**Benchmarking and quality controls in sequence based studies of intrahost diversity**

John T. McCrone1 and Adam S. Lauring1,2 \*

1 Department of Microbiology and Immunology, University of Michigan, Ann Arbor, MI 48109

2 Division of Infectious Diseases, Department of Internal Medicine, University of Michigan, Ann Arbor, MI 48109

\* Corresponding author

Adam S. Lauring

1150 W. Medical Center Dr.

MSRB1 Room 5510B

Ann Arbor, MI 48109-5680

[alauring@med.umich.edu](mailto:alauring@med.umich.edu)

# Introduction

The advent of next generation sequencing (NGS) has enabled detailed characterization of intrahost diversity and has revolutionized studies of viral evolution and molecular epidemiology. Until recently, investigations of viral diversity were limited by labor intensive sequencing approaches. However, it is now feasible to sequence patient-derived samples at sufficient read depth to detect rare single nucleotide variants (SNV), and the cost-effectiveness of NGS has led to an explosion of studies that quantify viral diversity within and between hosts (e.g.[@Andersen:2015jj], [@Grubaugh:2015by], [@Rogers:2015ij],[@Poon:2016gf]). Given the increasing contribution of NGS to the field of virology, it is important to reflect on the accuracy of these methods and the controls that are needed to ensure robust conclusions.

Because next generation sequencing of viral populations often produces tens of thousands of reads per genomic position, the sensitivity of these approaches has been assumed to be quite high (cite). However, patient-derived samples can differ in input titer by orders of magnitude (e.g. [@Lau:2013ima],[@Teunis:2015iw],[@Takeyama:2015jg]), and this variability can have drastic impacts on the sensitivity of detection for rare events. Furthermore, some library preparations result in variable sequencing coverage both within and between samples. These sources of ascertainment bias can be easily overlooked in NGS based studies of viral diversity.

The specificity of NGS is another factor to consider when designing an experiment using patient-derived samples. Each sequencing platform has inherent error profiles (cite), and sample collection, target amplification, and library preparation are additional processes whereby errors can be introduced and propagated. The next generation sequencing of RNA virus populations, is particularly error prone, and these errors must be accounted for to ensure accurate secondary analysis.

Many sample preparation protocols have been developed to control for errors in NGS-based studies. However, each approach has its own caveats that ultimately limit its application. Cirseq is an ingenious approach in which template RNA is sheared and circularized prior to reverse transcription [@Acevedo:2014ej]. Subsequent rolling circle cDNA synthesis produces tandem reads, generating a consensus sequence for each RNA fragment. While this method is likely to be highly sensitive for rare variant detection and can control for reverse transcription, PCR, and sequencing errors, the requirement for a large and relatively pure population of viral RNA limits its applicability to patient-derived samples [@Acevedo:2014df]. "Primer ID" methods require less input and target sequencing to the viral genome. This approach relies upon barcoded primers to construct consensus sequences for each cDNA template and control for PCR and sequencing errors [@Jabara:2011eo]. Because Primer ID methods require that each bar code be physically attached to a PCR product, they are most easily applied to small targeted regions of the genome. As such, they have limited application in whole genome sequencing.

An adapted form of sequence independent single primer amplification (SISPA) is an alternative approach that allows for whole genome sequencing and controls for errors propagated during library preparation [@Djikeng:2008cd]. In this method RT-PCR products are sheared and tagged with bar-coded random primers in a klenow reaction, prior to library preparation. SISPA controls for any errors that may arise during the library amplification such as PCR biases. This method has been used in conjunction with statistical algorithms to control for accuracy in studies of intrahost influenza diversity (Ghedin group citations). However, the bar-coding reaction used in SISPA is biased and can result in uneven coverage and sensitivity across the genome [@Rosseel:2013dn].

Given the above limitations, many statistical approaches have been developed to distinguish true variants from sequencing errors [@Gerstung:1js; @Isakov:2015gq; @Wilm:2012br; @Macalalad:2012fx; @Koboldt:2009dk]. In general, these algorithms calculate base-specific error rates from various metrics including but not limited to: mapping quality (MapQ), base quality (Phred), strand bias, and sequence context. True variants are identified as those with frequencies exceeding the expected error rate according to some predetermined statistical test. Despite being employed in many NGS based studies of viral diversity (e.g. ), few of these algorithms have been benchmarked using defined viral populations, and to our knowledge none have been tested under conditions that mimic those found in patient-derived samples. The accuracy of such algorithms in the context of NGS studies of patient-derived viral populations is largely unknown.

Here, we use genetically-defined populations of Influenza A virus and variable input titers to indirectly determine the accuracy of rare variant detection in patient-derived samples. We highlight the challenges that accompany NGS-based studies of viral diversity and include a few means for improved the accuracy. This work exemplifies the controls that should be run prior to any NGS based study of viral populations and provides a comprehensive data set for benchmarking other pipelines.

## Methods

*Viruses and cells*

### Madin-Darby canine kidney cells were provided by Arnold S. Monto (University of Michigan School of Public Health) and were maintained in Dulbecco's Modified Eagle Medium (DMEM, Invitrogen) with 10% fetal bovine serum (Gibco and HyClone), 25mM HEPES (Invitogen), and 0.1875% bovine serum albumin (Life Technologies). Influenza A/WSN/33(H1N1) virus was rescued from transfected cells using the 8 plasmid reverse genetic system containing the genomic segments (pHW181-188), a kind gift from Robert Webster (St. Jude's Children's Research Hospital) [@Hoffmann:2002ih]. A biological clone of influenza A/Puerto Rico/8/1934(H1N1) was obtained from ATCC (VR-1469) and the genomic segments were cloned into the pHW2000 reverse genetic system [@Pauly:2015eq].

### *Viral populations*

We extracted viral RNA from infected supernaturants using QIAamp Viral RNA kits (Qiagen 5204) and generated cDNA using Superscript III one-step with HiFi platinum Taq (Invitrogen 12574). PCR products were purified using the GeneJet PCR Purification Kit (ThermoFisher Scientific K0701) according to the manufacturers' instructions.

#### PR8-WSN33 population

For the experiment in Figure 2, WSN33 and PR8 viruses were plaque purified and passaged three times in MDCK cells. We then verified the sequences of the these viruses by Sanger sequencing. Two microliters of RNA template were used to generate cDNA in eight segment-specific one-step RT-PCR with of the following primers : PB2-Forward-JT (5'-GCAGGTCAATTATATTCAATATGGAAA-3'), PB2-Reverse-JT (5'-CAAGGTCGTTTTTAAACTATTCGACAC-3'), PB1-Forward-JT (5'-GCAGGCAAACCATTTGAATGG-3'), PB1-Reverse-JT (5'-CAAGGCATTTTTTCATGAAGGACAAG-3'), PA-Forward-JT (GCAGGTACTGATTCAAAATGGAAG-3'), PA-Reverse-JT (CAAGGTACTTTTTTGGACAGTATGG-3'), NA-Forward-JT 5'-(GCAGGAGTTTAAATGAATCCAAACC-3'), NA-Reverse-JT (5'-CAAGGAGTTTTTTGAACAAACTACTTG-3'), HA-Forward-JT (5'-GCAGGGGAAAATAAAAACAACCAAAAT-3'), HA-Reverse-JT (5'-CAAGGGTGTTTTTCCTTATATTTCTGAA-3'), NP-Forward-JT (5'-GCAGGGTAGATAATCACTCACAG-3'), NP-Reverse-JT (5'-CAAGGGTATTTTTCTTTAATTGTCGTACT-3'), M-Forward-JT (5'-GCAGGTAGATATTGAAAGATGAGTC-3'), M-Reverse-JT (5'-CAAGGTAGTTTTTTACTCCAGCTCT-3'), NS-Forward-JT (5'-GCAGGGTGACAAAGACATAATG-3'), NS-Reverse-JT (5'-CAAAGGGTGTTTTTTATTATTAAATAAGCTG-3'). Reaction conditions were (60 min), (2 min), followed by 30 cycles of (30 sec), (30 sec), and (3 min). Molar equivalents of each PCR product were pooled to generate reconstituted cDNA genomes of WSN33 and PR8. The WSN33 cDNA pool was then serially diluted into the PR8 cDNA pool to yield WSN33-PR8 mixtures in which WSN33 made up 5, 2.5, 1.25, 0.63, and 0.16% of the population. Seven hundred and fifty nanograms of each mixture were sheared to an average size of 300-400 bp using a Covaris S220 focused ultrasonicator with the following settings : Intensity: 4, Duty cycle: 10%, Burst/second: 200, Duration: 80 seconds. Sequencing libraries were prepared from these fragmented products using the NEBNext Ultra DNA library prep kit (NEB E7370L), Agencourt AMPure XP beads (Beckman Coulter A63881), and NEBNext multiplex oligonucleotides for Illumina (NEB E7600S). Pooled libraries were sequenced on an Illumina Miseq machine using 2 x 250 paired end reads. A clonal plasmid control library was prepared from 8 plasmids containing PR8 genomic segments. These plasmids were mixed to equal molarity, and cDNA was generated using a multiplex one step RT-PCR with the primers Uni12/Inf1 (5'-GGGGGGAGCAAAAGCAGG-3'), Uni12/Inf3 (5'-GGGGGAGCGAAAGCAGG-3'), and Uni13/Inf1 (5'-5CGGGTTATTAGTAGAAACAAGG-3') as in [@Anonymous:A25Ka2tY]. This library was prepared in identical fashion to the experimental populations and was sequenced in the same Miseq lane.

#### Experimental intrahost population

Twenty point mutants were generated in the WSN33 background using the pHW2000 reverse genetics system [@Hoffmann:2002ih] (Visher et al., manuscript in preparation). In short, we used overlap PCR mutagenesis to introduce the following mutations: HA\_T1583G, HA\_G1006T, HA\_G542T, M\_T861G, M\_A541C, NA\_G1168T, NA\_C454T, NP\_A454C, NP\_A1160T, NS\_G227T, NS\_A809G, PA\_T964G, PA\_T237A, PA\_A1358T, PB1\_G599A, PB1\_G1764T, PB1\_T1288A, PB2\_A1854G, PB2\_A440T, PB2\_A1167T. Viruses were rescued from transfected cells as in (Pauly)

We passaged the 20 WSN33 point mutants and the WSN33 WT once in MDCK cells and verified the identity and relative clonality of the mutants by sequencing each on an Illumina Miseq as above. We quantified the genome copy number of each supernant using a SuperScript III Platinum One-Step qRT-PCR kit (Invitrogen 11732) and univeral influenza A/B primer and probe sets [@Anonymous:tc]. Equal genome equivalents of each infected supernatant were mixed and diluted to generate a population with each mutant present at 5% frequency and a total concentration of 105 genomes per microliter. We diluted this mixture into WT WSN33 supernatant to create populations in which each mutant was present at 2, 1, 0.5,and 0.2% frequency all with a total concentration of 105 genomes per micoliter. These 5 populations were diluted serially into basal media to generate samples with total nucleic acid concentrations of 104 and 103 genomes per microliter. Viral RNA was extracted from these samples and cDNA was generated in a one step multiplex RT-PCR as above. The WT WSN33 sample (105 genomes per micoliter) was processed and sequenced in duplicate. We prepared libraries as before and used Quanti PicoGreen dsDNA quantification (ThermoFisher Scientific P7589) to quantify the concentration of each indexed library. We pooled equal quantities (by nanogram) of each indexed library, and removed adaptor dimers by gel isolation with the GeneJET Gel Extraction Kit (ThermoFisher Scientific K0691) prior to sequencing an on Illumina Hiseq 2500 with 2 x 125 paired end reads. A clonal control library was processed in an identical fashion starting from an equimolar mix of 8 plasmids containing the WSN33 genomic segments.

#### Duplicate populations

Fresh RNA was isolated from the 5%, 2%, 1%, and 0.5% 104 genomes per microliter samples and processed in duplicate as above, and these duplicates were sequenced with a plasmid control using 2 x 125 paired end reads on an Illumina Hiseq 2500 machine.

### *Sequence analysis*

Reads were aligned to either a PR8 or a WSN33 reference sequence using Bowtie2 [@Langmead:2012jh]. Alignments were sorted and PCR duplicates removed using Picard (<http://broadinstitute.github.io/picard/>). Variants were called using either DeepSNV [@Gerstung:1js] or Lofreq [@Wilm:2012br] and filtered using the Pysam module in python (<https://github.com/pysam-developers/pysam>) and custom R scripts. Bases with a Phred <30 were masked in the DeepSNV analysis. We connected all of these steps into an analysis pipeline using bpipe [@em:2012wd] which is available for download at <https://github.com/lauringlab/variant\_pipeline>. To save memory during SNV processing, only variants with p <0.9 were included in our ROC curve analysis, as the vast majority of true negatives are trivial to identify and have a p=1. For ease of viewing, and to account for this analysis artifact, we extended the ROC curves horizontally from the last observed change in sensitivity. All raw fastq files were submitted to the Sequence Read Archive (SRA) under accession number \_\_\_\_\_\_\_\_\_\_\_, and all commands required to generate the figures in the manuscript are available for anonymous download at <https://github.com/lauringlab/Benchmarking\_paper> . Finally an interactive shiny app of our benchmarking work can be downloaded at <https://github.com/lauringlab/benchmarking\_shiny>.

## Results

The ability to reliably identify single nucleotide variants (SNV) is integral to accurate NGS-based studies of viral diversity. The accuracy of any SNV calling pipeline can be described in terms of its sensitivity and specificity. Sensitivity is the proportion of true variants that are properly identified, and specificity is the proportion of true negatives that are properly identified. A closely related concept is the false positive rate, which is 1-specificity, and gives the ratio of false positive variants to true negatives. A sensitivity and specificity of 1 describe an assay with perfect accuracy, in which all the true variants are found and only true variants are found.

There is an obvious trade-off between sensitivity and specificity. Improved sensitivity often requires less stringent criteria in variant calling, but reduces specificity. Conversely, increased stringency can improve specificity but often reduces sensitivity. This relationship can be visualized using a receiver operating characteristic (ROC) curve (Figure 1). An ROC curve plots the sensitivity of an assay along the y-axis and 1-specificity, or false positive rate, along the x-axis. A variant calling pipeline must be tested against known data in order to construct an ROC curve. The outcomes can then be stratified according to a metric that quantifies the probability that a given variant is real, often a p value or quality score. In a controlled benchmarking experiment, all true variants are known, and the sensitivity and specificity can be calculated at different cut-offs (Figure 1A). These points are then used to construct the curve (Figure 1B). A perfect ROC curve in which all the true positives can be separated from all false positives is a right angle that follows the upper left perimeter of the plot.

### Initial accuracy

A comprehensive comparison of SNV calling approaches is beyond the scope of this work. Instead, we robustly benchmark one variant caller, DeepSNV, and highlight approaches for improving its accuracy. In doing so we demonstrate the importance of robustly benchmarking any variant calling method under the experimental conditions to which it is applied.

DeepSNV is a variant calling algorithm that uses a clonal, plasmid-derived, control to estimate local error rates across the genome [@Gerstung:1js]. Because it is clonal, the sequence of the control is known with a high degree of confidence, and any nonconsensus base is indicative of an error. Additionally, the control and experimental samples are processed together and are assumed to have identical noise characteristics. DeepSNV then applies a hierarchical binomial model at each genomic position and identifies true variants as those with frequencies significantly above the noise found in the plasmid control. Like many variant calling algorithms, the accuracy of DeepSNV was initially determined using samples that required minimal PCR amplification. However, its accuracy has not been tested when applied to whole genome sequencing of a viral population amplified by reverse transcription PCR (RT-PCR).

In our first benchmarking data set we created defined mixtures of two influenza strains, WSN33 and PR8. We plaque purified and expanded WSN33 and PR8 populations. Complementary DNA from both viruses were mixed serially such that WSN33 cDNA was present at frequencies of 5, 2.5, 1.25, 0.63 and 0.16% (Figure 2A). These viruses differ at 491 positions (primer sites used in RT-PCR were excluded from analysis), providing 491 true positives in each dilution. On plasmids subjected to limited PCR, DeepSNV identified known variants at 0.1% frequency with a sensitivity of 0.860 and specificity of 1.0 [@Gerstung:1js]. Under our experimental conditions and with a slightly more stringent p value (p=0.01 vs. p=0.05) we found a moderate reduction in sensitivity (0.830 for variants at 0.63% and 0.140 for variants at 0.16%). The specificity was above 0.998 in all dilutions. While this drop in specificity appears small, a false positive rate of 0.002 corresponds to 78 false positives when applied across the 39,171 potential variants present in the 13057 bp of the influenza genome.

### An experimental intrahost population

Although the initial benchmarking experiment validated our ability to accurately detect rare variants in influenza populations, the experiment was run under relatively artificial conditions. Patient-derived populations are typically less divergent than WSN33 and PR8 (cite Ghedin papers), and the number of viral genomes in patient-derived samples is much lower than found in cell culture. To mimic these conditions, we generated 20 viral clones, each with a single point mutation in the WSN33 background. We sequenced these mutants on the Illumina Miseq platform to account for any additional mutations that might have arisen between transfection and the passage 1 stock. Four additional mutations were found above 1% frequency (frequencies 1.2%-3.7%). We also determined the genome copy number of each stock using quantitative RT-PCR. We then mixed equal genome equivalents of these 20 viruses, to generate an artificial population with copies per microliter with each mutation present at 5% frequency. This population was serially diluted into a stock of WSN33, generating artificial populations with all 20 mutations present a 2, 1, 0.5, 0.2, and 0.1% frequency (Figure 3 A). We then serially diluted these populations into basal media to obtain mixtures with lower nucleic acid input. The range ( copies per microliter) matches the inputs typically found in patient-derived influenza virus samples (data not shown). We sequenced the samples on the Illumina Hiseq platform, and called variants using DeepSNV. We also processed and sequenced the WSN33 stock in duplicate to control for any mutations in the viral dilutent.

The 20 mutations present in our initial viral mixture were the only true positives considered in our analysis. Any SNV that was present at >1% frequency either in both duplicates of the wild type stock or any one of the viral clones was excluded from the analysis and considered neither a true positive nor a true negative.

In these populations with lower diversity and input titer, we maintained greater than 0.85 sensitivity for SNV at 1% frequency or higher. Despite a high depth of coverage (>10,000 reads per bp), our sensitivity was considerably lower for variants at or below 0.5% frequency (Figure 3B and C). Therefore, the drop in sensitivity is most likely due to the 1,000-fold decrease in nucleic acid concentration compared to the first data set, and the fact that library preparation requires a number of sampling steps that may limit detection. During library preparation, a disproportionate amount of the 2.0% sample was lost compared to the others, which likely accounts for the reduced sensitivity relative to the 1.0% sample.

The specificity was greater than 0.99 in all the samples with genomes per microliter. As above, while 0.99 specificity appears robust, a false positive rate of 0.01 results in over 200 false positive variants when applied to the 40,123 potential variants in the influenza genome. The false positives outnumber the true positives by 100-fold in these populations with realistic diversity and input. We were able to increased our specificity by applying a more stringent p-value cutoff. However, as shown by the ROC curves in Figure 3, this move towards the y axis markedly reduces sensitivity. Our data demonstrate that with moderate concentrations of input nucleic acid, even statistically significant p-values from a robust variant caller are not sufficient to accurately separate true from false positive variants.

### Additional filtering criteria

Many next generation sequencing studies utilize mapping quality (MapQ) and/or base quality (Phred) cut-offs to ensure that only the highest caliber sequencing data is used to call variants. Mapping quality measures the probability that a given read is mapped to the correct position in the genome while base quality estimates the likelihood that the base call by the sequencer is correct. In the above analysis we masked bases that had a Phred score less than 30 (0.001 probability of being incorrect) and did not apply any MapQ cut-offs. In our next analysis, we applied frequently employed cut-offs such as a MapQ of 20 and a Phred of 30 to our data. These criteria were unable to distinguish true from false positives in our samples (Figure 4A) and indicate that many false positives occur on well mapped reads with high quality base calls.

We further parsed our false variant calls by locating them within individual sequencing reads. It is well known that sequence quality drops near the end of a read (cite?), and we found that our false positives clustered at the termini of our paired end reads (Figure 4B). The average Phred score of these false positives was 37.1, further demonstrating that filtering on quality score alone is insufficient. In contrast, true positives were uniformly distributed across the reads resulting in an average read position near the middle of the read.

Based on these results, we applied a number of empirically determined cut-offs and markedly improved our specificity to >0.999 without sacrificing sensitivity (Figure 4C and 4D). For a variant to be considered in our analysis we required a mean mapping quality of 30, a mean Phred score of 35, and an average read position within the middle 50% of the read. Under these conditions we found 10 or fewer false positives in 4 of the 5 samples. Given this success we applied a number of other strategies to further increase our accuracy, including but not limited to: Benjamini Hochberg p-value correction, more stringent p values (<0.01) or frequency cut-offs (>0.2%), retention of duplicate PCR reads, trimming the ends of the influenza genome, and employing alternative statistical distributions to estimate the error rate in the control sample. None of these significantly improved our accuracy over the above quality and read-position criteria.The impact of various filtering criteria on our data can be visualized in an interactive shiny application available for download at xxx.

We also benchmarked the accuracy of our DeepSNV pipeline in estimating the frequency of the true positive variants (Figure 5). Although the median of the measured frequencies match the expected values, we found substantial spread in each sample and the overall the fit was quite poor ( R2=0.6508). The mean percent difference between the measured frequency and the expected was 41%. This error is likely due to stochastic bias (cite?) in the initial RT-PCR reaction and should be kept in mind when employing downstream analyses that depend on frequency measurements (e.g. variant fitness, haplotype reconstruction, Shannon entropy, and other diversity metrics).

### Alternative variant callers

DeepSNV is one of many variant callers that employ a combination of empiric and statistical approaches to model error rates. We asked whether the decreased accuracy observed in our data set was due simply to peculiarities specific to DeepSNV. We therefore analyzed our input populations using Lofreq, another variant caller that is commonly used in next generation sequencing studies, and has been reported to have perfect specificity [@Wilm:2012br]. Under our experimental conditions, Lofreq had marginally reduced sensitivity compared to DeepSNV when applied to the variant frequencies 1.0% but marginally increased sensitivity when applied to variant frequencies 1.0% (figure 6). The specificity of LoFreq was comparable to what we observed with DeepSNV in our high-input cell culture-derived populations (Figure 2), and better than DeepSNV in our initial implementation (Figure 3 prior to Phred, MapQ, and read position filtering). This increased specificity is most likely due to the fact that the LoFreq algorithm already takes MapQ and Phred scores into account when calling variants and has a stringent strand bias filter that removes many of the variants found only at one end of a paired-end read. However, even with these additional characteristics, the specificity of LoFreq was lower than our improved DeepSNV pipeline (compare Figure 6 and Figure 4), with over 40 false positives per sample. It appears that higher than expected false positive rates are not specific to DeepSNV and most likely plague many variant callers applied to patient-derived viral samples.

### Accuracy at lower input levels

Host-derived viral populations vary in copy number and titer by several orders of magnitude. This variability can be attributed to a variety of factors including, but not limited to: collection site, ease of of nucleic acid isolation, the presence of host nucleic acid, efficiency of library preparation, and host and viral factors. To ensure accuracy across a range of input levels, we diluted our experimental populations serially into basal media (Figure 3A) and identified variants using our modified DeepSNV analysis pipeline (Figure 7). As expected, our sensitivity was lower in populations with fewer genomes. For example, a variant with a 0.5% frequency in a genomes per microliter sample is expected to be present on only 700 genomes in the initial RT-PCR. Many of these will be lost due to bottlenecks in the amplification and library preparation process. We also found reduced specificity in samples with lower starting input. The increase in false positives is presumably due to a greater dependence on RT-PCR amplification and consequent propagation of errors. These data highlight the importance of controlling for input levels when comparing diversity across experimental samples.

In most cases, RT-PCR errors should be sporadic and randomly distributed across the amplified region. If RT-PCR errors are responsible for the reduced specificity found at lower input levels, they should be easily identified as variants present in only one of two RT-PCR reactions performed on the same RNA. To test this hypothesis, we sequenced duplicate RT-PCR reactions of the 5, 2, 1, and 0.5% variant frequency samples from our genomes per microliter collection. The duplicates were processed separately, but sequenced on the same lane of an Illumina Hiseq. We applied the previously mentioned quality cut-offs and required that a given variant be found in both duplicates. By analyzing samples in duplicate, we reduced the number of false positives in each sample to 10 or fewer resulting in a specificity of > 0.9998 (Figure 8). This increased specificity was not accompanied by a decreased sensitivity. In fact, we found a slight increase in sensitivity, most likely due to variability in library preparation. Thus, accurate analysis of low input samples can be achieved through duplicate RT-PCR and careful benchmarking experiments.

### Sub-optimal SNV identification confounds diversity measurement

NGS of intrahost populations is commonly used determine the interaction between host or environmental factors and viral diversity. Because measurements of viral diversity rely entirely on SNV identified in NGS data, they are very sensitive to the accuracy of these variant calls. To illustrate this problem, we calculated the diversity of our 105 genomes per microliter samples at each step of our benchmarking process using three complementary metrics (Table 1). Richness is the count of nonconsensus variants present in a population (often referred to as intrahost SNV). Shannon's Entropy is a diversity metric that accounts for both the number of variants present (richness) and their frequencies (evenness). Because our data are unphased (i.e. no haplotypes), we have reported the average entropy per nucleotide position (cite). The last metric, L1-norm, is a distance measurement that describes how similar two populations are to one another. Identical populations will have an L1-norm of 0. To mimic experimental conditions, we included all variants identified in each analysis regardless of whether or not subsequent benchmarking distinguished them as true or false positives. We found that the accuracy of the SNV calling method has a profound effect on measurements of diversity. It is clear from the richness measurements in Table 1 that the number of false SNV (i.e. the specificity) largely determines the accuracy of the down stream anlaysis. Thus, our adapted DeepSNV protocol, which was able to distinguish between true and false SNV with the highest accuracy, gave the most accurate measures of diversity, followed by Lofreq and the default version of DeepSNV.

## Discussion

Robust benchmarking is essential in NGS-based studies of viral diversity. Differences in experimental design and sample preparation can lead to wide variability in the accuracy of SNV identification. We that found input nucleic acid concentration, which can vary greatly in patient-derived samples, had a large impact on both the sensitivity and specificity of rare variant detection. At moderate levels of nucleic acid input we could improve accuracy by filtering putative SNV based on quality metrics and read position. We further improved our accuracy at low input levels by processing these samples in duplicate. While our quality cut-offs may not be universally applicable to all samples and variant callers, our data suggest that experimental design is critical for accurate SNV detection. These findings are important as few, if any, variant callers have been benchmarked under patient-derived conditions. Finally, we show that these inaccuracies in SNV calling drastically impact downstream analysis.

We chose DeepSNV for our studies, because it is one of the only variant callers that has been benchmarked on real sequencing reads in which all true positive variants and their frequencies were known *a priori* and independent of NGS (cite Gerstung). In our experimental conditions, we found decreased sensitivity and specificity when applied to cell culture-derived mixtures of WSN33 and PR8. The most striking drop in sensitivity came for variants found at 0.16%. Under these conditions we found true variants at a rate of 0.1426 compared the expected sensitivity of 0.860 for variants at 0.1%. This decrease is most likely due to the fact that we are looking for variants across 13057bp of the influenza genome where previous benchmarking focused on just 1,500 bp of the HIV pol gene. Therefore, our sequencing efforts were diluted across the influenza genome resulting in decreased read depth, which limits sensitivity. Additionally, Because we aligned all reads to the PR8 genome, many of the WSN33-derived reads may have not aligned as well, biasing the analysis against detection rare WSN33 SNV. We also observed a relatively modest decrease in specificity. Despite the use of bonferronni corrected p values, the drop in specificity is most likely due to multiple testing errors caused by applying the algorithm across 13057bp (cite).

A key strength of our study is that we applied DeepSNV to experimental populations that more closely mimic the diversity and viral genome inputs found in patient-derived samples of influenza virus. At a modest input concentration of 105 genomes per microliter and “default” DeepSNV settings, we found that false positives can outnumber true positives by a factor of 4. It should be noted that our specificity in all cases remained above 0.99. When applied across an entire viral genome, a specificity of > 0.9995 is required to reduce false positives to reasonable levels. This method undoubtedly introduced unavoidable PCR biases into our populations. Gerstung *et. al.* have shown that PCR amplification biases lead to increased variability in the subconsensus nucleotide counts (cite DeepSNV documentation). Increased variability in the test samples increases the likelihood that an erroneous SNV will exceed the error rate estimated in the plasmid control. DeepSNV attempts to control for such variability in part by estimating the dispersion of a sample, which allows the algorithm to incorporate greater variance into the null model. In fact, our 20 mutant libraries were more dispersed than those in cell culture, indicating a need for greater variance. However, the dispersion estimates are limited and the variability likely exceeded what can be controlled by this approach.

Because DeepSNV is somewhat agnostic toward the mapping quality and base quality of a given variant, we sought to improve our accuracy by identifying cut-offs that more effectively distinguished the true positives. Here we found that commonly used thresholds (MapQ >20, Phred >30) did little to improve accuracy. In our datasets, the distributions of average MapQ and Phred scores of putative SNV were clearly bimodal, with true SNV found in the higher of the two distributions. Our empirically determined quality thresholds were chosen to eliminate any putative SNV found in the lower of the two distributions. Importantly, we found no correlation between MapQ and Phred (p>x), suggesting that these metrics are independent and that the two cut-offs eliminate distinct sets of false positives. These cut-offs should be reproducible, as we have sequenced over 300 influenza libraries in 5 Hiseq runs and have observed consistent MapQ and Phred quality distributions. We have seen the same bimodal trend in libraries of poliovirus populations, but have applied a lower empirically determined MapQ cut-off. This difference is most likely due to differences in the genomic structures of the viruses. The sequence conservation at the termini of all 8 influenza genomic segments, decreases the MapQ of reads that map to these areas. Therefore, while our MapQ and Phred thresholds are robust in our system, they may need to be adjusted for use in others.

Even in the face of stringent MapQ and Phred cut-offs, we found many high-quality false positive SNV were identified only at the termini of paired end reads. We removed these by filtering putative SNV calls based on their average position in a paired end read. These false positives were found almost exclusively in regions of the genome that were enriched for read start sites. This enrichment may be a consequence of sequence context, the fragmentation process, and our size selection protocol. We suggest that there might also be a biological reason for this effect, as our PCR-amplified plasmid control samples did exhibit this bias. For example, defective interfering particles, which commonly arise during cell passage, contain truncated genomic segments and would only be present in infected supernatants and not a plasmid control. These "read position" false positives were reproducibly found across a number of libraries

While read filtering and trimming are common in NGS datasets, we have taken a slightly different approach in our analysis. In the initial SNV identification step we masked bases with Phred <30 but made no additional restrictions on the raw data. We only imposed additional quality restrictions after putative SNV - those that exceed the frequency expected given a plasmid control - were identified. Our approach treats variant nucleotides more stringently than consensus base calls. We do not think that this differential stringency introduces unnecessary bias, Because we identified specificity as the major problem in accurate SNV identification, stringent filtering of potential false positives seems appropriate. Furthermore, the vast majority of reads call a consensus base and our mean quality score thresholds would therefore not be expected to remove many consensus base calls from the analysis.

Frequency thresholds of 0.1 to 1% are an additional quality filter that is often applied to SNV after identification. We did not apply direct frequency thresholds in our analysis, as we found that arbitrary cut-offs limited sensitivity without improving specificity. Read depth, or coverage, is another metric that is often used in conjunction with frequency to ensure accurate SNV identification. Although we did not apply a direct coverage cut-off, DeepSNV has been reported to require a coverage of ten times the reciprocal of frequency for sufficient power to call SNV. Therefore, a coverage of 1,000x is needed to detect a variant at 1% frequency. In our analysis the lowest coverage for a true positive was 1795 (4.8% frequency) while the lowest coverage for a false positive variant was 966 (8.5% frequency). If a given dataset has variability in read depth across the genome, SNV at identical frequencies may be detected with differing sensitivity. The application of variant frequency or read depth thresholds would lead to severe ascertainment bias in subsequent analyses of diversity.

Few studies of intrahost diversity quantify or control for the number of genomes in a sample. We found that input copy number is a key factor in variant detection. Despite high accuracy at 105 genomes per microliter we observed a decrease in sensitivity in our 103 genomes per microliter samples. More importantly, this drop in sensitivity was accompanied by reduced specificity. At lower nucleic acid concentrations, our SNV pipeline relied more heavily on RT-PCR amplification, which tends to propagate errors that are otherwise indistinguishable from true positives in sequence data. We were able to limit these sporadic and random errors by processing low input samples in duplicate. Quantifying and controlling for RT-PCR errors in this way will allow us to accurately compare patient-derived samples that vary over a range of input.

Many variant callers are benchmarked on simulated datasets, plasmids, or PCR products, and may not have comparable sensitivity and specificity when applied to viral samples. Our goal is not to compare the strengths and weaknesses of a few algorithms, but rather to highlight how accuracy can be experiment-specific. We recognize that tsome variant callers may perform better, and that some may beto. Hsuggestsinherentthat essential. We have been able to greatly improve the accuracy of DeepSNV under our experimental conditions, and we are now equipped with an understanding of the limitations of our method.

Our study highlights previously under-recognized issues in the variant calling and suggests factors that should be considered in future studies of viral diversity. The need for target amplification, the structure of the viral genome, and variation in input genome copy number may lead to errors specific to given experiment. We have shown that these errors can have profound effects on downstream analyses and are especially important in comparative studies of intrahost diversity. In studying influenza, we have benefited from an easily accessible reverse genetics system that made for a straightforward approach to benchmarking. We have made our datasets and code available to the community so that others may benchmark their own pipelines or improve on our work. Next generation sequencing is clearly a powerful tool that is just beginning to unlock the complex dynamics of intrahost viral populations. NEED AM UPLIFTING FINAL SENTENCE.

**Acknowledgements**

We thank Pat Schloss, Mike Imperiale, and Robert Woods for helpful discussion and a critical reading of the manuscript. This work was supported by a Clinician Scientist Development Award from the Doris Duke Charitable Foundation and R01 AI118886, both to ASL. JTM was supported by the Michigan Predoctoral Training Program in Genetics (T32GM007544).

1. could duplicates have been pooled and sequenced as one sample?
2. Comparision to Lofreq and Consequences of errors
3. our benchmarking shows that DeepSNVs means of distinguishing true from false variants is partitially breaks down under conditions that mimic patient derived samples.
4. This can be overcome with the addition of specific quality filters.
5. We would not have known about these inaccuracies had we not run our benchmarking experiment.
6. Do other variant callers have similar inaccuracies.
7. Yes Lofreq, although more accurate than our unmodified DeepSNV pipeline, had decreased specificity than had previously been reported.
8. The initial Lofreq sensitivity and specificity metrics were based largely on a *in silico* data sets.
9. Lofreq did not display this high accuracy in under our experimental conditions.
10. While we did not robustly attempt to improve the Lofreq accuracy we did find that frequency threshold could improve its specificity.
11. Experimental conditions matter, regardless of variant calling method and so accurate study requires controls to be run.

# References