and designed the experiments. Y. Tellez, A. Nunez, E. M. Ryan, C. M. Malboeuf, J. Z. Levin, and N. J. Lennon performed the experiments. P. Parameswaran, P. Charlebois, M. R. Henn, and E. Harris analyzed the data. M. R. Henn, E. Harris, N. J. Lennon, A. Balmaseda, P. Charlebois, and P. Parameswaran contributed materials/reagents/analysis tools. P. Parameswaran, M. R. Henn, and E. Harris wrote the manuscript.

We declare no competing financial, personal or professional conflicts of interest.

REFERENCES

1. **Aaskov J, Buzacott K, Thu HM, Lowry K, Holmes EC.** 2006. Long-term transmission of defective RNA viruses in humans and Aedes mosquitoes. Science **311**:236–238.
2. **Aviles W, Ortega O, Kuan G, Coloma J, Harris E.** 2007. Integration of information technologies in clinical studies in Nicaragua. PLoS Med. **4**:1578–1583. doi:10.1371/journal.pntd.0000757.
3. **Balmaseda A, et al.** 2003. Diagnosis of dengue virus infection by detection of specific immunoglobulin M (IgM) and IgA antibodies in serum and saliva. Clin. Diagn. Lab. Immunol. **10**:317–322.
4. **Balmaseda A, et al.** 2006. High seroprevalence of antibodies against dengue virus in a prospective study of schoolchildren in Managua, Nicaragua. Trop. Med. Int. Health **11**:935–942.
5. **Balmaseda A, Sandoval E, Perez L, Gutierrez CM, Harris E.** 1999. Application of molecular typing techniques in the 1998 dengue epidemic in Nicaragua. Am. J. Trop. Med. Hyg. **61**:893–897.
6. **Balmaseda A, et al.** 2010. Trends in patterns of dengue transmission over 4 years in a pediatric cohort study in Nicaragua. J. Infect. Dis. **201**:5–14.
7. **Bennett SN, et al.** 2006. Molecular evolution of dengue 2 virus in Puerto Rico: positive selection in the viral envelope accompanies clade reintroduction. J. Gen. Virol. **87**:885–893.
8. **Bhardwaj S, Holbrook M, Shope RE, Barrett AD, Watowich SJ.** 2001.

Biophysical characterization and vector-specific antagonist activity of domain III of the tick-borne flavivirus envelope protein. J. Virol. **75**:4002–4007.

9. **Brockman W, et al.** 2008. Quality scores and SNP detection in sequencing-by-synthesis systems. Genome Res. **18**:763–770.
10. **Chao DY, et al.** 2005. Strategically examining the full-genome of dengue virus type 3 in clinical isolates reveals its mutation spectra. Virol. J. **2**:72.
11. **Colbert JA, et al.** 2007. Ultrasound measurement of gallbladder wall thickening as a diagnostic test and prognostic indicator for severe dengue in pediatric patients. Pediatr. Infect. Dis. J. **26**:850–852.
12. **Crill WD, CG.** 2004. Localization and characterization of flavivirus envelope glycoprotein cross-reactive epitopes. J. Virol. **78**:13975–13986.
13. **Crill WD, Delorey HHMJ, Chang GJ.** 2009. Humoral immune responses of dengue fever patients using epitope-specific serotype-2 virus-like particle antigens. PLoS One **4**:e4991. doi:10.1371/journal.pone.0004991.
14. **Crill WD, RJ.** 2001. Monoclonal antibodies that bind to domain III of dengue virus E glycoprotein are the most efficient blockers of virus adsorption to Vero cells. J. Virol. **75**:7769–7773.
15. **Descloux E, Cao-Lormeau VM, Roche C, De Lamballerie X.** 2009. Dengue 1 diversity and microevolution, French Polynesia 2001–2006: connection with epidemiology and clinics. PLoS Negl. Trop. Dis. **3**:e493. doi:10.1371/journal.pntd.0000493.
16. **Domingo E, Holland JJ.** 1997. RNA virus mutations and fitness for survival. Annu. Rev. Microbiol. **51**:151–178.
17. **Edgar RC.** 2004. MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res. **32**:1792–1797.
18. **Farci P, et al.** 2002. Early changes in hepatitis C viral quasispecies during interferon therapy predict the therapeutic outcome. Proc. Natl. Acad. Sci. U. S. A. **99**:3081–3086.
19. **Fernandez RJ, Vazquez S.** 1990. Serological diagnosis of dengue by an ELISA inhibition method (EIM). Mem. Inst. Oswaldo Cruz **85**:347–351.
20. **Ganeshan S, Dickover RE, Korber BT, Bryson YJ, Wolinsky SM.** 1997. Human immunodeficiency virus type 1 genetic evolution in children with different rates of development of disease. J. Virol. **71**:663–677.
21. **Gromowski GD.** 2007. Characterization of an antigenic site that contains a dominant, type-specific neutralization determinant on the envelope protein domain III (ED3) of dengue 2 virus. Virology **366**:349–360.
22. **Gubler DJ, Kuno G, Sather GE, Velez M, Oliver A.** 1984. Mosquito cell cultures and specific monoclonal antibodies in surveillance for dengue viruses. Am. J. Trop. Med. Hyg. **33**:158–165.
23. **Guzman MG, et al.** 1990. Dengue hemorrhagic fever in Cuba, 1981: a retrospective seroepidemiologic study. Am. J. Trop. Med. Hyg. **42**:179–184.
24. **Halstead SB.** 2007. Dengue. Lancet **370**:1644–1652.
25. **Halstead SB, Suaya JA, Shepard DS.** 2007. The burden of dengue infection. Lancet **369**:1410–1411.
26. **Hammond SN, et al.** 2005. Differences in dengue severity in infants, children, and adults in a 3-year hospital-based study in Nicaragua. Am. J. Trop. Med. Hyg. **73**:1063–1070.
27. **Harris E, et al.** 2000. Clinical, epidemiologic, and virologic features of dengue in the 1998 epidemic in Nicaragua. Am. J. Trop. Med. Hyg. **63**: 5–11.
28. **Henn MR, et al.** 2012. Whole genome deep sequencing of HIV-1 reveals the impact of early minor variants upon immune recognition during acute infection. PLoS Pathog. **8**:e1002529. doi:10.1371/journal.ppat.1002529.
29. **Holmes EC.** 2003. Patterns of intra- and interhost nonsynonymous variation reveal strong purifying selection in dengue virus. J. Virol. **77**:11296–11298.
30. **Holmes EC, Twiddy SS.** 2003. The origin, emergence and evolutionary genetics of dengue virus. Infect. Genet. Evol. **3**:19–28.
31. **Jerzak G, Bernard KA, Kramer LD, Ebel GD.** 2005. Genetic variation in West Nile virus from naturally infected mosquitoes and birds suggests quasispecies structure and strong purifying selection. J. Gen. Virol. **86**: 2175–2183.
32. **Joos B, et al.** 2005. Low human immunodeficiency virus envelope diversity correlates with low in vitro replication capacity and predicts spontaneous control of plasma viremia after treatment interruptions. J. Virol. **79**:9026–9037.
33. **Kuan G, et al.** 2009. The Nicaraguan pediatric dengue cohort study: study design, methods, use of information technology, and extension to other infectious diseases. Am. J. Epidemiol. **170**:120–129.
34. **Kumar P, Henikoff S, Ng PC.** 2009. Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm. Nat. Protoc. **4**:1073–1081.
35. **LaFleur C, et al.** 2002. HLA-DR antigen frequencies in Mexican patients with dengue virus infection: HLA-DR4 as a possible genetic resistance factor for dengue hemorrhagic fever. Hum. Immunol. **63**:1039–1044.
36. **Lai CY, et al.** 2008. Antibodies to envelope glycoprotein of dengue virus during the natural course of infection are predominantly cross-reactive and recognize epitopes containing highly conserved residues at the fusion loop of domain II. J. Virol. **82**:6631–6643.
37. **Lanciotti RS, Calisher CH, Gubler DJ, Chang GJ, Vorndam AV.** 1992. Rapid detection and typing of dengue viruses from clinical samples by using reverse transcriptase-polymerase chain reaction. J. Clin. Microbiol. **30**:545–551.
38. **Lee HY, Perelson AS, Park SC, Leitner T.** 2008. Dynamic correlation between intrahost HIV-1 quasispecies evolution and disease progression. PLoS Comput. Biol. **4**:e1000240. doi:10.1371/journal.pcbi.1000240.
39. **Leitmeyer et al.** 1999. Dengue virus structural differences that correlate with pathogenesis. J. Virol. **73**:4738–4747.
40. **Lennon NJ, et al.** 2010. A scalable, fully automated process for construction of sequence-ready barcoded libraries for 454. Genome Biol. **11**:R15. doi:10.1186/gb-2010-11-2-r15.
41. **Li D, et al.** 2011. Defective interfering viral particles in acute dengue infections. PLoS One **6**:e19447. doi:10.1371/journal.pone.0019447.
42. **Lin B, Murray PCJM, Wright PJ.** 1994. Localization of a neutralizing epitope on the envelope protein of dengue virus type 2. Virology **202**:885–890.
43. **Lin SR, et al.** 2004. Study of sequence variation of dengue type 3 virus in naturally infected mosquitoes and human hosts: implications for transmission and evolution. J. Virol. **78**:12717–12721.
44. **Lok SM, et al.** 2008. Binding of a neutralizing antibody to dengue virus alters the arrangement of surface glycoproteins. Nat. Struct. Mol. Biol. **15**:312–317.
45. **Macalalad AR, et al.** 2012. Highly sensitive and specific detection of rare variants in mixed viral populations from massively parallel sequence data. PLoS Comput. Biol. **8**:e1002417. doi:10.1371/journal.pcbi.1002417.
46. **Margulies M, et al.** 2005. Genome sequencing in microfabricated high-density picolitre reactors. Nature **437**:376–380.
47. **Messer WB, Harris GDE, Sivananthan K, de Silva AM.** 2003. Emergence and global spread of a dengue serotype 3, subtype III virus. Emerg. Infect. Dis. **9**:800–809.
48. **Midgley CM, et al.** 2011. An in-depth analysis of original antigenic sin in dengue virus infection. J. Virol. **85**:410–421.
49. **Modis Y, Ogata S, Clements D, Harrison SC.** 2003. A ligand-binding pocket in the dengue virus envelope glycoprotein. Proc. Natl. Acad. Sci. U. S. A. **100**:6986–6991.
50. **Modis Y, Ogata S, Clements D, Harrison SC.** 2004. Structure of the dengue virus envelope protein after membrane fusion. Nature **427**:313–319.
51. **Moreau I, Levis J, Crosbie O, Kenny-Walsh E, Fanning LJ.** 2008. Correlation between pre-treatment quasispecies complexity and treatment outcome in chronic HCV genotype 3a. Virol. J. **5**:78.
52. **Murcia PR, et al.** 2010. Intra- and interhost evolutionary dynamics of equine influenza virus. J. Virol. **84**:6943–6954.
53. **Nguyen TP, et al.** 2008. Protective and enhancing HLA alleles, HLA-DRB1*0901 and HLA-A*24, for severe forms of dengue virus infection, dengue hemorrhagic fever and dengue shock syndrome. PLoS Negl. Trop. Dis. **2**:e304. doi:10.1371/journal.pntd.0000304.
54. **OhAinle M, et al.** 2011. Dynamics of dengue disease severity determined by the interplay between viral genetics and serotype-specific immunity. Sci. Transl. Med. **3**:114ra128. doi:10.1126/scitranslmed.3003084.
55. **Oliphant T, et al.** 2006. Antibody recognition and neutralization determinants on domains I and II of West Nile Virus envelope protein. J. Virol. **80**:12149–12159.
56. **Potts JA, Rothman AL.** 2008. Clinical and laboratory features that distinguish dengue from other febrile illnesses in endemic populations. Trop. Med. Int. Health **13**:1328–1340.
57. **Prestwood TR, Sharar PDKL, Zellweger RM, Shresta S.** 2008. A mouse-passaged dengue virus strain with reduced affinity for heparan sulfate causes severe disease in mice by establishing increased systemic viral loads. J. Virol. **82**:8411–8421.
58. **Restrepo BN, et al.** 2008. Serum levels of cytokines in two ethnic groups with dengue virus infection. Am. J. Trop. Med. Hyg. **79**:673–677.
59. **Rey FA.** 2003. Dengue virus envelope glycoprotein structure: new insight

into its interactions during viral entry. Proc. Natl. Acad. Sci. U. S. A. **100**:6899–6901.

60. **Rey FA, Heinz FX, Mandl C, Kunz C, Harrison SC.** 1995. The envelope glycoprotein from tick-borne encephalitis virus at 2 A resolution. Nature **375**:291–298.
61. **Rosen L.** 1977. The emperor's new clothes revisited, or reflections on the pathogenesis of dengue hemorrhagic fever. Am. J. Trop. Med. Hyg. **26**: 337–343.
62. **Shrestha B, et al.** 2010. The development of therapeutic antibodies that neutralize homologous and heterologous genotypes of dengue virus type 1. PLoS Pathog. **6**:e1000823. doi:10.1371/journal.ppat.1000823.
63. **Sironen T, Kallio ER, Vaheri A, Lundkvist A, Plyusnin A.** 2008. Quasispecies dynamics and fixation of a synonymous mutation in hantavirus transmission. J. Gen. Virol. **89**:1309–1313.
64. **Sukupolvi-Petty S, et al.** 2010. Structure and function analysis of therapeutic monoclonal antibodies against dengue virus type 2. J. Virol. **84**: 9227–9239.
65. **Sukupolvi-Petty S, et al.** 2007. Type- and subcomplex-specific neutralizing antibodies against domain III of dengue virus type 2 envelope protein recognize adjacent epitopes. J. Virol. **81**:12816–12826.
66. **Sullivan DG, et al.** 2007. Hepatitis C virus dynamics during natural infection are associated with long-term histological outcome of chronic hepatitis C disease. J. Infect. Dis. **196**:239–248.
67. **Thai KT, et al.** 2012. High-resolution analysis of intrahost genetic diversity in dengue virus serotype 1 infection identifies mixed infections. J. Virol. **86**:835–843.
68. **Vasilakis N, Weaver SC.** 2008. The history and evolution of human dengue emergence. Adv. Virus Res. **72**:1–76.
69. **Vaughn DW, et al.** 2000. Dengue viremia titer, antibody response pattern, and virus serotype correlate with disease severity. J. Infect. Dis. **181**: 2–9.
70. **Vignuzzi M, Stone JK, Arnold JJ, Cameron CE, Andino R.** 2006. Quasispecies diversity determines pathogenesis through cooperative interactions in a viral population. Nature **439**:344–348.
71. **Volk DE, et al.** 2004. Solution structure and antibody binding studies of the envelope protein domain III from the New York strain of West Nile virus. J. Biol. Chem. **279**:38755–38761.
72. **Vu TT, et al.** 2010. Emergence of the Asian 1 genotype of dengue virus serotype 2 in Vietnam: in vivo fitness advantage and lineage replacement in south-east Asia. PLoS Negl. Trop. Dis. **4**:e757. doi:10.1371/journal.pntd.0000757.
73. **Wahala WM, Kraus AA, Haymore LB, Accavitti-Loper MA, de Silva AM.** 2009. Dengue virus neutralization by human immune sera: role of envelope protein domain III-reactive antibody. Virology **392**:103–113.
74. **Wang WK, Lin SR, Lee CM, King CC, Chang SC.** 2002. Dengue type 3 virus in plasma is a population of closely related genomes: quasispecies. J. Virol. **76**:4662–4665.
75. **Wang WK, Sung TL, Lee CN, Lin TY, King CC.** 2002. Sequence diversity of the capsid gene and the nonstructural gene NS2B of dengue-3 virus in vivo. Virology **303**:181–191.
76. **WHO.** 1997. Dengue haemorrhagic fever: diagnosis, treatment, prevention and control, 2nd ed. World Health Organization, Geneva, Switzerland.
77. **Yu Q, et al.** 2011. PriSM: a primer selection and matching tool for amplification and sequencing of viral genomes. Bioinformatics **27**:266–267.
78. **Yu S, et al.** 2004. Solution structure and structural dynamics of envelope protein domain III of mosquito- and tick-borne flaviviruses. Biochemistry **43**:9168–9176.