Clinical Chemistry 69:10
1174–1185 (2023)

Long-Read Sequencing with Hierarchical Clustering for Antiretroviral Resistance Profiling of Mixed Human Immunodeficiency Virus Quasispecies

Timothy Ting-Leung Ng [a,†] Junhao Su,[b,†] Hiu-Yin Lao,[a] Wui-Wang Lui,[b] Chloe Toi-Mei Chan,[a] Amy Wing-Sze Leung,[b] Stephanie Hoi-Ching Jim,[a] Lam-Kwong Lee,[a] Sheeba Shehzad,[a] Kingsley King-Gee Tam,[c] Kenneth Siu-Sing Leung,[c] Forrest Tang,[a] Wing-Cheong Yam,[c] Ruibang Luo [b,*] and Gilman Kit-Hang Siu[a,*]

BACKGROUND: HIV infections often develop drug resistance mutations (DRMs), which can increase the risk of virological failure. However, it has been difficult to determine if minor mutations occur in the same genome or in different virions using Sanger sequencing and short-read sequencing methods. Oxford Nanopore Technologies (ONT) sequencing may improve antiretroviral resistance profiling by allowing for long-read clustering.

METHODS: A new ONT sequencing-based method for profiling DRMs in HIV quasispecies was developed and validated. The method used hierarchical clustering of long amplicons that cover regions associated with different types of antiretroviral drugs. A gradient series of an HIV plasmid and 2 plasma samples was prepared to validate the clustering performance. The ONT results were compared to those obtained with Sanger sequencing and Illumina sequencing in 77 HIV-positive plasma samples to evaluate the diagnostic performance.

RESULTS: In the validation study, the abundance of detected quasispecies was concordant with the predicted result with the R^2 of > 0.99. During the diagnostic evaluation, 59/77 samples were successfully sequenced for DRMs. Among 18 failed samples, 17 were below the limit of detection of 303.9 copies/μL. Based on the receiver operating characteristic analysis, the ONT workflow achieved an F1 score of 0.96 with a cutoff of 0.4 variant allele frequency. Four cases were found to have quasispecies with DRMs, in which 2 harbored quasispecies with more than one class of DRMs. Treatment modifications were recommended for these cases.

CONCLUSIONS: Long-read sequencing coupled with hierarchical clustering could differentiate the quasispecies resistance profiles in HIV-infected samples, providing a clearer picture for medical care.

Introduction

AIDS, caused by HIV, is a significant public health concern that affected an estimated 38.4 million people worldwide in 2021 (1). Long-term antiretroviral (ARV) treatment is often required to combat this incurable disease. Although a wide range of ARV classes, such as protease inhibitors, nucleotide/nucleoside reverse transcriptase inhibitors (NRTI), nonnucleoside reverse transcriptase inhibitors (NNRTI), and integrase inhibitors are currently available for HIV treatment, the drug of choice could be highly limited by the development of drug resistance mutations (DRMs) during the course of treatment. Furthermore, the presence of mixed quasispecies with different drug resistance profiles could further complicate treatment (2). Previous studies showed that minority quasispecies of drug-resistant viruses with a frequency range of 0.07% to 2.5%, detected at baseline can rapidly outgrow and become the major virus population and subsequently lead to early therapy failure in treatment-naive patients who receive antiretroviral therapy regimens with a low genetic resistance barrier (2–5).

Sanger sequencing is commonly used to identify DRMs in HIV-infected patients (6–9). Since minor

[a]Department of Health Technology and Informatics, The Hong Kong Polytechnic University, Kowloon, Hong Kong SAR, China; [b]Department of Computer Science, The University of Hong Kong, Pokfulam, Hong Kong SAR, China; [c]Department of Microbiology, Queen Mary Hospital, The University of Hong Kong, Pokfulam, Hong Kong SAR, China.
*Address correspondence to: G.K-H.S., Department of Health Technology and Informatics, The Hong Kong Polytechnic University, Rm. Y930, 9/F, Lee Shau Kee Bldg., Hung Hom, Kowloon, Hong Kong SAR, China. E-mail: gilman.siu@polyu.edu.hk. R.L. at Department of Computer Science, The University of Hong Kong, Rm. 422, Chow Yei Ching Bldg., Pokfulam, Hong Kong SAR, China. E-mail: rbluo@cs.hku.hk.
†Timothy Ting-Leung Ng and Junhao Su contributed equally to this work.
Received March 9, 2023; accepted June 28, 2023.
https://doi.org/10.1093/clinchem/hvad108

variants, which are defined as mutations with low variant allele frequencies (VAFs), are usually masked by the dominant base signals, Sanger sequencing can only detect quasispecies with a VAF up to 30% of a population (10), resulting in incomprehensive ARV resistance information of the quasispecies within the host viral population. The variant analysis using Sanger sequencing also has limited ability to characterize heterozygous insertions and deletions (10, 11). Massively parallel sequencers, including next-generation sequencing (NGS) and third-generation long-read sequencing, are able to overcome these issues and are thus preferred in the study of HIV quasispecies.

NGS allows highly accurate sequencing (99.9% accuracy) that enables the identification of minor variants (12, 13), with the recommended allele frequency cutoff set at 0.2 by the WHO (14). With the PCR-tiling strategy, full coverage of the targeted genomic regions can be supported (15). However, the profiling of ARV resistance in HIV quasispecies may be hampered by the read-length restrictions, where mutations cannot be assigned to specific HIV quasispecies, greatly underestimating the complexity of a patient's ARV resistance profile. Additionally, for HIV with a small genome size (9 kb), the fixed sequencing capacity of NGS may not financially benefit laboratories with a low sample throughput.

Although long-read nanopore sequencing has comparatively lower read accuracy than Sanger sequencing and NGS, the use of nanopore sequencing for generating near-full-length HIV genomic sequences has been reported in other studies (16, 17). Long-read sequencing can reveal the linkage of one amino acid mutation to another found in the same quasispecies. With its low adoption cost and the compact size of the MinION sequencer, this technology is now under the spotlight for the clinical diagnosis of infectious diseases.

In this study, an ONT workflow and a novel bioinformatics software, ClusterV, were developed for detecting quasispecies in HIV-infected samples and the corresponding ARV resistance profile to each quasispecies. The performance of long-read nanopore sequencing coupled with hierarchical clustering in bioinformatics was investigated with reference to Sanger sequencing and Illumina sequencing.

Materials and Methods

SAMPLE COLLECTION

A total of 77 plasma samples, collected from 70 male and 7 female patients ages 18 to 68, were obtained from the Queen Mary Hospital in Hong Kong between 2002 and 2014 (Supplemental Table 1).

RNA EXTRACTION AND LONG REGION AMPLIFICATION

Frozen plasma sample (approximately 1.5 mL) was thawed on ice for 2 hours, then resuspended and centrifuged at 20 800*g* for 1.5 hours at 4°C. Supernatant was removed. Viral RNA in the pellet was extracted using QIAamp Viral RNA Kit, according to the manufacturer's instructions. Then, 8 μL of extracted RNA was subjected to DNA removal using ezDNase™, reverse transcription with LunaScript RT SuperMix (5X), and amplification targeting the long genomic region (NC_001802.1: 1413–7363, amplicon length: 5951 base pairs) using the reagents and cycling conditions described in Supplemental Table 2. Amplicons were purified with 0.5×AMPure XP beads (for nanopore sequencing) or with QIAquick PCR Purification Kit (for Illumina) and quantified with Qubit® dsDNA HS Assay Kits.

SANGER SEQUENCING

Protease and reverse transcriptase nucleotide sequences were determined by our in-house Sanger sequencing-based genotypic resistance test (18), whereas the sequences of integrase were characterized by another in-house Sanger sequencing protocol developed by our team (19).

NANOPORE SEQUENCING

For nanopore sequencing, libraries were constructed with SQK-LSK109, EXP-NBD104, and EXP-NBD114, following the official protocols of Native Barcoding Amplicons. At most 12 libraries were pooled and sequenced on ONT GridION for 48 hours using the SUP basecalling mode, with the minimum quality score for read filtering as 10 and the modified demultiplexing setting (trim_barcodes = "on", require_barcodes_both_ends = "on", detect_mid_strand_barcodes = "on", min_score = 85).

ILLUMINA SEQUENCING

For Illumina sequencing, 1 ng of amplicon was used for library preparation with the Nextera XT DNA Library Preparation Kit and IDT® for Illumina® DNA/RNA UD Indexes Set A, Tagmentation. The quality and quantity of the library were assessed with the Bioanalyzer and QIAseq Library Quant Assay Kit, respectively. Libraries were pooled and sequenced with the MiSeq Reagent Kit Nano V2 on the Illumina MiSeq System (250 × 2 cycles).

QUANTITATIVE REVERSE TRANSCRIPTION POLYMERASE CHAIN REACTION

The viral load (copies per μL) of each RNA sample was determined using the GeneSig Standard Real-time PCR

Detection Kit for HIV-1 and the Oasig Lyophilized OneStep qRT-PCR MasterMix Kit.

LIMIT OF DETECTION

Limit of detection (LOD) was defined as the minimum input viral concentration (copies/μL) required for having 0.9 probability to reach the average depth of coverage (DP). For LOD analysis, 2 DP cutoffs, 50× and 1500×, were employed. A minimum DP of 50× is sufficient for variant calling in a quasispecies with high confidence in nanopore sequencing data when using Clair-ensemble embedded in ClusterV (20). However, a DP of 1500× was recommended to detect 30 quasispecies with an abundance as low as 0.03. A sample that passed the DP was set as 1, while a sample that did not pass was set as 0. Logistic regression was used for estimating the LOD.

BIOINFORMATICS

ONT sequencing reads were used to iteratively cluster quasispecies and call variants using ClusterV (Supplemental information S1 and S2) to generate abundance, DRM profiles, and clinical reports for each quasispecies based on the input of the alignment files and Browser Extensible Data files (Supplemental Fig. 1). In brief, sequencing reads were mapped to the HIV reference genome NC 001802.1 (2.24-r1122) using Minimap2; reads with an defective genome or that failed to cover the targeted region were excluded. Then variants were called using Clair-ensemble and used as markers for iterating hierarchical clustering processes in order to identify HIV quasispecies within a sample. A consensus sequence for each quasispecies was generated based on its variants. Finally, an ARV resistance report was generated using SierraPy, based on the consensus sequences of all quasispecies found within a sample.

Illumina data was first aligned to HIV genome NC_001802.1 with bwa mem (v1.15.1) (21). Unmapped reads and secondary alignments were filtered with the SAMtools view function (with –F 3844) (22). Variant calling of the filtered alignments was performed using Clair3 (20) with Illumina mode (v0.1-r11, with –haploid_sensitive, --no_phasing_for_fa flag).

VALIDATION OF CLUSTERING PERFORMANCE IN THE ONT WORKFLOW

One HIV plasmid (pHIV-1_pr-V82A, provided by the European Virus Archive—Global), and 2 clinical samples (Sample ID: KB2061 and KB2979), each of which has a unique quasispecies with a median VAF > 0.9 (Supplemental Table 3), were used to evaluate the clustering performance of ClusterV. An in silico simulation data set was prepared by mixing the HIV plasmid and 2 clinical samples (Sample ID: KB2061 and KB2979) in various combinations (10:10:80, 33:33:33, 80:20:0, 50:50:0, 5:95:0). Also, a gradient series of HIV plasmid and KB2061 amplicons was prepared in triplicate with the following ratios: 95:5, 90:10, 85:15, and 80:20. The abundance of quasispecies determined by ClusterV was compared with the corresponding mixing ratio; R^2 was used for evaluating their linear relationship.

EVALUATION OF THE DIAGNOSTIC PERFORMANCE OF THE ONT SEQUENCING WORKFLOW

To evaluate the diagnostic performance of the ONT workflow, the reported amino acid mutations (including AVR resistance associated and non-AVR related mutations) and their associated genomic variants were compared with those reported in Sanger and Illumina sequencing, with the assistance of Integrative Genome Viewer (23). Illumina genomic variants with a VAF greater than 0.03 (13, 24–26) were considered valid. To determine the accuracy of ONT mutations, it was considered true if they were concordant with Sanger sequencing. In cases where there was a discrepancy with or no Sanger reference available, the ONT mutations were compared with Illumina results, and they were considered true if the results were concordant. If the ONT mutations were discordant or could not be validated with Illumina, they were considered false or inconclusive, respectively (Fig. 1).

Amino acid mutations exclusively found in Sanger but not in ONT were also validated with Illumina. Concordant mutations were considered true while discordant mutations were considered false (Fig. 1). The F1 score [2× (precision × recall) / (precision + recall)] was used for evaluating performance of variant calling and ARV resistance detection. To determine the VAF cutoff, an overall VAF of each amino acid mutation within a sample was calculated. Then a receiver operating characteristic (ROC) analysis was performed; the VAF with the highest sensitivity and the lowest false-positive rate was regarded as the cutoff value. Statistical analysis was performed with GraphPad Prism (v9.4.1) and Microsoft Excel.

Results

VALIDATION OF CLUSTERING PERFORMANCE IN THE ONT WORKFLOW

The number of detected quasispecies by ClusterV was concordant with the expected number of quasispecies in the gradient series and in silico simulation data set; the predicted abundance was in a linear relationship with the true abundance (Supplemental Table 4). The R^2 for HIV plasmid and KB2061 gradient series were both 0.996 (Supplemental Figs. 2A and B), while the R^2 for the in silico simulation data set was 0.9939 (Supplemental

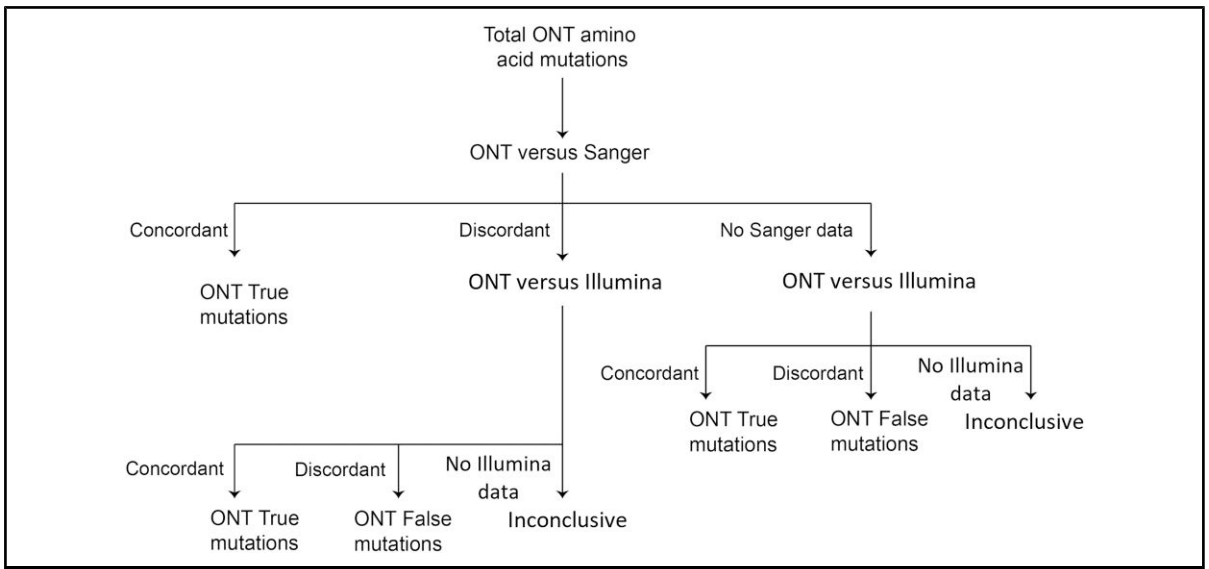

Fig. 1. A flowchart to demonstrate how amino acid mutations found in the ONT workflow were classified as true or false mutations.

Fig. 2C). The median VAF for variants found in each quasispecies was 0.89 or above (Supplemental Table 4).

THE VARIANT CALLING AND DIAGNOSTIC PERFORMANCE OF THE TARGETED SEQUENCING ONT WORKFLOW

Fifty-nine out of 77 plasma samples were successfully sequenced using the ONT workflow, while sequencing failures occurred with the remaining 18 samples due to low viral load (<250 copies/μL), except for one sample KB2992 (745.36 copies/μL). Sanger sequencing results covering protease and reverse transcriptase (RT) (amino acid position 1–401) were available for 54 samples, except for 5 samples (KB0097, KB0270, KB0548, KB0552, and KB0553) covering RT amino acid position 1 to 335; results for integrase were only available in 16 samples. Illumina data was not available in 16 samples with insufficient sample volume; 43 samples were sequenced by Illumina as a result.

One HIV quasispecies was found in 28/59 samples (47.45%), followed by 2 quasispecies in 10 samples (16.94%) and 3 quasispecies in 6 samples (10.17%), and more than 3 quasispecies were found in 19 samples (25.42%). Subtypes B (38.98%) and CRF01_AE (35.59%) were dominant in the sample set, followed by CRF07_BC (11.29%). Three samples, KB2974, KB2980, and KB2998, carried a mixture of subtypes (Supplemental Table 5A).

A total of 4104 amino acid mutations were found in 59 samples in the ONT workflow (Fig. 2A, Supplemental Table 5B). Of these, 2200 mutations and 271 mutations were respectively concordant and discordant with Sanger sequencing, whereas 1506 (36.7%) mutations could not be validated due to the unavailability of Sanger sequencing results. The remaining 127 (3.1%) mutations were classified as unknown amino acid mutations due to complicated insertion-deletion variants. Among the 271 discordant and 1506 nonvalidated mutations, 1153 (28%) and 222 (5%) mutations were concordant and discordant with Illumina, respectively, and 49 (1%) discordant with Sanger could not be validated with Illumina. The remaining 353 mutations (9%) could not be validated because of unavailable Sanger and Illumina results.

Additionally, 135 amino acid mutations were uniquely found in Sanger sequencing (Fig. 2B), with 27 of the mutations concordant with Illumina, 62 mutations discordant, and 46 mutations that could not be validated because of unavailable Illumina results. To conclude, the precision and recall of the ONT workflow were 93.79% (2200 + 1153)/(2200 + 1153 + 222) and 99.2% (2200 + 1153)/(2200 + 1153 + 27), respectively, and the F1 score was 0.964. A ROC curve was generated (Fig. 2E; Supplemental Table 6) with the area under the curve of 0.903. In this case, the VAF cutoff for ONT sequencing was 0.4 with the true mutation rate of 0.9072 and false mutation rate of 0.1712.

With the ONT workflow, 33 DRMs were found in 22 samples (Fig. 2C; Supplemental Table 7A). By comparing with the Sanger sequencing results, 26 out of 33 DRMs (78.8%) were considered as true mutations, 6 DRMs (18.2%) were discordant, and one DRM (3%) was inconclusive. When comparing the 6 discordant DRMs with Illumina, one was considered as a true mutation, 4 were considered as false mutations, and one was

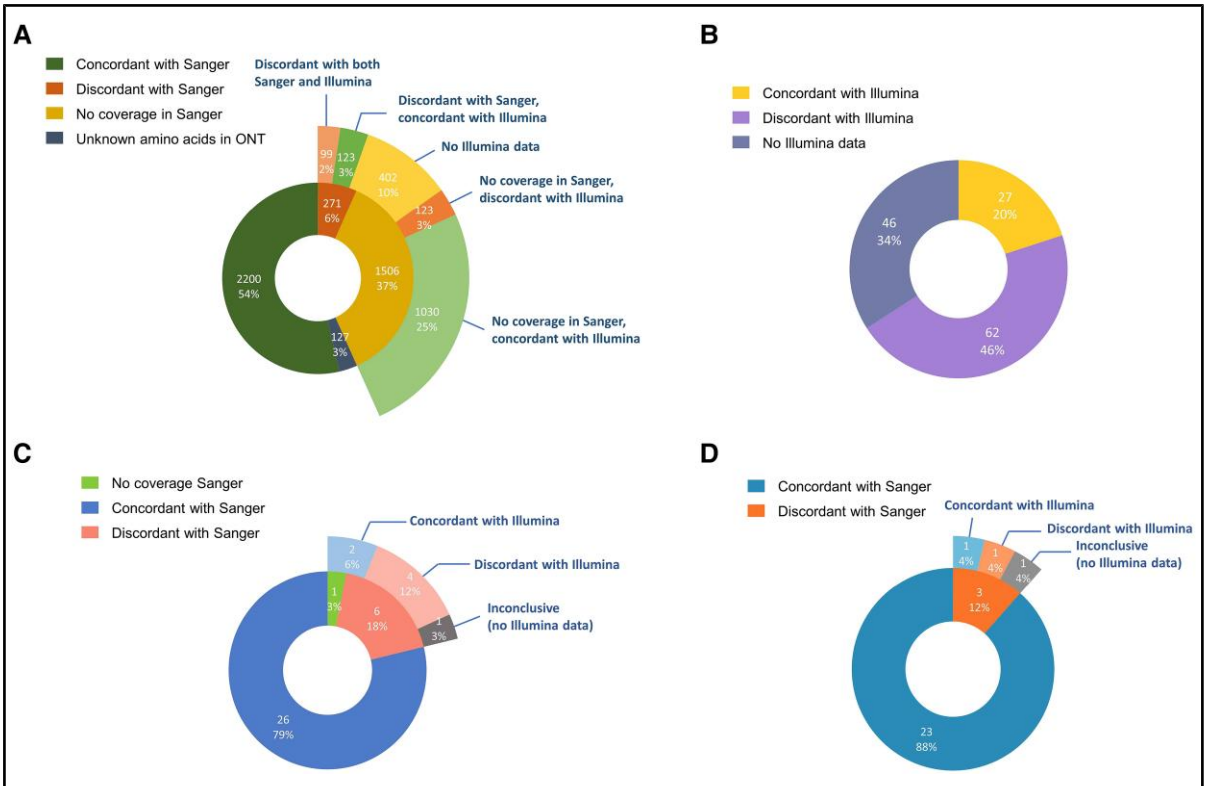

Fig. 2. (A), The distribution of the amino acid mutations (n = 4104) found in ONT and their agreements with Sanger and Illumina; (B), the agreement between unique amino acid mutations (n = 135) in Sanger and Illumina; (C), the distribution of AVR associated amino acid mutations (n = 33) found in ONT with the agreement with Sanger or Illumina before the cutoff (VAF = 0.4) and (D), after the cutoff; (E), The ROC analysis of variant calling performance in the ONT workflow. Each data point on the curve displayed sensitivity and false-positive rate. The overall VAF cutoff of 0.4 is highlighted in red.

Figure continued on next page

inconclusive. Additionally, the 1 DRM without a reference Sanger result was also considered as a true mutation since it was concordant with Illumina.

There were 3 DRMs with mixed alleles uniquely found in Sanger (Supplemental Table 7B). With reference to the results of Illumina sequencing, one DRM was considered a true mutation while the other 2 were considered false mutations. The precision and the recall of the diagnostic performance were 0.875 (28/32) and 0.965 [28/(28 + 1)], respectively, and the F1 score was 0.918.

With the cutoff of 0.4 from the ROC analysis, a total of 7 mutations with an overall VAF below the cutoff were excluded (Fig. 2D). The precision and the recall of the diagnostic performance were 0.96 (24/25) and 0.96 [24/(24 + 1)], and the F1 score was 0.96.

DETECTION OF THE MUTATIONS WITH THE QUASISPECIES

With the ONT workflow, 4 samples were found harboring different patterns of DRMs in different quasispecies. For example, all of the quasispecies (KB2070_1 and KB0270_2) in the sample KB0270 (Fig. 3A, B, and C) harbored mutation G73S in the protease gene, as well as mutations M41L, S68G, and M184V in the RT gene, conferring a potentially low-level protease inhibitors resistance (atazanavir, fosamprenavir, indinavir, saquinavir, and nelfinavir) and a low-level to high-level of NRTI resistance (abacavir, didanosine, emtricitabine, and lamivudine). However, mutation T215F was exclusively found in the quasispecies KB0270_2 (abundance = 32.27%), conferring an additional low-intermediate resistance level of NRTI (abacavir, azidothymidine, stavudine, didanosine, and tenofovir).

Another example, KB2987 (Fig. 3D, E, and F), mutation V106I in RT, conferring a potentially low resistance level of NNRTI (doravirine, etravirine, nevirapine, and rilpivirine), was found in 5 out of 7 quasispecies (KB2987_2, KB2987_3, KB2987_4, KB2987_5, and KB2987_7) with a total abundance of 63.2%. However, G190E, a minor mutation (overall VAF of 0.063) that conferred intermediate-high resistance of NNRTI (doravirine, efavirenz, nevirapine, rilpivirine,

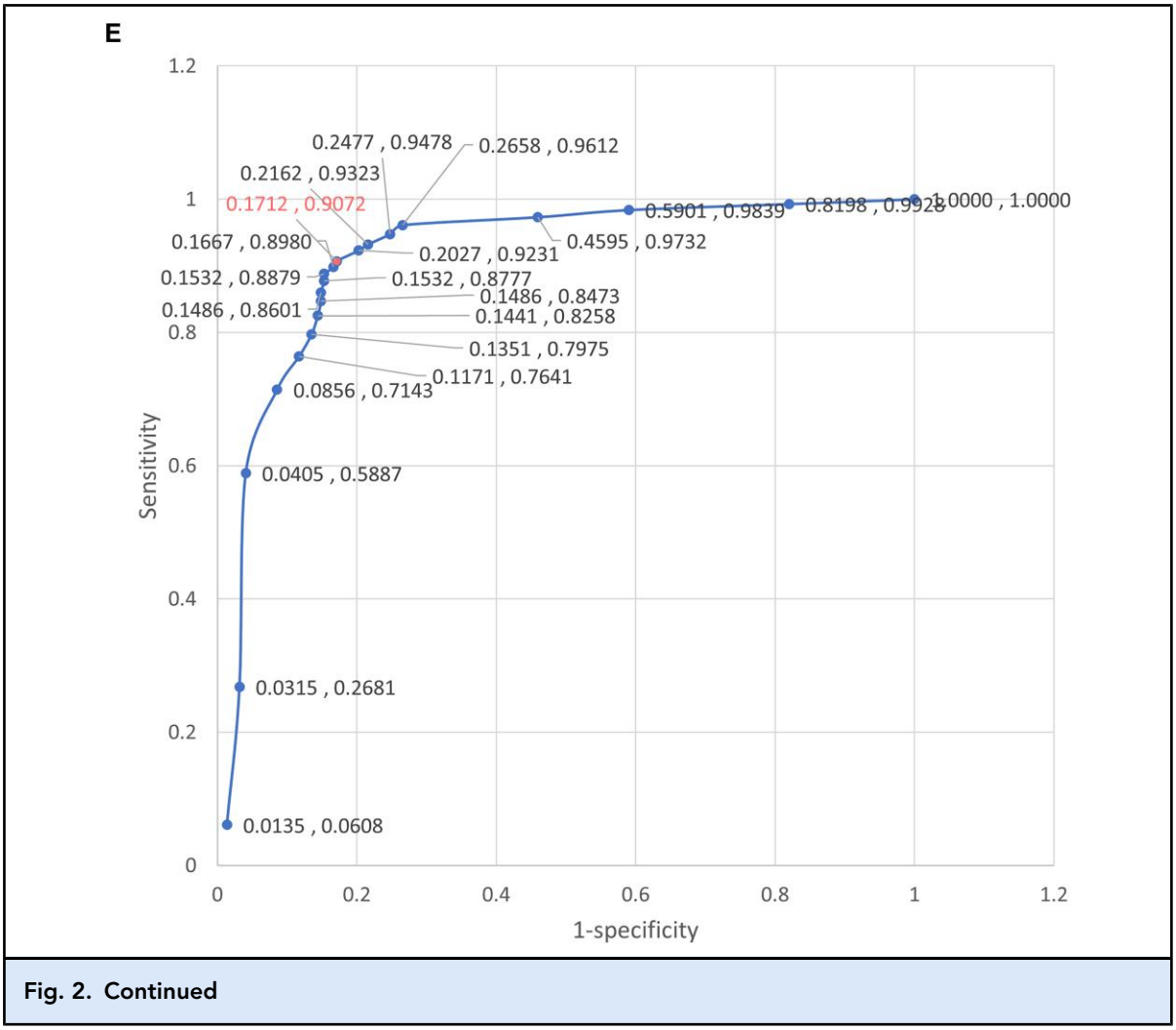

Fig. 2. Continued

and etravirine), was uniquely found in the quasispecies KB2987_6 (abundance = 6.79%). This minor mutation was also detected by the Illumina method.

LIMIT OF DETECTION

Two logistic regression models were constructed based on the 2 DP cutoffs: 50× and 1500× (Supplemental Fig. 3; Supplemental Table 8). The area under the curve for the cutoffs DP = 50× and DP = 1500× were 0.8795 and 0.853, respectively. By calculating the viral load (copies/μL) required for having a probability 0.9 of reaching the DP = 50× and DP = 1500×, the LODs were 303.9 copies/μL and 1930.6 copies/μL, respectively.

TIME TO REPORT AND OPERATION COST

For a batch of 12 plasma samples, 5 hours and another 3.5 hours were respectively required for viral RNA extraction and real-time PCR on the first working day. On the next day, the amplicons were purified and enzymatically converted to a library (3 hours). After the 48-hour sequencing, data analysis was performed on the fourth working day, which lasted for 2 hours generally (based on a computer with 2 12-core Intel Xeon Silver 4116 processors with 126GB RAM running in 10 threads). An additional 6 hours was required for a computer with 32 GB RAM, a CPU of Intel(R) Xeon(R) CPU E5-2678 v3 or equivalent, and a clock speed of 2.5 GHz or equivalent. Therefore, the time to report for this ONT workflow was 4 working days (Fig. 4). The running cost per sample was $120.93 (12 samples per flow cell) and $88.02 (24 samples per flow cell) (Supplemental Table 9).

Discussion

The clinical significance of detecting low-frequency DRMs in HIV remains controversial. Some studies have shown a marginal correlation between detecting

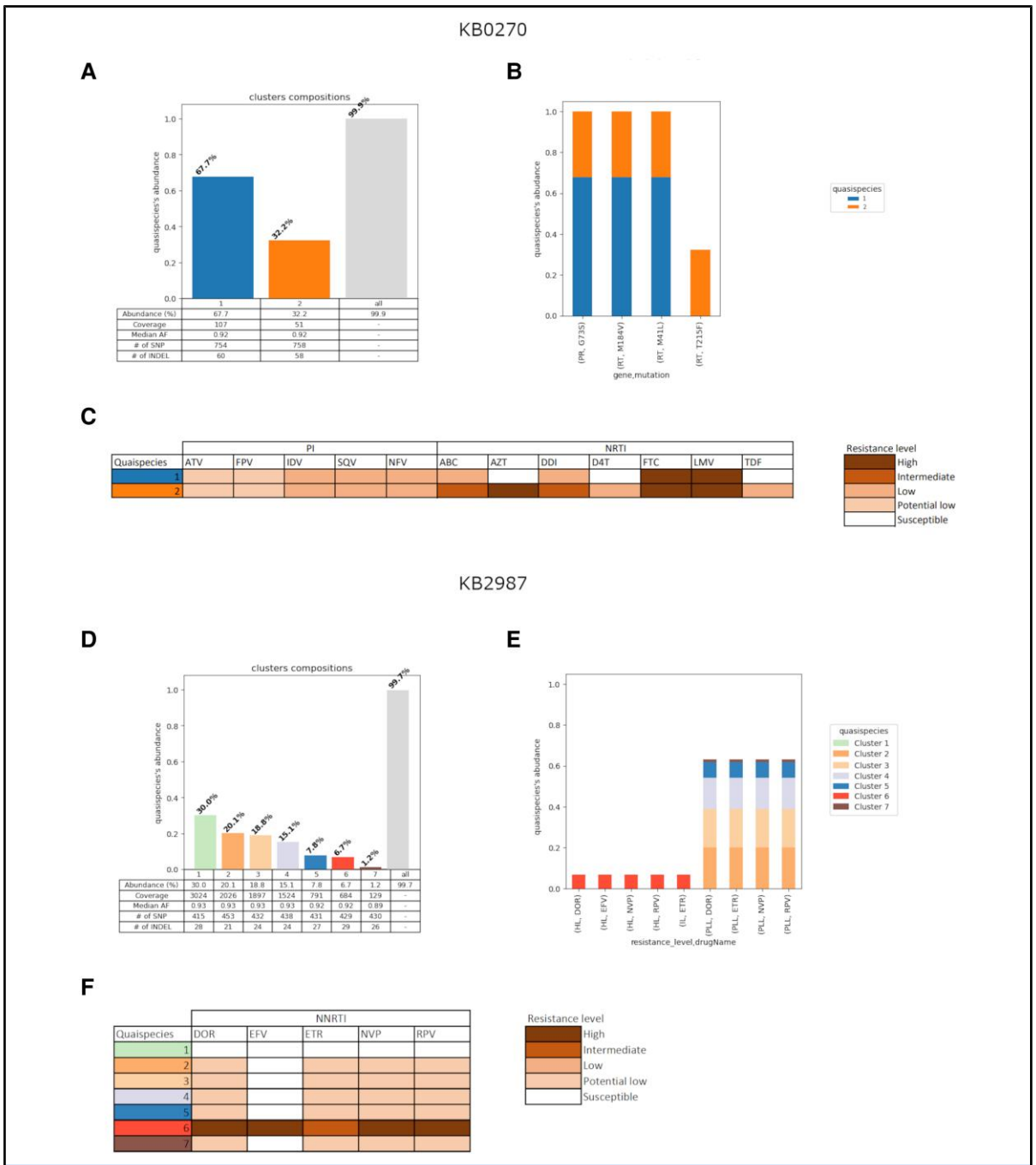

Fig. 3. Two examples demonstrating different DRM patterns in different quasispecies in the same samples. (A), the example in KB0270 with the abundance; (B), different resistance patterns and (C), resistance level found in different quasispecies; (D), the example in KB2987 with the abundance, (E), different resistance patterns, and (F), resistance level found in different quasispecies.

minor DRMs in treatment-naïve patients and the clinical failure of first-line ARV therapy (27–29). However, these studies detected low-frequency DRMs based on short sequence reads with a read length of no more than 150 bp. One common limitation of short-read sequencing is that researchers cannot analyze viral variants at the single genome level, making it difficult to determine whether these minor mutations are linked with

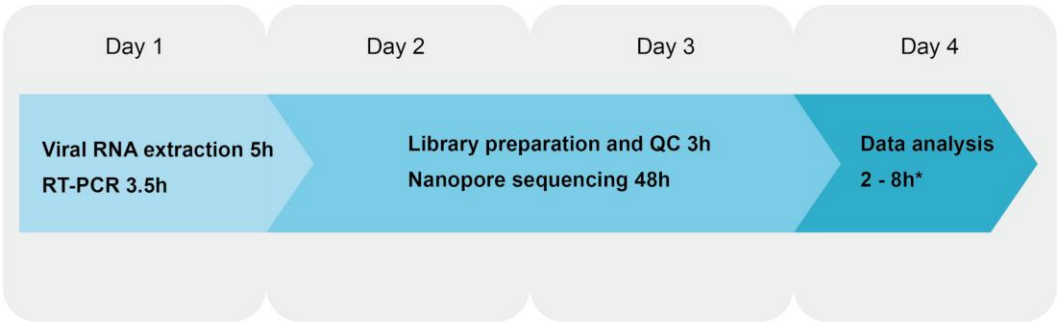

Fig. 4. The time to report utilizing the ONT workflow for direct AVR resistance detection in HIV.

each other on the same genome or occur separately in different quasispecies virions.

Linked dual-class resistance mutations occurring in a single genome have been previously associated with an increased risk of virological failure (30). As shown in Fig. 5, variants with a VAF of 0.33 could originate from the same genome, conferring quasispecies coresistance to multiple classes of ARV. In this scenario, the strain is likely to survive even under combined ARV therapy and become a predominant strain, eventually resulting in virological failure. Conversely, if the DRMs are separately carried by different viral particles, they can still be suppressed by the other effective ARV class under highly active antiretroviral therapy without requiring treatment modification. Therefore, whether low-frequency DRMs eventually lead to therapy failure depends on whether they are present on the same viral genome.

However, different classes of DRMs located distantly apart from each other cannot be revealed by Sanger and short-read sequencing, and their linkage cannot be determined. Our long-read sequencing (>6 kbp in length) coupled with hierarchical clustering can accurately link DNA mutations to different quasispecies, addressing the limitations of Sanger and Illumina sequencing. This provides a more precise prediction of the clinical outcome for samples with low-frequency DRMs. For treatment-naïve patients, detecting drug-resistant minorities before commencing treatment can guide the choice of ARV and reduce the chance of treatment failure.

In this study, 4 out of 59 (6.8%) samples were found to have quasispecies with DRMs. Notably, 2 cases harbored quasispecies with more than one class of DRMs, which could potentially lead to virological failure of first-line ARV therapy. In the highlighted example of KB0270, 2 quasispecies carrying DRMs associated with resistance to protease inhibitors and NRTI were identified. In another example, KB2897 was found to have a mutation that can cause intermediate- to high-level resistance to 5 NNRTIs in one of the quasispecies with low abundance. Treatment modifications were recommended for these cases. However, both patients were lost to follow-up and thus clinical outcome could not be determined.

In addition to quasispecies, long-read sequencing and clustering can identify distant subtypes and their corresponding drug resistance profiles in cases of HIV superinfection. HIV superinfection occurs when a patient is infected with one HIV strain and then becomes infected with another distant HIV strain. Some studies have reported the superinfection of susceptible HIV strains with resistant strains or vice versa (31, 32). In one special case, a patient infected with one resistant strain was then infected with another resistant strain (33). In this study, hierarchical clustering allowed the identification of mixed subtypes in some cases. For example, sample KB2974 was a mixture of Subtype B and Subtype CRF01_AE, while only Subtype CRF01AE was detected by Sanger sequencing. Another example, sample KB2980, was a mixture of Subtype CRF07_BC and Subtype CRF01_AE, while only CRF07_BC was identified by Sanger sequencing. Sample KB2998 was a mixture of Subtype CRF07_BC and Subtype B + C. Both of them could be detected by Sanger and ONT sequencing. These results supported the occurrence of HIV superinfection with different subtypes in these 3 cases. Using hierarchical clustering, the sequences of different subtypes could be separated and isolated into individual genomes for phylogenetic analysis, enabling epidemiological investigation and transmission tracing.

Another benefit of hierarchical clustering is to avoid misinterpreted amino acids caused by mismatching multiple nucleotide substitutions at mixed alleles in the same genetic code. Hierarchical clustering can cluster the reads by selecting variants with near VAF peaks prior to the final variant calling in different quasispecies (Supplemental Table 10). For example, mutation A71I (ATT encoding for isoleucine) in protease in sample KB2019 was reported in Sanger sequencing, but the genetic codes ACT (encodes for theronine) and GTT (encodes for valine) were observed in 2 different quasispecies using ONT (Fig. 6A). Another example is that V111M (ATG encoding for methionine)

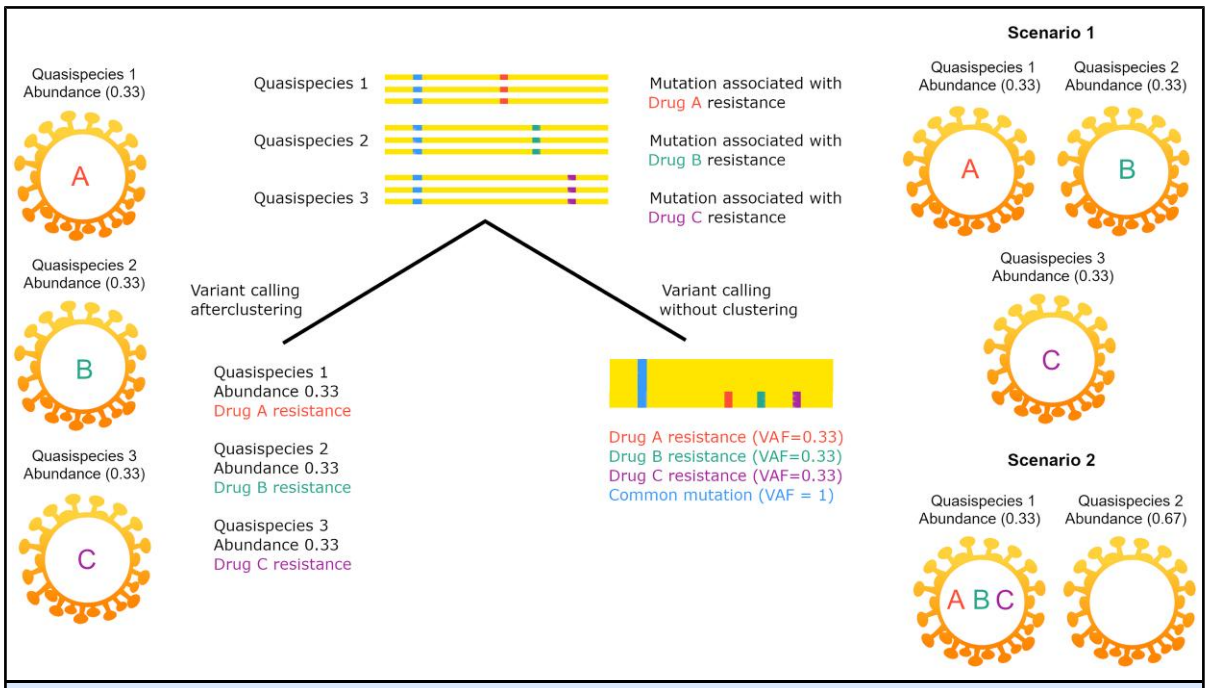

Fig. 5. Schematic diagram of long-read nanopore sequencing coupled with hierarchical clustering for the detection and assignment of low-frequency DRMs to different HIV quasispecies of a sample. The DRMs with VAFs of 0.33 could be attributed to 2 different scenarios. In the first scenario, 3 different DRMs are separately harbored in 3 different quasispecies. Under a standard treatment regimen with combined ARV, the quasispecies with only a single class of DRM could be suppressed by the other classes of drugs. In scenario 2, a quasispecies, which accounts for one-third of the viral population, harbors 3 DRMs on the same genome, conferring the viral strain coresistance to 3 different classes of ARV. The strain is likely to survive even under combined ARV therapy and become a predominant strain, eventually resulting in virological failure.

in RT in KB2987 was reported by Sanger sequencing; however, ATA (encodes for isoleucine) and GTG (encodes for valine) were found in different quasispecies of the sample (Fig. 6B).

Hierarchical clustering can be computationally demanding, especially when processing a large batch of long-read sequencing data and constructing a distance matrix. To simplify the process and reduce the computational burden, the clustering strategy employed in this study was based on multiple rounds of selecting variants with near VAF peaks rather than a distance matrix between the sequences. This approach minimized the need for constructing a distance matrix and simplified the hierarchical clustering process. The lower adoption cost of the nanopore sequencer and the simplified design of hierarchical clustering could be of benefit as a workflow in clinical centers that require short analysis time and low computer specifications. Furthermore, the reduced computational burden of our approach compared to traditional hierarchical clustering methods makes it more accessible for adoption in resource-limited settings. In addition, the simplified design of our workflow allows for efficient and rapid analysis of large amounts of data, making it an ideal solution for clinical applications where timely results are critical.

The studies of minor variant detection in a gradient series and an in silico simulation data set demonstrated that the ONT workflow was able to detect the true minor variants. A direct linear relationship between the reported abundance in the ONT workflow and the corresponding true abundance was obtained, though the reported abundance was generally slightly lower than the gradient ratios. However, from the ROC analysis with the 59 plasma samples, the false-positive rate was still high ($> = 0.2$) for the overall VAF of less than 0.35. As a result, a cutoff of 0.4 was set to keep the balance between the lower false-positive rate (<0.2) and the higher true positive rate (approximately 0.9). With this cutoff, the F1 score of ARV resistance detection was increased from 0.918 to 0.96 as the false mutations with overall VAF below the cutoff were filtered. Of note, 2 mutations with overall VAF above the cutoff were discordant with Sanger. The mutation N348I in KB2971 could not be validated with unavailable Illumina results, while

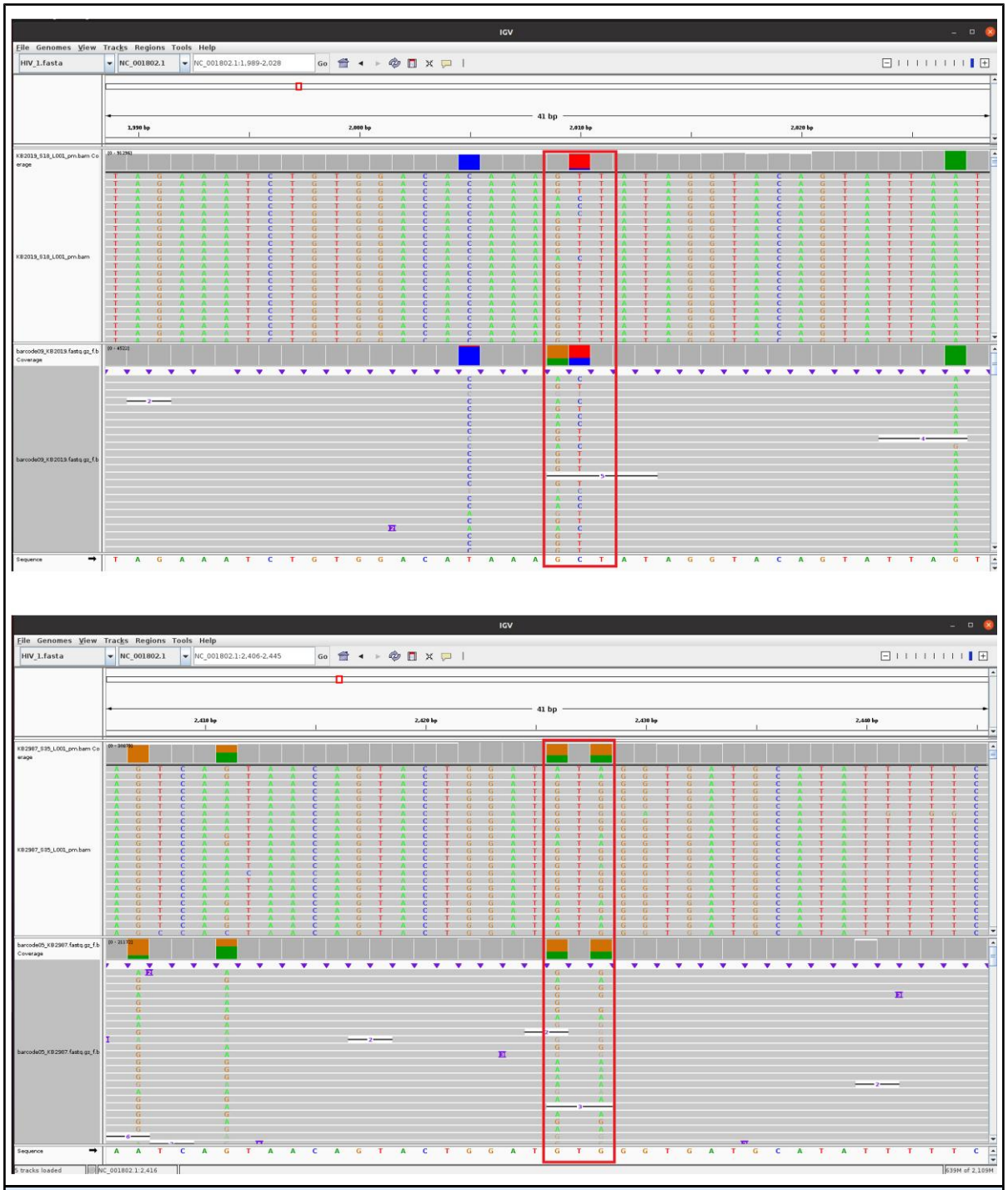

Fig. 6. Two examples (upper) KB2019 and (lower) KB2987 demonstrated how the amino acid mutations were misinterpreted in Sanger, comparing with the genetic codes found in ONT and Illumina (in red square).

mutation K70R in KB2016 was considered as false because of its discordance with Illumina. The reason was still unclear but possibly due to the low viral load in KB2016 and KB2971 (104.79 copies/μL and 98.79 copies/μL, respectively) that occasionally amplified only one of the quasispecies and dominated the overall VAF. To avoid missing

true mutations in the marginal overall VAF (<0.4), these mutations could be reported as possibly true and should be confirmed.

There were several limitations in this study: (*a*) longitudinal samples were not available, and temporal changes in the DRMs of each quasispecies could not be determined; (*b*) although Illumina was employed to validate the clustering performance, the VAF might be varied due to the inability to filter defective HIV genomes (34) in Illumina, while such genomes were removed in the ONT workflow; (*c*) the clustering performance was validated with the gradient series and in silico simulation data set, but validation with other long-read sequencing technologies such as Pacific Biosciences may be beneficial; (*d*) the F1 score might be vulnerable to any discordance within this cohort due to a small number of resistant samples; and (*e*) the ClusterV algorithm relies on accurate calling of variants. A somatic variant calling is more suitable for performing variant calling of quasispecies in HIV than the germline variant calling, However, a somatic small variant caller was not available for ONT data.

In conclusion, an ONT workflow was developed for detecting ARV resistance in HIV with direct sequencing of vRNA extracted from plasma samples. It reached an F1 score of 0.96 with the recommended overall VAF. This workflow also demonstrated that long-read sequencing technologies and hierarchical clustering could synergistically reveal the detailed resistance profile by linking the ARV resistance to different quasispecies. Overall, this approach provides a more comprehensive clinical picture to better enable medical decision-making.

Supplemental Material

Supplemental material is available at *Clinical Chemistry* online.

Nonstandard Abbreviations: ARV, antiretroviral; DRM, drug resistance mutations; LOD, limit of detection; NGS, next-generation sequencing; NRTI, nucleotide/nucleoside reverse transcriptase inhibitors; NNRTI, nonnucleoside reverse transcriptase inhibitors; ONT, Oxford Nanopore Technologies; ROC, receiver operating characteristic; RT, reverse transcriptase; VAF, variant allele frequency.

Author Contributions: *The corresponding author takes full responsibility that all authors on this publication have met the following required criteria of eligibility for authorship: (a) significant contributions to the conception and design, acquisition of data, or analysis and interpretation of data; (b) drafting or revising the article for intellectual content; (c) final approval of the published article; and (d) agreement to be accountable for all aspects of the article thus ensuring that questions related to the accuracy or integrity of any part of the article are appropriately investigated and resolved. Nobody who qualifies for authorship has been omitted from the list.*

Timothy Ting Leung Ng (Conceptualization-Equal, Data curation-Lead, Formal analysis-Lead, Methodology-Equal, Validation-Lead, Visualization-Equal, Writing—original draft-Lead, Writing—review & editing-Lead), Junhao Su (Conceptualization-Equal, Formal analysis-Supporting, Methodology-Equal, Software-Lead, Validation-Supporting, Visualization-Supporting, Writing—original draft-Supporting, Writing—review & editing-Supporting), Hiu-Yin Lao (Data curation-Equal, Formal analysis-Equal, Validation-Equal, Visualization-Equal, Writing—review & editing-Equal), Wui-Wang Lui (Conceptualization-Supporting, Methodology-Supporting, Software-Supporting, Validation-Supporting), Chloe Toi-Mei Chan (Data curation-Equal, Formal analysis-Equal, Validation-Supporting, Writing—review & editing-Equal), Amy Wing-Sze Leung (Conceptualization-Supporting, Methodology-Supporting, Software-Supporting, Supervision-Supporting, Writing—original draft-Supporting), Stephanie Hoi-Ching Jim (Data curation-Equal, Writing—original draft-Supporting), Lam-Kwong Lee (Data curation-Equal), Sheeba Shehzad (Data curation-Supporting), Kingsley King-Gee Tam (Data curation-Supporting, Resources-Supporting), Kenneth Siu Sing Leung (Data curation-Supporting, Resources-Supporting), Forrest Tang (Data curation-Supporting), WC YAM (Resources-Supporting, Supervision-Supporting), Ruibang Luo (Conceptualization-Equal, Project administration-Equal, Software-Equal, Supervision-Equal, Writing—review & editing-Supporting), and Gilman Kit Hang SIU (Conceptualization-Equal, Funding acquisition-Lead, Project administration-Lead, Supervision-Equal, Writing—review & editing-Equal)

Authors' Disclosures or Potential Conflicts of Interest: *Upon manuscript submission, all authors completed the author disclosure form.*

Research Funding: This study was supported by the AIDS Trust Fund of Hong Kong (Ref no. MSS 299 R); R.Luo received support from ONT and partial support from GRF (17113721).
Disclosures: None declared.

Role of Sponsor: The funding organization played no role in the design of study, choice of enrolled patients, review and interpretation of data, preparation of manuscript, or final approval of manuscript.

Data Availability: ClusterV is open-source software (BSD 3-Clause license), hosted by GitHub at https://github.com/HKU-BAL/ClusterV.

Acknowledgments: We would like to thank Professor Thomas Klimkait, Head of Research Group Molecular Virology, Department Biomedicine—Petersplatz, University of Basel, and European Virus Archive Global for the provision of HIV plasmid that was crucial for the validation of targeted sequencing workflow.

References

1. UNAIDS. Global HIV & AIDS statistics—fact sheet. https://www.unaids.org/en/resources/factsheet (Accessed January 2023).
2. Metzner KJ, Giulieri SG, Knoepfel SA, Rauch P, Burgisser P, Yerly S, et al. Minority quasispecies of drug-resistant HIV-1 that lead to early therapy failure in treatment-naive and -adherent patients. Clin Infect Dis 2009;48:239–47.
3. Hedskog C, Mild M, Jernberg J, Sherwood E, Bratt G, Leitner T, et al. Dynamics of HIV-1 quasispecies during antiviral treatment dissected using ultra-deep pyrosequencing. PLoS One 2010;5:e11345.
4. Kapoor A, Shapiro B, Shafer RW, Shulman N, Rhee SY, Delwart EL. Multiple independent origins of a protease inhibitor

resistance mutation in salvage therapy patients. Retrovirology 2008;5:7.
5. Kyeyune F, Gibson RM, Nankya I, Venner C, Metha S, Akao J, et al. Low-frequency drug resistance in HIV-infected Ugandans on antiretroviral treatment is associated with regimen failure. Antimicrob Agents Chemother 2016;60:3380–97.
6. Woods CK, Brumme CJ, Liu TF, Chui CK, Chu AL, Wynhoven B, et al. Automating HIV drug resistance genotyping with RECall, a freely accessible sequence analysis tool. J Clin Microbiol 2012;50:1936–42.
7. Kingwara L, Karanja M, Ngugi C, Kangogo G, Bera K, Kimani M, et al. From sequence data to patient result: a solution for HIV drug resistance genotyping with exatype, end to end software for pol-HIV-1 sanger based sequence analysis and patient HIV drug resistance result generation. J Int Assoc Provid AIDS Care 2020;19: 2325958220962687.
8. Arias A, Lopez P, Sanchez R, Yamamura Y, Rivera-Amill V. Sanger and next generation sequencing approaches to evaluate HIV-1 virus in blood compartments. Int J Environ Res Public Health 2018;15:8.
9. Manyana S, Gounder L, Pillay M, Manasa J, Naidoo K, Chimukangara B. HIV-1 drug resistance genotyping in resource limited settings: current and future perspectives in sequencing technologies. Viruses 2021;13:6.
10. Ode H, Matsuda M, Matsuoka K, Hachiya A, Hattori J, Kito Y, et al. Quasispecies analyses of the HIV-1 near-full-length genome with Illumina MiSeq. Front Microbiol 2015; 6:1258.
11. Hjelm LN, Chin EL, Hegde MR, Coffee BW, Bean LJ. A simple method to confirm and size deletion, duplication, and insertion mutations detected by sequence analysis. J Mol Diagn 2010;12:607–10.
12. Monaco DC, Zapata L, Hunter E, Salomon H, Dilernia DA. Resistance profile of HIV-1 quasispecies in patients under treatment failure using single molecule, real-time sequencing. AIDS 2020;34:2201–10.
13. Lee ER, Parkin N, Jennings C, Brumme CJ, Enns E, Casadella M, et al. Performance comparison of next generation sequencing analysis pipelines for HIV-1 drug resistance testing. Sci Rep 2020;10:1634.
14. World Health Organization. HIVResnet HIV drug resistance laboratory operational framework. Geneva (Switzerland): World Health Organization; 2020.
15. Yamaguchi J, Olivo A, Laeyendecker O, Forberg K, Ndembi N, Mbanya D, et al. Universal target capture of HIV sequences from NGS libraries. Front Microbiol 2018;9: 2150.
16. Link RW, De Souza DR, Spector C, Mele AR, Chung C-H, Nonnemacher MR, et al. HIV-Quasipore: a suite of HIV-1-specific nanopore basecallers designed to enhance viral quasispecies detection. Front Virol 2022;2:858375.
17. Wright IA, Delaney KE, Katusiime MGK, Botha JC, Engelbrecht S, Kearney MF, van Zyl GU. NanoHIV: a bioinformatics pipeline for producing accurate, near full-length HIV proviral genomes sequenced using the Oxford nanopore technology. Cells 2021;10:2577.
18. Chen JH, Wong KH, Li PC, Chan KK, Lee MP, To SW, Yam WC. In-house human immunodeficiency virus-1 genotype resistance testing to determine highly active antiretroviral therapy resistance mutations in Hong Kong. Hong Kong Med J 2012;18:20–4.
19. To SW, Chen JH, Wong KH, Chan KC, Ng HM, Wu H, et al. Performance comparison of an in-house integrase genotyping assay versus the ViroSeq Integra48, and study of HIV-1 integrase polymorphisms in Hong Kong. J Clin Virol 2013;58:299–302.
20. Zheng Z, Li S, Su J, Leung AW-S, Lam T-W, Luo R. Symphonizing pileup and full-alignment for deep learning-based long-read variant calling. Nat Comput Sci 2022;2:797–803.
21. Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics 2009;25:1754–60.
22. Danecek P, Bonfield JK, Liddle J, Marshall J, Ohan V, Pollard MO, et al. Twelve years of SAMtools and BCFtools. Gigascience 2021;10:2.
23. Robinson JT, Thorvaldsdottir H, Turner D. Mesirov JP. Igv.js: an embeddable JavaScript implementation of the Integrative Genomics Viewer (IGV). Bioinformatics 2023;39:1.
24. Dreyer V, Utpatel C, Kohl TA, Barilar I, Groschel MI, Feuerriegel S, Niemann S. Detection of low-frequency resistance-mediating SNPs in next-generation sequencing data of Mycobacterium tuberculosis complex strains with binoSNP. Sci Rep 2020;10:7874.
25. Spencer DH, Tyagi M, Vallania F, Bredemeyer AJ, Pfeifer JD, Mitra RD, Duncavage EJ. Performance of common analysis methods for detecting low-frequency single nucleotide variants in targeted next-generation sequence data. J Mol Diagn 2014;16:75–88.
26. Van Poelvoorde LAE, Delcourt T, Coucke W, Herman P, De Keersmaecker SCJ, Saelens X, et al. Strategy and performance evaluation of low-frequency variant calling for SARS-CoV-2 using targeted deep Illumina sequencing. Front Microbiol 2021;12:747458.
27. Gianella S, Delport W, Pacold ME, Young JA, Choi JY, Little SJ, et al. Detection of minority resistance during early HIV-1 infection: natural variation and spurious detection rather than transmission and evolution of multiple viral variants. J Virol 2011;85: 8359–67.
28. Dalmat RR, Makhsous N, Pepper GG, Magaret A, Jerome KR, Wald A, Greninger AL. Limited marginal utility of deep sequencing for HIV drug resistance testing in the age of integrase inhibitors. J Clin Microbiol 2018;56:12.
29. Maruapula D, Seatla KK, Morerinyane O, Molebatsi K, Giandhari J, de Oliveira T, et al. Low-frequency HIV-1 drug resistance mutations in antiretroviral naive individuals in Botswana. Medicine (Baltimore) 2022; 101:e29577.
30. Boltz VF, Shao W, Bale MJ, Halvas EK, Luke B, McIntyre JA, et al. Linked dual-class HIV resistance mutations are associated with treatment failure. JCI Insight 2019; 4:19.
31. Martin F, Lee J, Thomson E, Tarrant N, Hale A, Lacey CJ. Two cases of possible transmitted drug-resistant HIV: likely HIV superinfection and unmasking of pre-existing resistance. Int J STD AIDS 2016; 27:66–9.
32. Smith DM, Wong JK, Hightower GK, Ignacio CC, Koelsch KK, Daar ES, et al. Incidence of HIV superinfection following primary infection. JAMA 2004; 292:1177–8.
33. Brenner B, Routy JP, Quan Y, Moisi D, Oliveira M, Turner D, et al. Persistence of multidrug-resistant HIV-1 in primary infection leading to superinfection. AIDS 2004; 18:1653–60.
34. Kuniholm J, Coote C, Henderson AJ. Defective HIV-1 genomes and their potential impact on HIV pathogenesis. Retrovirology 2022;19:13.