Validation of multiplex PCR sequencing assay of SIV

Ryan V Moriarty
University of Wisconsin-Madison https://orcid.org/0000-0001-7100-1906

Nico Fesser
University at Buffalo - The State University of New York

Matthew S Sutton
National Institute of Allergy and Infectious Diseases

Vanessa Venturi
University of New South Wales

Miles P Davenport
University of New South Wales

Timothy Schlub
The University of Sydney School of Public Health

Shelby L O'Connor (✉ slfeinberg@wisc.edu)
University of Wisconsin-Madison https://orcid.org/0000-0003-0183-5010

Research

Keywords: SIV, multiplex PCR, deep sequencing, SNV detection

DOI: https://doi.org/10.21203/rs.3.rs-35062/v2

Abstract

Background: The generation of accurate and reproducible viral sequence data is necessary to understand the diversity present in populations of RNA viruses isolated from clinical samples. While various sequencing methods are available, they often require high quality templates and high viral titer to ensure reliable data.

Methods: We modified a multiplex PCR and sequencing approach to characterize populations of simian immunodeficiency virus (SIV) isolated from nonhuman primates. We chose this approach with the aim of reducing the number of required input templates while maintaining fidelity and sensitivity. We conducted replicate sequencing experiments using different numbers of quantified viral RNA (vRNA) or viral cDNA as input material. We performed assays with clonal SIVmac239 to detect false positives, and we mixed SIVmac239 and a variant with 24 point mutations (SIVmac239-24X) to measure variant detection sensitivity.

Results: We found that utilizing a starting material of quantified viral cDNA templates had a lower rate of false positives and increased reproducibility when compared to that of quantified vRNA templates. This study identifies the importance of rigorously validating deep sequencing methods and including replicate samples when using a new method to characterize low frequency variants in a population with a small number of templates.

Conclusions: Because the need to generate reproducible and accurate sequencing data from diverse viruses from low titer samples, we modified a multiplex PCR and sequencing approach to characterize SIV from populations from non-human primates. We found that increasing starting template numbers increased the reproducibility and decreased the number of false positives identified, and this was further seen when cDNA was used as a starting material. Ultimately, we highlight the importance of vigorously validating methods to prevent overinterpretation of low frequency variants in a sample.

Introduction

Characterizing the sequence diversity of RNA virus populations is an essential component of studying viral pathogenesis and transmission in individuals (Vignuzzi et al. 2006, Poirier and Vignuzzi 2017). This sequence data can be used to identify antiviral drug resistance mutations (Dudley et al. 2014), understand how viruses evolve (Lessler et al. 2016, Henn et al. 2012), and track virus transmission during epidemics (Hadfield et al 2018), such as the Ebola virus outbreak in West Africa in 2014, the Zika virus outbreak in Brazil in 2015 (Quick et al. 2017, Carroll et al. 2015), and the current SARS-CoV-2 outbreak (Fauver et al 2020, CDC 2020).

The accumulation of mutations in RNA viruses can impact their pathogenesis (Sanjuán and Domingo-Calap 2016). While many mutations can be deleterious or neutral, some are beneficial for virus proliferation, survival, or transmission (Henn et al. 2012, Zanini et al. 2017). Naturally elicited host immune responses that fail to eliminate replicating viruses select for variants that avoid immune detection (Henn et al. 2012). Drug resistance mutations can also accumulate when antiretroviral therapy does not fully suppress virus replication (Bangsberg et al 2007). Accurate detection of these variants in RNA virus populations can help determine whether therapeutic interventions eliminate or exacerbate mutations from the replicating virus population.

Sequencing RNA viruses requires the generation of viral cDNA, followed by amplification of either long (greater than 1000bp) or short (less than 400bp) DNA segments. Long amplicons are used to study distantly linked nucleotides on the same virus templates using Pacific Biosciences (Brese et al. 2018) or Oxford Nanopore instruments (Tyler et al. 2018, Oikonomopoulos et al. 2016), In contrast, Illumina technology can generate sequence data from shorter viral segments with higher throughput, better fidelity, and improved efficiency (Adey et al. 2010). While each approach has advantages and disadvantages, the desire to acquire sequence data with newer assays often trumps taking the time to perform experiments required to validate their sensitivity and reproducibility.

Our goal was to implement a multiplex PCR approach, similar to those used for Ebola (Arias et al 2016), ZIKV (Quick et al 2017), and SARS-CoV-2 (Fauver et al 2020), to improve the reproducibility and sensitivity of sequencing SIV derived from plasma with low virus titers or cell-associated vRNA isolated from different tissues. SIV dynamics are frequently studied in nonhuman primates (Harris et al. 2013, Gambhira et al. 2014, Hassounah et al. 2014), but the samples collected from animals with interesting biological phenotypes often have a low virus titer. With an ongoing emphasis on understanding the dynamics of SIV replication in nonhuman primates (Kumar et al. 2016), we aimed to determine if the multiplex method could be applied to SIV to improve the characterization of virus populations with improved sensitivity and reproducibility.

We developed a multiplex PCR approach to amplify and sequence SIV. To validate this method, we sequenced different numbers of vRNA and viral cDNA templates of clonal SIVmac239, as well as variable ratios of two clonal SIV strains differing at 24 nucleotide positions. We found improved sensitivity and reproducibility of variant calling when normalizing to the number of viral cDNA templates added to the reaction when compared to the number of vRNA templates added to the reaction. By validating the SIV multiplex sequencing method here, we identify the strengths and limitations of this method, which are essential for defining the usability of any new technique.

Results

Design of a multiplex PCR assay for SIV.

Candidate multiplex primers for SIV were designed in Primal Scheme, a tool developed by Quick et al. (2017). Each primer set was tested individually and then pooled such that the amplicon products would not overlap with each other (Figure 1A, Table 1). The most 5’ primer binds just upstream of the start codon for gag, and has an identical sequence to the 3’ LTR, so it is not necessary to amplify the 5’ LTR as well. Primers generated by PrimalScheme are selected based on Primer3 software, as described in Quick et al 2017. Primer pools were tested to verify that individual primer pairs would generate amplicons spanning the entire viral genome when combined. Final primer pair concentrations, corresponding sequences, and positions relative to SIVmac239 (Accession: M33262) reference can be found in Table 1.

We first isolated vRNA from a stock of clonal SIVmac239. For Method 1, we quantified the vRNA stock and then diluted it to 10^6 vRNA templates per reaction. Serial dilutions of quantified vRNA were converted to viral cDNA by reverse-transcription. The multiplex PCR was performed on the viral cDNA. For Method 2, we prepared total cDNA from 10^7 vRNA templates of each stock. We then quantified the viral cDNA with a qPCR reaction specific for *gag*. The quantified viral cDNA was diluted to 10^6 cDNA *gag* copies per reaction and the multiplex PCR was then performed. After multiplex PCR for either Method 1 or 2, 75ng of each pool of PCR products were combined into a single tube to generate a 150ng DNA pool containing all the generated PCR amplicons. This amplicon library was then tagged using an Illumina TruSeq kit, and sequenced on an Illumina MiSeq.

Detection of false positives in clonal SIVmac239

We first sequenced clonal SIVmac239 to determine the frequency of false positives when using either Method 1 or 2. We used serially diluted 100% SIVmac239 vRNA or viral cDNA for this part of the project. For each replicate using Method 1, new cDNA was prepared and then multiplex PCR and sequencing were performed. These experiments were performed in triplicate. For each replicate using Method 2, the same prepared cDNA was used for all of the multiplex PCR reactions. These experiments were performed in duplicate.

FASTQ sequences were examined using a modified version of a custom pipeline previously used to analyze multiplex PCR ZIKV sequences (Dudley et al. 2017), and uploaded to a Docker container in order to ensure reproducibility. Using this tool, we randomly subsampled up to 2000 reads per amplicon across each data set and mapped them to SIVmac239 (Accession: M33262), as described in the Materials and Methods. Amplification of each PCR product does not occur equally, so by subsampling up to 2000 reads, we could attempt to informatically normalize the depth of coverage, while not oversampling any one single amplicon. VarScan (https://sourceforge.net/projects/varscan/) was then used to identify nucleotides present in the virus population that were different from the reference at a frequency of 1% or greater and had a depth of coverage of at least 1800 nucleotides, or 90% of our maximum subsampled depth. SNPeff (Cingolani et al. 2012) was used to annotate variants and their effect on each coding sequence. Any single nucleotide variant (SNV) present at a frequency of 1% or greater and with a depth of coverage of at least 1800 nucleotides was categorized as a false positive for our analysis. These thresholds are more conservative than the 3% cutoff and 400x coverage required by Grubaugh et. al (2019).

We began by assessing false positives present in sequences generated by Method 1. The average rate of false positives in a single replicate was related to the number of input templates, with samples containing 10^3 input copies having a higher average rate of false positives at 1.13×10^{-2} false positives per nucleotide, and samples containing 10^6 input copies having a lower average rate of 2.6×10^{-3} false positives per nucleotide (Figure 2A, left panel, closed circles, $p < 0.0001$, Tukey's multiple comparisons test). We then determined whether the rate of false positives declined when considering two or more replicates. We found there was not a significant copy-dependent decrease in false positives when we used two replicates compared to a single replicate (Figure 2A, left panel, open circles, $p = 0.83$, Tukey's multiple comparisons test).

We investigated the individual nucleotide positions where we detected false positives in multiple replicates when using Method 1 (Figure 2B). We found 11 positions with false positives at all input copy numbers, with 4 being insertions at nucleotide positions 1254,1480, 5428, and 7396, and 9 being substitutions at nucleotide positions 6181, 6186, 6188, 6190, 6192, 6201, 6205, 6207, and 6713. We found the median false positive frequency did not depend on the number of input copies ($p = 0.92$, Kruskal-Wallis) (Figure 2B). Each of the insertions occurred in a poly-A region containing 6 consecutive adenines. Although the chemistry of Illumina sequencing does not lead to the same errors in homopolymers that are notorious in other sequencing platforms (Bently et al. 2008), there can still be PCR-based errors in homopolymeric regions (McInerney et al. 2014, Potapov and Ong 2017). All substitutions, aside from position 6713, were present within a stretch of 27 nucleotides that are adjacent to a primer binding site. These SNVs are contained within an overlap region between Amplicons 20 and 21. Notably, these variants were present in Amplicon 21, but not Amplicon 20, suggesting that Amplicon 21 may be more prone to the incorporation of PCR-based substitutions than Amplicon 20. While unfortunate, inaccuracies in variant reporting is not an uncommon phenomenon at the ends of amplicons and has been reported previously (Schirmer et al 2016, McCoy et al 2014). We also observed that when using a different analysis pipeline that does not normalize coverage across the genome through subsampling, these variants were not reported in the vcf file, highlighting the importance of validating the analysis methods prior to calling variants as true variants. However, we felt that the benefit of standardizing variant calling with normalized coverage across the genome outweighed the complexity associated with variabilities related to relative oversampling of individual amplicons.

We then used the same metrics to identify false positives using Method 2 (Figure 2C). Similar to Method 1, the average number of false positives per nucleotide in at least one replicate was related to the number of input templates, with 10^3 input cDNA templates having an average of 1.05×10^{-2} false positives per nucleotide and 10^6 input cDNA templates having an average of 1.52×10^{-3} false positives per nucleotide (Figure 2A, right panel, closed circles, $p < 0.001$, Tukey's multiple comparisons test). When only including false positives detected in at least two replicates, there was no difference in the rate of false positives between 10^3 and 10^6 cDNA templates (Figure 2A, right panel).

We identified 9 nucleotide positions with false positives in at least one replicate of all input template levels using Method 2 (Figure 2C). All of the false positives detected by Method 2 were also detected by Method 1. Since these false positives are present in nearly every sample and this is a clonal virus stock (Figures 2B and 2C), it is likely an artifact of the method rather than true variants, highlighting the importance of validating novel methods with virus stocks of known composition. Additionally, it is important to understand the effects of nucleotide sequence and primer binding sites on false positive detection, as primer slippage may be a confounding factor.

To help determine if the rate of false positives was related to coverage depth, we calculated the frequency of nucleotide sites that had sufficient coverage (a nucleotide depth of at least 1800) for our cDNA and vRNA data sets. There was no significant difference between the percentage of bases with at least 1800x coverage using Method 1 or 2 (Method 1 mean = 76.27% nucleotides over 1800, Method 2 mean = 75.03% nucleotides over 1800; p = 0.95, Mann-Whitney, data not shown), indicating that the differences in false positive frequency are more likely a result of starting template than coverage alone.

Detection of genome-wide variants using multiplex SIV sequencing

We then examined the sensitivity and reproducibility of detecting individual SNVs in SIV by Methods 1 and 2. We used two stock viruses, SIVmac239 and SIVmac239-24x, that differed at 24 nucleotides throughout the entire viral coding sequence (Figure 1b, Table 2). Viral RNA was isolated from these two stocks and quantified with a *gag* qPCR assay. We proceeded with Method 1 by mixing the two stocks of vRNA to a total number of 10^6 copies at the following SIVmac239 to SIVmac239-24x ratios: 100:0, 95:5, 90:10, 75:25, 50:50, and 0:100 (Figure 3A). Each mixture of vRNA was serially diluted to 10^5, 10^4, and 10^3 templates per 11 ul. We also tested Method 2 by first preparing viral cDNA from 10^7 vRNA templates of each of the two stocks, quantifying viral cDNA, and then mixing the cDNA templates to a total of 10^6 templates in the same ratios as the vRNA templates were mixed (Figure 3B). The same quantified vRNA or viral cDNA mixtures were used for the entire experiment.

The remaining multiplex PCR procedures were performed for the different numbers of input templates and for each of the individual ratios. PCR products were tagged, and sequencing was performed on the Illumina MiSeq. FASTQ reads were mapped to SIVmac239 and the frequencies of each individual SNV relative to SIVmac239 were determined as described for the clonal SIVmac239 data.

We compared the observed to the expected variant frequencies for all 24 positions in the genome for both Methods 1 and 2. We generated a linear regression for each number of input templates (Figure 4A) to determine if the relationship between the expected and observed SNV frequency was the same. We did not find a significant difference when we compared the slopes for all four linear regression lines with either Method 1 (p = 0.069, Figure 4A) or Method 2 (p = 0.185, Figure 4B). Notably, all of these data sets had an SNV present at position 9110 (Figure 4A and 4B, open circles) that was consistently detected inaccurately. While there did appear to be a slight increase in observed variant frequency when compared to expected variant frequency, site 9110 was a clear outlier in the data sets (Figure 4C).

To further understand how the number of templates and the type of quantified starting material affects the reproducibility of the detected SNV frequencies, we compared the observed frequencies of each of the 24 individual SNVs across all the data sets. We found that when using Method 2, there was less variability in variant frequencies across the number of input templates when compared to using Method 1. (Figure 5A-F). This observation is consistent with data indicating that the process of reverse transcription is inefficient and variable (Bustin et al. 2015), such that when 10^3 vRNA input templates are used in the assay, it is unlikely that there are actually 10^3 viral cDNA templates available for subsequent PCR. For both input types, it was not surprising that as the number of templates increased, the SNV frequencies tended to be more consistent and reproducible across the genome.

Detection of variants within biological samples

To ensure the that multiplex PCR can be used with biological samples, we sequenced vRNA isolated from plasma and a lymph node (LN) of an SIV+ cynomolgus macaque at the same time point. We diluted quantified vRNA to a starting number of 10^3, 10^4, and 10^5 input templates. Each sample was sequenced in triplicate. Again, we required a variant to be present at a frequency of 1% or greater and with a sample depth of 1800 to be considered a true variant. While we hoped to readily detect the same variants in sequences generated from all three numbers of input templates, we detected a substantial amount of amplicon dropout for the samples starting with 10^3 vRNA templates. Still, we confidently obtained sufficient coverage over at least 80% of the genome from the 10^4 and 10^5 vRNA template data sets, and used these sets for further analysis.

We found 24 SNVs were present in at least 2 out of 3 replicates in each data set (Figure 6). However, as a result of insufficient coverage in some of the replicates, none of the variants shown in Figure 6 were present in all three replicates of each data set. Notably, the frequencies of these 24 SNVs were similar across samples. Similar to what was shown in figure 2B, there was considerable variation in the nucleotide region between 6181 and 6205. Ultimately, we show that variants are detectable consistently at a minimum of 10^4 vRNA input templates, and that variants are similar between the LN and plasma in a given animal, which is consistent with prior reports (Immonen et al, 2020, Vanderford et al 2011).

Discussion

The goal of this study was to adapt a multiplex PCR and sequencing approach (Quick et al. 2017) to sequence SIV from low quality starting material. This would include occasions where SIV is present at low titer or as partially degraded vRNA. Recognizing that different sequencing methods have their limitations, we set out to validate this approach in a series of assays described in this study.

We first sequenced clonal SIVmac239 to determine the false positive rate using both Methods 1 and 2. We found 4 nucleotide sites (1254, 2480, 5428, and 7396) in homopolymeric regions with consecutive adenines where false positive indels were detected. We predict that these insertions were introduced during the PCR step. We also found 8 individual false positives in the stock that were attributed to substitutions consistently present in the same 27-nucleotide region of Amplicon 21, but not in the adjacent Amplicon 20. We hypothesize that these substitutions are specific to the generation of Amplicon 21 and the analysis pipeline, rather than actually being real substitutions.

We also found that sequencing replicates reduced the detection of false positives, particularly when there are low numbers of input templates, consistent with previous results (Grubaugh et al 2019). While we realize that there are not always enough resources available to sequence a sample in duplicate, our data highlights that caution should be taken when interpreting data from a single assay of a sample with low virus titer. Importantly, the process of validating a method with a known clonal virus stock is key to distinguishing between false positives, sequencing error, and true variants. Without doing the validation assays in this study, it would be impossible to know the benefits and technical limitations of using the multiplex PCR approach to sequence virus isolated from animals infected with SIVmac239. Detecting these method-dependent systematic errors by characterizing false positives in a clonal stock is important so that investigators using this method can perform the assay with knowledge of which variants are real and which are technical artifacts.

By mixing SIVmac239 and SIVmac239-24x, we detected variants at a frequency of 5% with as few as 1000 input copies. We opted for this conservative threshold because we already knew that there were some false positives detected when a threshold of 1% was used (Figure 2), and the most relevant variants accumulate over time to a higher frequency. Thus, detection of variants at a frequency of <5% was less critical for broad analyses of SIV population diversity. Future studies that require more sensitive variant detection could address whether variants present between 1% and 5% can be accurately detected.

For these mixing studies, we chose two viruses with variants scattered throughout the genome, with at least one variant present in each gene. This let us determine whether we could effectively detect variants throughout the genome and across a large number of PCR amplicons generated by either Method 1 or 2. We were surprised to find it difficult to interpret the SNV frequency at position 9110. This site lies in a region dense with adenines and guanines which may contribute to some inconsistencies as a result of PCR slippage or misincorporation of nucleotides during PCR amplification (Pfeiffer et al 2018). In addition, the forward primer for Amplicon 33 is one nucleotide different from its complementary sequence in SIVmac239-24x due to the modified nucleotide 9110 present in the SIVmac239-24x sequence. While we did trim primers computationally, this would not prevent PCR error from occurring. As a result, some SIVmac239-24x templates may not be amplified as efficiently because of a single nucleotide difference, which may also lead to amplicon dropout and skewed results.

Throughout our study, we compared the results obtained using Methods 1 and 2, which used quantified vRNA and quantified viral cDNA, respectively. Reverse transcription is inefficient (Bustin et al 2015), so we wanted to determine if there were fewer false positives and more consistent detection of SNVs when quantified cDNA was used as the starting material rather than vRNA. We found the observed SNV frequencies were more similar to expected frequencies when quantified cDNA was used as a starting template (Figure 4b) and, not surprisingly, when increased numbers of vRNA or cDNA templates were used. Even though our quantification of viral cDNA was based only on the copies of *gag*, we found that using quantified viral cDNA as the input improved the reproducibility of variant detection when we mixed two clonal virus inocula at predefined ratios, even when using only 10^3 quantified templates. This observation further raises concerns that using quantified vRNA as starting material gives an overestimation of the number of vRNA templates that are actually converted to cDNA and amplified to yield the reported sequence data.

By testing this method with biological samples, we provide evidence for the use of SIV multiplex PCR in animal studies as a way to characterize viral populations *in vivo*. Unfortunately, we were unable to fully amplify the biological samples with the lowest input titer. This may be a consequence of host RNA or DNA that may have reduced the efficiency of primer binding. Nonetheless, we could still amplify from samples containing 10^4 input vRNA templates isolated from plasma or host cells. We detected variants with similar frequencies that passed our threshold in two of three replicates in each data set (Figure 6). The variants that were not present in all three replicates typically did not have enough coverage in all three replicates to pass our threshold, which is stringent compared to other similar analyses (Grubaugh et al 2019). Importantly, we found the variant frequencies were similar between tissues and plasma, consistent with previous reports (Immonen et al, 2020, Vanderford et al 2011), and that a biologically relevant amount of virus (10^4 total vRNA input templates) is sufficient to provide reproducible variant detection across nearly all of the SIV genome.

Overall, we found that the multiplex PCR approach could be successfully used to generate genome wide sequences of SIV, but our results strongly imply that any new sequencing and analysis methods be validated before using them widely to characterize variant frequency in a virus population. While it was possible to generate sequence data from 10^3 vRNA templates, the use of quantified cDNA was more consistent. Further, although this method could be used to successfully detect SNVs across the genome, we found there were key features in the viral genome that affected the accuracy of the multiplex PCR approach. Thus, while the multiplex PCR method has many advantages for deep sequencing virus populations, validation experiments and visualization of the output alignments are essential for correct data reporting, as expected for any sequencing approach.

Conclusions

Our initial goal of this study was to generate a sequencing approach that was able to characterize viral population diversity with low input templates. Multiplex PCR has been used to accurately sequence other viruses, including Zika (Quick et al. 2017), Dengue, and Chikungunya (Kafetzopoulou et al. 2018), at titers between 10^3 and 10^6 vRNA copies per mL, and most recently with SARS-CoV-2 (CDC et al 2020, Fauver et al 2020) and we were hoping this would extend to SIV. However, many publications fail to state the viral input titer when describing their sequencing methods. We learned that increasing numbers of input SIV templates and utilization of quantified cDNA as a starting material improved reproducibility of variant calling. Further, our data suggests that the multiplex PCR and sequencing approach may not be as sensitive at low numbers of input templates for SIV, when compared to other using low numbers of templates for other viruses. We additionally were able to demonstrate that this model may be used with biological samples and at biologically relevant levels of circulating virus. Most importantly, our study demonstrates the need to validate new sequencing approaches because the same method may not be viable for sequencing all viruses with the same sensitivity and reproducibility. We now understand the limitations of the assay so that experiments can be designed to maximize the likelihood of success and minimize the overinterpretation of data.

Methods

***Primer design*:** Primers were designed using Primal Scheme, as previously described by Quick et al, 2017. FASTA files of SIVmac239 and three consensus sequences of virus populations isolated from animals infected with SIVmac239 were used as the foundation for the Primal Scheme tool. 37 primer pairs (Table 1) were generated to span the entire SIV genome. The lengths of the resulting amplicons ranged from 285bp to 397bp, with an average length of 351bp. The number of overlapping nucleotides for each amplicon ranged from 40bp to 149bp, with an average length of 100bp. Primer pairs were split into two pools to ensure that the amplicons generated within each pool would not overlap. Primer sequences, pools, and concentrations can be found in Table 1. Final concentration of Pool 1 was 35uM and Pool 2 was 24uM.

Isolation of vRNA for sequencing: SIVmac239 and SIVmac239-24x vRNA were isolated from clonal virus stocks. Briefly, 1ml of each virus stock was centrifuged at 13,000rpm for 30 seconds to pellet any cells that were present. Plasma vRNA was obtained from an animal described in {Ellis et al., 2020, #20499}, and lymph node (LN) vRNA was isolated from a necropsy tissue homogenate by adding TriZol and conducting a standard phenol-chloroform RNA extraction. The supernatant of each sample was transferred to a 1.5mL Eppendorf tube and spun at 13,000rpm for 1 hour at 4C to concentrate virus particles. After spinning the sample, we removed all the supernatant, except 200ul of liquid, so as not to disturb the viral pellet. The vRNA was then extracted using the Qiagen MinElute vRNA extraction kit, according to manufacturer’s protocols (Qiagen). Prior to elution, 25uL of Buffer AVE was added directly to the MinElute Column membrane and incubated for 5 minutes.

Preparation of viral cDNAs: The vRNA isolated from the SIVmac239 and SIVmac239-24x virus stocks were each diluted to 10^6 copies/11ul in nuclease-free water. They were mixed at SIVmac239:SIVmac239-24x ratios of 100:0, 95:5, 90:10, 75:25, 50:50, and 0:100. These mixtures were diluted 1:10 in nuclease-free water to generate vRNA template concentration dilution series of 10^6, 10^5, 10^4, and 10^3 templates per 11uL. From each mixture, we used 11ul of vRNA and performed cDNA synthesis using SuperScript IV Reverse Transcriptase (Invitrogen), according to manufacturer’s protocol. For experiments where quantified viral cDNA was used as starting material, approximately 10^7 viral templates were used for cDNA synthesis using SuperScript IV Reverse Transcriptase (Invitrogen), according to manufacturer’s protocol. Viral cDNA and vRNA was then quantified using a *gag* qPCR assay as previously described (Cline et al. 2005).

Multiplex PCR reactions: Each tube of viral cDNA generated from the virus stocks or biological samples was split equally, such that 10uL of viral cDNA was PCR amplified with the two separate primer pools. Amplification was performed with the Q5 polymerase and the following reaction conditions: 98^{o}C for 30 seconds, 35 cycles of 95^{o}C for 15 seconds and 65^{o}C for 5 minutes, and then cooled to 4^{o}C. Products were verified using a 1% agarose gel and were quantified using the Qiagen High Sensitivity DNA kit (Thermo Fisher).

Library Preparation and Sequencing: After the two amplicon pools were generated, 75ng of each pool was mixed to generate a total of 150ng DNA. This pool of PCR products was tagged with the Illumina TruSeq Nano HT kit, according to the manufacturer's protocol (Illumina). Following tagging and purifying, the libraries were quantified using the Qiagen High Sensitivity DNA kit. The quality of each library was characterized with a High Sensitivity DNA kit (Agilent) on an Agilent Bioanalzyer. If unligated adapter dimers were detected at 140bp, an additional bead clean up step was performed. The average tagged library size was approximately 503bp (range 491 to 512). Tagged libraries were pooled at equimolar concentrations and diluted so that the final concentration of DNA molecules per run was 10pM. This diluted pool and 10pM PhiX were denatured with 0.2N sodium hydroxide for 5 minutes at room temperature. Denatured PhiX was then added to the pool at a final frequency of 10 percent. Each pool was loaded at 10pM concentration onto a 500-cycle v2 MiSeq cartridge and sequenced.

***Data analysis*:** FASTQ reads were demultiplexed and then processed using a modified pipeline from David O'Connor's lab, called the Zequencer, with the initial scripts available at https://bitbucket.org/dholab/ and referenced in Dudley et al 2017. All scripts used, and their documentation, can be found on our github repository, https://github.com/SLO-Lab/SIV_MultiplexPCR. Briefly, reads were trimmed, merged, and normalized using bbtools (https://jgi.doe.gov/data-and-tools/bbtools/) and Seqtk (https://github.com/lh3/seqtk). A FASTA file was generated that contained the nucleotide reference sequences for all 37 amplicons, as they would exist in SIVmac239 (Accession: M33262). Up to 2000 merged reads that mapped at low sensitivity to each of the 37 reference amplicons were extracted from the data set. These reads were then aligned to SIVmac239 using NovoAlign (http://www.novocraft.com/products/novoalign/). A pileup file was generated from the BAM alignment. Variants with a frequency of 1% or higher were called by VarScan (https://sourceforge.net/projects/varscan/) and annotated by SNPeff (Cingolani et al. 2012). VCF files were processed and analyzed in R(v3.6.1). Variants with a sample depth less than 1800 were discarded to reduce bias as a result of poor sample depth. Position 9609 codes for a stop codon in the *nef* protein in the M33262 Genbank reference for SIVmac239, but our stock virus is SIVmac239-nef-open, which has a T to G transversion at this position, converting the stop codon (TAA) to a glutamate (GAA) amino acid.

List Of Abbreviations

HIV – Human immunodeficiency virus; SIV – Simian immunodeficiency virus; ZIKV – Zika virus; vRNA – Viral RNA; SNV – single nucleotide variant

Declarations

Ethics approval: Not applicable.

Consent to publication: not applicable.

Availability of data and materials: code used to generate data can be found on the lab's GitHub page (see Data analysis section).

Competing interests: All authors read and approved the manuscript and declare no competing interests.

Funding: Research reported in this publication was supported by the Office Of The Director, National Institutes of Health under Award Number P51OD011106 to the Wisconsin National Primate Research Center, University of Wisconsin-Madison. This work was also supported by an Australian National Health and Medical Research Council (NHMRC) grants 1080001 and 1052979 (to MPD) and NHMRC Career Development Fellowship 1067590 (to VV).

Contributions: The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health. RVM and SLO wrote the manuscript. Primer design and validation by NF and MSS. Data analysis was aided by MPD, TS, and VV, and conducted by RVM and SLO.

Acknowledgements: We would like to thank the Virology Services unit at the Wisconsin National Primate Research Center for quantifying SIV vRNA and cDNA, which allowed us to dilute our samples to predetermined levels. We would also like to thank Josh Quick and Nick Loman for helping us with primer design.

References

1. Adey, A., Morrison, H. G., Asan, Xun, X., Kitzman, J. O., Turner, E. H. et al. (2010). Rapid, low-input, low-bias construction of shotgun fragment libraries by high-density in vitro transposition. *Genome Biol*, *11*(12), R119.
2. Bustin, S., Dhillon, H. S., Kirvell, S., Greenwood, C., Parker, M., Shipley, G. L. et al. (2015). Variability of the reverse transcription step: practical implications. *Clin Chem*, *61*(1), 202-212.
3. Carroll, M. W., Matthews, D. A., Hiscox, J. A., Elmore, M. J., Pollakis, G., Rambaut, A. et al. (2015). Temporal and spatial analysis of the 2014-2015 Ebola virus outbreak in West Africa. *Nature*, *524*(7563), 97-101.
4. Cingolani, P., Platts, A., Wang, L. L., Coon, M., Nguyen, T., Wang, L. et al. (2012). A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w1118; iso-2; iso-3. *Fly (Austin)*, *6*(2), 80-92.
5. Cline, A. N., Bess, J. W., Piatak, M., & Lifson, J. D. (2005). Highly sensitive SIV plasma viral load assay: practical considerations, realistic performance expectations, and application to reverse engineering of vaccines for AIDS. *J Med Primatol*, *34*(5-6), 303-312.
6. Dudley, D. M., Bailey, A. L., Mehta, S. H., Hughes, A. L., Kirk, G. D., Westergaard, R. P. et al. (2014). Cross-clade simultaneous HIV drug resistance genotyping for reverse transcriptase, protease, and

integrase inhibitor mutations by Illumina MiSeq. *Retrovirology*, *11*(1), 122.

7. Dudley, D. M., Newman, C. M., Lalli, J., Stewart, L. M., Koenig, M. R., Weiler, A. M. et al. (2017). Infection via mosquito bite alters Zika virus tissue tropism and replication kinetics in rhesus macaques. *Nat Commun*, *8*(1), 2096.
8. Gambhira, R., Keele, B. F., Schell, J. B., Hunter, M. J., Dufour, J. P., Montefiori, D. C. et al. (2014). Transmitted/founder simian immunodeficiency virus envelope sequences in vesicular stomatitis and Semliki forest virus vector immunized rhesus macaques. *PLoS One*, *9*(10), e109678.
9. Grubaugh, N. D., Gangavarapu, K., Quick, J., Matteson, N. L., De Jesus, J. G., Main, B. J. et al. (2019). An amplicon-based sequencing framework for accurately measuring intrahost virus diversity using PrimalSeq and iVar. *Genome Biol*, *20*(1), 8.
10. Harris, M., Burns, C. M., Becker, E. A., Braasch, A. T., Gostick, E., Johnson, R. C. et al. (2013). Acute-phase CD8 T cell responses that select for escape variants are needed to control live attenuated simian immunodeficiency virus. *J Virol*, *87*(16), 9353-9364.
11. Hassounah, S. A., Mesplède, T., Quashie, P. K., Oliveira, M., Sandstrom, P. A., & Wainberg, M. A. (2014). Effect of HIV-1 integrase resistance mutations when introduced into SIVmac239 on susceptibility to integrase strand transfer inhibitors. *J Virol*, *88*(17), 9683-9692.
12. Henn, M. R., Boutwell, C. L., Charlebois, P., Lennon, N. J., Power, K. A., Macalalad, A. R. et al. (2012). Whole genome deep sequencing of HIV-1 reveals the impact of early minor variants upon immune recognition during acute infection. *PLoS Pathog*, *8*(3), e1002529.
13. Kafetzopoulou, L. E., Efthymiadis, K., Lewandowski, K., Crook, A., Carter, D., Osborne, J. et al. (2018). Assessment of metagenomic Nanopore and Illumina sequencing for recovering whole genome sequences of chikungunya and dengue viruses directly from clinical samples. *Euro Surveill*, *23*(50).
14. Kumar, N., Chahroudi, A., & Silvestri, G. (2016). Animal models to achieve an HIV cure. *Curr Opin HIV AIDS*, *11*(4), 432-441.
15. Lessler, J., Chaisson, L. H., Kucirka, L. M., Bi, Q., Grantz, K., Salje, H. et al. (2016). Assessing the global threat from Zika virus. *Science*, *353*(6300), aaf8160.
16. McInerney, P., Adams, P., & Hadi, M. Z. (2014). Error Rate Comparison during Polymerase Chain Reaction by DNA Polymerase. *Mol Biol Int*, *2014*, 287430.
17. Oikonomopoulos, S., Wang, Y. C., Djambazian, H., Badescu, D., & Ragoussis, J. (2016). Benchmarking of the Oxford Nanopore MinION sequencing for quantitative and qualitative assessment of cDNA populations. *Sci Rep*, *6*, 31602.
18. Poirier, E. Z., & Vignuzzi, M. (2017). Virus population dynamics during infection. *Current opinion in virology*, *23*, 82-87.
19. Potapov, V., & Ong, J. L. (2017). Examining Sources of Error in PCR by Single-Molecule Sequencing. *PLoS One*, *12*(1), e0169774.
20. Quick, J., Grubaugh, N. D., Pullan, S. T., Claro, I. M., Smith, A. D., Gangavarapu, K. et al. (2017). Multiplex PCR method for MinION and Illumina sequencing of Zika and other virus genomes directly from clinical samples. *Nat Protoc*, *12*(6), 1261-1276.

21. Quick, J., Loman, N. J., Duraffour, S., Simpson, J. T., Severi, E., Cowley, L. et al. (2016). Real-time, portable genome sequencing for Ebola surveillance. *Nature*, *530*(7589), 228-232.
22. Sanjuán, R., & Domingo-Calap, P. (2016). Mechanisms of viral mutation. *Cell Mol Life Sci*, *73*(23), 4433-4448.
23. Tyler, A. D., Mataseje, L., Urfano, C. J., Schmidt, L., Antonation, K. S., Mulvey, M. R. et al. (2018). Evaluation of Oxford Nanopore's MinION Sequencing Device for Microbial Whole Genome Sequencing Applications. *Sci Rep*, *8*(1), 10931.
24. Vignuzzi, M., Stone, J. K., Arnold, J. J., Cameron, C. E., & Andino, R. (2006). Quasispecies diversity determines pathogenesis through cooperative interactions in a viral population. *Nature*, *439*(7074), 344-348.
25. Zanini, F., Puller, V., Brodin, J., Albert, J., & Neher, R. A. (2017). <i>In vivo</i> mutation rates and the landscape of fitness costs of HIV-1. *Virus Evol*, *3*(1), vex003.
26. Schirmer, M., D'Amore, R., Ijaz, U.Z. *et al.*Illumina error profiles: resolving fine-scale variation in metagenomic sequencing data. *BMC Bioinformatics* **17,** 125 (2016). https://doi.org/10.1186/s12859-016-0976-y
27. McCoy RC, Taylor RW, Blauwkamp TA, Kelley JL, Kertesz M, et al. (2014) Illumina TruSeq Synthetic Long-Reads Empower *De Novo*Assembly and Resolve Complex, Highly-Repetitive Transposable Elements. PLOS ONE 9(9): e106689. https://doi.org/10.1371/journal.pone.0106689
28. Bangsberg, D. R., Kroetz, D. L., & Deeks, S. G. (2007). Adherence-resistance relationships to combination HIV antiretroviral therapy. *Current HIV/AIDS Reports*, *4*(2), 65.
29. CDC, C. O. V. I. D.-1. R. T., Jorden, M. A., Rudman, S. L., Villarino, E., Hoferka, S., Patel, M. T. et al. (2020). Evidence for Limited Early Spread of COVID-19 Within the United States, January-February 2020. *MMWR Morb Mortal Wkly Rep*, *69*(22), 680-684
30. Fauver, J. R., Petrone, M. E., Hodcroft, E. B., Shioda, K., Ehrlich, H. Y., Watts, A. G. et al. (2020). Coast-to-coast spread of SARS-CoV-2 during the early epidemic in the United States. *Cell.*
31. Hadfield, J., Megill, C., Bell, S. M., Huddleston, J., Potter, B., Callender, C. et al. (2018). Nextstrain: real-time tracking of pathogen evolution. *Bioinformatics*, *34*(23), 4121-4123.
32. Arias, A., Watson, S. J., Asogun, D., Tobin, E. A., Lu, J., Phan, M. V. T. et al. (2016). Rapid outbreak sequencing of Ebola virus in Sierra Leone identifies transmission chains linked to sporadic cases. *Virus Evol*, *2*(1), vew016.
33. Pfeiffer, F., Gröber, C., Blank, M., Händler, K., Beyer, M., Schultze, J. L. et al. (2018). Systematic evaluation of error rates and causes in short samples in next-generation sequencing. *Sci Rep*, *8*(1), 10950.
34. Ellis, AL, Balgeman AJ et al. (2020). MAIT Cells Are Functionally Impaired in a Mauritian Cynomolgus Macaque Model of Siv and Mtb Co-Infection. *PLoS Pathog*5 e1008585.
35. Immonen TT, Camus C, Reid C, Fennessey CM, Del Prete GQ, Davenport MP, Lifson JD, Keele BF. (2020) Genetically barcoded SIV reveals the emergence of escape mutations in multiple viral

lineages during immune escape. Proc Natl Acad Sci U S A. doi: 10.1073/pnas.1914967117. PMID: 31843933; PMCID: PMC6955354.

36. Vanderford TH, Bleckwehl C, Engram JC, Dunham RM, Klatt NR, Feinberg MB, Garber DA, Betts MR, Silvestri G. (2011) Viral CTL escape mutants are generated in lymph nodes and subsequently become fixed in plasma and rectal mucosa during acute SIV infection of macaques. PLoS Pathog. doi: 10.1371/journal.ppat.1002048. PMID: 21625590; PMCID: PMC3098234.

Tables

Due to technical limitations the Tables are available as a download in the Supplementary Files.

Figures

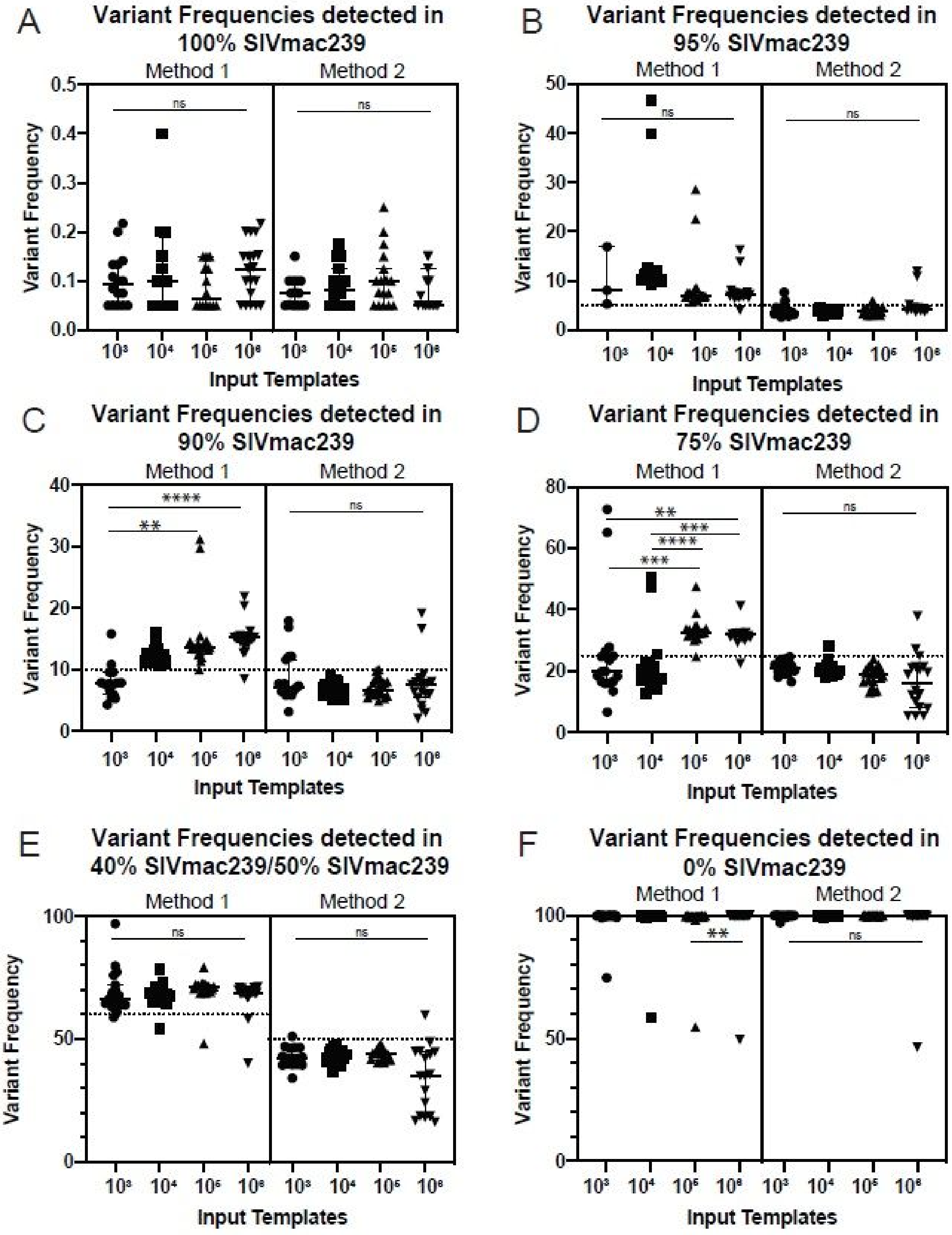

Figure 5

A-F) Observed variant frequency of SIVmac239-24x SNPs in varying SIVmac239:SIVmac239-24x ratios by vRNA (left) or cDNA (right) input templates. Dotted line indicates expected variant frequency. Significance determined by Kruskal-Wallis and designated by asterisks. $p < 0.05$ (*), $p < 0.01$ (**), $p < 0.001$ (***), $p < 0.0001$ (****).

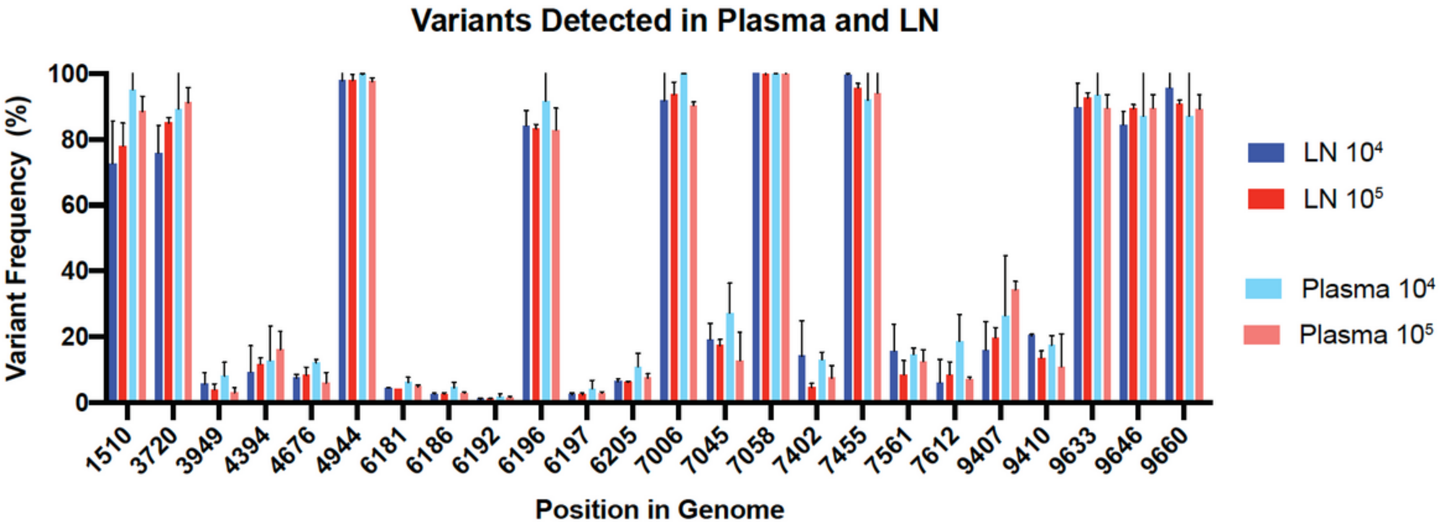

Figure 6

A) Variant frequency of SNVs identified in at least two of three replicates for that input copy and sample type data set. vRNA input template numbers denoted by colors, with darker shades representing LN and lighter colors representing the plasma. Bars indicate 95% CI. All variants shown are present at a frequency of 1% or greater and have a nucleotide depth of at least 1800.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- MoriartyVJResubmissionTable1.pdf