Sequencing Complete Genomes of RNA Viruses With MinION Nanopore: Finding Associations Between Mutations

Noam Harel*, Moran Meir*, Uri Gophna, Adi Stern

School of Molecular Cell Biology and Biotechnology, Tel-Aviv University, Tel-Aviv, Israel

* These authors contributed equally

One of the key challenges in the field of RNA virus genetics is the inference of full haplotypes from next generation sequencing data. The MinION Oxford Nanopore sequencer allows sequencing very long reads, with the potential of sequencing the complete genomes of RNA viruses in individual reads. However, MinION suffers from high error rates, rendering the detection of true viral mutations very difficult. Here we propose a new statistical approach to differentiate between true mutations and sequencing errors from direct RNA sequencing using MinION. Our strategy relies on the assumption that sequencing errors will be dispersed randomly along sequencing reads, and hence will not be associated with each other, whereas real mutations will display a non-random pattern of association with other mutations. We demonstrate our approach using direct RNA sequencing data from an evolved population of the MS2 bacteriophage, whose genome length of 3,569 base pairs makes it ideal for MinION. Our results pinpointed several mutations in the phage genome that were corroborated using parallel Illumina sequencing, allowing us to reconstruct associations between mutations. Our approach is amenable to long read sequencing data from any organism for accurate detection of *bona fide* mutations.

Introduction

One of the major challenges of viral genetics today is the characterization of intra-host genetic diversity in a population of viruses. This diversity is generated by a variety of different mechanisms. In particular in RNA viruses with high mutation rates, nucleotide substitutions and point insertions or deletions (indels) are highly prevalent (*1*). Longer indels may also occur (*2*), as may events of genetic recombination. The resulting viral genetic diversity has significant implications for epidemiology, molecular and serologic diagnosis, pathogenesis, and therapeutic management (*3-6*). Thus, new molecular methods are continuously developed to improve the characterization of this high intra-host viral diversity. The availability of second-generation DNA sequencing technologies (*7*), with the Illumina platform currently at the forefront, has made the sequencing of virus genomes conventional, allowing in depth characterization of viral genomic variability.

Illumina-based sequencing allows the detection of minor variants that the standard Sanger-based method often missed. However, Illumina short-read sequencing technology all share one major limitation, the short length of each read, which typically ranges between 75 and 600 bp (for a paired end read). This means that a complete viral genome sequence cannot be obtained in a single read, impairing the ability to link distant mutations in an individual genome and making it difficult to identify recombinants. Another Illumina limitation is that RNA cannot be sequenced directly. During library preparation, RNA is transcribed into cDNA and amplified by PCR. This creates multiple problems that have been extensively discussed but not resolved (*8*): first, PCR may introduce errors during early stages of amplification that will be carried on to later stages (*9-11*). Second, some molecules may be preferentially amplified over others, a term known as PCR bias. Third, PCR and reverse transcription reactions can often induce recombination between different sequences (*12-14*). Together, these problems make the inference of haplotypes from PCR-based libraries that are sequenced with Illumina extremely limited.

Currently, single-molecule third-generation sequencing systems, such as Oxford Nanopore technologies, provide a promising alternative for sequencing full-length single viral genomes(*15*). In fact, these technologies now allow directly sequencing either DNA or RNA. The long reads afforded by these methods have the potential to allow for the inference of the entire genome of a typical RNA virus, whose genome is often less than 10,000 bp (*16-19*). However, one of the major shortcomings of the third generation technologies are their relatively high error rates, with the proportion of errors on a

read often exceeding 10% (*20, 21*). This high error rate makes the detection of true single-nucleotide variants difficult. Here, we devised a simple statistical procedure that allows weeding out real mutations from technical errors. We focused our analysis on the MS2 bacteriophage, an extremely small (3,569 bases) and fast evolving +ssRNA virus that is highly amenable to direct sequencing with Oxford Nanopore MinION. We sequenced virus populations in parallel using both MinION and Illumina, allowing us to corroborate the inferences of our method. This then allowed us to directly infer relationships between mutations and to deduce the entire genome sequence of viruses in the population.

Results

We set out to perform direct RNA sequencing of two evolved populations of MS2 using the two technologies. Two replicate populations were originally evolved for fifteen serial passages at 37°C (denoted as 37A and 37B) and sequenced deeply using an accurate next generation sequencing protocol (Methods) (*22*). This revealed several mutations that dramatically rose in frequency across time (Fig. 1). However, due to the short read nature of the sequencing it was impossible to infer whether these mutations resided on the same genome. The similar pattern of mutation frequencies across time led us to suspect that several mutations were linked; however, we could not rule out that similar mutation frequencies occurred simply by chance or due to similar independent selection pressures.

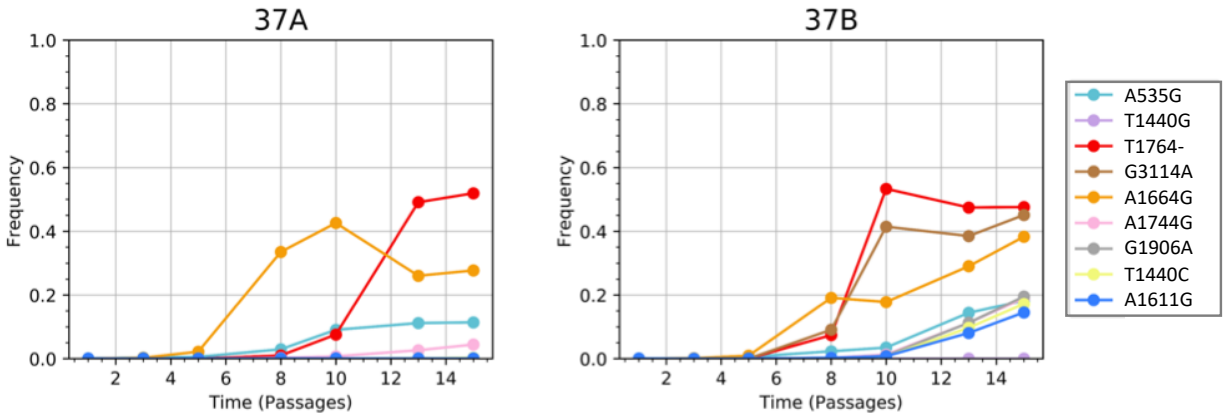

Figure 1. Mutation frequencies across time using Illumina sequencing. Shown are the frequencies of nine mutations (at eight loci, since two mutations are found at position 1440) detected above 10% in two replicate populations of evolved MS2 phages.

We next sequenced the population of RNA viruses from passage 15 of both replicas using Oxford Nanopore's MinION. Importantly, we employed direct RNA sequencing, without using reverse transcription or PCR amplification, and without any shearing of the genomes. The only requirement for library preparation is the ligation of an adaptor to the 3' of the RNA genome, allowing the 3' to enter the sequencing pore. Two independent sequencing runs were performed for each replica, denoted as p15-37A and p15-37B.

Read Lengths and Alignment

A total of 417,000 and 105,000 reads were produced for the p15-37A and p15-37B runs, respectively. In order to map the reads to the MS2 reference genome we ran our computational pipeline (Methods) (*22*), which infers the proportion of each point mutation (A, C, G, T, or "-") at each position in the genome. Over 99.9% of the reads were mapped to the reference. Coverage plots are shown in Fig. 2, and demonstrate that sequencing often terminated before it reaches the 5' end. Nevertheless, approximately 15% of the reads (corresponding to 8,000-10,000 reads) covered the entire MS2 genome.

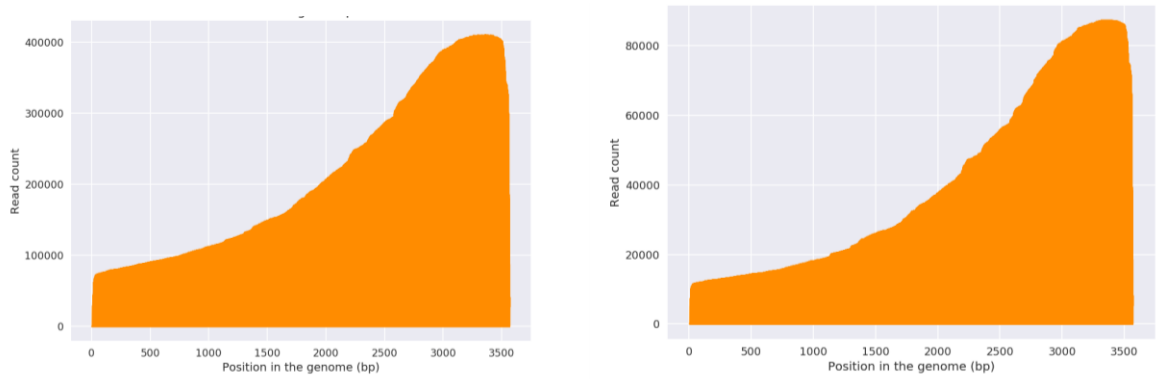

Figure 2. Coverage plots of MinION direct RNA sequencing of MS2 in p15-37A (left) and p15-37B (right). The biased slope toward the right highlights the distribution of read lengths, where a large proportion of shorter reads (~1000 bp) was produced by MinION, and a smaller fraction reaching the full length of MS2 (~ 3,569 bp).

Distribution of observed variants

We next focused on the frequency of an observed variant, defined here as any base called differently from what is present in the reference sequence, at any position. We expect such variants to be the sum

of two independent processes: real biological mutations derived from mutation and selection in the phage populations, and technical process errors. Comparing between the variant distributions observed for the same population sequenced using Illumina versus MiniON, it is evident that MiniON suffers from a very high technical error rate (Fig. 3). Notably, the number of variants exceeding a frequency of 5% in the Illumina sequencing is very small. This allows us to infer that the vast majority of MinION variant frequencies are technical errors, and further allows us to roughly estimate the various types of error rates using this platform, for our experiment (Table 1). Notably, we observed that the point deletion and point insertion rates together exceed the substitution rate, reinforcing previous observations (*23, 24*).

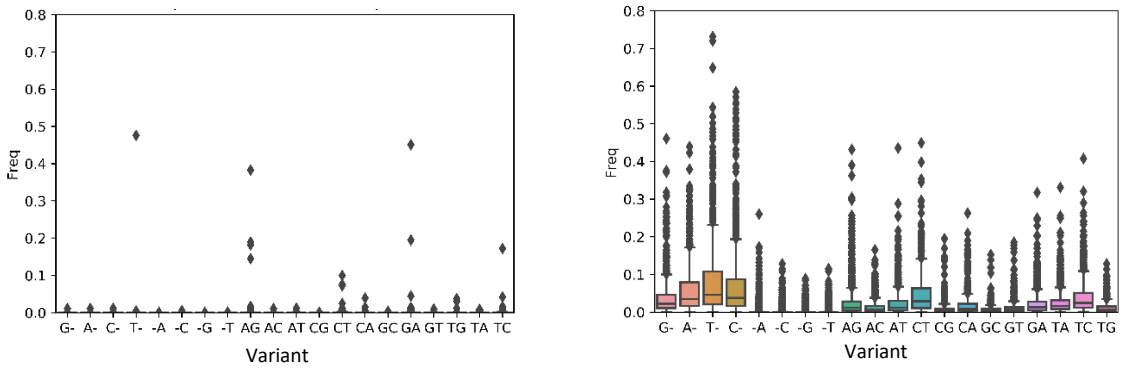

Figure 3. Variant distributions from p15-37B populations sequenced using Illumina Miseq (left) and MinION (right). Variants are broken down by the type of mutation they cause, substitutions or point deletions. "X-"indicates a single base deletion, "-X" indicates a single base insertions and "XY" indicates X replaced by Y (where X and Y are any of the four bases).

Table 1. Estimated error rates for MinION sequencing. Shown are mean values for each type of error, based on p15-37B.

Type of error	**Mean rate**
Substitution	0.070366
Insertion	0.031773
Deletion	0.06807
Total Error Rate (sum)	**0.170209**

Relationship between variants

One of the main goals of MinION sequencing, in particular in the context of RNA virus evolutionary experiments, is the detection of haplotypes and shared mutations: viral genomes that share the same mutations. However, the original MinION results showed a vast abundance of variants on the same genome: between 10 and 20% of the bases on each read were supposedly shared mutations, yet more likely these were technical errors.

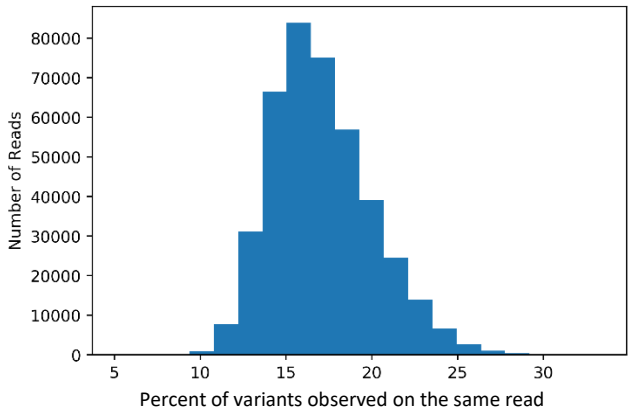

Figure 4. The distribution of the percent of variants observed on the same MinION read.

We thus sought a strategy to weed out the technical errors from the real mutations. We first calculated the conditional probabilities of observing one variant given another variant observed on the same read. When observing the pattern of conditional probabilities (Fig. S1), we noted varying patterns. Some variants co-occurred more or less randomly with all other variants, whereas some variants displayed a non-random pattern. This led us to conclude that random technical errors are expected to display a different pattern than real biological mutations: we expect technical errors to be associated randomly with any other mutation, whereas real biological mutations are expected to be non-randomly associated with each other. Thus, real mutations will tend to be either present with some other mutations on the same read, or *not* present with some other mutations on the same read. Both these properties (tendency to be present or not present with other mutations) reflect non-random association between mutations.

One of the most commonly used methods to test for associations between two properties is the chi-square test: here we use this test to see whether the observed joint variant counts deviate from what is

expected when variants are counted independently. To this end, each variant was classified as either wild-type (WT) or non-WT, based on whether it was identical or not to the reference genome. Notably, this led to $\binom{3569}{2} \approx 6 \cdot 10^6$ different tests. As a first shot, we were less interested in finding all the statistically significant associations but rather to detect the strongest associations. To this end, we plotted the chi-square statistic along the genome for each site (Fig. 5). Accordingly, it was evident that a small number of variants stood out as bearing very high chi-square values. When looking into the pairs with the highest chi-square statistics, it was evident that a given variant may tend either to co-occur or *not* to co-occur with another variant. Notably, all of these variants except one were variants detected by Illumina as *bona fide* high-frequency mutations, indicating that our association approach has the power to resolve real mutations from technical errors based on the MinION data alone.

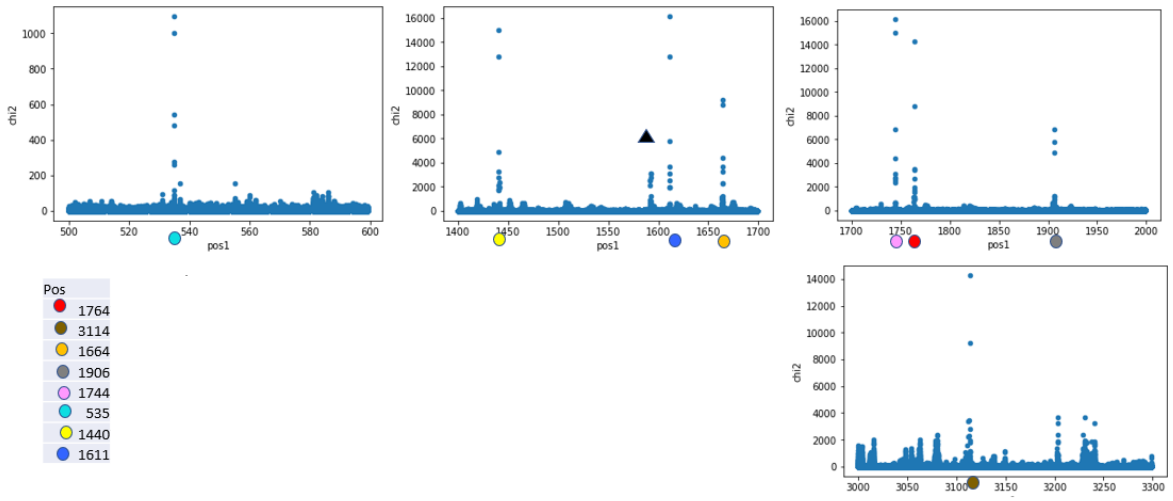

Figure 5. Chi-square statistics plotted along four windows in the genome for the p15-37B minION sequencing. Each dot represents the chi-square statistic between a given variant and all other variants at sites 40 bp or more apart. For most sites in the genome, the chi-square statistic with all other variants is low. However, some sites show an elevation of the chi-square statistic with many other variants. In fact the variants with the highest chi-square values are exactly at the eight loci identified as presenting at high frequencies using Illumina sequencing (inset box at the bottom left, Fig. 1). One exception (marked as a triangle) is discussed in the text. Colors of mutations correspond to those shown in Fig. 1.

Finally, our association test allowed to focus on the real biological mutations and resolve linkage among the different mutations (Fig. 6). Accordingly, two mutations with a very similar temporal pattern (T1764-

and G3114A) in 37B were indeed found to most often reside on the same genomes. On the other hand, a cluster of five mutations with a very similar temporal pattern in 37B (A1611G, A1744G, T1440C, G1906A, A535G) was found to likely represent three different clusters: the first three mutations cluster together and most often reside on the same genomes. The fourth mutation G1906A is most often on its own but sometimes overlaps the previous cluster, and A535G is independent of the other four. All five mutations are sometimes found together with A1664G. Finally, A1664G and T1764-, both of which rose to high frequencies in both replicas but showed a different temporal pattern, were found to be mutually exclusive in both replicas.

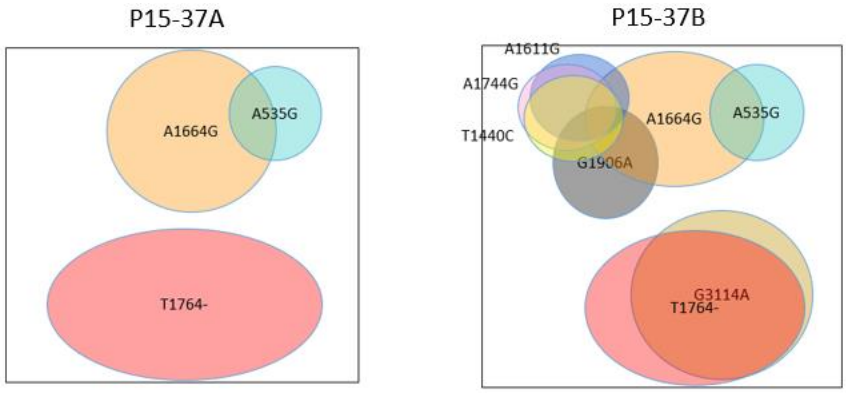

Figure 6. Venn diagram illustrating the relationships between mutations found in p15-37A (left) and p15-37B (right). Colors of mutations correspond to those shown in Fig. 1.

Discussion

We have developed here a simple and intuitive approach to detect real mutations from a background of abundant technical errors present during MinION populations sequencing. Our approach is based on the notion that errors will be randomly dispersed along the reads, whereas real mutations tend to reside on specific genetic backgrounds. We applied our method to sequencing data from an evolved population of phages where Illumina sequencing was available, allowing us to corroborate our findings. Strikingly, all the high frequency mutations observed in the 37B data (>10%) were easily picked up using our statistical association approach. The same was true for the 37A line (Fig. S2), suggesting a very high sensitivity of the method. However, we also detected additional loci using our association approach: one additional locus came up in the 37B line (black triangle in Fig. 5, locus 1593), and a different locus in the 37A line (Fig. S2, locus 1766). In the MinION data a variant at a frequency of 5% was detected in both these loci, yet the Illumina data showed essentially no mutations (frequency < 0.01). One possibility is that these

variants may represent a form of RNA editing, which may not be picked up by Illumina sequencing but is picked up by MinION(*25*). Future research is required to resolve this.

Originally, when observing the time-series data in Fig. 1, as a first approximation it seemed likely to assume that mutations with a similar temporal pattern would be mutations shared on the same genomes. Accordingly, we had hypothesized that two clusters of mutations (T1764-/G3114A and A1611G/ A1744G/T1440C/G1906A/A535G) would be on the same genomes. Notably we were correct for one cluster but incorrect for the other cluster, illustrating the utility of MinION to resolve the relationships among mutations, and its advantage for differentiating variants with mutations displaying similar temporal patterns.

We anticipate that due to its ease of use and advantages listed above, namely direct RNA sequencing and the ability to sequence extremely long reads, MinION will be increasingly valuable in the field of RNA virus genetics and in additional diverse fields such as cancer genetics and microbiology. The approach we suggest herein is simple and transferable to any organism, and as such we hope it will be a useful addition to the genetics toolbox of researchers out there.

Methods

The MS2 bacteriophage was subject to fifteen serial passages at 37C in two biological replicas (hereby denoted as 37A and 37B). We then determined the population frequency of each mutation through repeated whole genome deep sequencing using our newly developed accurate sequencing protocol, based on Illumina Miseq (*22*). In parallel to sequencing a control plasmid bearing the MS2 genome was also sequenced. The control plasmid identified 27 genomic positions that showed extreme variability and were therefore excluded from downstream analysis.

Illumina library preparation

The reverse transcription reaction was performed using SuperScript® III Reverse Transcriptase (Thermo Scientific) according to manufacturer instructions with some modifications. Briefly, the reverse transcription reaction was performed on 1µl of RNA phage (500 ng) with 1 µl of dNTP mix (10 mM), 1 µl of R4 primer (TGGGTGGTAACTAGCCAAGCAG) (10 µM) and 10 µl of sterile distilled water. The mixture was incubated for 10 minutes at 65°C followed by incubation on ice for 7 minutes. After brief centrifugation, 4 µl of 5X First-Strand Buffer were added along with 1 µl DTT (0.1 M), 1 µl of RNase OUT

(Thermo Scientific) and 1 µl of SuperScript™ III RT (200 units/µl). The mixture was incubated for 60 minutes at 55°C followed by inactivation for 15 minutes at 70°C. The last step was adding 1 µl of RNase H to the reaction according to manufacturer instructions. cDNA from the reverse transcription reaction was directly used as a template for the PCR amplification of the full MS2 genome in three overlapping fragments. PCR reactions were performed using "Phusion High-Fidelity DNA- polymerase" (Thermo Scientific) according to manufacturer instructions with the primers: 1F (GGGTGGGACCCCTTTCGG), 1R (TTTTTCTAGAGAGCCGTTGCCT),2F(GGCCCAAATCTCAGCCATGC),2R(CGTGTCTGATCCACGGC),3F(GGCACAA GTTGCAGGATGCA),3R(TGGGTGGTAACTAGCCAAGCAG). After PCR, the three amplicons were purified with a PCR clean up kit (Promega). Purified amplicons were quantified with Qubit assays (Q33216, Life Technologies), diluted and pooled in equimolar concentrations. The Illumina Nextera XT library preparation protocol and kit were used to produce DNA libraries, according to manufacturer instructions with some modifications. Briefly, the tagmentation reaction was performed with 0.8 ng/µl of DNA, 10 µl TD Buffer, 5 µl ATM enzyme and up to 5 µl DDW. The mixture was incubated for 5 minutes at 55°C and then directly used as a template for 50 µl PCR reaction using "Phusion High-Fidelity DNA- polymerase" (Thermo Scientific) according to manufacturer instructions with the index primers from the Nextera XT kit. After PCR, a double sized selection was performed using Ampure beads to remove short and long library fragments, since 350bp fragments were required. We collected the supernatant and read 2 µl from each sample in the TapeStation (Agilent high sensitives D1000) to verify fragment size.
Libraries were sequenced on an Illumina Miseq using the 2X250 MiSeq reagent kit (Illumina, MS-102-2003) for paired-end reads. Bioinformatics processing of the data was performed using the AccuNGS pipeline (*22*). Briefly, this pipeline is based on (a) mapping the reads to the reference genome using BLAST, (b) searching for variants that appear on both overlapping reads, (c) calling variants with a given Q-score threshold and inferring their frequency. All libraries attained a mean coverage of ~10,000 reads/base

Long-read Oxford Nanopore MinION sequencing

The Oxford Nanopore MinION was used to sequence the MS2 RNA directly. We sequenced 2 samples (p15-37A and p15-37B) in two separate runs. We prepared direct RNA libraries according to manufacturer library prep instructions with some modifications. We altered the supplied reverse transcriptase adapter (RTA)(*15*), which has a T10 overhang, to specifically target the MS2 genome with 23 nucleotides complementary to the MS2 conserved 3' end. The ligation reaction was performed with 500 ng RNA in 9 µl, 100 nM costume adapter in 1 µl, 10 µl of NEBNext Quick Ligation Buffer and 1.5 µl of T4 DNA Ligase enzyme (NEB). The mixture was incubated for 10 minutes at 65°C followed by incubation

on ice for 2 minutes and then directly used as a template for the next step of the library prep which is cDNA synthesis according to manufacturer instructions. The cDNA synthesis step was performed in order to maintain the RNA fragments integrity during MinION library prep and sequencing. The library was cleaned up each time using 2 µl of AMPure XP DNA beads per 1 µl of sample and we added 3 µl of RNase OUT (Thermo Scientific) to protect the RNA. The RNA was directly sequenced on the MinION nanopore sequencing device using a FLO-MIN107 flow cell equipped with the R9.5 chemistry. The MinKNOW control software version 1.14.1 was used and was allowed to proceed for 48 hours. The basecalling was performed locally by the Miniknow software as well, and the data was written out in the FASTQ format. Reads were filtered with the MiniKNOW default cutoff of a minimum average qscore of 7. The AccuNGS pipeline (*22*) described above was next applied to the data in order to determine variant frequencies. BLAST parameters were modified to %ID=60 and e-value threshold= 1E-07, to allow the highly variable minion reads to map.

References

1. R. Sanjuán, M. R. Nebot, N. Chirico, L. M. Mansky, R. Belshaw, Viral Mutation Rates. **84**, 9733-9748 (2010).
2. V. V. Rezelj, L. I. Levi, M. J. C. o. i. v. Vignuzzi, The defective component of viral populations. **33**, 74-80 (2018).
3. V. Sauvage *et al.*, Early minion™ nanopore single-molecule sequencing technology enables the characterization of hepatitis B virus genetic complexity in clinical samples. **13**, e0194366 (2018).
4. E. M. Volz, K. Koelle, T. J. P. c. b. Bedford, Viral phylodynamics. **9**, e1002947 (2013).
5. J. A. Johnson *et al.*, Minority HIV-1 drug resistance mutations are present in antiretroviral treatment–naïve populations and associate with reduced treatment efficacy. **5**, e158 (2008).
6. O. G. Pybus, A. J. N. R. G. Rambaut, Evolutionary analysis of the dynamics of viral infectious disease. **10**, 540 (2009).
7. S. Goodwin, J. D. McPherson, W. R. J. N. R. G. McCombie, Coming of age: ten years of next-generation sequencing technologies. **17**, 333 (2016).
8. R. J. Orton *et al.*, Distinguishing low frequency mutations from RT-PCR and sequence errors in viral deep sequencing data. **16**, 229 (2015).
9. V. Potapov, J. L. J. P. o. Ong, Examining sources of error in PCR by single-molecule sequencing. **12**, e0169774 (2017).
10. L. Clarke, C. Rebelo, J. Goncalves, M. Boavida, P. J. M. P. Jordan, PCR amplification introduces errors into mononucleotide and dinucleotide repeat sequences. **54**, 351 (2001).
11. C. Quince *et al.*, Insight into biases and sequencing errors for amplicon sequencing with the Illumina MiSeq platform. *Nucleic Acids Research* **43**, e37-e37 (2015).
12. C. Waugh *et al.*, A general method to eliminate laboratory induced recombinants during massive, parallel sequencing of cDNA library. **12**, 55 (2015).
13. G. Fang, G. Zhu, H. Burger, J. S. Keithly, B. J. J. o. v. m. Weiser, Minimizing DNA recombination during long RT-PCR. **76**, 139-148 (1998).
14. A. Meyerhans, J.-P. Vartanian, S. Wain-Hobson, DNA recombination during PCR. *Nucleic Acids Research* **18**, 1687-1691 (1990).
15. M. W. Keller *et al.*, Complete genome direct RNA sequencing of influenza A virus. 300384 (2018).
16. D. R. Garalde *et al.*, Highly parallel direct RNA sequencing on an array of nanopores. **15**, 201 (2018).
17. M. Jain, H. E. Olsen, B. Paten, M. J. G. b. Akeson, The Oxford Nanopore MinION: delivery of nanopore sequencing to the genomics community. **17**, 239 (2016).
18. J. Quick *et al.*, Multiplex PCR method for MinION and Illumina sequencing of Zika and other virus genomes directly from clinical samples. **12**, 1261 (2017).
19. J. Wang, N. Moore, Y.-M. Deng, D. Eccles, R. Hall, MinION nanopore sequencing of an influenza genome. **6**, (2015).
20. S. Goodwin *et al.*, Oxford Nanopore sequencing, hybrid error correction, and de novo assembly of a eukaryotic genome. **25**, 1750-1756 (2015).
21. T. Laver *et al.*, Assessing the performance of the Oxford nanopore technologies MinION. **3**, 1-8 (2015).
22. M. Gelbart *et al.*, AccuNGS: detecting ultra-rare variants in viruses from clinical samples. 349498 (2018).
23. M. Jain *et al.*, Improved data analysis for the MinION nanopore sequencer. **12**, 351 (2015).
24. A. Magi, B. Giusti, L. Tattini, Characterization of MinION nanopore data for resequencing analyses. *Briefings in Bioinformatics* **18**, 940-953 (2016).

25. H. Liu *et al.*, Accurate detection of m6A RNA modifications in native RNA sequences. 525741 (2019).