Grubaugh et al. Genome Biology (2019) 20:8
https://doi.org/10.1186/s13059-018-1618-7

Genome Biology

METHOD Open Access

An amplicon-based sequencing framework for accurately measuring intrahost virus diversity using PrimalSeq and iVar

Nathan D. Grubaugh[1,2*], Karthik Gangavarapu[1*], Joshua Quick[3], Nathaniel L. Matteson[1], Jaqueline Goes De Jesus[3,4], Bradley J. Main[5], Amanda L. Tan[6], Lauren M. Paul[6], Doug E. Brackney[7], Saran Grewal[8], Nikos Gurfield[8], Koen K. A. Van Rompay[9], Sharon Isern[6], Scott F. Michael[6], Lark L. Coffey[5], Nicholas J. Loman[3] and Kristian G. Andersen[1,10]

Abstract

How viruses evolve within hosts can dictate infection outcomes; however, reconstructing this process is challenging. We evaluate our multiplexed amplicon approach, PrimalSeq, to demonstrate how virus concentration, sequencing coverage, primer mismatches, and replicates influence the accuracy of measuring intrahost virus diversity. We develop an experimental protocol and computational tool, iVar, for using PrimalSeq to measure virus diversity using Illumina and compare the results to Oxford Nanopore sequencing. We demonstrate the utility of PrimalSeq by measuring Zika and West Nile virus diversity from varied sample types and show that the accumulation of genetic diversity is influenced by experimental and biological systems.

Keywords: Viral sequencing, Amplicon sequencing, Intrahost evolution, Zika, West Nile, SNP calling

Background

RNA viruses, including HIV, influenza, West Nile, and Zika, pose significant threats to public health worldwide. Part of this burden stems from their ability to rapidly evolve within hosts [1]. Generation of intrahost genetic diversity allows virus populations to evade host immune responses [2–4], alter the severity of disease [5], and adapt to changing environments [6, 7]. Studying virus populations, both within naturally infected hosts and during experimental evolution, can therefore lead to breakthroughs in our understanding of virus-host interactions and novel approaches for outbreak response [8–11].

In many cases, however, accurately measuring intrahost RNA virus diversity using deep sequencing remains a significant challenge. Multiple factors, such as virus titer, sample preparation, sequencing errors, and computational inferences, can bias measures of genetic diversity [12–16]. Moreover, for many clinical samples, low ratios of viral to host RNA often necessitate enrichment of viral nucleic acid to recover sufficient templates for deep sequencing [17]. This is especially true for Zika virus, where low viremias (< 1000 copies/μL of RNA) are often detected during natural and experimental infections [18–21]. PCR amplification of virus nucleic acid is a common approach to overcome this challenge [4, 22, 23], although it can introduce biases by altering the composition of intrahost genetic variants [14, 24]. Therefore, to ensure accuracy, comprehensive validation of deep sequencing approaches should accompany diversity measures from biological samples.

We previously developed a multiplex primer design tool ("Primal Scheme") coupled to a laboratory protocol ("PrimalSeq") to sequence RNA viruses directly from clinical samples in a way that is cheap, accurate, and scalable under resource-limited conditions [17]. Versions of PrimalSeq have been used to sequence the majority of Zika virus genomes from the epidemic in the Americas [19, 25–27], yellow fever virus in Brazil [28], and West Nile virus in the USA [29]. PrimalSeq has also been used to characterize Zika virus during infection of

* Correspondence: nathan.grubaugh@yale.edu; gkarthik@scripps.edu
Nathan D Grubaugh and Karthik Gangavarapu are co-first authors.
Nicholas J Loman and Kristian G Andersen are co-senior authors.
[1]Department of Immunology and Microbiology, The Scripps Research Institute, La Jolla, CA 92037, USA
Full list of author information is available at the end of the article

non-human primates [21, 30]. While PrimalSeq was shown to be superior to other methods for obtaining consensus sequences [25, 26], it has yet to be validated for measuring intrahost diversity.

In this study, we benchmarked PrimalSeq for sequencing diverse virus populations, highlighting its limitations and providing recommendations for accurately measuring intrahost single-nucleotide variants (iSNVs) from both Illumina and Oxford Nanopore data. We used these results to develop comprehensive laboratory protocols and a computational tool (iVar), and further tested PrimalSeq to characterize Zika virus populations generated from cell culture, mosquito, non-human primate, and human clinical samples. We demonstrate the utility of PrimalSeq for other viruses by designing an amplicon scheme for West Nile virus and measuring genetic diversity from field-collected mosquitoes and birds. Our data show that virus diversity can be significantly impacted by the experimental and biological systems, and we provide a framework to uncover the underlying mechanisms. PrimalSeq and iVar provide a scalable platform for viruses other than Zika and West Nile that can be applied to discover ecological, epidemiological, and immunological drivers of virus evolution in a variety of systems.

Results

Virus concentration and sequencing depth impact intrahost variant calling

We previously developed Primal Scheme (Quick et al.; primal.zibraproject.org), a multiplex primer design tool for amplicon-based sequencing of RNA virus genomes directly from clinical samples [17]. Our Zika virus PrimalSeq protocol generates 35 overlapping amplicons of ~ 400 base pairs from two multiplexed PCR reactions, an approach similar to "RNA jackhammering," which was developed to sequence HIV [31]. The process of PCR amplification to generate sufficient templates for high-throughput sequencing, however, may bias the measurements of intrahost virus diversity through differential amplification efficiencies for divergent virus haplotypes present in a population [32, 33].

Given the potential biases that may be introduced during PCR amplification, we sought to assess the accuracy of PrimalSeq for iSNV detection. Through a series of experiments, we determined that at least 1000 RNA virus copies are needed to detect iSNVs at greater than 3% frequency, when sequenced to a coverage depth of at least 400× (i.e., the number of sequenced nucleotides per targeted genome position, Fig. 1). To set up these experiments, we created genetically diverse Zika virus populations by mixing two divergent virus strains: Zika virus #1 isolated from Puerto Rico in 2015 (Genbank KX087101) and virus #2 isolated from Cambodia in 2010 (Genbank KU955593). Using gold-standard untargeted metagenomic sequencing [34], we determined that there were 159 fixed consensus sequence differences between the viruses located throughout the 10,807 nucleotide genome of virus #1 and #2 (Fig. 1a).

For our initial evaluation, we selected three of the 35 Zika virus primer sets that flanked at least five variable genome positions (amplicons 5, 24, and 33). We then made two sets of mixed virus populations: (1) altering the ratios of mixed viruses, while keeping the overall input concentration constant at 1000 virus RNA copies (Fig. 1b) and (2) maintaining a constant ratio of 14% of virus #2 while altering the input concentrations of virus RNA used for cDNA synthesis (Fig. 1c). For each test, we measured the frequencies of the 18 iSNVs between virus #1 and #2 (Fig. 1a). We generated the amplicons independently three times and sequenced each using the Illumina MiSeq platform ("technical replicates").

We found that the measured mean iSNV frequencies were accurate from populations containing 50%, 25%, 14%, 7%, and 3% of virus #2 (Fig. 1b). At 1.5% of virus #2, the standard deviation of our measured mean iSNV frequency (0.2–1.2%) fell below the expected frequency, indicating that we could not measure the true iSNV frequency at that dilution (Fig. 1b). This demonstrates that the lower limit of accurate iSNV detection for PrimalSeq in this scenario is between 1.5 and 3%. When we altered input concentrations of a population containing 14% virus #2 from 100,000 to 10 virus RNA copies (10-fold serial dilutions), we found that the variances of measured frequencies became significantly higher from concentrations containing 100 or less copies (Levene's test for variance, $p < 0.05$; Fig. 1c). Therefore, input virus concentrations can dramatically alter iSNV detection, as others have also discovered [12]. We conclude that a minimum of 1000 virus RNA copies should be used with PrimalSeq to accurately measure iSNVs greater than 3% frequency.

Sequencing coverage depth is another important factor for iSNV detection [13], so we sought to define the level of sequencing coverage needed to accurately measure iSNVs. From our samples containing 1000 virus RNA copies with 97% virus #1 and 3% virus #2, we sequenced the targeted genome regions in triplicates to a coverage depth of ~ 3000×. We randomly downsampled these datasets to generate coverage depths of 1000, 600, 400, 200, 100, and 50× (Fig. 1d). We found that the variances of iSNV frequencies became significantly higher from coverage depths lower than 400× (i.e., at 200, 100, and 50×; Levene's test for variance, $p < 0.05$; Fig. 1d). Thus, we conclude that a minimum sequencing coverage depth of 400× is required to maintain iSNV measurement accuracy at the lower limit of frequency detection (3%) and input concentration (1000 virus RNA copies).

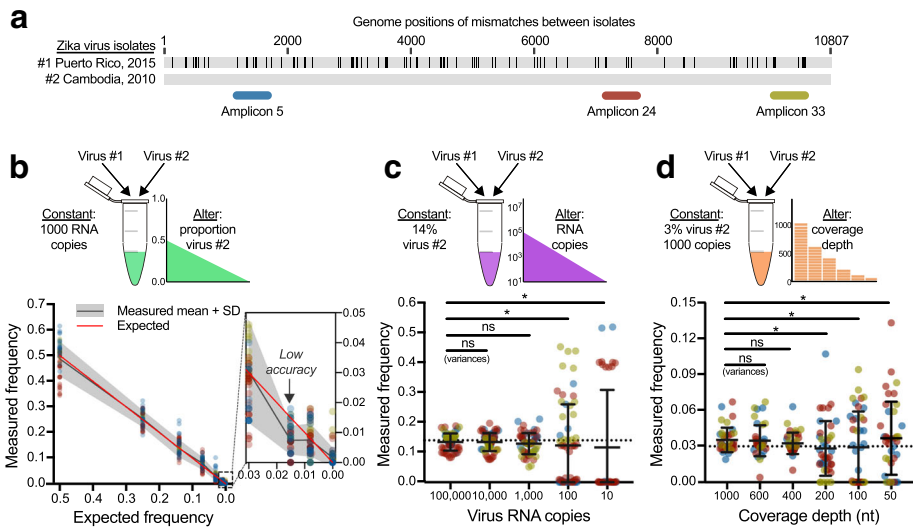

Fig. 1 Measurement intrahost variant frequencies are more accurate at high frequencies and are susceptible to input concentrations and coverage depths. **a** We created genetically diverse virus populations by mixing two Zika virus isolates with 159 consensus nucleotide differences to test the effects of PCR amplification prior to sequencing to measure intrahost single-nucleotide variant (iSNV) frequencies. For these initial experiments, we amplified three ~ 400 bp regions of the Zika virus genome using primers without any mismatches to either of the mixed virus (shown as amplicons 5, 24, and 33). "Amplicon 5" contains 5 iSNV sites, "amplicon 24" contains 8 iSNV sites, and "amplicon 33" contains 5 iSNV sites. **b** We created virus populations containing 50%, 25%, 14%, 7%, 3%, 1.5%, and 0.8% virus #2 to test the impact of PCR amplification prior to sequencing on measuring ranges of iSNV frequencies. The data points represent individual iSNVs amplified and sequenced in triplicate from each population (colored by amplicon 5, 24, or 33 as shown in **a**. **c** We 10-fold serially diluted a mixed population containing 14% of virus #2 (expected, dotted line) from 100,000 to 10 copies to test the effects of input concentrations on accurate iSNV measurements. **d** We randomly downsampled the datasets generated from 1000 input virus RNA copies containing 3% virus #2 to set coverage depths (sequenced nucleotides [nt] per genome position) to determine the minimum coverage needed to yield accurate iSNV measurements. For **c** and **d**, the Levene's test was used to assess equality among variances of iSNV measurements from each coverage depth (ns, not significant; *, $p < 0.05$). Data shown as means with standard deviations

Primer mismatches impact intrahost variant frequency measurements

A concern for using PCR-based sequencing protocols for measuring intrahost virus diversity is the potential for primer mismatches to alter PCR efficiency that can bias iSNV frequency measurements. Indeed, we found that primer mismatches, especially those close to the 3′ end, can alter iSNV frequencies in the generated amplicons (Fig. 2). To make this assessment, we used (1) our Zika virus PrimalSeq strategy and (2) a mix containing 90% of virus #1 and 10% of virus #2 ("Mix10%"). Of the 159 consensus nucleotide differences between virus #1 and #2, 24 resulted in mismatches within the primer regions of 20 of the oligos used to generate 18 of the 35 PCR amplicons (Fig. 2a, Additional file 1: Table S1). In addition, at least one nucleotide difference occurred within each of the 18 amplicons (outside of the primer-binding region), which allowed us to assess the influence of primer mismatches on iSNV frequency measurements across the Zika virus genome.

We amplified the Mix10% virus population (1000 RNA copies) independently three times and sequenced each replicate to a minimum coverage depth of 1000× using the Illumina MiSeq. We measured iSNV frequency at each site and calculated the mean iSNV frequency of all iSNVs within an amplicon to estimate the computed virus #2 haplotype frequency (Fig. 2b). We found that iSNV frequencies measured from amplicons without primer mismatches were significantly closer to the expected value of 10% than amplicons with one or more mismatches in the primer regions (Welch's t test, $p < 0.05$, Fig. 2c). Moreover, we found that mismatches closer to the 3′ end of the primer were more likely to lead to inaccurate frequency measurements (Pearson r, $p < 0.05$, Fig. 2d). Overall, our data demonstrate that the accuracy of intrahost virus diversity measures is highly impacted by primer mismatches during PCR. Thus, when iSNVs are detected from amplicons with mismatches in the primer binding sites, the resulting diversity data from those amplicons should be interpreted with caution.

Removal of false positive intrahost variants with replicate sequencing

Measurements of virus intrahost genetic diversity are sensitive to PCR and sequencing errors [12, 14, 16]. These factors, combined with others such as virus concentration and sequencing coverage (Fig. 1), can lead to erroneous iSNV detection (i.e., false positives) and bias measures of genetic diversity. To improve accurate iSNV detection, we examined the distribution of false positive iSNV calls and

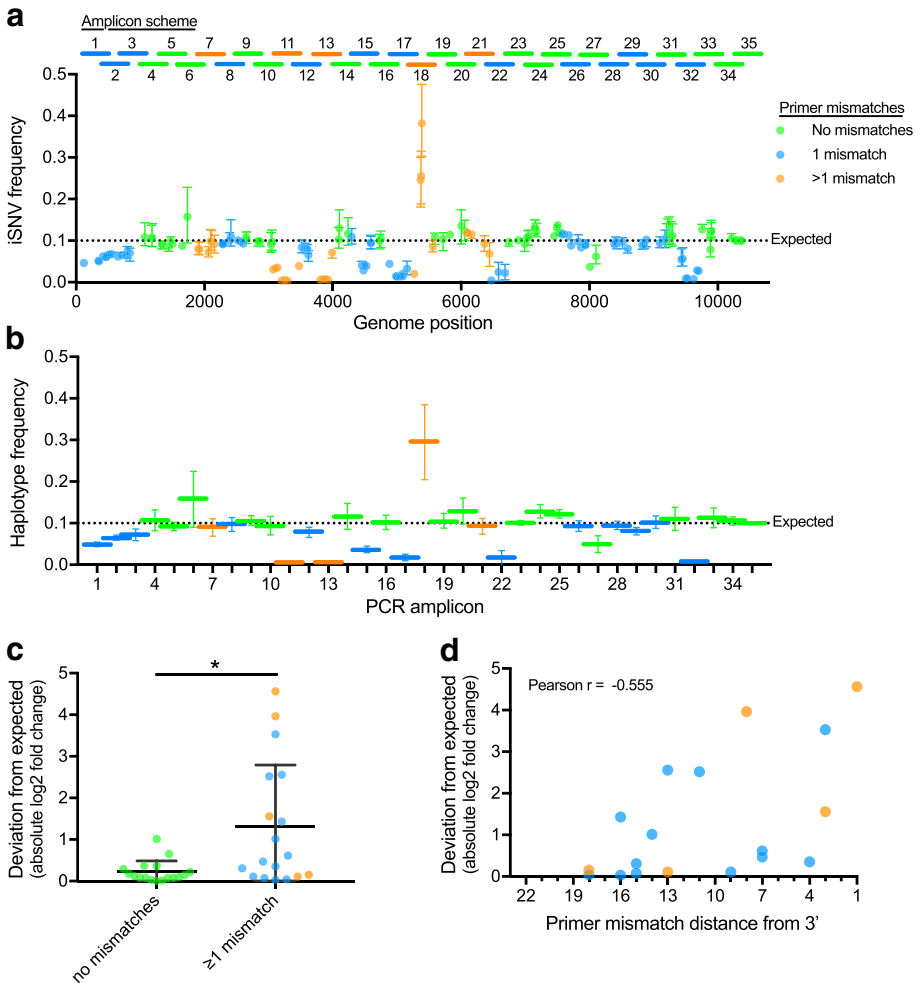

Fig. 2 Measures of intrahost variant frequencies are sensitive to primer mismatches. **a** To assess the impacts of primer mismatches on accurately measuring intrahost single-nucleotide variants (iSNVs), we sequenced a mixed Zika virus population using 35 overlapping PCR amplicons (see "Amplicon scheme" above panel). The virus population contained 10% virus #2 (Expected) and 1000 virus RNA copies were amplified and sequenced in triplicate. The amplicons and iSNVs are colored according to the number of mismatches in the primer sequences used to generate that amplicon. Data shown as means and ranges. **b** To account for unequal iSNV sites within each amplicon, the iSNV frequencies on each amplicon were averaged to produce a haplotype frequency for virus #2 mixed at 10% (Expected). Data shown as means and ranges. **c** We calculated the deviations between the measured and expected virus #2 haplotype frequencies (absolute value of the $\log_2$ fold change) to assess the bias introduced during PCR of amplicons containing primer mismatches to virus #2 (*, Welch's *t* test, $p < 0.05$). Data shown as means and standard deviations. **d** We plotted the deviations from expected haplotype frequencies by the distance of mismatches from the 3′ end of the primer to investigate the impact of mismatch location. If more than one mismatch was present on a primer pair (orange), the data is shown using the closest mismatch to the 3′ end. Mismatches closer to the 3′ end of the primer are more likely decrease the accuracy of iSNV or haplotype measurements from that amplicon (correlation by Pearson *r*, $p < 0.05$). Data shown as the mean from all three replicates

investigated methods to remove them during analysis. We found that (1) the distribution of false positive iSNVs more closely matched the profile of sequencing errors than PCR errors and that (2) the majority of false iSNV > 3% could be removed by replicate sequencing (Fig. 3).

To investigate false positive iSNV calls, we amplified our Mix10% virus population (1000 RNA copies) individually three times and sequenced each on the Illumina MiSeq. We limited our analysis to only those amplicons with perfect primer matches. Within these regions, we analyzed 54 sites with expected 10% iSNVs (true positives) and 4173 sites that were invariable in our mixed virus population. We considered any iSNVs detected > 0.1% frequency at the invariable sites as false positives. We found that on average, 631 of the 4173 expected invariable sites (16%) had false positive iSNVs at a > 0.1% frequency cutoff. We observed false positive iSNVs on every amplicon but found that they were unevenly distributed across the 250 nucleotide long Illumina reads (Fig. 3a, b). Specifically, we found that virus genome sites covered by sequencing reads at positions > 150 nucleotides had significantly more false positives (Wilcoxon test, $p < 0.05$; Fig. 3a insert) at

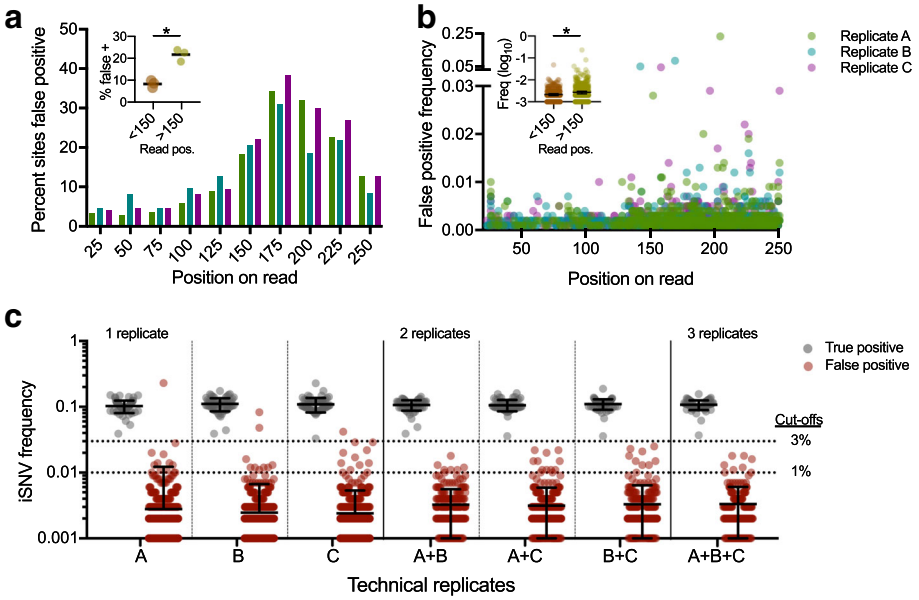

Fig. 3 False positive intrahost variants caused by sequencing errors can be removed by technical replicates and frequency cutoffs. We sequenced mixed Zika virus population containing 10% of virus #2 in triplicate, limited our analysis to the regions only covered by perfect PCR primer matches, and removed sites with intrahost single-nucleotide variant (iSNV) detected at > 1% frequency in either of the Zika virus isolates. This left us with 61 true positive (10% frequency) iSNV sites and 3940 sites not expected to be variable to investigate false positives (> 0.1% frequency). **a** The locations of false positives on the sequencing read position were mapped and shown as the distribution within 25 nt bins by percent of sites with false positive iSNV calls. Each color represents data from an independent replicate. Inset: read positions > 150 nt had a significantly higher false positive rate than positions < 150 nt (*, Wilcoxon test, $p < 0.05$). **b** The iSNV frequencies from each false positive were also plotted by position on the sequencing read. Each color represents data from an independent replicate. Inset: False positive iSNV frequencies were significantly higher at read positions > 150 nt than < 150 nt (*, Mann-Whitney test, $p < 0.05$). **c** True and false iSNVs were plotted by frequency for each individual replicate (A, B, and C) and combined as technical duplicates and triplicates showing the mean frequencies of iSNVs only found in all replicates. Data shown as means and standard deviations. The line indicates the proposed cutoff at 3% based on removing false positives from the replicate data while still in the range of high accuracy (Fig. 1b)

significantly higher frequencies (Mann-Whitney test, $p < 0.05$; Fig. 3b insert) than positions < 150 nucleotides into the read. Our data therefore show that false positive iSNVs are not evenly distributed across the Illumina sequencing reads, and that the profiles are consistent with published Illumina error rates [16]. We found that the occurrence and iSNV frequency of false positives, decreased during the last 50 nucleotides of the Illumina reads (positions 200–250, Fig. 3a, b), which is due to overlapping (not merged) reads during paired-end sequencing of 400 bp amplicons (data not shown).

Knowing the general distribution of false positive iSNVs, we sought to remove them post sequencing. Based on previous investigations [12, 35, 36], we proposed to remove false positive iSNVs by (1) amplifying and sequencing each sample as technical replicates (at least twice) and (2) only calling iSNVs detected in all replicates. Using our Mix10% virus population, we analyzed each replicate in isolation or in combination and calculated the mean iSNV frequencies (Fig. 3c). From individually sequenced replicates, we found 1–2 false positive iSNVs per sample were within the frequency distribution of our true iSNVs (Fig. 3c, panel "1 replicate"), demonstrating that a simple frequency cutoff will either leave false positive or remove true positive iSNVs. When considering replicates in combination, however, we found that the percent of sites with a false positive iSNV call (above 0.1%) dropped from ~ 16% (Fig. 3c "1 replicate") to ~ 9% (Fig. 3c "2 replicates"). More importantly, we found that all of the false iSNVs that passed the duplicate filter had frequencies below the 3% limit of accurate iSNV measurements (Fig. 1b). This allowed us to use a secondary frequency cutoff (3%) to remove the remainder of the false positives, while maintaining all of the true (10%) iSNVs. We found that the addition of a third technical replicate only resulted in a moderate reduction of sites with false iSNVs above 0.1% (9 to 6%) and did not help us to decrease the frequency cutoff filter (Fig. 3c "3 replicates"). Using pseudo replicates (i.e., using the same replicate more than once, instead of using technical replicates) to filter variants, we found that this did not lead to an improvement in eliminating false positives (Additional file 2: Figure S1). This finding shows that the elimination of false positives when using technical replicates is not due to an apparent increase in

sequencing coverage, but rather is due the independent nature of each replicate sequencing library. We conclude that PrimalSeq can be used for accurate iSNV detection above 3% when using at least two technical replicates.

The accuracy of PrimalSeq is comparable to metagenomic sequencing

The gold standard for virus sequencing is untargeted metagenomics—sampling all RNA or DNA present in a sample [34]. Compared to amplicon-based sequencing, metagenomic sequencing uses random priming and does not select for specific RNA sequences and is thus less biased. In addition, for virus sequencing in resource-limited conditions, the Oxford Nanopore MinION is gaining popularity due to its portability and low instrument cost [19, 28, 37], which is enabling real-time outbreak tracking [38]. To compare iSNV calling accuracy across platforms and methods, we evaluated iSNV measurements using either: (1) metagenomic Illumina sequencing, (2) PrimalSeq with Illumina sequencing, and (3) PrimalSeq with Oxford Nanopore sequencing. We found that PCR amplification leads to more variable true iSNV frequency measurements, but that the overall accuracy of PrimalSeq is comparable to metagenomic sequencing (Fig. 4). PrimalSeq using Nanopore sequencing can be used to detect iSNVs, but, as expected, high error rates [39, 40] makes it difficult to differentiate between true and false iSNVs (Fig. 4).

To compare PrimalSeq and metagenomic sequencing approaches, as well as platforms (Illumina or Nanopore), for iSNV measurement accuracy, we generated triplicate sequencing libraries from our Mix10% virus population (1000 RNA copies). The triplicate amplicon libraries were sequenced using both Illumina MiSeq and Oxford Nanopore MinION platforms, while the metagenomics libraries were sequenced using the MiSeq (Fig. 4a). We found that compared to metagenomic sequencing, mean true positive iSNV frequencies generated from PrimalSeq were not significantly different when using the Illumina platform (Welch's t test, $p > 0.05$; Fig. 4b, c). The variances of individual iSNV frequencies, however, were significantly higher using PCR amplification (Levene's test for variance, $p < 0.05$; Fig. 4c). Compared to Illumina sequencing, the mean iSNV frequencies measured using the Oxford Nanopore MinION were not significantly different (Welch's t test, $p > 0.05$; Fig. 4b, c), however, variances of individual iSNV frequencies were significantly higher using Nanopore (Levene's test for variance, $p < 0.05$; Fig. 4c). To determine the discriminatory power of calling true negative and false positive iSNVs, we analyzed iSNV frequencies > 3% (cutoff determined in Fig. 3c) from 4173 sites expected to be invariable (i.e.,

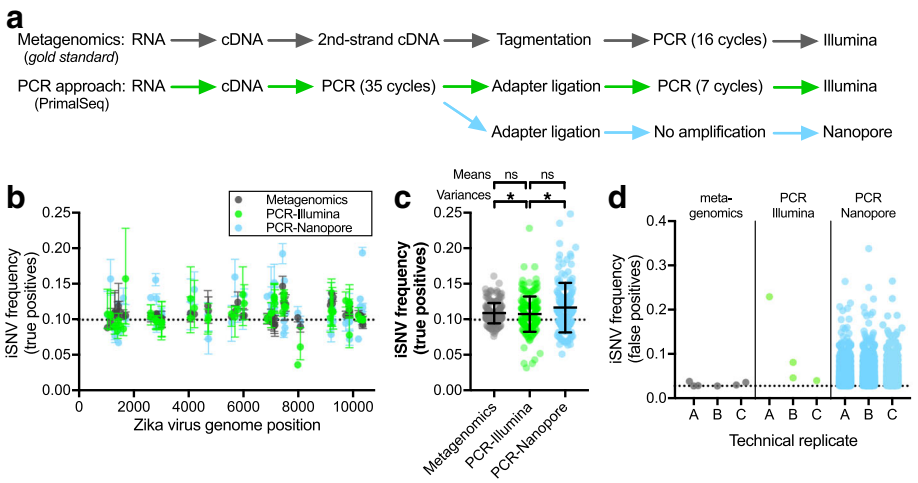

Fig. 4 PCR amplification prior to sequencing leads to similar overall measurements of genetic diversity. **a** We compared our PrimalSeq that enriches for specific virus sequences to the current 'gold standard' for measuring intrahost genetic diversity, metagenomics; and we compared sequencing the amplicons using the Illumina and Oxford Nanopore platforms. The schematic outlines the general workflow for all approaches. **b** We sequenced our mixed Zika virus population (1000 virus RNA copies) containing 10% virus # 2 (Expected) in triplicate using both approaches and platforms to compare the accuracy of measuring known intrahost single-nucleotide variants (iSNVs). We only analyzed regions of the Zika virus #1 and #2 genomes (Fig. 1a) that were perfect matches to the PCR primer sequences, leaving 61 iSNV sites. Data shown as mean and range of triplicate tests. **c** We combined the frequency measurements for each iSNV site and replicate ($n = 183$) to compare the accuracy between the two approaches and platforms. Dashed line shows the expected true iSNV frequencies at 10%. Data shown as means and standard deviations. The mean frequencies were not significantly different (ns, Welch's t test, $p > 0.05$), but the variances were not equal (*, Levene's test, $p < 0.05$). **d** We analyzed the frequency of false positive iSNVs > 3% (cutoff determined in Fig. 1c) from each sequencing method and technical replicate ("A, B, C") from 4173 sites that are expected to be true negatives. From our metagenomics and PCR-Illumina sequencing data, the same false positive iSNVs > 3% frequency are not found in multiple technical replicates, however, many are found in the PCR-Nanopore replicates (see Fig. 5). Dashed line shows the iSNV cutoff at 3%

true negatives). The > 3% false positives using PCR amplification (1–2 per replicate) were similar to metagenomics (1–3 per replicate) when the libraries were sequenced on the Illumina platform (Fig. 4b), and importantly, none of the same > 3% false positives were found in multiple replicates (see also Fig. 3c). More than 700 false positive iSNVs > 3%, however, were detected in each PCR amplified library sequenced on the Nanopore platform (Fig. 5d). These findings show that while PCR amplification leads to more variable results for individual iSNVs, the overall measured diversity and false positive rates are comparable to metagenomic sequencing. Furthermore, we show that iSNV frequencies become even more variable when sequencing using the Oxford Nanopore platform, which is consistent with its higher error rate [39, 40].

High intrahost variant false discovery rate using Oxford Nanopore sequencing

To further explore Oxford Nanopore sequencing for measuring intrahost virus diversity, we examined if we could (1) differentiate between mixed genotypes within a virus population and (2) computationally remove false positive iSNVs. Though the mean iSNV frequencies measured from Nanopore were not significantly different from Illumina (Fig. 4) and we could assign reads to the correct haplotype (i.e., virus #1 or #2), we found it difficult to differentiate between true and false positive iSNVs (Fig. 5). For this evaluation, we used our Nanopore data generated from the Mix10% Zika virus population (Fig. 4a).

First, using a reference database containing virus #1 and #2, plus two divergent Zika viruses (French Polynesia, 2007; Uganda, 1947), we determined if we could differentiate between virus haplotypes in a mixed population. We found that 92.38% of the aligned reads mapped to virus #1 and 7.35% mapped to virus #2, the roughly expected 90%:10% proportions (0.27% of reads mapped erroneously to haplotypes not present in the mixture). Overall, the results indicate that nanopore sequencing reads are useful for identifying highly divergent haplotypes within a mixture—as might be expected for some co-infections [41]—despite a high error rate. This approach, however, will be less useful for detecting co-infections if the divergence between the haplotypes is small or the haplotypes are unknown.

To attempt to differentiate between true and false positive iSNVs, we limited our analysis to regions only covered by perfect primer matches and analyzed 54 true positive and 4173 true negative sites, as we did above for the Illumina data (Fig. 3). We filtered the sequencing data using technical replicates and a 3% frequency cutoff, which we demonstrated above could be used to remove false iSNV calls in our Illumina data (Fig. 3c). Using these filters, we found that > 17% of the 4173 invariant sites had false iSNV calls in the Nanopore data, even when including all three replicates (Fig. 5a). This is

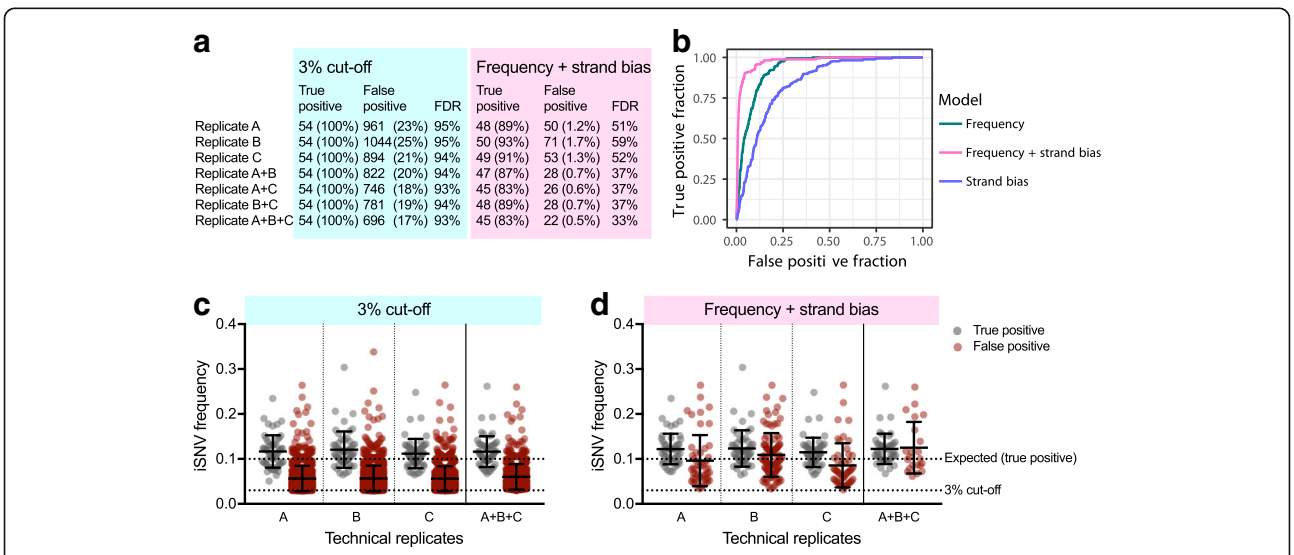

Fig. 5 High false discovery rates of intrahost variants using Nanopore sequencing. **a** iSNV false discovery rates (FDR) from Oxford Nanopore sequencing data. We analyzed 54 true positive and 4173 true negative sites, and determined the proportion of true and false positive iSNV calls from datasets containing 1, 2, or 3 technical replicates using either a 3% frequency cutoff or a logistic regression of iSNV frequency and strand bias. **b** A receiver operator characteristic (ROC) curve showing a logistic regression model that incorporates allele frequency and strand bias as features and the presence or absence of a iSNV as the response variable. The model was trained and tested using a 10-fold cross validation scheme. The model was performed using a frequency and strand bias threshold alone, and combining the two features. Post-filter iSNV frequencies of true and false positive calls using **c** a 3% cutoff or **d** a logistic regression of iSNV frequency and strand bias. Data shown as the means and 95% confidence intervals

because the majority of the false positive iSNVs had measured frequencies as high, or higher, than the 10% true positives (Fig. 5c), leading to a > 93% false discovery rate (Fig. 5a). To investigate if the false discovery rate could be reduced using additional data within the sequencing reads, we trained a logistic regression model incorporating iSNV frequency and strand bias [42] as features, and the presence or absence of known iSNVs as the response variable (Fig. 5b). Based on this analysis, we found that using a frequency and strand bias filter resulted in a higher true or false iSNV discriminatory power, as shown by its greater area under the curve, than the two features independently (Fig. 5b). Using this filter for individual replicates, we were able to reduce the number of false positive iSNVs from ~ 900 to 1000 (3% cutoff) to ~ 50–70 (frequency + strand bias) and the false discovery rate from ~ 95 to ~ 55% (Fig. 5a). By including replicate sequencing (either 2 or 3), we could further reduce the false discovery rate to < 40% (Fig. 5a). Despite this significant reduction, the remaining false positive iSNVs still had high frequencies in our dataset (~ 5–25%, Fig. 5d). It should be noted that because we are comparing two divergent viruses, the false discovery rates will likely increase when sequencing virus populations with fewer true iSNVs; however, applying the frequency and strand bias filter will still provide higher true or false iSNV discriminatory power. These findings show that estimating intrahost virus genetic diversity using the Oxford Nanopore platform will require additional technological and computational innovations for anything other than simple scenarios of co-infections with diverse virus haplotypes.

Accurate analysis of amplicon-based sequencing data using iVar

Using the above validation experiments, we generated a comprehensive experimental protocol (Additional file 3) and an open source computational tool, iVar (*i*ntrahost *v*ariant *a*nalysis of *r*eplicates; github.com/andersen-lab/ivar) to accurately analyze data from amplicon-based sequencing (Fig. 6). Our framework should be compatible with any PCR-based sequencing approach, but was specifically designed for use with PrimalSeq on the Illumina platform. For the experimental protocol, we added the following recommendations to PrimalSeq to measure intrahost virus diversity: (1) start with at least 1000 RNA copies of the virus (Fig. 1c), (2) prepare each sample as technical duplicates (Figs. 3c and 4c), and (3) sequence to a depth of at least 400× (Fig. 1d).

Our computational package, iVar, contains functions broadly useful for viral amplicon-based sequencing that cannot be accomplished using currently existing tools.

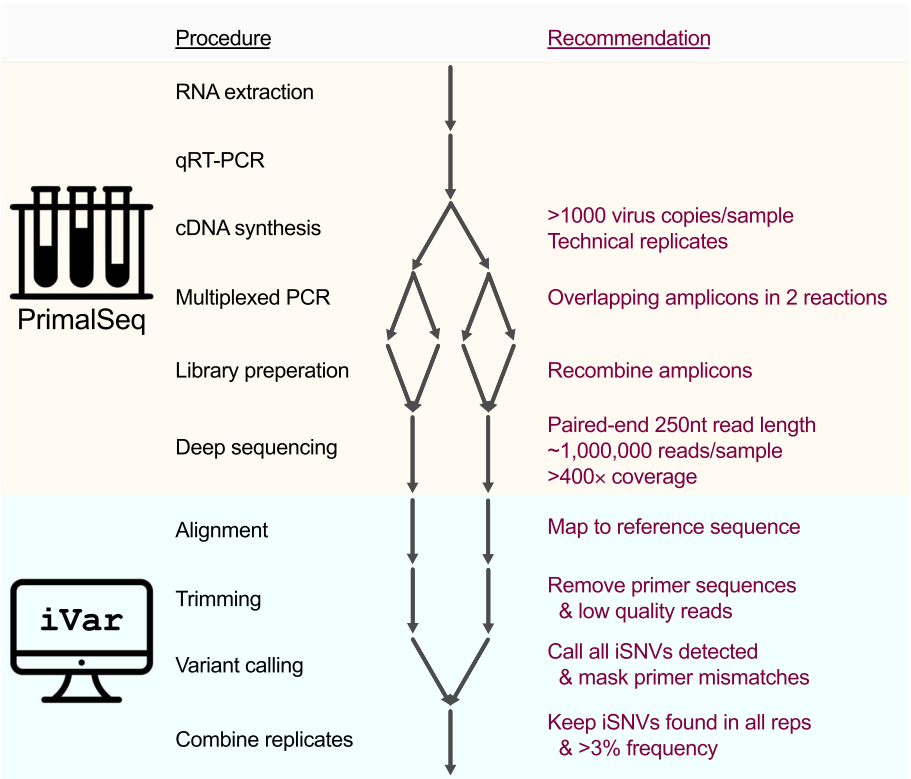

Fig. 6 Experimental and computational workflow for measuring intrahost virus diversity using PrimalSeq

We programmed iVar in C++ with minimal dependencies, and created the following functions to accurately call intrahost variants and generate virus consensus genomes from sequencing data across multiple replicates: (1) trimming of primers and low-quality bases, (2) consensus calling of virus sequences, (3) intrahost variant calling of iSNVs, insertions, and deletions, followed by a filtering step that uses variants called across multiple replicates to exclude false positives, and (4) identification of mismatches in primer sequences and exclusion of the corresponding reads from alignment files. We also created two pipelines using iVar to call intrahost variants from samples with or without known reference sequences, and prepared with or without technical replicates (with no limit on the number of technical replicates). When using iVar in combination with PrimalSeq, from our empirically-derived data we found that the following guidelines produced robust and reproducible results: (1) only call intrahost variants detected in two or more technical replicates greater than 3% frequency (Figs. 1a and 3c) and (2) remove reads from amplicons with mismatched primers to normalize comparisons of intrahost populations (Fig. 2).

We incorporated several functions into iVar for accurate intrahost variant calling that are currently not available in other software packages (Fig. 6). First, iVar removes primer sequences from aligned reads in an input BAM file, based on a BED file with primer positions. This allows iVar to accurately trim primer sequences irrespective of potential mismatches in the aligned region of the sequencing reads and primer sequences. Following the trimming of primer sequences, iVar uses a sliding window approach to remove low quality bases based on phred score thresholds that can be specified by the user. During the trimming process, iVar stores primer sequences that were trimmed off as auxiliary data for each read in each input BAM file. Second, for virus consensus sequence generation, iVar uses the output of mpileup taking into account ambiguous nucleotides and a minimum threshold for base coverage that can be specified by the user. Third, to detect iSNVs, deletions, and insertions, iVar uses the output of mpileup taking into account a minimum threshold for base quality and a minimum threshold for variant frequency. iVar then uses the intrahost variants called across multiple technical replicates to exclude variants that may have been introduced into individual replicates due to amplification, library preparation, and/or sequencing errors (Fig. 3c). Fourth, to identify primer sequences with mismatches, iVar calls variants on an alignment of primer sequences and identifies those with mismatches to the reference. Reads with auxiliary data that matches these identified primers are selectively removed from the alignment. This ensures that varying primer binding efficiency will not bias the frequency of the intrahost variants called with iVar (Fig. 2). Thus, iVar provides an inclusive software package that integrates a set of critical functions for accurate primer and quality trimming, consensus calling, and intrahost variant detection from data generated using amplicon-based sequencing, including PrimalSeq.

We benchmarked iVar against the pre-existing tools VarScan2 [43], MAFFT [44], Geneious [45], Trimmomatic [46], and cutadapt [47] to validate the trimming, consensus sequence generation, and intrahost variant calling functions in iVar. We found that iVar performed as well as, or better than, each of these tools (Additional file 2: Figures S4-S6). We used two simulated datasets and two clinical Zika virus samples sequenced using PrimalSeq to validate iSNV calling, and found an almost perfect correlation between iVar and VarScan2 (Spearman's $\rho = 1$; Additional file 2: Figure S4). We also found zero nucleotide differences in the consensus sequences called using iVar and Geneious at all four thresholds (0%, 25%, 50%, and 90%) and across the four datasets (Additional file 2: Figure S5). We found that iVar was better than cutadapt at trimming primer sequences in amplicon-based sequencing datasets (Additional file 2: Figure S6). This is because iVar uses primer positions specified in a BED file to soft clip the primer regions after alignment, whereas cutadapt trims sequencing reads by comparing the primer nucleotide sequence with the nucleotides at the 5′ end of each read, before alignment. As a result, iVar was able to trim sequencing reads that might not start, or end, exactly at the beginning of the primer sequence (Additional file 2: Figure S6). Since cutadapt uses the actual primer sequences, which are assumed to be anchored at the 5′ or 3′ end to do the trimming, it misses these cases. We trimmed the length of the longest primer sequence (22 bp) from the 5′ end of all the sequenced reads using the "HEADCROP" option in Trimmomatic. This approach, however, is crude and will result in a loss of 22 bp from the 5′ end of all sequenced reads (Additional file 2: Figure S6). Thus, iVar contains functionality that is critical for performing primer and quality trimming, consensus calling, and variant calling from datasets generated using amplicon-based protocols.

PrimalSeq and iVar can be used to measure intrahost virus genetic diversity from primary samples

There are many sequencing options available to measure intrahost virus diversity from cell culture stocks (e.g., [36]) or infected animals with high titers (e.g., [34]). These approaches, however, often do not generate sufficient data when there is high host background RNA, or low virus copies, as is the case for many viruses, including Zika, during human and primate infections [18–21]. Using our

validated PrimalSeq and iVar framework (Fig. 6), and by using Primal Scheme to create a new multiplexed primer set to amplify West Nile virus, we thoroughly evaluated our approaches to measure intrahost diversity from 36 Zika virus and West Nile virus populations (Fig. 7). These samples came from a variety of laboratory and field-derived sample types and were each amplified, sequenced, and analyzed in duplicate (Additional file 1: Table S2, Fig. 7). To account for the influence of virus concentration on our measures of genetic diversity (Fig. 1c), all Zika virus and West Nile virus samples were normalized at 1000 and 10,000 RNA copies, respectively. We omitted (i.e., "masked") from comparative genomic analysis regions of the virus genomes with iSNV mismatches in the primer binding regions (~ 2 per sample) or with sequencing coverage depth < 400× (Additional file 1: Table S2, Fig. 7, gray bands).

To demonstrate the types of analyses that can be performed with PrimalSeq and iVar, we compared the mosquito- and vertebrate-derived virus samples using several measures of intrahost diversity (Fig. 8). We measured genetic richness (the number of iSNV sites; Fig. 8a), complexity (uncertainty associated with randomly sampling an allele; Fig. 8b), and distance (the sum of all iSNV frequencies; Fig. 8c) of iSNVs > 3% frequency. We did not analyze masked regions, so that only high confidence regions of the genome from all samples within the experiment were compared (Fig. 7). We found that Zika virus genetic complexity and distance was significantly higher from populations derived from primate (Hela) cells than *Aedes aegypti* (Aag2) cells (Fig. 8). In vivo, however, our findings were reversed. Zika virus genetic richness and complexity were significantly higher in *Ae. aegypti* bodies than primate (rhesus macaque) plasma (Fig. 8). Furthermore, we found that the distribution of iSNV frequencies of the Zika virus populations was similar across different in vivo infections (Fig. 8d). This finding indicates that the increased Zika virus diversity in mosquitoes was driven by more 3–20% iSNVs that were also common in macaques, and not by a few additional high frequency iSNVs. We found that from both Zika and West Nile virus field samples, genetic diversity was not significantly different between virus populations isolated from their mosquito vectors (*Ae. aegypti* or *Culex*) or vertebrate hosts (humans or birds), though the mosquito samples were more variable (Fig. 8). We also compared measuring intrahost virus genetic diversity using a 3% iSNV cutoff (Fig. 8) to using a 5% iSNV cutoff (Additional file 2: Figure S2) and found the results mostly comparable, though some of the vector-host comparisons were no longer significant. Overall, our data suggest that virus diversity is highly dependent on the experimental and biological systems, and future uses of PrimalSeq will help identify the mechanisms underlying these evolutionary differences (Fig. 8e).

Discussion

Understanding how RNA viruses evolve within hosts can help lead to breakthroughs in medicine and biology. Rapid developments in sequencing technologies are facilitating more research into these areas, yet usage of unvalidated tools and systematic biases can dramatically limit the utility of the results [12–16]. To address these concerns, we developed PrimalSeq [17] and validated it across different platforms and sample types. In our development of iVar we show that PrimalSeq can be used to accurately measure intrahost virus diversity from different viruses and samples with amplicon-based sequencing. Based on our experimental validations, we provide a detailed laboratory protocol (Additional file 3), and suggest the following best practices for measuring intrahost virus diversity when using amplicon-based sequencing approaches:

1. Start with at least 1000 RNA copies of the virus for the initial cDNA synthesis step.
2. Prepare the RNA from virus populations for sequencing in duplicate.
3. Sequence each library to a depth of at least 400× at each genome position using the Illumina platform.
4. Only call iSNVs greater than 3% frequency that are detected in both replicates (a lower frequency may be achievable with higher RNA quantities).
5. For multi-sample comparisons of genetic diversity, omit genome regions amplified with primers that contain iSNVs within the binding sites.

Several factors can alter the accuracy of measuring intrahost virus diversity. In particular, we found that input virus concentrations, sequencing coverage depths, and primer mismatches can have profound effects on iSNV estimations. Using the recommendations above, however, we could consistently and accurately detect iSNVs at 3% frequency and higher. We predict that the lower limit of iSNV detection can be improved with a higher effective sampling depth (i.e., more input virus and deeper coverage) [13].

Given no primer mismatches to the virus sequences, we found that measures of iSNVs frequencies from PrimalSeq were nearly as accurate as an untargeted metagenomics approach [34]. Because iVar remove the primer sequences from downstream analysis and use overlapping amplicons, frequency measures of iSNVs within the primer regions themselves are not skewed. Instead, iSNVs within primer regions can alter the measured frequencies of other iSNVs within that particular amplicon. In these cases, results should be interpreted with caution, and we incorporated a step in iVar to mask out such regions for comparative analyses. It is plausible that using primers with degenerate nucleotides at mismatched iSNV positions could help alleviating this bias [33].

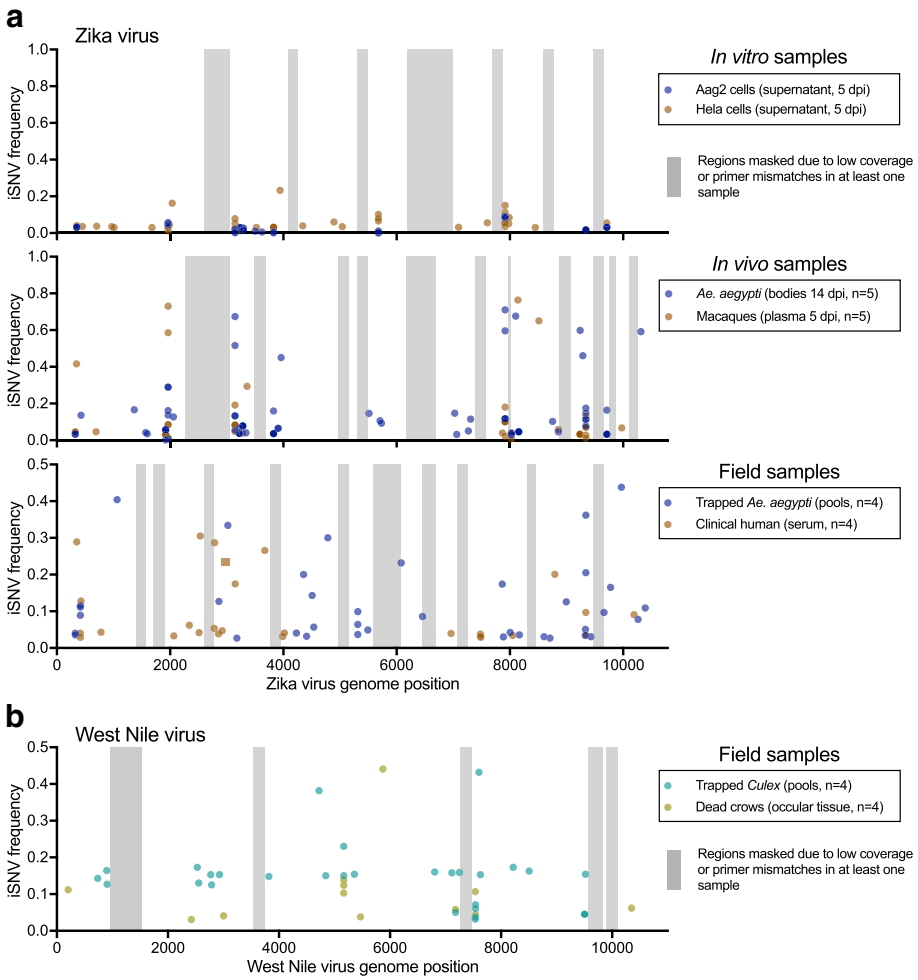

Fig. 7 PrimalSeq can be used to measure intrahost variants from a variety of sample types. **a** We sequenced technical duplicates of Zika virus populations (1000 virus RNA copies each) to identify intrahost single-nucleotide variants (iSNVs) > 3% within each sample. In vitro and in vivo samples were generated using Zika virus strain PRVABC59 (isolated from Puerto Rico, 2015) during infection of *Ae. aegypti* Aag2 cells (derived from embryos), human HeLa cells (derived from cervical epithelial cells), *Ae. aegypti* mosquitoes (orally infected), and Indian origin rhesus macaques (subcutaneously infected). For the in vitro and in vivo samples, where the reference population sequence is known, the iSNV frequencies were calculated by change in frequency from pre- to post-infection. Field Zika virus samples from pooled *Ae. aegypti* and human clinical samples were collected from Florida during the 2016 Zika virus outbreak. **b** *Culex* mosquitoes and dead American crows were collected from San Diego County, CA, during 2015 to sequence West Nile virus from field samples (10,000 virus RNA copies each). The iSNV frequencies from the field samples are the minor allele frequencies (maximum frequency = 0.5) because the reference virus sequence was not known. For both (**a** and **b**), analysis was limited to regions of the genome with > 400× coverage depth in the protein coding sequence and we masked amplicons with primer mismatches from our analysis (gray regions) for direct comparisons of intrahost genetic diversity

False iSNV calls significantly influence measurements of intrahost virus diversity [12]. We found that the positive association of false positive iSNVs with sequencing read lengths better fit the profiles of Illumina sequencing errors, rather than PCR errors [16, 48]. In fact, we estimate that the Illumina MiSeq error rate (~ 0.9% [16]) is ~ 60× greater than the error rates during PCR in our approach (~ 0.02% [49]). Therefore, PrimalSeq likely does not add significantly more error, and by extension false iSNVs, than what was already inherent to the Illumina sequencing platform. Indeed, we found that PrimalSeq was comparable to PCR-free metagenomic sequencing in estimating intrahost virus diversity.

The ease and portability of Oxford Nanopore technologies, particularly the MinION, are revolutionizing the way we sequence viruses, including its use in near real-time outbreak tracking [19, 28, 37]. Our data indicate, however, that the Nanopore platform is not yet adequate for detection of minor alleles and measures of intrahost diversity. While it may provide value in tracking frequency changes of known iSNVs over time, we found that the high error rates (10–15% [39, 40]) makes it difficult to differentiate between true and false iSNV

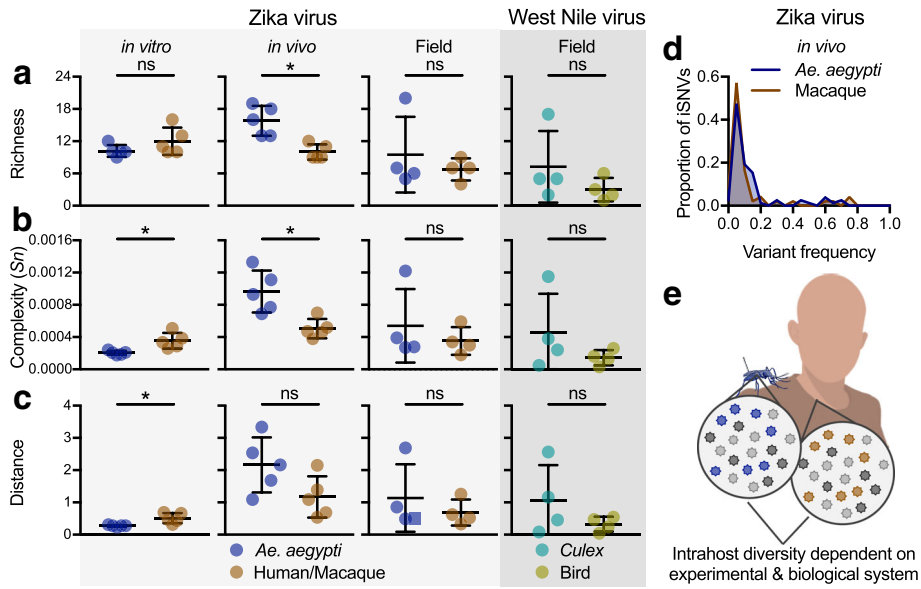

Fig. 8 Intrahost virus genetic diversity is dependent on the experimental and biological system. Variants called from Zika and West Nile virus populations derived from in vitro, in vivo, and field studies (Fig. 6) were used to compare intrahost virus diversity from mosquito vectors (*Ae. aegypti* and *Culex* species) and vertebrate hosts (primates or birds). We compared **a** richness (the number of intrahost single-nucleotide variant [iSNV] sites; Fig. 7a), **b** complexity (uncertainty associated with randomly sampling an allele, measured by Shannon entropy [*Sn*]), and **c** distance (the sum of all iSNV frequencies). The mosquito and vertebrate-derived populations were compared using unpaired Mann-Whitney rank tests (ns, not significant; *, $p < 0.05$). Data shown as mean and standard deviation. **d** The proportion of Zika virus iSNVs detected in the *Ae. aegypti* and rhesus macaque in vivo samples were distributed by frequency. Bin width is 0.05. **e** Our combined data suggests that intrahost virus diversity is dependent upon the experimental system (i.e., in vitro, in vivo, or field samples)

calls. We found that stringent post-filtering, such as combining iSNV frequencies and strand bias across comparing replicate samples, significantly reduce false positive iSNV calls, but there is still a high false discovery rate. Effectively using Nanopore sequencing for intrahost virus diversity measurements will require higher sequencing accuracy and base calling, exploitation of co-occurring variants (i.e., haplotyping) [50], or utilization of different molecular approaches, including the $1D^2$ method (where template and complementary strands of each fragment are sequenced) [40], tandem repeat consensus techniques [36, 51] or unique molecular identifiers [52].

For viruses that utilize multiple hosts, like mosquito-borne viruses, being able to compare results from many samples types is critically important. A lack of standardization, however, means that the field does not yet have a consensus to whether the mosquito vector or the vertebrate host contributes the most to virus genetic diversity [53–62]. The development of PrimalSeq and iVar allows for such measurements to be performed across diverse environments, sample types, and experimental designs. Using PrimalSeq, for example, we found that in vitro Zika virus diversity was significantly greater in human cells, when compared to mosquito cells. However, we found that these results were reversed during in vivo studies. Furthermore, we did not detect significant differences in field-collected mosquito and vertebrate samples for both Zika and West Nile virus. A caveat for the field samples, however, is that we do not know the reference sequence and cannot account for consensus-changing mutations introduced through intrahost bottlenecks and genetic drift [63–65]. In addition, a limitation for all of our samples is that we can only compare diversity measurements from iSNVs greater than 3% frequency, and iSNVs below this threshold may be important for the virus population structure and phenotype [5, 11, 66]. Even still, our incongruent results among experimental designs help to explain why there is still debate about the relative impact of vectors and hosts on virus evolution, and further use of PrimalSeq will help to resolve these issues.

Conclusions

We demonstrate that PrimalSeq can accurately measure intrahost virus genetic diversity if properly validated. We benchmarked our highly multiplexed and streamlined amplicon-based sequencing method using a series of experiments with mixed virus populations, developed an all-inclusive computational analysis tool (iVar), and showcase its utility by measuring intrahost virus diversity from cells, mosquitoes, primates, birds, and humans. Furthermore, using our free online primer designer, Primal Scheme (primal.zibraproject.org) [17], PrimalSeq can be

modified for use with a wide range of viruses. Overall, our detailed laboratory and computational approaches presented here can reveal important insights about intrahost virus evolution directly from clinical or experimental samples in a way that is cheap, accurate, and scalable.

Methods

Mixed virus populations

Zika virus RNA from isolates PRVABC59 (Puerto Rico 2015, Genbank KX087101, "virus #1") and FSS13025 (Cambodia 2010, Genbank KU955593, "virus #2") were quantified by qRT-PCR (as previously described [26]). The consensus sequences from PRVABC59 and KX087101 were determined using untargeted metagenomics (see below) and a strict > 99% majority nucleotide threshold at each site. Sites that were mixed (i.e., containing an iSNV > 1% frequency) were not used to evaluate iSNVs at known frequencies (Fig. 1). Using quantified virus RNA copies, the two viruses were mixed to achieve the desired total RNA copies (one half required amount because 2 μL of RNA was used for cDNA) and ratios of PRVABC59:FSS13025. Metagenomic sequencing of a 10:1 mixed virus population (i.e., 10% FSS13025) was used to verify our mixing approach (Fig. 4). Each mixed virus population was sequenced in triplicate using the metagenomic and amplicon approaches described below.

Laboratory-infected cells, mosquitoes, and primates

Zika virus was collected from in vitro and in vivo experiments to compare intrahost diversity between mosquitoes (*Ae. aegypti*) and primates (humans and macaques, Additional file 1: Table S2). All in vitro and in vivo experiments were conducted using Zika virus isolate PRVABC59 (Puerto Rico, 2015, KX087101). All Zika virus RNA was quantified by qRT-PCR, as described [26].

Aag2 (derived from *Ae. aegypti* embryos [67]) and HeLa (derived from human cervical epithelial cells, ATCC CCL-2) cells were infected using a multiplicity of infection of 0.01 and supernatant was harvested 5 days post infection. Both cell lines were maintained using Minimal Essential Medium (Sigma-Aldrich) supplemented with 10% (*v*/*v*) fetal bovine serum, L-glutamine, sodium bicarbonate, and antibiotics (penicillin and streptomycin). Aag2 and HeLa cells were incubated with 5% CO_2 at 27 °C and 37 °C, respectively.

Ae. aegypti mosquitoes were infected with Zika virus as previously described [68]. In brief, colonized mosquitoes originating from Los Angeles, California, in 2016 feed on viremic mice inoculated with 5 $\log_{10}$ Vero plaque-forming units of Zika virus (PRVABC59). At 14 days post infection, individual mosquitoes were collected and homogenized. Viral RNA was extracted from 50 μL of mosquito homogenate using the using the MagMax Viral RNA Extraction Kit and eluted 50 μL of elution buffer (Buffer EB, Qiagen). Indian origin rhesus macaques (*Macaca mulatta*) were inoculated subcutaneously with 3 $\log_{10}$ Vero plaque-forming units of Zika virus (PRVABC59) and plasma was collected 5 days post infection, as described [69, 70]. RNA was extracted from at least 300 μL of rhesus macaque plasma using the MagMax Viral RNA Extraction Kit and was eluted in 60 μL of elution buffer. RNA extracts from laboratory infected mosquitoes and macaque plasma used for this study had been thawed previously at least one time.

Field-collected mosquitoes and clinical samples

Clinical and entomological samples were collected during the 2016 Florida Zika virus outbreak [26] to compare intrahost Zika virus diversity between naturally infected humans and mosquitoes (Additional file 1: Table S2). Human clinical samples were obtained for diagnostic and surveillance purposes and excess human sera were used for this study. RNA was extracted using the RNAeasy kit (Qiagen) and eluted into 50–100 μL using the supplied elution buffer. Entomological samples were collected by the Miami-Dade Mosquito Control for surveillance of Zika virus activity. *Ae. aegypti* mosquitoes were collected using BG-Sentinel mosquito traps (Biogents AG) and sorted into pools of up to 50 females per trap. The pooled mosquitoes were stored in RNAlater (Invitrogen), RNA was extracted using the RNAeasy kit (Qiagen), and Zika virus RNA was quantified by qRT-PCR [26]. RNA from Zika virus positive pools used in this study contained 13–39 individual mosquitoes; however, considering that ~ 1 in 1600 were infected [26], it is highly unlikely that any pool contained > 1 infected mosquito.

Culex quinquefasciatus mosquitoes (up to 50 per trap) and dead American crows were collected by the San Diego County Vector Control Program during 2015. RNA was extracted using the RNAeasy kit (Qiagen) and screened for the presence of West Nile virus RNA using standard qRT-PCR.

Quantification of virus RNA copies

Zika virus RNA copies were quantified using a qRT-PCR assay targeting the NS5 protein coding region of the genome using the BioRad One-step qRT-PCR for probes kit. In a 20-μL reaction, 2 μL of virus RNA was added to 10 μL of iTaq universal probes reaction mix, 0.5 μL of iScript RT, 6 μL of nuclease-free water, 0.5 μL of the forward primer (5′-AGTGCCAGAGCTGTGTGTAC-3′; genome positions 9007–9027), 0.5 μL of the reverse primer (5′-TCTAGCCCCTAGCCACATCT-3′; genome positions 9097–9117), and 0.5 μL of the 6-FAM labeled probe (5′-GGCAGCCGCGCCATCTGGT-3′; genome positions 9078–9096). The reactions were then amplified on a thermocycler with the following conditions: 50 °C

for 10 min, 95 °C for 3 min, and followed by 40 cycles of 95 °C for 10 s and 57 °C for 10 s (fluorescence read at the end of the 57 °C step). To calculate the number of virus template copies using standard curves, we include 10-fold dilutions of partial Zika virus RNA genomes spanning the primer sites (10^7 to 10^0 copies per reaction). The Zika virus RNA standards were constructed by PCR amplifying a 848 bp segment of the Zika virus NS5 protein coding region (genome positions 8644 to 9492) with the following primers: forward containing a T7 promoter region (5′-TAATACGACTCACTATAGG GAGATCAGGCTCCTGTCAAAACCC-3′; underlined = T7 promoter sequence; genome positions 8644–8664) and reverse primer (5′-AGTGACAACTTGTCCGCTC C-3′; genome positions 9472–9492). The amplified cDNA was converted into RNA to be used as standards using the Invitrogen MEGAscript T7 Transcription Kit.

The accuracy of measuring virus RNA copies by targeting one small genome region, in this case positions 9007–9117, is dependent on relatively equal proportions of the virus genome present in the sample. To address this, we used untargeted metagenomic sequencing of our 1000 RNA copy stocks of virus #1, virus #2, and three replicates of the Mix10% population. The normalized coverage shows that depth is consistent across the virus genome (Additional file 2: Figure S3). The normalized coverage changes are consistent among virus samples and replicates, suggesting that coverage depth is more dependent on intrinsic factors of the virus genome influence replication efficiency (i.e., GC content [71]) rather than significant RNA degradation leading to the loss of a fraction of the virus genome. Hence, we are confident that our qRT-PCR results are relatively informative for determining the virus RNA copy numbers across the whole genome.

For all Zika virus samples, 1000 virus RNA copies were used for sequencing, unless otherwise specified (e.g., Fig. 1c). For all West Nile virus sample, 10,000 virus RNA copies were used. Normalizing input copy numbers allowed us to more accurately compare sequencing results.

PCR amplification of the virus genomes

Virus RNA (2 μL) was reverse transcribed into cDNA using Invitrogen SuperScript IV VILO (20 μL reactions). Virus cDNA (2 μL) was amplified in 35× ~ 400 bp fragments from two multiplexed PCR reactions using Q5 DNA High-fidelity Polymerase (New England Biolabs) using the conditions previously described [17]. For the data shown in Fig. 1, the mixed Zika virus populations were amplified in one multiplexed reaction containing primer sets 5, 24, and 33. A detailed protocol can be found in Additional file 3 and the Zika and West Nile virus primers can be found in Additional file 1: Tables S3 and S4, respectively.

Amplicon-based Illumina sequencing

A detailed protocol for our amplicon-based sequencing methods can be found in Additional file 3. Protocol updates will be released online at http://grubaughlab.com/open-science/amplicon-sequencing/ [72] and https://andersen-lab.com/secrets/protocols/ [73]. Virus amplicons from the two multiplex PCR reactions (above section) were purified using Agencourt AMPure XP beads (Beckman Coulter) and combined (25 ng each) prior to library preparation. The libraries were prepared using the Kapa Hyper prep kit (Kapa Biosystems, following the vendor's protocols but with one fourth of the recommended reagents) and NEXTflex Dual-Indexed DNA Barcodes (BIOO Scientific, diluted to 250 nM). Agencourt AMPure XP beads (Beckman Coulter) were used for all purification steps. The libraries were quantified and quality-checked using the Qubit (Thermo Fisher) and Bioanalyzer (Agilent). Paired-end 250 nt reads were generated using the MiSeq V2 500 cycle or V3 600 cycle kits (Illumina).

Untargeted metagenomic Illumina sequencing

We followed the general outline of a previously developed protocol for untargeted sequencing of the mixed viral populations [34]. In brief, cDNA was generated as described for the amplicon-based methods. Second-strand cDNA was generated using *Escherichia coli* DNA ligase and polymerase (New England Biolabs). The cDNA was purified by Agencourt AMPure XP beads (Beckman Coulter) prior to library preparation using Nextera XT (Illumina) following the vendor's protocols, but with less reagents. Specifically, for tagmentation (12.5 μL reaction), we concentrated our cDNA to 4 μL using a DNA speedvac and used 5 μL of Tagment DNA Buffer (one half recommended) and 1 μL of Amplicon Tagment Mix (one fifth recommended). After incubation, the reaction was stopped using 2.5 μL of Neutralize Tagment Buffer (one half recommended). The libraries were indexed and amplified using one half of the Nextera PCR reagents and primers in a 25-μL reaction. Agencourt AMPure XP beads (Beckman Coulter) were used for the final purification step (purified twice at a ratio of 0.7:1 beads to sample). The libraries were quantified and quality-checked using the Qubit (Thermo Fisher) and Bioanalyzer (Agilent). Paired-end 251 nt reads were generated using the MiSeq V2 500 cycle kit. The paired-end reads were aligned to a provided reference genome using BWA [73], the reads were quality trimmed (Phred quality score < 20) using Trimmomatic [46], and iSNVs were called based on frequency from the bam files using Geneious v9.1.5 [45]. No other iSNV filters, such as strand bias, were used to better compare to the amplicon-based Illumina data.

Illumina data processing and variant calling using iVar

We developed an open source software package to process virus sequencing data and call iSNVs from technical replicates, iVar (intrahost variant analysis from replicates), and detailed documentation can be found at github.com/andersen-lab/ivar [74]. The tool is licensed under GNU General Public License v3.0 and the source code is available at github.com/andersen-lab/ivar [74]. The version of the code used in this paper is available at https://doi.org/10.5281/zenodo.2471612. A list of iVar commands and their brief descriptions are provided in Additional file 1: Table S5, and additional details about the options are available in the documentation and can also be accessed in the help menu distributed with the tool [74]. iVar was used to write two pipelines for calling iSNVs from samples with or without the known reference sequence (i.e., experimental and field-collected samples, respectively; Fig. 9). The software package was written in C++ and has two dependencies—HTSlib (github.com/samtools/htslib) and Awk (cs.princeton.edu/~bwk/btl.mirror/). Awk is generally available on most unix based operating systems and HTSlib has only one dependency—zlib (zlib.net/). The output of `mpileup` command from the widely used SAMtools [75] was used by the package to call iSNVs and generate a consensus sequence from an alignment. The computational pipeline was divided in four main sections: (1) alignment, (2) trimming, (3) constructing nucleotide matrices and scan for iSNVs within primer regions, and (4) statistical comparisons between replicate datasets. The paired-end reads were aligned to a provided reference genome using BWA [76]. The paired-end reads were not merged at any point in this process. The primer sequences were trimmed from the reads using a BED file, with the primer positions, followed by quality trimming. iSNVs above the frequency cutoff of 3% were then called using the `mpileup` command from SAMtools (maximum coverage depth cutoff was set to 0["-d 0"], making the limit on depth the maximum limit of an signed 32 bit integer (i.e., 2,147,483,647). In addition to the frequency threshold, Fisher's exact test was used to determine if the frequency of the iSNV is significantly higher than the mean error rate at that position (see contingency table, Additional file 1: Table S6). Based on the iSNVs called, mismatches in the primer sequences were identified and the reads from the amplicon were removed. The variant calling step were repeated to remove any influence of the primer mismatch on iSNV frequencies.

Validation of iVar

iVar was validated against existing tools. (Additional file 1: Table S7, Additional file 2: Figure S4-S6). We validated iSNV calling in iVar against the `mpileup2snp` and `mpileup2indel` commands in VarScan2 (v2.3.9) [43] using four datasets—two simulated datasets and two clinical Zika virus samples, sequenced using PrimalSeq. We ran both tools with no thresholds and with a quality threshold of 20 and a minimum frequency threshold of 3%. We validated the consensus calling in iVar against the consensus calling available in Geneious (v11.1.4) [45] using four datasets—two simulated datasets and two clinical Zika virus samples, sequenced using PrimalSeq. We did the consensus calling at four different thresholds—0% (majority), 25%, 50% (strict), and 90%. We counted the mismatches between the resulting consensus sequences from iVar, Geneious and the reference sequence by performing a multiple sequence alignment using MAFFT (v7.388) [44]. While counting mismatches, we ignored mismatches when one sequence had a gap and the other had a "N," since in either case. We validated primer trimming and quality trimming in iVar against anchored adapter trimming in cutadapt (v1.16) [47] and against Trimmomatic [46]. iVar uses a sliding window starting from the 5′ end and checks if the average quality within the window drops below the threshold. As soon as the quality drops below the threshold, it trims the sequence by soft clipping the read. This is different from the algorithm cutadapt uses to do quality trimming, but is similar to the sliding window approach used by Trimmomatic. The data and code used for the validation are at github.com/andersen-lab/paper_2018_primalseq-ivar [77].

Oxford Nanopore sequencing and analysis

Using the same PCR amplicons used for amplicon-based Illumina sequencing, we sequenced three replicates of the mixed Zika virus population (90% virus #1, 10% virus #2) using the Oxford Nanopore GridION sequencer. Native, 1D barcode libraries (SQK-NSK007, Oxford Nanopore Technologies, UK) were prepared according to previously published methods [17], with three amplification replicates corresponding to barcodes 1, 2 and 3. The pooled sequencing library was sequenced on an R9.4 version flowcell (FLO-MIN106, Oxford Nanopore Technologies, UK). Reads were basecalled using Albacore 2.3.1 using the command-line read_fast5_basecaller.py -c r94_450bps_linear.cfg -i fast5 -o fastq -r -t 12. Reads were subsequently demultiplexed with Porechop 0.2.3_seqan2.1.1 using default (lenient) settings (github.com/rrwick/Porechop). A total of 2.4 million reads were generated which after alignment and trimming covered 95.66% of the reference genome (Genbank KX087101). For the purposes of assigning to genotypes (i.e., unique virus haplotypes), reads were assigned to individual strains using BWA-MEM [78] against a custom reference database comprising four Zika virus genomes: Genbank KX087101 (virus #1), KU955593 (virus #2), EU545988 (an Asian lineage virus isolated in 2007) and

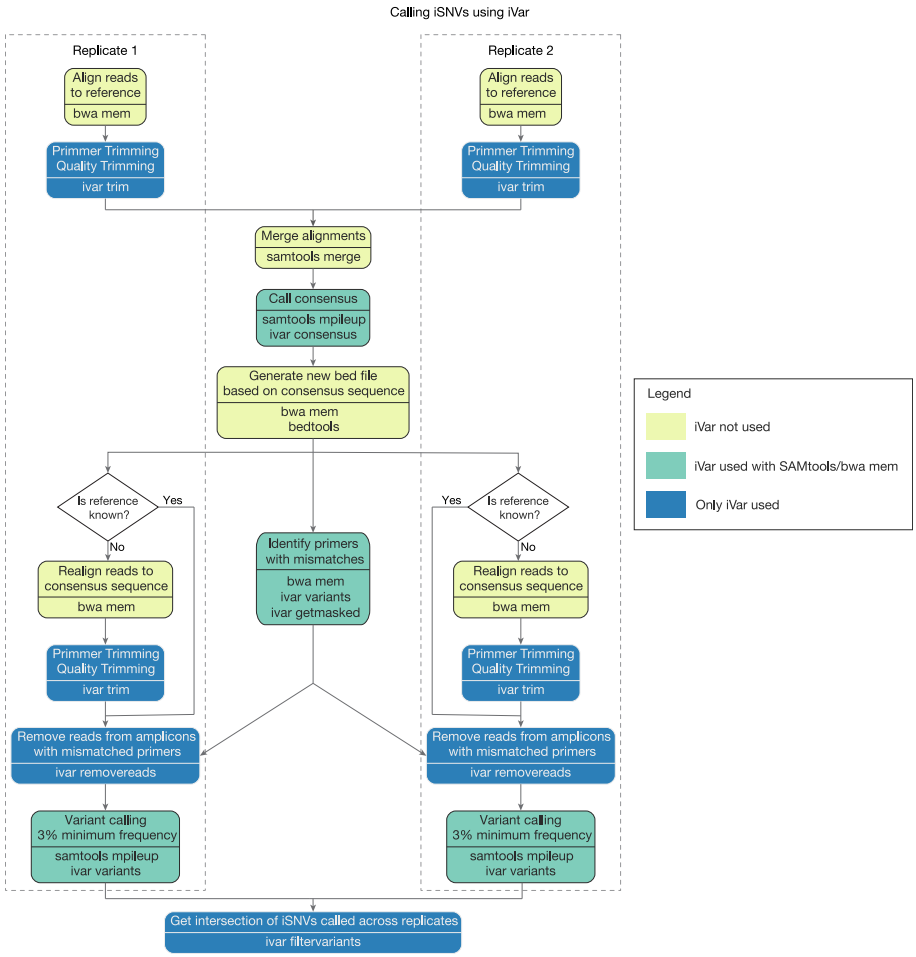

Fig. 9 Overview of iVar pipeline. iVar was used to construct two pipelines for calling intrahost single-nucleotide variants (iSNVs) from samples with and without a known reference sequence. The nodes in the chart are colored based on the usage of iVar, bwa, and SAMtools at each step. For samples with a known reference sequence, the primer sequences are trimmed from the sequenced reads, followed by quality trimming. A consensus sequence for the sample is called by merging the aligned BAM files from each replicate. The primer sequences are then aligned to this consensus sequence and mismatches are identified by iVar after performing variant calling on the aligned primers. The reads corresponding to the mismatched primers are removed from the aligned BAM file of each replicate to ensure that any bias introduced in the iSNV frequencies is removed. The iSNVs are then called for each replicate, individually, with a minimum frequency threshold of 3% and an intersection of the iSNVs across all the replicates are considered to be the "true" iSNVs. For samples with an unknown reference sequence, the iSNVs cannot be called directly using a reference sequence. In this case, after generating the consensus sequence, reads from each replicate are aligned back to this consensus sequence and these realigned BAM files are used for the same subsequent steps as in the case of samples with a known reference sequence

NC_012532 (MR766, an African lineage virus). Counts for each assignment were retrieved, ignoring multi-mapping reads using the shell command bwa mem -x ont2d | samtools view -h -F 256 - | samtools view -h -F 2048 - | cut -f 3 | sort | uniq -c. Next, each replicate was aligned to the PRVACB59 reference genome with BWA-MEM using setting -x ont2d. Primer binding sites and any residual adaptor sequence were masked in the resulting BAM alignment using the align_trim script from the Zibra pipeline [17]. Allele frequencies and putative iSNVs (ignoring insertions or deletions) were extracted from BAM files using a Python script freqs.py (included in the accompanying code repository: github.com/nickloman/zika-isnv [79]). This script utilizes the pileup functionality of samtools via the pysam Python interface module (github.com/pysam-developers/pysam). Only predicted variants with more than 10 supporting forward and 10 supporting reverse reads were considered. The logistic regression model was trained and tested under a 10-fold cross validation scheme using the train function with the parameters method = "glm" and family = "binomial" from the caret (github.com/topepo/caret/) library in R. Class probabilities for the ROC curve were captured from the same function and plotted using ggplot2 (github.com/tidyverse/ggplot2).

Diversity metrics

Virus iSNVs > 3% were used to genetically characterize the populations derived from in vitro, in vivo, and field samples (Figs. 6 and 7). For this purpose, insertions and deletions were not analyzed. For the Zika virus in vitro and in vivo samples, where the ancestral PRVABC59 sequence was known, the iSNV frequencies were calculated by the difference between the ancestral (pre-infection) and derived (post-infection) frequencies (e.g., if variant X was detected at 5% in the ancestral and 10% in the derived population, the iSNV frequency was listed as 5%). For the Zika and West Nile virus field samples, where the ancestral virus sequence was not known, the derived iSNV frequencies were used for the population genetic analysis. Richness was calculated by the total number of iSNV sites per population (Fig. 7a). Complexity, uncertainty associated with randomly sampling an allele, was calculated at each site using Shannon entropy:

$$Sn = \frac{-(p \times \ln(p)) + ((1-p) \times \ln(1-p))}{\ln(2)},$$

where p is the iSNV frequency and the mean Sn from all evaluated sites within the virus genome was used to determine the population complexity (Fig. 7b). Distance was calculated by the sum of all of the iSNV frequencies per population (Fig. 7c).

Additional files

Additional file 1: Table S1. Location of PCR primer mismatches to divergent Zika viruses. **Table S2.** Laboratory and field-collected Zika and West Nile virus samples used in this study and sequencing statistics. **Table S3.** Primer sequences for tiled amplification of Zika virus. **Table S4.** Primer sequences for tiled amplification of West Nile virus. **Table S5.** A list of commands and descriptions available in iVar. **Table S6.** Contingency table for Fisher's exact test to determine if the frequency of the iSNV is significantly higher than the mean error. **Table S7.** The list of commands in iVar and the tool that was used for validation. (XLSX 39 kb)

Additional file 2: iSNV frequencies calculated using pseudo replicates (**Figure S1.**), sensitivity of measuring intrahost diversity at 5% (**Figure S2.**), metagenomic sequencing coverage depth (**Figure S3.**), and validation of iVar for intrahost single-nucleotide variant calling (**Figure S4.**), consensus calling (**Figure S5.**), and trimming (**Figure S6.**). (PDF 7109 kb)

Additional file 3: Laboratory protocol for generating sequencing libraries for measuring intrahost virus genetic diversity. (PDF 198 kb)

Acknowledgements

We thank Barney Graham (VRC; NIAID/NIH) for support with the non-human primate studies, the Florida Department of Health for providing clinical samples, Miami-Dade Mosquito Control for providing collected *Ae. aegypti* pools, San Diego County Vector Control Program for providing West Nile virus samples, and Glenn Oliveira, Mark Zeller, Refugio Robles, Emily Spencer, Dylan Grubaugh, and Sophie Taylor for technical support.

Funding

NDG was supported by NIH training grant 5T32AI007244-33. JQ is supported by a grant from the NIHR Surgical Reconstruction and Microbiology Research Centre (SRMRC). KKAVR is supported by the Office of Research Infrastructure Programs/OD (P51OD011107) to CNPRC and NIH R21AI129479. SI and SFM are supported by NIH NIAID R01AI099210. LLC was supported by startup funds from the UC Davis Department of Pathology, Microbiology and Immunology and the Pacific Southwest Regional Center of Excellence for Vector-Borne Diseases funded by the U.S. Centers for Disease Control and Prevention (Cooperative Agreement 1U01CK000516). BJM was supported by Abt Associates and a consortium of vector control districts in California: Coachella Valley, Orange County, Greater Los Angeles County, San Gabriel Valley, West Valley, Kern, Butte County, Tulare, Sacramento-Yolo, Placer, and Turlock. The rhesus macaque studies were supported by NIH 1R21AI129479–01 & Supplement, California National Primate Research Center pilot research grant P51OD011107 and FDA HHS HHSF223201610542P. NJL is supported by a Medical Research Council Bioinformatics Fellowship as part of the Cloud Infrastructure for Microbial Bioinformatics (CLIMB) project. KGA is a Pew Biomedical Scholar, and is supported by NIH NCATS CTSA UL1TR001114, NIAID contract HHSN272201400048C, NIAID R21AI137690, NIAID U19AI135995, and The Ray Thomas Foundation.

Availability of data and materials

All additional files can be found at github.com/andersen-lab/paper_2018_primalseq-ivar [77] and raw sequencing files can be found at console.cloud.google.com/storage/browser/andersen-lab_project_ivar-primalseq. The laboratory protocols generated from this study can be found in Additional file 3. Our computational tool, iVar, is licensed under an open source license compliant with OSI (GPL-3.0), is installable via bioconda ("conda install ivar"), and the source code is available at github.com/andersen-lab/ivar [74]. The version of the code used in this paper is available at https://doi.org/10.5281/zenodo.2471612. Protocol updates and additional primer schemes can be found at grubaughlab.com/open-science/amplicon-sequencing/ [72] and andersen-lab.com/secrets/protocols/ [73]. The validation analyses from this study can be found in Additional file 2, github.com/andersen-lab/ivar-validation/, github.com/nickloman/zika-isnv [79], NCBI Bioproject PRJNA438514 (illumina data) [80], and ENA project PRJEB30574 (nanopore data).

Authors' contributions

The study was conceived and coordinated by NDG, KG, NJL, and KGA. The samples were provided by BJM, ALT, LMP, DEB, SG, NG, KKAVR, SI, SFM, and LLC. Library preparation and sequencing was performed by NDG, JGDJ, and JQ. The variant calling pipeline (iVar) was designed and built by KG and NLM. The data was analyzed and interpreted by NDG, KG, NJL, and KGA. The manuscript was written by NDG, KG, NJL, and KGA with input from all co-authors. All authors read and approved the final manuscript.

Ethics approval and consent to participate

Research on human subjects was conducted in compliance with existing regulations relating to the protection of human subjects and was evaluated and approved (#IRB-15-6664) by the Institutional Review Board/Ethics Review Committee at The Scripps Research Institute. Clinical samples were obtained from the Florida Department of Health (DOH) and Antibody Systems Inc. Samples collected in Florida were collected under a waiver of consent granted by the Florida DOH Human Research Protection Program. The work received a non-human subjects research designation (category 4 exemption) by the Florida DOH since this research was performed with leftover clinical diagnostic samples involving no more than minimal risk. Hence, written informed consent was not obtained. All samples were de-identified prior to receipt by the study investigators. The experimental methods used comply with the Helsinki Declaration. Research involving Indian origin rhesus macaques was conducted at the California National Primate Research Center, and experimental infections of mice upon which *Ae. aegypti* mosquitoes fed were performed at the University of California, Davis, School of Veterinary Medicine. Both institutes are fully accredited by the Association for the Assessment and Accreditation of Laboratory Animal Care International. Animals were cared for in accordance with the National Research Council Guide for the Care and Use of Laboratory Animals and the Animal Welfare Act. Animal experiments were approved by the Institutional Animal Care and Use Committee of UC Davis (protocols #19211 and #19695 for rhesus macaques, protocol #19404 for mice). All macaques samples used in this study were from approved studies [70]; and none were generated specifically for this work.

Consent for publication

Not applicable.

Competing interests

NJL has received travel and accommodation expenses from Oxford Nanopore Technologies to attend meetings, and an honorarium to speak at an internal company meeting. NJL has previously received free-of-charge reagents and consumables in support of research projects from Oxford Nanopore Technologies. The other authors declare that they have no competing interests.

Publisher's Note

Author details

[1]Department of Immunology and Microbiology, The Scripps Research Institute, La Jolla, CA 92037, USA. [2]Department of Epidemiology of Microbial Diseases, Yale School of Public Health, New Haven, CT 06510, USA. [3]Institute of Microbiology and Infection, University of Birmingham, Birmingham B15 2TT, UK. [4]Laboratory of Experimental Pathology, Gonçalo Moniz Institute, Oswaldo Cruz Foundation, Salvador, Bahia, Brazil. [5]Department of Pathology, Microbiology and Immunology, University of California, Davis, CA 95616, USA. [6]Department of Biological Sciences, College of Arts and Sciences, Florida Gulf Coast University, Fort Myers, FL 33965, USA. [7]Department of Environmental Sciences, The Connecticut Agricultural Experiment Station, New Haven, CT 06504, USA. [8]Department of Environmental Health, San Diego County Vector Control Program, San Diego, CA 92123, USA. [9]California National Primate Research Center and Department of Pathology, Microbiology and Immunology, University of California, Davis, CA 95616, USA. [10]Scripps Research Translational Institute, La Jolla, CA 92037, USA.

Received: 4 August 2018 Accepted: 26 December 2018
Published online: 08 January 2019

References

1. Holland J, Spindler K, Horodyski F, Grabau E, Nichol S, VandePol S. Rapid evolution of RNA genomes. Science. 1982;215:1577–85 researchgate.net.
2. Hensley SE, Das SR, Bailey AL, Schmidt LM, Hickman HD, Jayaraman A, et al. Hemagglutinin receptor binding avidity drives influenza A virus antigenic drift. Science. 2009;326:734–6.
3. Henn MR, Boutwell CL, Charlebois P, Lennon NJ, Power KA, Macalalad AR, et al. Whole genome deep sequencing of HIV-1 reveals the impact of early minor variants upon immune recognition during acute infection. PLoS Pathog. 2012;8:e1002529.
4. Parameswaran P, Wang C, Trivedi SB, Eswarappa M, Montoya M, Balmaseda A, et al. Intrahost selection pressures drive rapid dengue virus microevolution in acute human infections. Cell Host Microbe. 2017;22:400–10.e5.
5. Vignuzzi M, Stone JK, Arnold JJ, Cameron CE, Andino R. Quasispecies diversity determines pathogenesis through cooperative interactions in a viral population. Nature. 2006;439:344–8.
6. Stapleford KA, Coffey LL, Lay S, Bordería AV, Duong V, Isakov O, et al. Emergence and transmission of arbovirus evolutionary intermediates with epidemic potential. Cell Host Microbe. 2014;15:706–16.
7. Stern A, Yeh MT, Zinger T, Smith M, Wright C, Ling G, et al. The evolutionary pathway to virulence of an RNA virus. Cell. 2017;169:35–46.e19.
8. Grubaugh ND, Andersen KG. Experimental evolution to study virus emergence. Cell. 2017;169:1–3.
9. Worby CJ, Lipsitch M, Hanage WP. Shared genomic variants: identification of transmission routes using pathogen deep-sequence data. Am J Epidemiol. 2017;186:1209–16.
10. Poirier EZ, Vignuzzi M. Virus population dynamics during infection. Curr Opin Virol. 2017;23:82–7.
11. Dolan PT, Whitfield ZJ, Andino R. Mapping the evolutionary potential of RNA viruses. Cell Host Microbe. 2018;23:435–46.
12. McCrone JT, Lauring AS. Measurements of intrahost viral diversity are extremely sensitive to systematic errors in variant calling. J Virol. 2016;90:6884–95.
13. Illingworth CJR, Roy S, Beale MA, Tutill H, Williams R, Breuer J. On the effective depth of viral sequence data. Virus Evol. 2017;3:vex0303. Available from: https://academic.oup.com/ve/article/3/2/vex030/4629376?searchresult=1 .
14. Zanini F, Brodin J, Albert J, Neher RA. Error rates, PCR recombination, and sampling depth in HIV-1 whole genome deep sequencing. Virus Res. 2017;239:106–14.
15. Iyer S, Casey E, Bouzek H, Kim M, Deng W, Larsen BB, et al. Comparison of major and minor viral SNPs identified through single template sequencing and pyrosequencing in acute HIV-1 infection. PLoS One. 2015;10:e0135903.
16. Schirmer M, Ijaz UZ, D'Amore R, Hall N, Sloan WT, Quince C. Insight into biases and sequencing errors for amplicon sequencing with the Illumina MiSeq platform. Nucleic Acids Res. 2015;43:e37.
17. Quick J, Grubaugh ND, Pullan ST, Claro IM, Smith AD, Gangavarapu K, et al. Multiplex PCR method for MinION and Illumina sequencing of Zika and other virus genomes directly from clinical samples. Nature Protoc. 2017;12:1261–76.
18. Magnani DM, Rogers TF, Beutler N, Ricciardi MJ, Bailey VK, Gonzalez-Nieto L, et al. Neutralizing human monoclonal antibodies prevent Zika virus infection in macaques. Sci Transl Med. 2017;9. Available from:. https://doi.org/10.1126/scitranslmed.aan8184.
19. Faria NR, Quick J, Claro IM, Thézé J, de Jesus JG, Giovanetti M, et al. Establishment and cryptic transmission of Zika virus in Brazil and the Americas. Nature. 2017; Available from:. https://doi.org/10.1038/nature22401.
20. Barzon L, Pacenti M, Franchin E, Lavezzo E, Trevisan M, Sgarabotto D, et al. Infection dynamics in a traveller with persistent shedding of Zika virus RNA in semen for six months. Euro Surveill. 2016;21. Available from. https://doi.org/10.2807/1560-7917.
21. Dudley DM, Newman CM, Lalli J, Stewart LM, Koenig MR, Weiler AM, et al. Infection via mosquito bite alters Zika virus tissue tropism and replication kinetics in rhesus macaques. Nat Commun. 2017;8:2096.
22. Zanini F, Brodin J, Thebo L, Lanz C, Bratt G, Albert J, et al. Population genomics of intrapatient HIV-1 evolution. Elife. 2015;4. Available from:. https://doi.org/10.7554/eLife.11282.
23. Moncla LH, Zhong G, Nelson CW, Dinis JM, Mutschler J, Hughes AL, et al. Selective bottlenecks shape evolutionary pathways taken during mammalian adaptation of a 1918-like avian influenza virus. Cell Host Microbe. 2016;19:169–80.
24. Varghese V, Wang E, Babrzadeh F, Bachmann MH, Shahriar R, Liu T, et al. Nucleic acid template and the risk of a PCR-induced HIV-1 drug resistance mutation. PLoS One. 2010;5:e10992.
25. Metsky HC, Matranga CB, Wohl S, Schaffner SF, Freije CA, Winnicki SM, et al. Zika virus evolution and spread in the Americas. Nature. 2017;66:366.
26. Grubaugh ND, Ladner JT, Kraemer MUG, Dudas G, Tan AL, Gangavarapu K, et al. Genomic epidemiology reveals multiple introductions of Zika virus into the United States. Nature. 2017;90:4864.
27. Black A, Potter B, Dudas G, Feldstein L, Grubaugh ND, Andersen KG, et al. Genetic characterization of the Zika virus epidemic in the US Virgin Islands: bioRxiv; 2017. p. 113100. [cited 2017 May 11]. Available from: http://biorxiv.org/content/early/2017/03/03/113100.abstract.
28. Faria NR, Kraemer MUG, Hill S, de Jesus JG, de Aguiar RS, Iani FCM, et al. Genomic and epidemiological monitoring of yellow fever virus transmission potential. Science. 2018;361:894-99. Available from. http://science.sciencemag.org/content/361/6405/894.
29. Hepp CM, Cocking JH, Valentine M, Young SJ, Damian D, Sheridan K, et al. Phylogenetic analysis of West Nile Virus in Maricopa County, Arizona: Evidence for dynamic behavior of strains in two major lineages in the American Southwest. PLOS ONE. 2018;13:e0205801 Available from: https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0205801.
30. Magnani DM, Rogers TF, Maness NJ, Grubaugh ND, Beutler N, Bailey VK, et al. Fetal demise and failed antibody therapy during Zika virus infection of pregnant macaques. Nat Commun. 2018;9:1624.
31. Worobey M, Watts TD, McKay RA, Suchard MA, Granade T, Teuwen DE, et al. 1970s and "Patient 0" HIV-1 genomes illuminate early HIV/AIDS history in North America. Nature. 2016 [cited 2016 Oct 26]; Available from:. https://doi.org/10.1038/nature19827.
32. Sipos R, Székely AJ, Palatinszky M, Révész S, Márialigeti K, Nikolausz M. Effect of primer mismatch, annealing temperature and PCR cycle number on 16S rRNA gene-targetting bacterial community analysis. FEMS Microbiol Ecol. 2007;60:341–50.
33. Krehenwinkel H, Wolf M, Lim JY, Rominger AJ, Simison WB, Gillespie RG. Estimating and mitigating amplification bias in qualitative and quantitative arthropod metabarcoding. Sci Rep. 2017;7:17668.
34. Matranga CB, Andersen KG, Winnicki S, Busby M, Gladden AD, Tewhey R, et al. Enhanced methods for unbiased deep sequencing of Lassa and Ebola RNA viruses from clinical and biological samples. Genome Biol. 2014;15:519.

35. Robasky K, Lewis NE, Church GM. The role of replicates for error mitigation in next-generation sequencing. Nat Rev Genet. 2014;15:56–62.
36. Acevedo A, Andino R. Library preparation for highly accurate population sequencing of RNA viruses. Nat Protoc. 2014;9:1760–9.
37. Quick J, Loman NJ, Duraffour S, Simpson JT, Severi E, Cowley L, et al. Real-time, portable genome sequencing for Ebola surveillance. Nature. 2016;530:228–32.
38. Gardy JL, Loman NJ. Towards a genomics-informed, real-time, global pathogen surveillance system. Nat Rev Genet. 2018;19:9–20.
39. Jain M, Koren S, Miga KH, Quick J, Rand AC, Sasani TA, et al. Nanopore sequencing and assembly of a human genome with ultra-long reads. Nat Biotechnol. 2018;36:338–45.
40. Rang FJ, Kloosterman WP, de Ridder J. From squiggle to basepair: computational approaches for improving nanopore sequencing read accuracy. Genome Biol. 2018;19:90.
41. Tardif KD, Simmon KE, Kommedal O, Pyne MT, Schlaberg R. Sequencing-based genotyping of mixed human papillomavirus infections by use of RipSeq software. J Clin Microbiol. 2013;51:1278–80.
42. Nielsen R, Paul JS, Albrechtsen A, Song YS. Genotype and SNP calling from next-generation sequencing data. Nat Rev Genet. 2011;12:443–51.
43. Koboldt DC, Zhang Q, Larson DE, Shen D, McLellan MD, Lin L, et al. VarScan 2: somatic mutation and copy number alteration discovery in cancer by exome sequencing. Genome Res. 2012;22:568–76.
44. Katoh K, Standley DM. MAFFT: iterative refinement and additional methods. Methods Mol Biol. 2014;1079:131–46.
45. Kearse M, Moir R, Wilson A, Stones-Havas S, Cheung M, Sturrock S, et al. Geneious Basic: an integrated and extendable desktop software platform for the organization and analysis of sequence data. Bioinformatics. 2012;28:1647–9.
46. Bolger AM, Lohse M, Usadel B. Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics. 2014;30:2114–20.
47. Martin M. Cutadapt removes adapter sequences from high-throughput sequencing reads. EMBnetjournal. 2011;17:10–2.
48. Lee DF, Lu J, Chang S, Loparo JJ, Xie XS. Mapping DNA polymerase errors by single-molecule sequencing. Nucleic Acids Res. 2016;44:e118.
49. Potapov V, Ong JL. Correction: examining sources of error in PCR by single-molecule sequencing. PLoS One. 2017;12:e0181128.
50. Ebler J, Haukness M, Pesout T, Marschall T, Paten B. Haplotype-aware genotyping from noisy long reads: bioRxiv; 2018. p. 293944. [cited 2018 Jul 13]. Available from: https://www.biorxiv.org/content/early/2018/04/03/293944.abstract
51. Li C, Chng KR, Boey EJH, Ng AHQ, Wilm A, Nagarajan N. INC-Seq: accurate single molecule reads using nanopore sequencing. Gigascience. 2016;5:34 gigascience.biomedcentral.com.
52. Kinde I, Wu J, Papadopoulos N, Kinzler KW, Vogelstein B. Detection and quantification of rare mutations with massively parallel sequencing. Proc Natl Acad Sci U S A. 2011;108:9530–5.
53. Vasilakis N, Deardorff ER, Kenney JL, Rossi SL, Hanley KA, Weaver SC. Mosquitoes put the brake on arbovirus evolution: experimental evolution reveals slower mutation accumulation in mosquito than vertebrate cells. PLoS Pathog. 2009;5:e1000467.
54. Lin S-R, Hsieh S-C, Yueh Y-Y, Lin T-H, Chao D-Y, Chen W-J, et al. Study of sequence variation of dengue type 3 virus in naturally infected mosquitoes and human hosts: implications for transmission and evolution. J Virol. 2004;78:12717–21.
55. Sim S, Aw PPK, Wilm A, Teoh G, Hue KDT, Nguyen NM, et al. Tracking dengue virus intra-host genetic diversity during human-to-mosquito transmission. PLoS Negl Trop Dis. 2015;9:e0004052.
56. Coffey LL, Vignuzzi M. Host alternation of chikungunya virus increases fitness while restricting population diversity and adaptability to novel selective pressures. J Virol. 2011;85:1025–35.
57. Jerzak GVS, Brown I, Shi P-Y, Kramer LD, Ebel GD. Genetic diversity and purifying selection in West Nile virus populations are maintained during host switching. Virology. 2008;374:256–60.
58. Jerzak G, Bernard KA, Kramer LD, Ebel GD. Genetic variation in West Nile virus from naturally infected mosquitoes and birds suggests quasispecies structure and strong purifying selection. J Gen Virol. 2005;86:2175–83.
59. Ciota AT, Jia Y, Payne AF, Jerzak G, Davis LJ, Young DS, et al. Experimental passage of St. Louis encephalitis virus in vivo in mosquitoes and chickens reveals evolutionarily significant virus characteristics. PLoS One. 2009;4:e7876.
60. Grubaugh ND, Smith DR, Brackney DE, Bosco-Lauth AM, Fauver JR, Campbell CL, et al. Experimental evolution of an RNA virus in wild birds: evidence for host-dependent impacts on population structure and competitive fitness. PLoS Pathog. 2015;11:e1004874.
61. Grubaugh ND, Fauver JR, Rückert C, Weger-Lucarelli J, Garcia-Luna S, Murrieta RA, et al. Mosquitoes transmit unique West Nile virus populations during each feeding episode. Cell Reports. 2017;19:709–18.
62. Nelson CW, Sibley SD, Kolokotronis S-O, Hamer GL, Newman CM, Anderson TK, et al. Selective constraint and adaptive potential of West Nile virus within and among naturally infected avian hosts and mosquito vectors. Virus Evol. 2018;4:vey013.
63. Grubaugh ND, Weger-Lucarelli J, Murrieta RA, Fauver JR, Garcia-Luna SM, Prasad AN, et al. Genetic drift during systemic arbovirus infection of mosquito vectors leads to decreased relative fitness during host switching. Cell Host Microbe. 2016;19:481–92.
64. Lequime S, Fontaine A, Ar Gouilh M, Moltini-Conclois I, Lambrechts L. Genetic drift, purifying selection and vector genotype shape dengue virus intra-host genetic diversity in mosquitoes. PLoS Genet. 2016;12:e1006111.
65. Forrester NL, Guerbois M, Seymour RL, Spratt H, Weaver SC. Vector-borne transmission imposes a severe bottleneck on an RNA virus population. PLoS Pathog. 2012;8:e1002897.
66. Acevedo A, Brodsky L, Andino R. Mutational and fitness landscapes of an RNA virus revealed through population sequencing. Nature. 2014;505:686–90.
67. Lan Q, Fallon AM. Small heat shock proteins distinguish between two mosquito species and confirm identity of their cell lines. Am J Trop Med Hyg. 1990;43:669–76.
68. Main BJ, Nicholson J, Winokur OC, Steiner C, Riemersma KK, Stuart J, et al. Vector competence of Aedes aegypti, Culex tarsalis, and Culex quinquefasciatus from California for Zika virus. PLoS Negl Trop Dis. 2018;12:e0006524.
69. Coffey LL, Pesavento PA, Keesler RI, Singapuri A, Watanabe J, Watanabe R, et al. Zika virus tissue and blood compartmentalization in acute infection of rhesus macaques. PLoS One. 2017;12:e0171148.
70. Coffey LL, Keesler RI, Pesavento PA, Woolard K, Singapuri A, Watanabe J, et al. Intraamniotic Zika virus inoculation of pregnant rhesus macaques produces fetal neurologic disease. Nat Commun. 2018;9:2414.
71. Ekblom R, Smeds L, Ellegren H. Patterns of sequencing coverage bias revealed by ultra-deep sequencing of vertebrate mitochondria. BMC Genomics. 2014;15:467.
72. Grubaugh ND. Amplicon sequencing: Grubaugh Lab. [cited 2018 Dec 19]. Available from: http://grubaughlab.com/open-science/amplicon-sequencing/
73. Andersen KG. Protocols: Andersen Lab. [cited 2018 Dec 19]. Available from: https://andersen-lab.com/secrets/protocols/
74. Gangavarapu K, Andersen KG. iVar: Github. [cited 2018 Dec 19]. Available from: https://github.com/andersen-lab/ivar
75. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, et al. The sequence alignment/map format and SAMtools. Bioinformatics. 2009;25:2078–9.
76. Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics. 2009;25:1754–60.
77. Gangavarapu K, Grubaugh ND, Andersen KG. Additional files and data for iVar and PrimalSeq: Github. [cited 2018 Dec 19]. Available from: https://github.com/andersen-lab/paper_2018_primalseq-ivar
78. Li H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM: arXiv; 2013. Available from: http://arxiv.org/abs/1303.3997
79. Loman NJ. Additional files and data for calling iSNV using MinION sequencing: Github. [cited 2018 Dec 19]. Available from: https://github.com/nickloman/zika-isnv
80. Grubaugh ND. BioProject: PRJNA438514: NCBI. [cited 2018 Dec 19]. Available from: https://www.ncbi.nlm.nih.gov/bioproject/?term=PRJNA438514