DEVELOPMENT OF AMPLICON SEQUENCING-BASED PROTOCOLS FOR DIRECT DETECTION OF ANTIMICROBIAL RESISTANCE IN *MYCOBACTERIUM TUBERCULOSIS* AND HUMAN IMMUNODEFICIENCY VIRUS FROM CLINICAL SAMPLES

NG TING LEUNG TIMOTHY

PhD

2023

The Hong Kong Polytechnic University

Department of Health Technology and Informatics

Development of amplicon sequencing-based protocols for direct detection of antimicrobial resistance in *Mycobacterium tuberculosis* and Human Immunodeficiency Virus from clinical samples

Ng Ting Leung Timothy

A thesis submitted in partial fulfillment of the requirements for the degree of

Doctor of Philosophy

December 2022

CERTIFICATE OF ORIGINALITY

I hereby declare that this thesis is my own work and that, to the best of my knowledge and belief, it reproduces no material previously published or written, nor material that has been accepted for the award of any other degree or diploma, except where due acknowledgement has been made in the text.

__________________________ (Signed)

Ng Ting Leung Timothy

1. Abstract

Direct sequencing of clinical specimens is now a trending approach for antibiotic resistance detection in target pathogenic microorganism as it can greatly reduce the time to report to a few working days that favors the choice of the appropriate regimens and the management of resource and samples. To further understand the process in developing sequencing workflows for this application, two target sequencing workflows: direct antimicrobial resistance (AMR) detection in *Mycobacterium tuberculosis* (*MTB*) in sputum samples and direct antiretroviral (ARV) resistance detection in human immunodeficiency virus 1 (HIV) in plasma samples, were successfully developed. The target sequencing workflow for AMR detection in MTB achieved 100% agreement between nanopore sequencing by Oxford Nanopore Technologies (ONT) and Illumina next generation sequencing (NGS) for the *MTB* DNA contents in samples above the limit of detection (LOD), while the target sequencing workflow for AVR resistance detection in HIV achieved high F1 score 0.918, or even 0.96 with a threshold from ROC analysis. The hierarchical clustering used in HIV sequencing workflow could even provide a detailed AVR resistance profile by associating the AVR resistance-associated amino acid mutation patterns with different quasispecies in the same samples. The success of these workflows proved the working principle of direct resistance detection in clinical specimens that requires only a few working days to report. Other than exploring the power of sequencing technologies, challenges for the workflow development were also highlighted. Both index misassignment and background nasal/oral flora (especially in sputum samples with low *MTB* gDNA content) can cause contamination that can lead to false results. The recommendations in this study including the choice of the index set, and

the employment of decoy strategy can minimize the impact of these issues. The above findings can be a reference for the future drug resistance workflow development for other infectious diseases.

Publications and presentations

Publications arisen from this study

1) **Ng TTL,** Su J, Lao HY, Lui WW, Chan CTM, Leung AWS, Jim SHC, Lee LK, Shehzad S, Tam KKG, Leung KSS, Tang F, Yam WC, Luo RB, Siu GKH. Development of Long-Read Sequencing Protocol and Hierarchical Clustering Pipeline for Antiretroviral Resistance Profiling in Clinical Samples with Mixed Human Immunodeficiency Virus Quasispecies. *Clin Chem*. 2023 (Submitted and Under review).

2) Tafess K, **Ng TTL**, Lao HY, Leung KSS, Tam KKG, Rajwani R, Tam STY, Ho LPK, Chu CMK, Gonzalez D, Sayada C, Ma OCK, Nega BH, Ameni G, Yam WC, Siu GKH. Targeted-Sequencing Workflows for Comprehensive Drug Resistance Profiling of Mycobacterium tuberculosis Cultures Using Two Commercial Sequencing Platforms: Comparison of Analytical and Diagnostic Performance, Turnaround Time, and Cost. *Clin Chem*. 2020 Jun 1;66(6):809-820. doi: 10.1093/clinchem/hvaa092. PMID: 32402055.

3) Leung KS, Tam KK, **Ng TT**, Lao HY, Shek RC, Ma OCK, Yu SH, Chen JX, Han Q, Siu GK, Yam WC. Clinical utility of target amplicon sequencing test for rapid diagnosis of drug-resistant *Mycobacterium tuberculosis* from respiratory specimens. *Front Microbiol*. 2022 Sep 9;13:974428. doi: 10.3389/fmicb.2022.974428. PMID: 36160212; PMCID: PMC9505518.

Co-authored articles

1) Leung KS, **Ng TT**, Wu AK, Yau MC, Lao HY, Choi MP, Tam KK, Lee LK, Wong BK, Man Ho AY, Yip KT, Lung KC, Liu RW, Tso EY, Leung WS, Chan MC, Ng YY, Sin KM, Fung KS, Chau SK, To WK, Que TL, Shum DH, Yip SP, Yam WC, Siu GK. Territorywide Study of Early Coronavirus Disease Outbreak, Hong Kong, China. Emerg Infect Dis. 2021 Jan;27(1):196-204. doi: 10.3201/eid2701.201543. PMID: 33350913; PMCID: PMC7774584.
2) Mok BW, Liu H, Deng S, Liu J, Zhang AJ, Lau SY, Liu S, Tam RC, Cremin CJ, **Ng TT**, Leung JS, Lee LK, Wang P, To KK, Chan JF, Chan KH, Yuen KY, Siu GK, Chen H. Low dose inocula of SARS-CoV-2 Alpha variant transmits more efficiently than earlier variants in hamsters. Commun Biol. 2021 Sep 20;4(1):1102. doi: 10.1038/s42003-021-02640-x. PMID: 34545191; PMCID: PMC8452646
3) Lao HY, **Ng TT**, Wong RY, Wong CS, Lee LK, Wong DS, Chan CT, Jim SH, Leung JS, Lo HW, Wong IT, Yau MC, Lam JY, Wu AK, Siu GK. The Clinical Utility of Two High-Throughput 16S rRNA Gene Sequencing Workflows for Taxonomic Assignment of Unidentifiable Bacterial Pathogens in Matrix-Assisted Laser Desorption Ionization-Time of Flight Mass Spectrometry. J Clin Microbiol. 2022 Jan 19;60(1):e0176921. doi: 10.1128/JCM.01769-21. Epub 2021 Nov 17. PMID: 34788113; PMCID: PMC8769742.
4) Junhao Su, Wui Wang Lui, YanLam Lee, Zhenxian Zheng, Gilman Kit-Hang Siu, **Timothy Ting-Leung Ng**, Tong Zhang, Tommy Tsan-Yuk Lam, Hiu-Yin Lao, Wing-Cheong Yam, Kingsley King-Gee Tam, Kenneth Siu-Sing Leung, Tak-Wah Lam, Amy Wing-Sze Leung, Ruibang Luo, Evaluation of Mycobacterium Tuberculosis enrichment in metagenomic samples using ONT adaptive sequencing and amplicon sequencing for

identification and variant calling, bioRxiv preprint doi: https://doi.org/10.1101/2022.12.17.520855.

Presentations

1) **Timothy Ting-Leung Ng** and Gilman KH Siu, The journey to rapid antibiotic resistance detection in Mycobacterium tuberculosis with direct sequencing of sputum samples: the benefits, the precautions, and the solution, Poster presentation, Poster presentation, International Conference and Expo on Applied Microbiology (ICAM 2022), 17/6/2022 - 18/6/2022, virtial event
2) **Timothy Ting-Leung Ng** and Gilman KH Siu, The journey to rapid antibiotic resistance detection in Mycobacterium tuberculosis with direct sequencing of sputum samples: the benefits, the precautions, and the solution, Poster presentation, Australian Society for Microbiology Annual Scientific Meeting 2022, 11/7/2022 - 14/7/2022, Sydney and virtual event

Acknowledgement

It is hard to believe that I had an opportunity to be a PhD candidate and continued my journey in academic research. I would like to thank Dr. Gilman KH Siu for this treasurable opportunity and his guidance throughout this journey. I would also like to thank our collaborator Dr. Ruibang Luo from Department of Computer Science in the University of Hong Kong and his colleagues Dr. Amy Leung, Edward Lui, and Junhao Su for assisting the data analysis and sharing research ideas. Also, it is my pleasure to with our collaborator Dr. Kingsley Tam and Dr. Kenneth Leung in Dr. WC Yam's team in Department of Microbiology in the University of Hong Kong for the support in clinical samples, and Professor Thomas Klimkait, Head of Research Group Molecular Virology, Department Biomedicine – Petersplatz, University of Basel, and European Virus Archive Global (EVAG), for the provision of HIV plasmid, that is crucial for the development of target sequencing workflow for HIV. I would like to express my appreciation to my teammates Hiuyin Lao, Chloe Chan, and Stephanie Jim for their support in the research. Their kindness in providing me with suggestions helped me to overcome many challenges within this research period. It was my pleasure to discuss science and share research ideas with you all.

Finally, I am extremely grateful for the support from my lovely wife Jane Fung and adorable daughter Natalie Ng for their love and support for my PhD studies. They are always my inspiration and motivation to keep me walking forward in my research, and my career.

May your life be filled with love, warmth, joy, and good luck. God bless you all.

Contents

1. Abstract 4

Contents 10

Abbreviations 12

2. Introduction 16

2-1. Tuberculosis (TB) 17

2-2. Anti-TB drugs and antimicrobial resistance 19

2-2-1. Isoniazid (INH) 19

2-2-2. Rifampicin (RIF) 20

2-2-3. Ethambutol (EMB) 21

2-2-4. Pyrazinamide (PZA) 22

2-2-5. Streptomycin (STR) 23

2-2-6. Bedaquiline (BDQ) 23

2-2-7. Fluoroquinolones (FQ) 24

2-2-8. Second line injection drugs (SLIDs) (kanamycin (KAN), capreomycin (CAP), and amikacin (AMK)) 25

2-2-9. Linezolid (LZD) 26

2-2-10. Summary 27

2-3. Human immunodeficiency virus and acquired immunodeficiency syndrome (HIV/AIDS) 28

2-4. Antiretroviral therapy (ART) and drug resistance 29

2-4-1. Nucleoside RT inhibitors (NRTI) 29

2-4-2. Non-nucleoside reverse-transcriptase inhibitors (NNRTI) 31

2-4-3. Protease inhibitors (PI) 31

2-4-4. Integrase Strand Transfer Inhibitor (INSTI) 32

2-4-5. Summary 33

2-5. Adoption of sequencing technologies to detect the drug resistance detection in clinical laboratories – the motivation of this study 34

2-5-1. Direct AMR detection in MTB in sputum samples 34

2-5-2. Direct ARV resistance detection in HIV in plasma samples 36

2-6. The objectives 38

3. Materials and methods 39

3-1. Targeted sequencing workflow for direct AMR detection in MTB in sputum samples 39

3-2. Targeted sequencing workflow for direct ARV resistance detection in HIV in plasma samples 51

4. Results 60

4-1. Targeted sequencing workflow for direct AMR detection in MTB in sputum samples 60

4-1-1. Assay optimization 60

4-1-2. The potential source of index misassignment in ONT and NGS 65

4-1-3. Background nasal/oral flora interference 68

4-1-4. Limit of Detection 83

4-1-5. The diagnostic accuracy of the drug resistance detection workflow 86

4-1-6. Time to report and operation cost 92

4-2. Targeted sequencing workflow for direct ARV resistance detection in HIV in plasma samples 94

4-2-1. Validation of clustering performance in the ONT workflow 100

4-2-2. The variant calling and diagnostic performance of the targeting ONT sequencing workflow 104

4-2-3. Detection of the mutations in the quasispecies 109

4-2-4. Limit of detection 111

4-2-5. Time to report and operation cost 113

5. Discussion 115

5-1. Target sequencing workflow for direct AMR detection in MTB in sputum samples 115

5-2. Targeted sequencing workflow for direct ARV resistance detection in HIV in plasma samples 119

5-3. Lessons from the target sequencing workflow development 128

5-3-1. The adoption of sequencing technologies to drug resistance detection in clinical laboratories 128

5-3-2. Direct whole genome sequencing versus target sequencing 128

5-3-3. Contamination - the potential source of false results 129

5-3-4. The database – the key element in sequencing workflow development 130

5-3-5. The room for improvement in the workflow development 131

6. Conclusion 133

7. References 134

Abbreviations

Abacavir (ABC)

Acquired immunodeficiency syndrome (AIDS)

Alignment score (AS)

Amikacin (AMK)

Aminoacyl transfer RNA (tRNA)

Amprenavir (APV)

Antimicrobial resistance (AMR)

Antiretroviral (ARV)

Antiretroviral therapy (ART)

Atazanavir (ATV)

Bedaquiline (BDQ)

Bictegravir (BIC)

Binary alignment map (BAM)

Browser Extensible Data (BED)

Cabotegravir (CAB)

Capreomycin (CAP)

Clofazimine (CFZ)

Coenzyme A (CoA)

Complementary DNA (cDNA)

Cycloserine (CS)

Darunavir (DRV)

Decaprenylphosphate (DP)

Decaprenylphosphoryl-beta-D-5-phosphoribose (DPPR)

Decaprenylphosphoryl-D-arabinose (DPA)

Delamanid (DLM)

Didanosine (ddI)

Dolutegravir (DTG)

Doravirine (DOR)

Efavirenz (EFV)

Elvitegravir (EVG)

Emtricitabine (FTC)

Envelope glycoprotein (ENV)

Ethambutol (EMB)

Ethionamide (ETO)

Etravirine (ETR)

Expanded Human Oral Microbiome Database (eHOMD)

Extensively drug-resistant TB (XDR TB)

Fatty acid synthesis type II system (FASII)

Fluoroquinolones (FQ)

Genomic DNA (gDNA)

Highly active antiretroviral therapy treatment (HAART)

HIV Drug Resistance Database (HIVDB)

Human immunodeficiency virus (HIV)

Imipenem–cilastatin (IPM-CLN)

Indinavir (IDV)

Integrase (INT)

Integrase Strand Transfer Inhibitor (INSTI)

Isoniazid (INH)

Kanamycin (KAN)

Lamivudine (3TC)

Levofloxacin (LFX)

Limit of detection (LOD)

Linezolid (LZD)

Mapping quality (MAPQ)

Meropenem (MPM)

Minimum inhibition concentration (MIC)

Moxifloxacin (MFX)

Multiple drug resistance tuberculosis (MDR-TB)

Mycobacterium tuberculosis (MTB)

Mycolic acid (MA)

Nelfinavir (NFV)

Nevirapine (NVP)

Next generation sequencing (NGS)

Nicotinamide adenine dinucleotide (NADH)

Non-nucleoside reverse-transcriptase inhibitors (NNRTI)

Non-template control (NTC)

Nucleoside RT inhibitors (NRTI)

Oxford Nanopore Technologies (ONT)

P-aminosalicylic acid (PAS)

Phenotypic drug susceptibility test (pDST)

Phosphoribose diphosphate (pRpp)

Polymerase chain reaction (PCR)

Pretomanid (PA)

Protease inhibitors (PI)

Pyrazinamide (PZA)

Raltegravir (RAL)

Receiver operating characteristic (ROC)

Reverse transcriptase (RT)

Ribosomal ambiguity (RAM)

Rifampicin (RIF)

Rifampicin resistant tuberculosis (RR-TB)

Rilpivirine (RPV)

RNA polymerase (RNAP)

Saquinavir (SQV)

Second-line injection drugs (SLIDs)

Stavudine (d4T)

Streptomycin (STR)

Susceptibility test (ST)

Tenofovir (TDF)

Terizidone (TRD)

Thymidine analog mutations (TAMs)

Tipranavir (TPV)

Tuberculosis (TB)

Unique dual index (UDI)

Variant allele frequency (VAF)

Viral RNA (vRNA)

Zidovudine (AZT)

2. Introduction

In clinical microbiology, the adoption of sequencing technology to clinical services, including pathogen identification and antimicrobial resistance (AMR), is currently a research interest worldwide. For example, the *16S rRNA* gene is one major marker for bacterial species identification. Sequencing reads are mapped to the reference *16S rRNA* sequences in a database [1]. The bacteria species is called when a read is mapped with the highest mapping score. Alternatively, fragmented genomic DNA sequencing reads are mapped to databases such as RefSeq, covering bacteria, fungi, viruses, and even AMR associated mobile plasmids [2]. A profile of identified species and AMR are reported.

Another application is the detection of known chromosomal mutations that are associated with the AMR gain in target bacteria or viruses. Variants are called after sequencing reads are mapped to the target reference genome. Those variants associated with AMR in a database are annotated. Tuberculosis (TB) and acquired immunodeficiency syndrome (AIDS) caused by human immunodeficiency virus (HIV) are two examples demonstrating how sequencing technologies are applied to detect drug resistance, revealing the current drug resistance distribution within a community, and clinically provide a reference for medical prescriptions.

2-1. Tuberculosis (TB)

Tuberculosis is a chronic disease caused by *Mycobacterium tuberculosis* (*MTB*). In the year 2020, there were around 5.8 million newly reported cases, and pulmonary TB accounted for 82 percent (4.8 million) of the cases. Though the number of newly reported cases declined by 18% compared with the year 2019, the percentage of rifampicin resistant tuberculosis (RR-TB) and multiple drug resistance tuberculosis (MDR-TB) among the pulmonary TB cases climbed from 61% (in the year 2019) to 71% (in the year 2020). With the economic impact of coronavirus pandemic in the year 2019 and the rising global price index, the decline in family income may discourage both TB diagnosis and treatment, especially for low-income families, and then it may worsen the TB incidence [3].

Multiple lines of antibiotics are available to combat *MTB* [4]. Once a patient has been confirmed to be TB-positive, an immediate 6-month standard of first line drug treatment phase is initiated. A combination of isoniazid (INH) and rifampicin (RIF), pyrazinamide (PZA), and either ethambutol (EMB) or streptomycin (STR) is prescribed for two months, followed by the extended prescription of INH and RIF for four months. Extension of the treatment may be required until the patient is completely cured or until the release of the drug susceptibility test (ST) results, in the case of suspected drug resistant TB. For rifampicin-resistant (RR) and multidrug-resistant (MDR) TB, there are two major regimens recommended by WHO (World Health Organization): the traditional longer treatment regimen (18 – 20 months) and the recently recommended shorter treatment regimen (an oral regime, 9 – 12 months), covering a combination of partial first ling drugs and the second line drugs. On the longer regimen, the patients are prescribed with all drugs in Group A (levofloxacin (LFX) or moxifloxacin (MFX) (members of fluoroquinolones (FQ), BDQ, and linezolid (LZD)), combined with at least one of the drugs in Group B (CFZ and cycloserine (CS) or terizidone (TRD)), while drugs in Group C (EMB, PZA, delamanid (DLM), Imipenem–cilastatin (IPM-CLN) or meropenem (MPM), AMK, ETO, PTO, and P-aminosalicylic acid (PAS) are used as reserve if any key drugs in Group A and Group B is unavailable or inappropriate. In the newly recommended shorter regimen (an oral regime), oral-based bedaquiline (BDQ) is one key drug used as a replacement for some second-line injection drugs (SLIDs) such as kanamycin (KAN), capreomycin (CAP), and amikacin (AMK), combining with LFX or MFX and other drugs such as ethionamide (ETO), EMB, INH, PZA, and clofazimine (CFZ). A higher cure rate is reported with the shorter regime than with the longer regime, and the shorter regime is thought to be favorable to treatment coverage, patient follow-up, and so TB management. However, it is limited to patients with exposure to the second line drug treatment for less than one month, lower disease severity, and no report of FQ-resistance in *MTB*. For MDR/RR-TB patients with the additional resistance to FQ, or even the additional resistance to at least one of the SLIDs (XDR-TB), a regimen composed of BQ, pretomanid (PA), and LZD, commonly known as the BPAL regime (6 – 9 months), is introduced. In the study by the NIX-TB Trial Team, a high positive rate of treatment outcome for BPAL (90%, 98 out of 109 patients) meant that the patients were cured with the culture conversion at 6 months after the end of the BPAL treatment [5]. Of which, the positive treatment outcomes for patients with XDR-TB and MDR-TB were 89% (63 out of 71 patients) and 92% (35 out of 38 patients), respectively. Of note, different degrees of adverse effects associated with

LZD were commonly found in the patients [5], reduced dosage may be an option. Careful patient support and management for handling these adverse effects are recommended.

2-2. Anti-TB drugs and antimicrobial resistance

2-2-1. Isoniazid (INH)

Isoniazid is a key first line drug that is widely used for treating TB. The full working mechanism is still being studied as it involves several pathways. One important mechanism is the inhibition of mycolic acid (MA) synthesis, a long fatty acid chain that is a key component of the thick cell wall in *MTB* [6]. The thick cell wall provides *MTB* protection from hydrophilic antibiotics and macrophage invasion. After the pro-INH is converted to active INH, it is then converted to an isonicotinic acyl radical with catalase-peroxidase encoded by the gene *katG*. This radical is then combined with the reduced form of nicotinamide adenine dinucleotide (NADH) and finally becomes the INH-NADH adduct. The adduct binds to the active site of 2-trans-enoyl-acyl carrier protein reductase (encoded by the gene *inhA*) and hinders the access of the original substrate, trans-2-enoyl-ACP, and so blocks the MA synthesis in the fatty acid synthesis type II (FASII) system.

Genomic mutations in the related genes are associated with INH resistance. For example, a highly prevalent genomic mutation led to an amino acid change at position 315 from serine to threonine (S315T) in *katG,* which confers *MTB* INH resistance by reducing its affinity for INH [7, 8]. Another typical mutation is C-15T in the promoter region of the gene *inhA,* which enhances the expression of 2-trans-enoyl-acyl carrier protein reductase [9]. Either of these two mutations alone may not necessarily totally impair the treatment efficiency with the high dose of INH (15–20 mg/kg), but the combination of these two mutations may lead to clinically ineffective doses [10]. Also, mutations (G-7A, A-10C, and G-12A) in the *furA-katG* intergenic region can downregulate the *katG* expression and so confer INH resistance [11].

2-2-2. Rifampicin (RIF)

Rifampicin is another key first-line drug that is usually prescribed with INH. The binding of RIF to RNA polymerase (RNAP) inhibits RNA synthesis by blocking RNA elongation beyond the 2nd or 3rd nucleotide in the RNAP beta subunit [12]. RIF contains several atoms that are responsible for binding to the RNAP beta subunit, which is encoded by the gene *rpoB*. Some known mutations on *rpoB* cause amino acid changes and weaken the binding affinity to RIF, and so confer *MTB* RIF resistance. These mutations are distributed in the 81-bp hotspot region (the codon range between 507 and 533) of gene *rpoB*, and they account for at least eighty percent of the reported RR cases [13]. Of which, genomic mutations leading to the amino acid change H445Y and S450L (H526Y and S531L in *Escherichia coli* respectively) confer high-level *MTB* high level resistance to RIF. The reported occurrence in codon 450 can be noticeably high at 60% - 80%, while the occurrence in codon 445 may largely vary from 3% to 30%, based on studies of cohorts of RIF resistant isolates held in China, Georgia, and Angola [14-16]. Interestingly, these two mutations also add to the *MTB* fitness burden. Compared with the wild-type strain, a reduced growth rate in hypoxia or poor nutrient conditions was observed in *MTB* carrying the H445D mutation [17]. Also, the mutation S450L could lead to a reduced growth rate, and such a cost could be compensated with secondary mutations in genes *rpoA*, *rpoB*, and *rpoC* [18]. Transcriptomic analysis done by Xu et al. suggested the *rpoC* mutation might rescue the impairment of oxidative respiratory pathway caused by the *rpoB* mutation for *MTB* exposed to RIF [19]. This may partially explain how the fitness cost is compensated with mutations on other RNAP subunits such as *rpoA* and *rpoC*.

2-2-3. Ethambutol (EMB)

Ethambutol is a key first line drug that is usually used with INH, RIF, and sometimes second line drugs. The mechanism of action is the binding of EMB to the active site of arabinosyltransferase EmbB of the EmbA-EmbB complex and the active site of the EmbC-EmbC complex, which inhibits the arabinose transfer that is essential for the synthesis of the special cell wall complex in *MTB* [20]. A recent study suggested another mechanism by which EMB enhanced the DNA binding of a repressor protein encoded by gene *Rv0273c* (EtbR) to *inhA* promoter, repressed the expression of *inhA*, enhanced the INH susceptibility [21]. Amino acid change at codon 306 of *EmbB* gene decreases the binding affinity of EMB to the active site without significant change in arabinose transferase activity, and so it increases the minimum inhibition concentration (MIC) by two to four folds [22], while mutation on codon 406 confers *MTB* mild EMB resistance [23]. Analysis of the prevalence of EMB resistance showed that the majority of the EMB resistant samples (80% to 90%) carried mutations in *embB*. Of which, the mutation frequency on codon 306 (including M306V and M306I) ranged from 30% - 75%, while the mutation frequency on codon 406 was much lower at less than 6% in EMB resistant samples [22, 24, 25].

Gene *ubiA* encoding 5-phospho-alpha-d-ribose-1-diphosphate: decaprenyl-phosphate 5-phosphoribosyltransferase is also associated with the EMB resistance. This enzyme converts phosphoribose diphosphate (pRpp) and decaprenylphosphate (DP) to decaprenylphosphoryl-beta-D-5-phosphoribose (DPPR), which is further converted to decaprenylphosphoryl-D-arabinose (DPA) in the downstream process within the DPA pathway. The DPA is a substrate for EmbB in cell wall synthesis [26]. Point mutations on codons 188, 237, 240, and 249 included in the transmembrane domain led to overexpression of *ubiA* and then increased the DPA level. DPA competes with EMB for the active sites, resulting the MTB EMB resistance [27].

2-2-4. Pyrazinamide (PZA)

Pyrazinamide is a drug that is commonly used in first- and second- line treatment regimens. The drug is converted to pyrazinoic acid (POA) by hydrolysis with pyrazinamidase or nicotinamidase encoded by the gene *pncA*. The mechanism of action of POA is still under exploration. One proposed mechanism is the binding of POA to the aspartate decarboxylase encoded by *PanD,* which promotes the *PanD* degradation. This prevents aspartate from being converted to beta-alanine, thereby suppressing downstream Coenzyme A synthesis (CoA) [28]. The high prevalence of mutations on *pncA* accounts for the PZA resistance. Excessive hydrogen bonds caused by mutations D8G, S104R, and C138Y lead to the rigid binding site and hinder the conversion of pyrazinamidase to its active form [29]. Another important mutation on codon 57, such as H57D, affects the coordination of the iron (II) ion in the active site region and inactivates the pyrazinamidase [30].

2-2-5. Streptomycin (STR)

Streptomycin is a well-known antibiotic that is used to treat a wide range of infectious diseases, including tuberculosis and plague. After binding to several nucleotides in 16S ribosomal RNA (gene *rrs*) and several amino acid residues in S12 protein encoded by gene *rpsL* in ribosomal subunit S30, this causes conformational change of the helices in the decoding site and then codon mismatch to aminoacyl transfer RNA (tRNA), and so intervenes in the normal protein synthesis [31]. Mutations in gene *rrs* were associated with STR resistance. On the other hand, mutations in gene *rpsL* lead to hyperaccurate phenotypes by destabilizing the ribosomal ambiguity (RAM) state that compensates for the stabilization caused by STR, and so the translation returns to normal [32]. The mutation frequency in *rrs* for STR resistance cases ranged from 16% to 44%, while the mutation frequency ranged from 31% to 63% in *rpsL*, according to the studies in China and Iran [33, 34].

2-2-6. Bedaquiline (BDQ)

Bedaquiline was approved for TB treatment in 2012 and is now a key antibiotic for combating MDR-TB. It shows promising bactericidal results even against the non-replicating MTB [35]. BDQ binds to the c-unit of F-ATP synthase in mycobacteria and hinders the rotation of the c-unit, which blocks the ion exchange between the periplasm and cytoplasm in the electron transport chain

[36, 37]. The ATP synthesis is eventually shut down, which slows down the cellular activities, and finally leads to cell death. However, mutations at codon 63 alone at the gene *atpE* encoding the c-unit of F-ATP synthase can raise the MIC level of MDR-TB culture from 0.05 ug/ mL to the higher level (4-8 ug/ mL), while the mutations at codon 83 and mutations at gene *Rv0678* encoding a repressor of MmpS5-MmpL5 efflux pump only raise the MIC level to around 0.5 ug/ mL [38]. Despite the fact that studies in South Africa and China reported a low percentage of BDQ resistance cases out of the total number of TB cases at the time (199/3005, 7% in South Africa, 6/1603, 0.4% in China) [39, 40], a warning of several cases harboring spontaneous mutations associated with BDQ resistance prior to BDQ treatment was reported [41]. Mutations at *Rv0678* are dominant among the BDQ resistant cases. More statistical information is required for close monitoring of the spread of BDQ resistant strains, which is necessary to safeguard the bactericidal power against MDR-TB.

2-2-7. Fluoroquinolones (FQ)

Among the antibiotics in the fluoroquinolone (FQ) family, levofloxacin (LFX) or moxifloxacin (MFX) are commonly used for treating MDR-TB. These antibiotics inhibit the gyrase complex (an enzyme type of Type II topoisomerases unique in bacteria) from resealing DNA double-strand breaks and form the complex with DNA that further blocks the transcription in the DNA transcription fork, which is toxic to the bacteria and causes cell death [42]. One proposed mechanism is the formation of a water-metal ion bridge between the gyrase and the antibiotics. Mutation D94G in gene *gyrA* strongly disrupts bridge formation and so confers MTB resistance [43]. Mutations D500A, N538T, T539P, and E540V in gene *gyrB* may be associated with low level

FQ resistance [44]. Mutations A90V and D94G in gene *gyrA* are highly frequent in FQ resistant cases, while mutations in gene *gyrB* or double mutations in both *gyrA* and *gyrB* are less common [45-47].

2-2-8. Second line injection drugs (SLIDs) (kanamycin (KAN), capreomycin (CAP), and amikacin (AMK))

Capreomycin is a bactericidal antibiotic used for treating MDR-TB. The mechanism of action is still under exploration. One study suggested it inhibited protein synthesis by blocking the interaction between L10 and L12 and lowering the GTPase activity of Elongation Factor G (EF-G) [48]. It was also proposed that CAP could bind to the ribosome 30S decoding site, which is sandwiched between 50S and 30S rRNA, inhibiting tRNA translocation and, eventually, protein synthesis [49, 50]. Mutations A1401G, C1402T, and G1484T in gene *rrs* are associated with CAP resistance [51]. Also, the binding of CAP to 70S ribosome requires the methylation of cytidine at positions 1409 on 16S and 2158 on 23S rRNA with (cytidine 1920-2'-O)-methyltransferase encoded by gene *tlyA*. Mutations such as N236K at gene *tlyA* lead to structural changes that impair the methylation activity, and so confer MTB resistance to CAP [52]. Based on several studies worldwide [53-56], mutations at gene *rrs* are highly frequent (49.3% - 84.3%) for CAP-resistant cases. Though the high mutation frequency of a non-synonymous mutation A33G is reported in India and Thailand [53, 56], its irrelevance to AMR is suggested. Other novel mutations are rare, and their AMR association is pending confirmation [56].

Kanamycin (KAN) was removed from the World Health Organization's List of Essential Medicines in 2019 because of its severe adverse effects, whereas Amikacin (AMK), an antibiotic synthesized from KAN, is still in use. AMK forms a complex at the A-site of the decoding region in the 30S ribosome subunit as well as the RNA region at the GC pairs C1404–G1497 and G1405–G1496 [57], and then causes the improper reading of mRNA and finally inhibits protein synthesis. Mutations within the RNA region at gene *rrs* are associated with AMK resistance. On the other hand, overexpression of enhanced intercellular survival protein (eis) can acetylate and inactivate AMK [58]. Over transcription of mRNA transcripts is caused by mutations such as -10 in the promoter region, which slightly increases the MIC (from 0.5 to 3 ug/ mL) [59]. Also, the expression of gene *eis* is also regulated by a transcriptional regulator encoded by gene whiB7. Mutations at the untranslated region (UTR) of whiB7 increase its mRNA level and the subsequent expression level of *eis* [60]. A few studies showed mutation frequency at gene *rrs* was dominant in AMK resistance cases (70% - 83%), whereas the mutations at the eis promoter region was less common (0.02-0.17%) (51-53, 59). The mutation at the gene whiB7 UTR region is rare [53, 61].

2-2-9. Linezolid (LZD)

Linezolid is one important antibiotic for treating MDR-TB and even XDR-TB. LZD binds to the ribosomal peptidyl transferase center (PTC) surrounded by the 23S ribosomal RNA in the 50S ribosome subunit and blocks the positioning of aminoacyl tRNA that is essential for peptide transfer [62]. Mutations at the nucleotide position in 23S ribosomal RNA (gene *rrl*), such as 2062 and 2576, are associated with resistance to this antibiotic. Another identified mutation, Cys154Arg, at gene *rplC* encoding 50S ribosomal protein L3, is also associated with resistance to

LZD as a loop of this peptide protrudes to the PTC of 23S ribosomal RNA [63, 64]. In the studies held in South Africa, China, and Moscow (Russia), both hot spot mutations at protein position 154 at gene rplC (around 20% - 90%) and mutations at gene *rrl* (around 12.7% - 30%) contributed to the LZD resistance [65-67].

2-2-10. Summary

While the introduction of second-line and new antibiotics helps to combat the rising prevalence of RR-TB and MDR-TB, antibiotic resistance is a concern. Rapid drug resistance detection tests are necessary to provide the patients with appropriate regimens in order to suppress the growth and spread of drug resistant TB.

With the ongoing research in understanding the mechanisms of action and resistance, more AMR associated mutations have been explored and confirmed. These mutations can serve as genetic markers for revealing the AMR profile, which allows for better choices of regimens. Phenotypic drug susceptibility test (pDST) is considered the gold-standard for detecting AMR in MTB. However, the time to a clinical report is too long (in terms of months) because of the slow growing properties of *MTB*. During this waiting period, patients may miss out on the best regimens. Nucleic acid amplification assays such as Xpert® MTB series offer quick genotypic results within a single working day, but the coverage of genetic markers is limited. With the decreasing running cost for sequencing technologies, many studies reported the adoption of sequencing technologies to detect drug resistance in clinical isolates or even clinical specimens. On the other hand, they can be used for public health surveillance that monitors the evolution and spread of

AMR (especially for second line drugs and new antibiotics) within a community. For example, the discovery of a high proportion (73.6%) of pan-XDR and XDR strains (meaning they also possessed FQ resistance) in MDR cases in Mumbai raises the specter of failure in second-line drug regimens [68].

2-3. Human immunodeficiency virus and acquired immunodeficiency syndrome (HIV/AIDS)

HIV/AIDS is a chronic immunodeficiency disease caused by HIV transmission through the exchange of body fluid such as bloods, breast milk, and secretions from the sex organs (including semen). According to the Joint United Nations Programme on HIV/AIDS (UNAIDS), they estimated that a total of 38.4 million people were infected with HIV, around 650,000 infected people died of HIV-related diseases, and there were around 1.5 million new cases in 2021. Approximately 28.7 million people (~74.7%) were being treated with antiretroviral therapy (ART).

The envelope glycoprotein (ENV) of HIV specifically binds to the surface receptors (CD4 receptor, CCR5, or CXCR4) on the CD4+ T cells. After the binding and fusion of the viral envelope into the cell membrane, the capsid containing two copies of viral RNA along with key enzymes such as reverse transcriptase (RT) and integrase (INT) is delivered to the cytoplasm. The HIV vRNA is converted to vDNA with its RT. This is one of the major targets of antiretroviral (ARV) drugs. The vDNA is then transferred into the nucleus and incorporated into the host cell genome with the aid of viral INT. The INT is another major target for ARV drugs. Essential materials (HIV protein chains and enzymes) and the viral genome are generated by transcription and translation using host cell machinery. The materials and the vRNA are then assembled at the host cell surface, forming a bud that becomes an immature HIV. Finally, immature HIV is converted to mature HIV with viral protein chains (such as *gag* and *gag-pol*) broken down with HIV protease (PR), and the new generation of mature HIV leaves the host cell. The protease is also the major target of ARV drugs. Since HIV/AIDS is not curable, the objective of ART is to slow down viral replication by interrupting the key processes in the HIV life cycle.

2-4. Antiretroviral therapy (ART) and drug resistance

2-4-1. Nucleoside RT inhibitors (NRTI)

Nucleoside RT inhibitors (NRTI) are nucleoside analogs (after the activation by adding phosphate groups with the intracellular kinase) but lack the 3’-hydroxyl group at the 2’-deoxyribosyl moiety. As a result, they compete with natural nucleotides and prevent the formation of a 3’-5’-phosphodiester bond in growing DNA chains during reverse transcription, and so the viral replication fails [69]. For pyridine analogs, zidovudine (AZT) and stavudine (d4T) are analogous to thymine, while lamivudine (3TC) and emtricitabine (FTC) are analogous to cytosine. For purine analogs, didanosine (ddI) and tenofovir (TDF) are analogous to adenosine, while abacavir (ABC) is analogous to guanosine.

Mutations at the genomic region encoding RT are associated with different levels of resistance to different NRTIs. One mechanism, called the mechanism of discrimination, is the exclusion of NRTI from the natural dNTPs. For example, M184VI causes high-level resistance to 3TC and FTC as it changes the geometry of the highly conserved YMDD motif [70] that is part of the active site of polymerase in RT [71]. Another Q151M complex mutation changes the conformation of the dNTP-binding pocket but overcomes the fitness cost of slow viral replication, conferring HIV high-level resistance to AZT, DDI, ABC, and d4T [72]. Another mechanism, called the mechanism of excision, is the pyrophosphorolysis of thymine analogs in the 3'-terminal end of blocked vDNA. Eventually, the thymine analog is released, and so the DNA synthesis can be resumed [73]. Thymidine analog mutations (TAMs) at several positions (M41L, D67N, K70R, L210W, T215F/Y, and K219Q/E) are associated with AZT and d4T resistance.

Interestingly, mutations such as M184VI and K65R alone confers high level resistance to some NRTIs, but it can increase the susceptibility to other NRTIs. Mutation M184VI confers high-level resistance to FTC and 3TC but increases the sensitivity to AZT and d4T. Mutation K65R confers high-level resistance to d4T, ddI, TDF, FTC, and 3TC, but greatly increases the sensitivity to AZT. However, the increased sensitivity can be compensated with other mutations such as Q151M. Compensation mutations at various sites, such as the Q151M complex (for example, F77L, F116Y, and Q151M), increase resistance to almost all commonly used NRTI. This means that the final resistance level of different NRTIs should be based on consideration of combination of the mutations (if any) instead of individual mutations.

2-4-2. Non-nucleoside reverse-transcriptase inhibitors (NNRTI)

Different from NRTIs, NNRTIs bind to the RT and open a hydrophobic pocket close to the active site of polymerase. This causes the confirmation change in the YMDD loop at the active site, the primer gripping site, and the base of the p66 thumb that is unfavorable to proper primer binding, and so inhibits viral replication [74, 75]. Doravirine (DOR), etravirine (ETR), efavirenz (EFV), nevirapine (NVP), and rilpivirine (RPV) are examples of NNRTIs. Since amino acid positions 100 – 108, 179-190, 227, 229, 234, 318 in p66 and 138 in p51 are responsible for the structure of this hydrophobic cavity, mutations at these positions and some nearby amino acid positions are associated with NNRTI resistance. Mutations at these positions usually lead to high-level resistance to different NNRTIs.

2-4-3. Protease inhibitors (PI)

HIV protease is used for cleaving *gag* and *gag-pol* to generate the envelope glycoproteins and enzymes for new generation virions. The protease inhibitors saquinavir (SQV), nelfinavir (NFV), and darunavir (DRV) bind to several amino acids at the active site, including the catalytic Asp25, with different hydrogen bonds and van der Waals force networks, and so they inhibit the protein chain cleavage [76, 77]. It is also proposed that DRV can additionally block the dimerization of PR by binding it to the PR monomer. Multiple mutations V32I, L33F, I54M, and I84V, on the other hand, confer DRV resistance [77]. Even the I84V alone confers low to high levels of resistance to various ARV drugs such as indinavir (IDV), atazanavir (ATV), tipranavir (TPV), and DRV, as amino acid changes result in a lower van der Waals force between the PI and PR [78]. The mutation

G48T and L89M cause a wider opening that is unfavorable to the van der Waals force binding to PR, resulting in high-level resistance to SQV, amprenavir (APV), and NFV [79]. In short, mutations at positions within or close to the active site are usually associated with PI resistance.

2-4-4. Integrase Strand Transfer Inhibitor (INSTI)

HIV integrase is responsible for genome integration of vDNA into host genome. The integrase cleaves two or three nucleotides from the 3' end of the vDNA (called 3'-processing), followed by the cleavage of the double strand host chromosomal DNA and the immediate strand transfer to the 5' end of one strand of host chromosomal DNA. The DNA polymerase closes the DNA gap (a few nucleotides) caused by HIV INT on the other strand. The 5' overhangs of vDNA are removed, and finally, the vDNA is integrated into the host genomic DNA with the aid of DNA ligase [80]. INSTI binds to the active site containing 3'-processed vDNA in INT and displaces the cleaved chromosomal DNA. Finally, the strand transfer by INT is retarded [81]. Elvitegravir (EVG) and raltegravir (RAL) are the first generation INSTIs. Bictegravir (BIC), cabotegravir (CAB), dolutegravir (DTG) are second generation INSTI as BIC and DTG can enhance the binding by reaching N117 and G118 with their oxazinane/oxazepane rings [81, 82].

The well known mutation Y143 causes the loss of π–π stacking interactions to RAL and the side chain of Y143, and so confers high-level RAL resistance [82, 83]. Also, mutations Q148R, N155H, E92Q, and T66I are associated with high-level EVG resistance found in failed EVG-containing regimen patients [84]. The second-generation drug DTG is not affected by the RAL resistance associated mutations Y143R and N155H or the EVG resistance associated mutations T66I and E92Q [82]. That is why it is such a powerful drug for combating AIDS/HIV. One mutation, G118R, has been linked to intermediate-level DTG and BIC resistance, but it is associated with reduced viral replication capacity, which may explain why this DTG resistance associated mutation is so rare [85]. Another mutation, R263K, may cause intermediate-level resistance to DTG but also lower the efficiency of strand transfer and 3’-processing [86].

2-4-5. Summary

AIDS/HIV is not curable because of the integration of vDNA into the host genome. Also, the high error rate of HIV RT leads to a high mutation rate and favors drug resistance development. According to recent studies in Guanxi (China), Kazakhstan, Uganda, and the USA, NNRTI resistance in newly diagnosed or treatment-initiated patients is higher than NRTI and PI [87-90], meaning that the resistance is being transmitted within the community. The INSTI resistance is still the lowest (<1%), but the prevalence may be significant decades later [91]. As a result, close monitoring of the spread of resistant strains and understanding more about the mechanism of resistance are required for public health management and drug discovery. Like TB, with sequencing technologies, drug resistance profiles can assist clinicians in optimizing the ART for individual patients.

2-5. Adoption of sequencing technologies to detect the drug resistance detection in clinical laboratories – the motivation of this study

2-5-1. Direct AMR detection in MTB in sputum samples

Benefiting from the power of collecting comprehensive genetic information with nucleic acid amplification and sequencing technologies (including next generation sequencing (NGS) by Illumina and long-read sequencing by Pacific Bioscience (PacBio) and Oxford Nanopore Technologies(ONT), direct sequencing of *Mycobacterium tuberculosis* (MTB) genomic DNA in clinical samples for rapid antibiotic resistance detection has become a research interest. Previous studies from other research teams proved the feasibility of using whole genome sequencing (WGS) for antibiotic resistance detection in sputum samples from pulmonary tuberculosis (TB) [92, 93]. Compared with the gold standard phenotypic drug susceptibility test (pDST), the time to clinical report by sequencing could be greatly reduced from weeks to a few days, given the high concordance of drug resistance results for both pDST and WGS of clinical isolates.

WGS can generate sequencing data for clinical antibiotic resistance reports as well as epidemiological analysis. In principle, however, the whole genome data size should be much larger than targeted sequencing data. Besides, the highly varied amount of MTB DNA in the mucosal content and the genomic DNA contamination from humans and other nasal or oral microbiota in sputum samples may result in a higher demand for sequencing depth in order to reach a sufficient depth of coverage for downstream analysis [94]. This strategy may raise the sequencing cost and the data storage requirement, which may not favor routine testing in clinical laboratories.

Target sequencing is one method to enrich the sequencing data in the region of interest. In our previous study, the working principle of target sequencing was proven to provide high agreement between ONT, NGS, and pDST in clinical isolates [95]. However, the minimum time for a clinical report is still around two weeks because of the long incubation period of *MTB*. A target sequencing workflow for direct sequencing of sputum samples is required to further shorten the time to clinical report. Studies from our team and other research teams proved the working principle of direct target sequencing for AMR detection in respiratory samples [96-99], but precautions should be highlighted before the adoption of the sequencing technologies in clinical laboratories.

Direct sequencing of sputum samples can be difficult due to contamination with genomic DNA from humans and nasal/oral flora. A few antibiotic resistance-associated housekeeping genes, such as RNA polymerase subunit B (*rpoB*), 16S ribosomal RNA (*rrs*), 23S ribosomal RNA (*rrl*), and DNA gyrase B (*gyrB*) exist in MTB and common nasal and oral microbiota, such as *Staphylococcus sp.* and *Corynebacterium sp.* [100]. Limited studies cover the possible interference of these genes from nasal/oral microbiota to the variant calling performance and, therefore, the specificity in antibiotic resistance prediction. More work is required for confirming the possible interference from this background nasal/oral flora.

2-5-2. Direct ARV resistance detection in HIV in plasma samples

The presence of DRMs in minor quasispecies can be the consequence of the mutation in the original strain, or the infection of a new distant resistant strain in the case of superinfection. Minor quasispecies carrying DRMs can lead to subsequent treatment failure in some cases [101-104]. During the treatment period, the resistant minor quasispecies outgrows the other quasispecies and subsequently lead to treatment failure. Also, other studies suggested quasispecues carrying dual-class DRMs was associated with the higher risk of treatment failure [105]. With the long-read sequencing and hierarchical clustering, not only the drug resistance results by the major DRMs can be detected, the minor quasispecies and the corresponding drug resistance profiles can also be revealed that provides more genetic information for the follow-up.

Sanger sequencing is one common technology used for ARV resistance detection [106, 107]. Due to the limited read length, separated DNA amplification in multiple regions is required for the full coverage of the genome of interest. Such laboriousness does not favor the handling of large sample batches. Also, it does not support mixed variants at low variant frequencies (for example, <0.2) [108] and misses drug resistance associated minor variants. Next generation sequencing (NGS) is one popular sequencing technology because of its high sequencing capacity and high base accuracy (typically 99.99%). It is used for variant calling, including low variant allele frequency variants. The PCR-tiling strategy can overcome the limited readlength in NGS (such as SARSCoV2), but it lacks the association between variants and quasispecies in a sample.

Oxford Nanopore Technologies (ONT) nanopore sequencing provides long-read sequencing readlength and a flexible sequencing batch size. The super-accurate (SUP) basecalling model with Flowcell R9.4.1 pushes the raw read accuracy to 98% above. With its low adoption cost and small size, the MinION sequencer is now being used in clinical diagnosis of infectious diseases. Not only does the long-read sequencing allow variant calling, it also allows the clustering of sequencing reads that clearly reveals the ARV resistance profile, including the association between the variants and the quasispecies in a sample.

2-6. The objectives

For AMR resistance detection in MTB:

1) A direct target sequencing workflow is designed that shortens the time to clinical reports to a few working days.
2) To explore the potential source of contamination and any precautions within the sequencing workflow, including the possible interference of nasal/oral flora, and to provide the solution.

For ARV resistance detection in HIV:

1) Using a hierarchical clustering strategy, a direct target sequencing workflow is designed to provide a comprehensive AMR resistance profile by associating ARV resistance with the quasispecies in a sample.

Finally, with these two examples of direct sequencing workflow development, the strategies, including the use of the advantages of targeted sequencing, the precautions, and the solutions, can be a reference for sequencing workflow development for drug resistance detection in other diseases.

3. Materials and methods

3-1. Targeted sequencing workflow for direct AMR detection in MTB in sputum samples

Sample collection and preparation

A validation cohort consisting of 13 clinical isolates (5 from Asella Hospital, Ethiopia, and 8 from Queen Mary Hospital, Hong Kong) and 37 TB-negative samples (from Queen Mary Hospital, Hong Kong) was used for the evaluation of non-TB interference filtering and the performance of drug resistance detection. Among the clinical isolates, H37RV was selected for preparing a 10-fold dilution spike-in series in triplicate (n=24), while three clinical isolates (026, 069, and 150B) were used for preparing a 4-fold dilution series in triplicate (n=31). All the dilution series were used for determining the limit of detection (LOD) and the evaluation of non-TB interference filtering. A testing cohort of 130 TB-positive clinical specimens collected from Queen Mary Hospital was also recruited to assess the performance of non-TB interference and drug resistance detection (see below).

All the samples were liquified and decontaminated. The genomic DNA was extracted with the AMPLICOR® respiratory specimen preparation kit, followed by 1x bead-based purification (AMPure XP beads) with an elution volume of 50 uL nuclease free water. The colony forming unit (CFU) of MTB in spike-in samples and TB-positive clinical specimens was estimated with *IS6110* quantitative polymerase chain reaction (PCR) [109]. The *IS6110* PCR Ct value was adjusted by adding 15 cycles in the first round of nested PCR.

Primer design

A total of 19 pairs of primers were designed to amplify 19 validated drug resistant genes in MTBC (Table 1a). For the primer set of Nanopore sequencing, universal sequences 5’TTTCTGTTGGTGCTGATATTGC-[forward primer]3’ and 5’ACTTGCCTGTCGCTCTATCTTC-[reverse primer]3’ were added to the primers in order to incorporate the barcode sequences to the amplicons in the subsequent barcoding PCR. The 19 pairs of primers were divided into Pool 1 (10 pairs) and Pool 2 (9 pairs) (Table 1b) based on the GC content and the coverage region of the amplicons.

Table 1a) The primer set Version 2 was designed for amplifying the 19 regions of interest in the MTB genome. The primer set, Version 1, was listed in Supplementary Table 2.

Sr #	Gene/ locus	Rv numbering/Locus tag	Forward Primer	Forward Primer length	Forward Primer region	Reverse Primer	Reverse Primer length	Reverse Primer region	Targeted region	Amplicon length
1	rpsA	Rv1630	**ACCGAGTTTGTCCAGCGTGT**	**20**	**1833451 - 1833470**	**GAACGCGATCAGCTGCAAGA**	**20**	**1834988 - 1835007**	**1833451 - 1835007**	**1557**
2	katG structural gene	Rv1908c	**TCCACTTCACCTTGCCACTG**	**20**	**2154044 - 2154063**	**GCAGATGGGGCTGATCTACG**	**20**	**2155424 - 2155443**	**2154044 - 2155443**	**1400**
3	rpoB	Rv0667	CAAAGTTCCTCGAATAACTCCGTACCCG	28	759861 - 759888	GGCGGTCAGGTACACGATCT	20	761368 - 761387	759861 - 761387	1527
4	MabA-inhA promoter, inhA structural	Rv1483(mabA), Rv1484 (inhA)	CGTACACGTCTTTATGTAGCGCGACATA	28	1673200 - 1673227	CATCGAAGCATACGAATACGCCGAGATG	28	1674564 - 1674591	1673200 - 1674591	1392
5	rrs	Rvnr01	CGTGGCCGTTTGTTTTGTCAGGATATTT	28	1471777 – 1471804	GGCTCTCGCCCACTACAGAC	20	1473438 - 1473457	1471777 - 1473457	1681
6	ubiA	Rv3806c	CATACAGCAGATACGTCCACGCTGTC	26	4268632 - 4268657	GACACGCCAAGTCAACTGAGCTTTCC	26	4270060 - 4270085	4268632 - 4270085	1254
7	rrl	Rvnr02	CCCGTAACTTCGGGAGAAGG	20	1475576 - 1475595	TTTGTATGTTCGGCGGTGTCCTACTTTT	28	1476992 - 1477019	1475576 - 1477019	1444
8	gyrB	Rv0005	CTGACCATCAACCTGACCGACGAGAG	26	5849-5874	TCGTGTCTGTCATCTATTCCTCGTTTGC	28	7287 - 7314	5849 - 7314	1466
9	embB	Rv3795	CGACCACGCTGAAACTGCT	19	4247154 - 4247172	AAAGATTGTGCTGACTGTGATCCCGTC	27	4248503 - 4248529	4247154 - 4248529	1376

10	tlyA	Rv1694	CAATGACCATCGATCCTGACCAGATCC	27	1917754 - 1917780	CCCTTTTCCAGACTGACTTCGTTGAGC	27	1919216 - 1919242	1917754 - 1919242	1489
11	FurA-KatG intergenic	Rv1909c (FurA)	CATTTCGTCGGGGTGTTCGTCCATAC	26	2155129 - 2155154	GGGAGTCATATTGTCTAGTGTGTCCTCT	28	2156584 - 2156611	2155129 - 2156611	1483
12	whiB7	Rv3197A	CGAGAAGAACTACGACCTCCTGTTGC	26	3568004-3568029	CGGATCTGTAACAACGAGCTGAACACTT	28	3569375 - 3569402	3568004 - 3569402	1399
13	pncA	Rv2043c	GTAGCTCATCCTCGCCTAAAGTCATTGT	28	2288057 - 2288084	GTTGTATCAACGGTGGTAATGCACTTCG	28	2289590 - 2289617	2288057 - 2289617	1561
14	gyrA	Rv0006	GCAAACGAGGAATAGATGACAGACACGA	28	7287 - 7314	CTGGGTGGTGAAGAACAGGATCAAATCG	28	8993 - 9020	8993 - 9020	1734
15	rplC	Rv0701	GCTACCGACTGAGAAGAACGTGTATTGC	28	800609 - 800636	GATGACCACCAGCACCTGTTTACGTTCT	28	801926 - 801953	800609 - 801953	1345
16	Rv0678	Rv0678	ATTTCACAAAGCAGTAGGTCAGGGCATC	28	778485 - 778512	GAGAATCCACAACCGCTTCGATCCAGAT	28	779814 - 779841	778485 - 779841	1357
17	eis promoter	Rv2416c	TCCTGTGGATGGGTGATGATGCTGATTC	28	2714515 - 2714542	GGAAAACTTGTTCTGGTCCAACGGG	25	2715726 - 2715750	2714515 - 2715750	1236
18	rpsL	Rv0682	GGTCGCTAGAGTCATTAGTTGGCCCTAA	28	780567 - 780594	AGTTAGCTGTCTATCACTGTCGGTTTGC	28	782430 - 782457	780567 - 782457	1891
19	atpE	Rv1305	GAACCGGTCGCAACTTATTCTTCCAATG	28	1460180 - 1460207	TCGCCACACCAGATAAACGATGACC	25	1461878 - 1461902	1460180 - 1461902	1722

Table 1b) The assignment of primers to Pool 1 and Pool 2 for multiplex PCR.

Pool 1	Volume (uL)	1000			
Primer	Genomic location ('000)	Stock molarity (uM)	Volume required for primer mix (for each primer)	Molarity in primer mix (uM)	H2O volume (uL)
gyrA	7	100	5	0.5	900
katG	2154	100	5	0.5	
pncA	2288	100	5	0.5	
rpsL	781	100	5	0.5	
rrs	1472	100	5	0.5	
rpsA	1833	100	5	0.5	
rpoB	760	100	5	0.5	
MabA	1673	100	5	0.5	
eis	2714	100	5	0.5	
atpE	1461	100	5	0.5	

Pool 2	Volume (uL)	1000			
Primer	Genomic location ('000)	Stock molarity (uM)	Volume required for primer mix (for each primer)	Molarity in primer mix (uM)	H2O volume (uL)
whiB7	3568	100	5	0.5	910
ubiA	4268	100	5	0.5	
rrl	1475	100	5	0.5	
gyrB	5	100	5	0.5	
tlyA	1917	100	5	0.5	
FurA	2155	100	5	0.5	
rplC	800	100	5	0.5	
Rv0678	778	100	5	0.5	
embB	4247	100	5	0.5	

Multiplex PCR

Each sample was amplified with pool 1 and pool 2 primers separately. The reaction mixture was prepared by mixing 5ul of DNA, 12.5ul of Platinum™ Multiplex PCR Master Mix (Thermo Fisher Scientific, Waltham, MA, USA), 0.5ul of GC enhancer for pool 1 primers or 3ul of GC enhancer for pool 2 primers, 3.75ul of pool 1 or pool 2 primers (0.5uM) and each reaction was filled up to 25ul with nuclease-free water. The PCR conditions were 95°C for 4min, 40 cycles of 95°C for 30sec, 63°C for 1.5min and 72°C for 2min, final extension at 72°C for 10min and hold at 4°C. The PCR products were purified with 0.4x AMPure XP beads and eluted in 20ul nuclease-free water. The DNA were quantified by Qubit 2.0 Fluorometer (Thermo Fisher Scientific, Waltham, MA, USA) using Qubit™ 1X dsDNA HS Assay Kit (Thermo Fisher Scientific, Waltham, MA, USA).

Nanopore sequencing

The library was prepared by following the official protocols of the PCR barcoding (96) amplicons kit (SQK-LSK109 and EXP-PBC096) from Oxford Nanopore Technologies (ONT). In brief, 0.5 nM amplicon from multiplex PCR of each sample was taken to barcoding PCR, followed by 0.65x AMPure XP bead-based purification. A batch of 12 or 24 samples was pooled for end-repairing and adapter ligation. A final 50-fmole library was sequenced for 8 or 16 hours (respectively for 12 and 24 samples), using the flow cell FLO-MIN106 R9.4.1 and the sequencer MinION. High accuracy (HAC) basecalling mode was selected on the MinKNOW software.

Next generation sequencing (NGS)

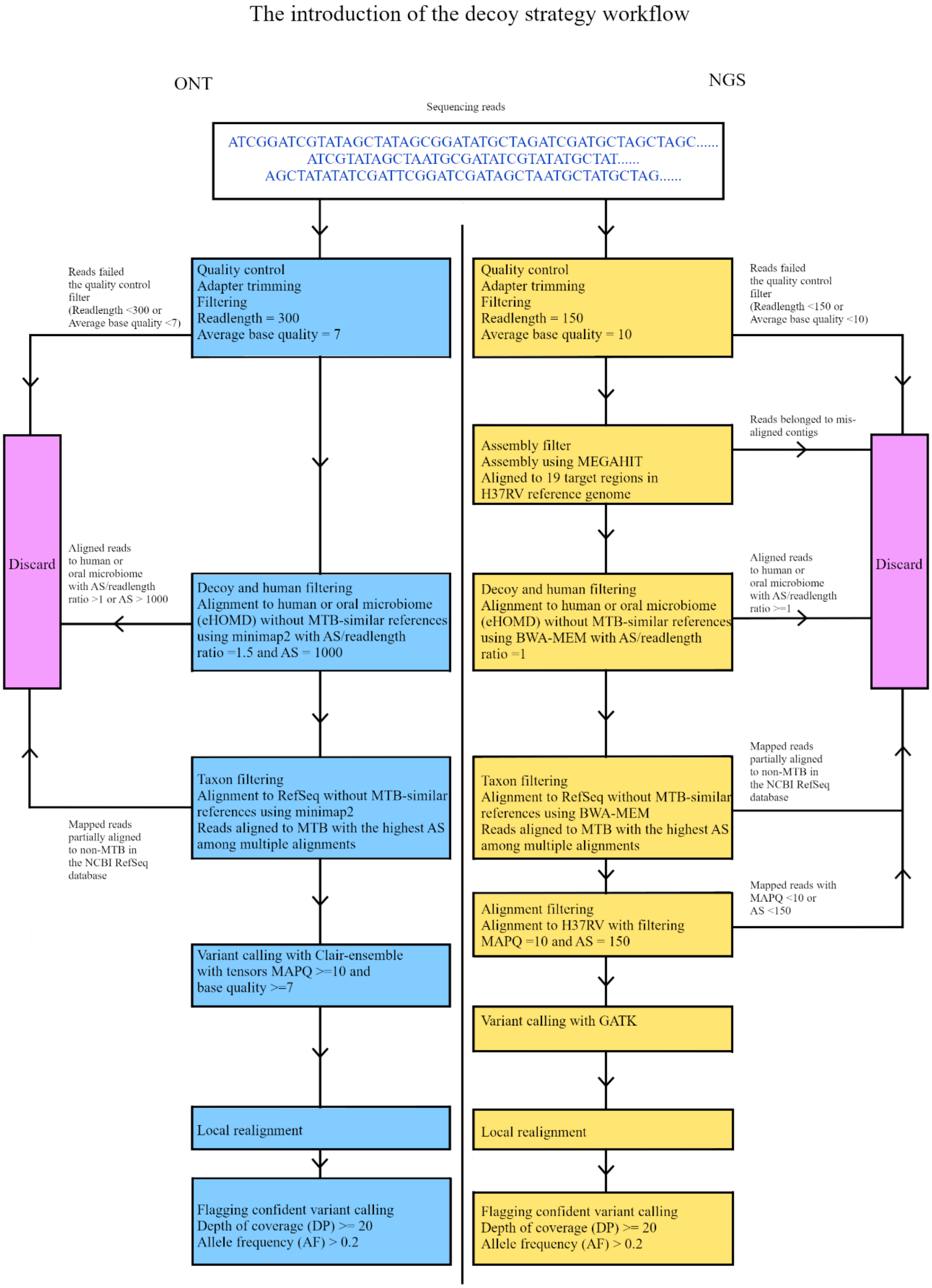
Library preparation was performed using the NEBNext® Ultra™ II FS DNA Library Prep Kit for Illumina and NEBNext® Multiplex Oligos for Illumina® (96 Unique Dual Index Primer Pairs) (NEB, Ipswich, Massachusetts, USA) according to the manufacturer’s protocol. Briefly, the DNA input for each sample (from multiplex PCR) was 100 ng. The DNA was enzymatically fragmented for 15 minutes, followed by the ligation with the adapter (1.5 uM). Then, a double bead-based size selection targeting the library size at 320-470 bp was performed. The libraries were enzymatically indexed and amplified in a 5-cycle PCR. The quantity of the libraries was estimated by real-time PCR using LightCycler® 480 Instrument II (Roche, Basel, Switzerland) and QIAseq™ Library Quant Assay Kit (Qiagen, Hilden, Germany). The quality of libraries was measured by 2100 Bioanalyzer system (Agilent, Santa Clara, CA, USA) using a High Sensitivity DNA Kit (Agilent, Santa Clara, CA, USA). The libraries were normalized and pooled into 4 nM and were subsequently diluted and denatured. Finally, a 10 pM pooled library spiked with 10% of 10pM PhiX Control Kit v3 (Illumina, San Diego, California, USA) was sequenced on the MiSeq system (Illumina, San Diego, California, USA) in the setting of 2 X 250 cycles, using MiSeq Reagent Nano Kit v2 (500-cycles) (Illumina, San Diego, California, USA).

Bioinformatics

Figure 1 summarizes the details of bioinformatic analysis, known as the decoy strategy. Briefly, NGS data preprocessing was done using the amplicon filtering module in MegaPath (source code available at https://github.com/HKU-BAL/MegaPath) [110]. Using the module, adaptors in sequencing reads were trimmed using BBMap (v36.28)[111]. Sequencing reads shorter than 150 bp or with an average base quality score lower than 10 were filtered. As a step to remove short-read alignment ambiguities, filtered sequencing reads were aligned to the regions of interest in the *Mycobacterium tuberculosis* H37Rv (NC_000962.3) reference genome using BWA-MEM (version 0.7.17-r1188)[112]. The aligned reads were assembled into contigs using MEGAHIT (version 1.2.9) [113] with the option --presets meta-sensitive. In the decoy filter, sequencing reads that belonged to contigs that were aligned to another region or non-amplification regions were removed. In addition, reads exactly matched to the decoy database containing the expanded Human Oral Microbiome Database (eHOMD) (version 9.03) or the human (GRCh38) reference genome were removed. The remaining reads were mapped to H37Rv again. In the taxon filter, reads partially aligned to the non-*M. tuberculosis* sequences in the NCBI RefSeq database were removed. Also, alignments with low mapping quality (MAPQ) < 10 or alignment score (AS) < 150 were filtered out. Genome Analysis Toolkit (GATK, version 4.1.9.0) [114] was used for variant calling, with two options enabled, including --read-filter PrimaryLineReadFilter and --max-reads-per-alignment-start 0. Local realignment was done at the called positions to achieve a precise variant allele frequency (VAF) estimation. Variants with coverage < 20-fold and AF < 0.2 were flagged as less confident.

ONT data preprocessing was done using the amplicon filtering module in MegaPath-Nano (source code available at https://github.com/HKU-BAL/MegaPath-Nano). MegaPath-Nano is an extension of MegaPath for ONT long-read [115]. Using the module, like NGS, the adapter was trimmed and demultiplexed using Porechop (version 0.2.4, source code available at https://github.com/rrwick/Porechop) with options --barcode_threshold 85, --require_two_barcodes, and --discard_middle. Trimmed sequencing reads shorter than 300 bp or an average base quality score below 7 were filtered out. The remaining reads were aligned to the regions of interest in the *Mycobacterium tuberculosis* H37Rv (NC_000962.3) reference genome with minimap2 (v2.17-r941)[116] for long-read alignment. In the decoy filter, reads aligned to eHOMD decoy database or human reference with AS ≥ 1000 or AS-to-read-length-ratio ≥ 1.5 were removed. Like in NGS, the remaining reads that were partially aligned to non-M. tuberculosis sequences from the NCBI RefSeq database were also removed in the taxon filter. Clair-ensemble (version 1.1, source code available at https://github.com/HKU-BAL/ECNano)[117] was used for variant calling. In each position to be called, Clair-ensemble was set to ignore the low-quality (i.e., base quality score < 7) bases. Like the post-processing of GATK output, local realignment was likewise performed at the called positions. Variants with coverage < 20-fold and VAF < 0.2 were flagged as less confident.

Fig. 1) The introduction of the decoy strategy.


The introduction of the decoy strategy workflow
ONT
NGS
Sequencing reads
ATCGGATCGTATAGCTATAGCGGATATGCTAGATCGATGCTAGCTAGC......
ATCGTATAGCTAATGCGATATCGTATATGCTAT......
AGCTATATATCGATTCGGATCGATAGCTAATGCTATGCTAG......
Quality control
Adapter trimming
Filtering
Readlength = 300
Average base quality = 7
Reads failed the quality control filter (Readlength <300 or Average base quality <7)
Discard
Decoy and human filtering
Alignment to human or oral microbiome (cHOMD) without MTB-similar references using minimap2 with AS/readlength ratio =1.5 and AS = 1000
Aligned reads to human or oral microbiome with AS/readlength ratio >1 or AS > 1000
Taxon filtering
Alignment to RefSeq without MTB-similar references using minimap2
Reads aligned to MTB with the highest AS among multiple alignments
Mapped reads partially aligned to non-MTB in the NCBI RefSeq database
Variant calling with Clair-ensemble with tensors MAPQ >=10 and base quality >=7
Local realignment
Flagging confident variant calling
Depth of coverage (DP) >= 20
Allele frequency (AF) > 0.2
Quality control
Adapter trimming
Filtering
Readlength = 150
Average base quality = 10
Reads failed the quality control filter (Readlength <150 or Average base quality <10)
Assembly filter
Assembly using MEGAHIT
Aligned to 19 target regions in H37RV reference genome
Reads belonged to mis-aligned contigs
Decoy and human filtering
Alignment to human or oral microbiome (cHOMD) without MTB-similar references using BWA-MEM with AS/readlength ratio =1
Aligned reads to human or oral microbiome with AS/readlength ratio >=1
Discard
Taxon filtering
Alignment to RefSeq without MTB-similar references using BWA-MEM
Reads aligned to MTB with the highest AS among multiple alignments
Mapped reads partially aligned to non-MTB in the NCBI RefSeq database
Alignment filtering
Alignment to H37RV with filtering MAPQ =10 and AS = 150
Mapped reads with MAPQ <10 or AS <150
Variant calling with GATK
Local realignment
Flagging confident variant calling
Depth of coverage (DP) >= 20
Allele frequency (AF) > 0.2

Evaluation of index assignment in nanopore sequencing and NGS

For ONT, the performance of index assignment with default demultiplexing setting in MinKNOW (index score = 60) and the stringent setting (index score = 85 or barcode threshold = 0.85, two barcodes at both ends, and discard middle barcode) in Porechop (v0.2.4) and MinKNOW were compared. Non-template control (NTC) samples (N=13) with expected zero mapped reads were included in sequencing batches. The percentage of mapped reads and the average depth of coverage per covered position for NTC obtained from different demultiplexing methods were calculated.

For NGS, libraries of twelve clinical specimens were prepared with NEBNext® Multiplex Oligos for Illumina® (Dual Index Primer Set 1) (includes 8 i5 index primers and 12 i7 index primers) and NEBNext® Multiplex Oligos for Illumina® (96 Unique Dual Index Primer Pairs) (UDI). Libraries with the same adapter set were sequenced in one batch, meaning that libraries with multiplex oligos and UDI were separately sequenced in two runs. The empty positions with expected zero sequencing reads and mapped reads were used for the evaluation. The percentage of the mapped reads and the average depth of coverage per covered position for empty positions were calculated.

Limit of detection (LOD)

Spike-in samples with known adjusted *IS6110* PCR Ct values (n = 55) were defined as a validation set, which was taken to logistic regression analysis with Prism software. The log average DP in an individual target region larger than or equal to 1.7 (equivalent to the average DP 50) was set as 1, or it was set as 0. A linear equation was constructed for each target region with the log odds against the adjusted *IS6110* Ct values. Then the adjusted *IS6110* Ct values having the probability of reaching the log average DP 50 in that target region were calculated. The LOD of target region was defined as the adjusted *IS6110* PCR Ct value required to have the 0.9 probability of reaching the log average DP 1.7 in that region. The LOD of the first line drug was the lowest LOD (in adjusted *IS6110* Ct value) among the target regions *katG*, *furA-KatG* intergenic regions, *mabA-inhA* promoter, *rpoB*, *ubiA*, *embB*, *rpsA*, and *pncA*. The LOD of the overall panel was the lowest LOD among all 19 target regions.

The performance of drug resistance detection

The variants found in ONT and NGS were annotated as resistant or susceptible by matching the variants to the database containing validated AMR associated genes and mutations, including the deduced ones (Supplementary Table 3). The variant allele frequency (VAF) of 0.2 and DP of 20 were the cutoff values for classifying a genotypic result as valid, a genotypic result with its values below either of these cutoff values was classified as uncertain. Also, with the high base accuracy of NGS, the agreement between ONT and NGS was used to evaluate the variant calling performance in ONT.

3-2. Targeted sequencing workflow for direct ARV resistance detection in HIV in plasma samples

Sample collection

Seventy-seven plasma samples were collected from hospital in Hong Kong between Year 2002 – Year 2014 (Supplementary Table 4). Seventy samples were male, and seven samples were female. The age ranged from 18 to 68 (average was 37). On the other hand, HIV plasmid pHIV-1_pr-V82A was acquired from the European Virus Archive - Global (EVAG). It was used as a control for gradient studies as well as *in-silico* simulation datasets (please see below). The reference genome sequence was available on EVAG.

RNA extraction, long region amplification and library preparation

Around 1.5 mL of frozen plasma was thawed on ice for 2 hours before being resuspended and centrifuged at 4°C for 1.5 hours at 14,000 rpm. The supernatant was discarded without disturbing the pellet.

Viral RNA was extracted from plasma samples by following the official protocols of QIAamp Viral RNA Kits. Host genomic DNA in the total nucleic acid sample (8 uL) was removed with ezDNase™. The reaction sample was taken to reverse transcription with LunaScript RT SuperMix (5X). Target genomic region (NC_001802.1: 1413 – 7363, amplicon length: 5951 base pairs) was amplified with the and cycling condition shown in Table 2. The amplicon was purified with 0.5 X AMPure XP beads (for nanopore sequencing) and quantified with Qubit® dsDNA HS Assay Kits.

Table 2) The PCR reagents and the cycling condition used for amplifying the viral genome region of interest for downstream sequencing.

Primer	Sequence
Forward primer	GCAAGRGTTTTGGCBGARGCAATGAG
Reverse primer	GCCCATAGTGCTTCCTGCTGCTCCCAAGAACC

Component	Volume (uL)	Final concentration
2X Platinum SuperFi II PCR Master Mix	25	1X
10 uM Forward primer	2.5	0.5 uM
10 uM Reverse primer	2.5	0.5 uM
Template cDNA (directly from reverse transcription reaction)	20	
Total volume	50	

Cycling condition

Cycle step	Temperature	Time	
Heat activation	98°C	2 minutes	
Denaturation	98°C	10 seconds	35 cycles
Annealing	60°C	10 seconds	
Extension	72°C	4 minutes 30 seconds	
Final extension	72°C	5 minutes	
Hold	4°C	∞	

Quantitative reverse transcription polymerase chain reaction (qRT-PCR)

The viral load (copies per uL) of each sample was quantified by following the official protocol of the genesig Standard Real-time PCR Detection Kit for HIV-1 and oasig lyophilised OneStep qRT-PCR MasterMix Kit. Briefly, 5 uL of extracted viral RNA (vRNA) and an HIV positive control template were enzymatically converted to complementary DNA (cDNA). Next, the cDNA was amplified in 50 cycles, and the vRNA of each sample was quantified with the linear standard curve.

The ONT workflow - nanopore sequencing

The library was prepared using the following Native Barcoding Amplicons (with SQK-LSK109, EXP-NBD104, and EXP-NBD114). Briefly, amplicons (for each sample) were enzymatically converted to blunt-end DNA and then ligated to barcode adapters, followed by 1 X AMPure XP bead purification. Barcoded amplicons were pooled and normalized to reach the recommended input 100 - 200 femtomole. The pooled barcoded amplicons were further ligated to Adapter Mix II. After the post-ligation cleanup with 0.5 X AMPure XP beads, the pooled library was again quantified with Qubit® dsDNA HS Assay Kits. A library of approximately fifty femtomoles was subjected to 48-hour sequencing on GridION using the SUP basecalling mode, a minimum quality score for read filtering of 10, and a modified demultiplexing setting (trim_barcodes="on", require_barcodes_both_ends="on", detect_mid_strand_barcodes="on", min_score=85).

Next generation sequencing (NGS)

One nanogram of purified 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 of the library was checked with High Sensitivity DNA Assay on Bioanalyzer, and the library was quantified with a QIAseq Library Quant Assay Kit. The final pooled library was sequenced with MiSeq Reagent Kit Nano V2 on the Illumina MiSeq System (250 X 2 cycles).

Bioinformatics

Sequencing reads from ONT were used to iterate quasispecies clustering and variant calling with a software package, called ClusterV, which was designated for providing abundance, variant calling, and clinical reports down to each quasispecies, based on the input of the alignment files and Browser Extensible Data (BED) files. Briefly, the sequencing reads were mapped to the HIV reference genome NC_001802.1 with Minimap2 (2.24-r1122)[116]. Sequencing reads with large INDELs or that failed to cover the targeted region in binary alignment map (BAM) files were excluded. The variants were called with Clair-ensemble (a variant calling tool for targeting sequencing with high depth of coverage) and were used as markers for iterating hierarchical clustering processes to find the quasispecies in a sample. After the iterative clustering processes, consensus sequences for each quasispecies were generated based on their variants. The final consensus sequences for all quasispecies in a sample were submitted to HIV Drug Resistance Database (HIVDB) [118] via SierraPy (a package to interact with HIVDB Sierra GraphQL Webservice) to generate the ARV resistance report.

Validation of clustering performance in the ONT workflow

To evaluate the clustering performance of ClusterV in the ONT method, one HIV plasmid and two clinical samples (Sample ID: KB2061 and KB2979) were employed. These samples contained only one quasispecies with a distinct amino acid mutation pattern (Supplementary Table 5) and a median VAF greater than 0.9. Briefly, an *in-silico* simulation dataset was prepared by mixing the HIV plasmid and two 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 in triplicate was prepared by mixing the amplicons of HIV plasmid amplicons and Sample ID: KB2061 in the following gradient ratios: 95:5, 90:10, 85:15, and 80:20. The R square was used to assess the linear relationship between ClusterV's predicted abundance of quasispecies and the true abundance in the corresponding ratios.

Evaluation of the diagnostic performance of the targeted ONT sequencing workflow

To evaluate the variant calling performance, the amino acid mutations reported in the ONT workflow were compared with those reported in Sanger sequencing. In some samples, genomic variants called from NGS were used to validate mutations found in ONT but inconsistent with Sanger sequencing or could not be validated without available Sanger sequencing results. Genomic variants reported in NGS with VAF > 0.03 were considered valid based on the recommendation in several studies [119-122]. The overall VAF of the amino acid mutations was calculated by the summation of abundance multiplied by the called VAF in all quasispecies of a sample. Briefly, the amino acid mutations (including AVR resistance associated and other mutations) found in ONT were compared with those found in Sanger (Fig. 2). A mutation called ONT was considered true if it was concordant with Sanger. If it was discordant with Sanger or the corresponding Sanger result was not available, it would be validated with the NGS. In this case, the ONT mutation was considered true if it was concordant with NGS. A mutation was considered false if it was discordant with both Sanger and NGS. If both Sanger and NGS were not available, the ONT mutation was inconclusive or uncertain. On the other hand, amino acid mutations exclusively found in Sanger, but not in ONT, were validated with NGS. Those mutations that were concordant with NGS were considered true, whereas those that were discordant ones were considered false. A receiver operating characteristic (ROC) curve was plotted to determine the VAF against the percentage of true mutations found in ONT. An overall VAF threshold was then determined when the threshold led to the optimum true mutation rate and false mutation rate. Finally, the F1 score (2 X (precision X recall) / (precision + recall) was used to assess variant calling and diagnostic performance with and without the overall VAF threshold. The statistical analysis was performed with GraphPad Prism (v9.4.1) and Microsoft Excel.

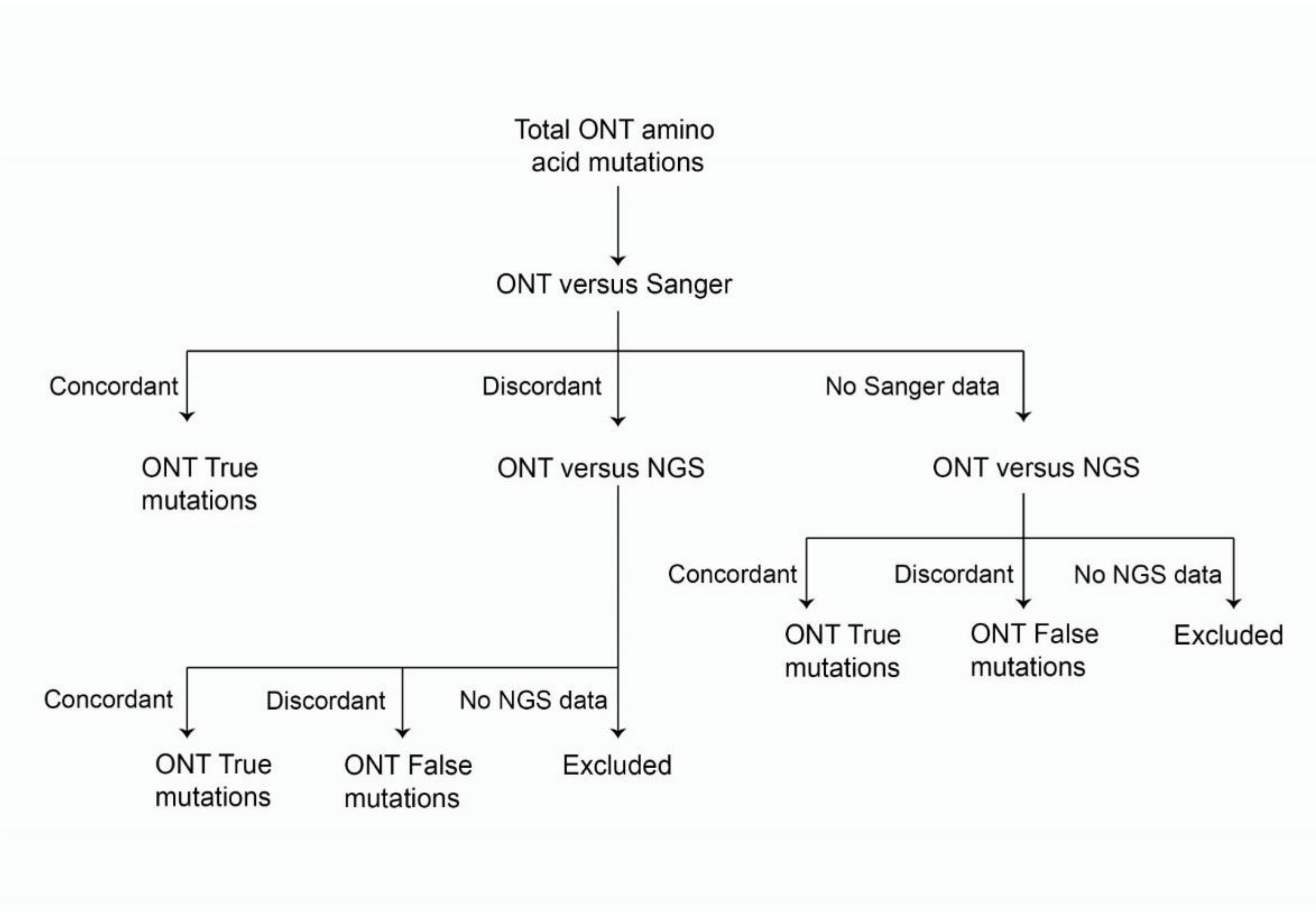


Fig. 2) A flowchart demonstrated how amino acid mutations found in ONT workflow were classified as true or false mutations.

Limit of detection (LOD)

The limit of detection was defined as the viral load required for having a 0.9 probability to reach average depth of coverage (DP). Two DP cutoffs, 50X (minimum) and 1500X (recommended), were used for LOD calculation. A sample with DP that passed the cutoff was set as 1 or 0. A logistic regression was constructed with the odds against the viral load. The AUC was used for evaluating the model, and the equation of this model was used for calculating the LOD.

4. Results

4-1. Targeted sequencing workflow for direct AMR detection in MTB in sputum samples

4-1-1. Assay optimization

To confirm the amplicon length of each targeted genomic region in Primer Set Version 2, amplicon (from amplification with clinical isolate 1033B) were taken to gel electrophesis (1.5% agarose in 1X Tris/Borate/EDTA buffer). The band size of each genomic regions met the expected corresponding amplicon length (Fig. 3).

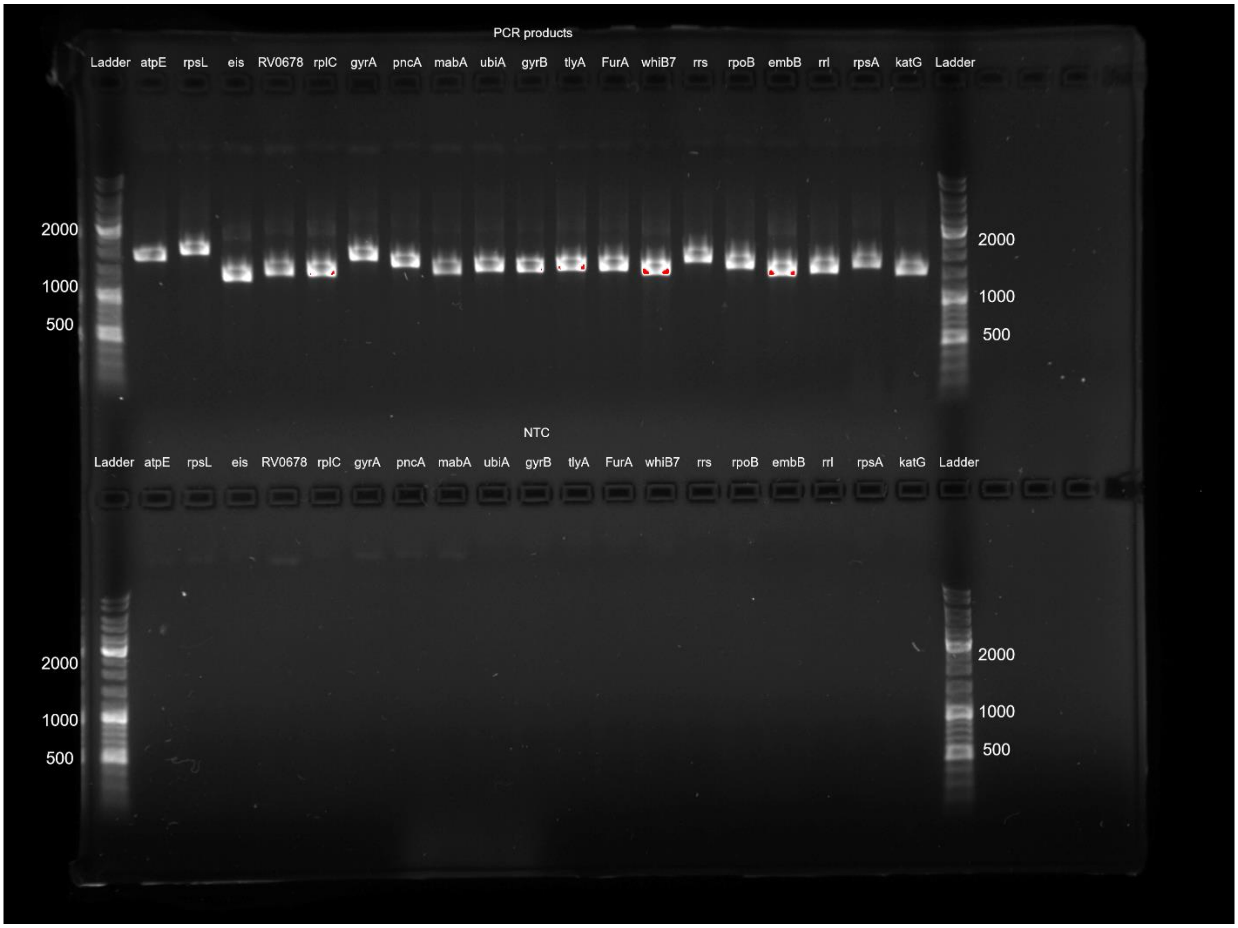


Fig. 3) Gel electrophoresis of PCR products of 19 targeted genomic regions. All the band size were within 1000 bp and 2000 bp, and their band positions met the expected corresponding amplicon length.

To avoid the overlapping region between targeted genomic regions and to minimize enzymatic bias to shorter amplicon length, primer sets targeting six genomic regions with longer amplicon length (*rpsA, katG, rrs, pncA, gryA, rpsL*, and *atpE*) was grouped in Pool 1, and the rest of thirteen genomic regions were grouped in Pool 2. This preliminary version was called V2_first_version.

To evaluate the DP distribution of each targeted genomic region, the DP of individual targeted genomic region was compared with the average DP of 19 targeted genomic regions. Briefly, the percentage of DP to the average DP was calculated by dividing the DP of the individual targeted region by the average DP, then the absolute DP distance was obtained by subtracting the percentage by one, only the different was kept in absolute value to avoid any negative numbers.

Absolute DP distance of individual targeted genomic region = | (DP of individual targeted genomic region / average DP of 19 targeted genomic regions) -1|

Six samples of H37RV spike-in (ID1_power0 in duplicate, ID1_power1 in duplicate, and ID1_power2 in duplicate) were included in the evaluation the DP distribution of Primer Set V2_first_version. The average absolute DP difference of these six samples was high in *eis promoter* (0.5758), *embB* (0.8419), *gyrB* (0.46), *katG* (0.411), *pncA* (0.4768), *rplC* (0.5491), *rpsA*

(0.5482), *rpsL* (0.563), *rrs* (0.5155), and *whiB7* (0.5091), while the average absolute DP difference was remained low in other targeted genomic regions (ranging from 0.15 to 0.3) (Supplementary Table 6) . This suggested enzymatic bias might happen within these 19 targeted genomic regions.

Subsequently, a newer version of primer set V2 (that was the final version of Primer Set V2 mentioned in Table 1a) was developed (called V2_final in this session). Three primer sets targeting *rpoB*, *MabA*, and *eis promoter* were moved from Pool 2 to Pool 1.

Six spike-in samples (ID1_power0, ID1_power1, ID1_power2, 069_1, 069_2, and 069_3) were used for the evaluation the DP distribution (Fig. 4). The average absolute DP difference of V2_final was reduced to below 0.4 (except embB with absolute DP difference of 0.4025). Of which, the absolute DP difference of gyrA, rplC, rpsA, rpsL, and whiB7 (P value <0.05). This concluded that the DP distribution of V2_final was more even than V2_first_version.

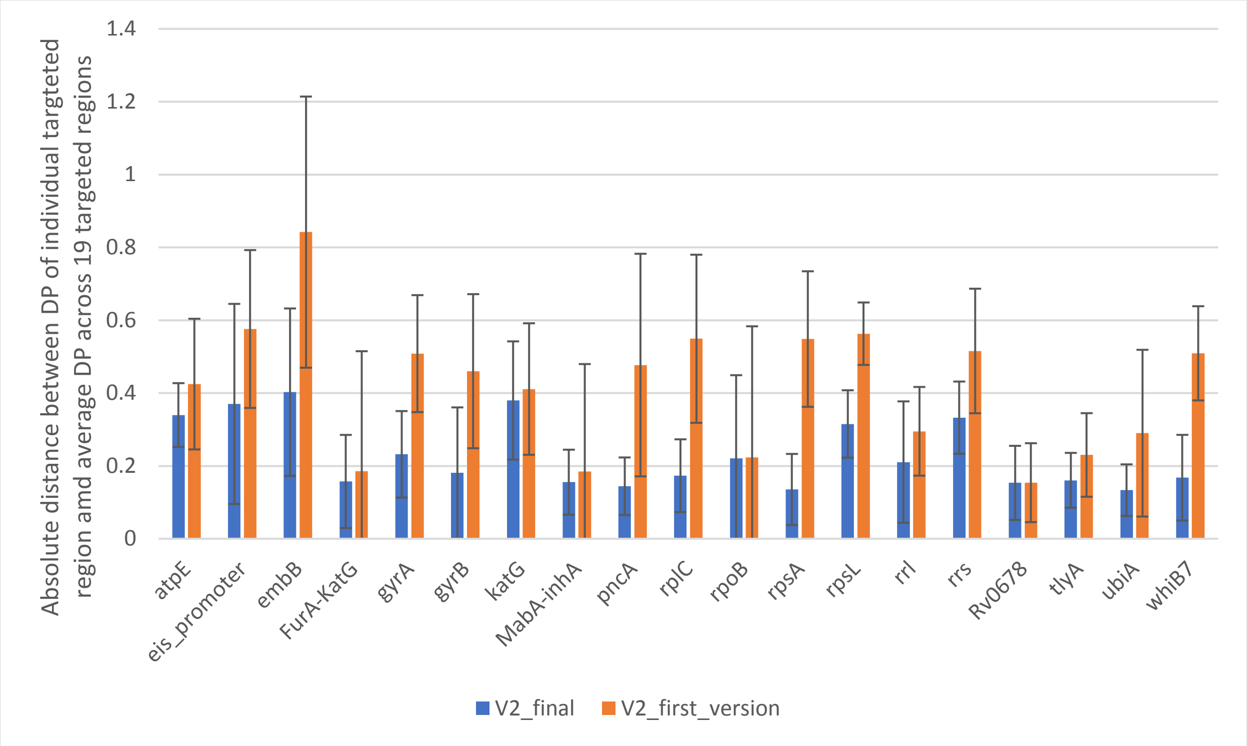


Fig. 4) The average absolute distance between DP of individual targeted regions and average DP across 19 targeted regions in Primer Sets V2_first_version (N = 6) and V2_final (N=6). The error bar represented the standard deviation.

4-1-2. The potential source of index misassignment in ONT and NGS

Index misassignment in sequencing was thought to be the source of cross-contamination between samples. Some reads from one index (sample) “leaked” to other indexes (samples). To study how index misassignment could cause cross contamination between clinical samples and empty samples within a sequencing batch, the number of mapped reads and the DP were examined in the empty samples in NGS and non-template control (NTC).

NTC samples were included in thirteen separate ONT sequencing batches. In average, 1.65% of sequencing reads demultiplexed with the default setting in MiniKNOW (index percent identity 60%) were mapped to the H37RV reference genome. The average DP per covered position was 8.59.

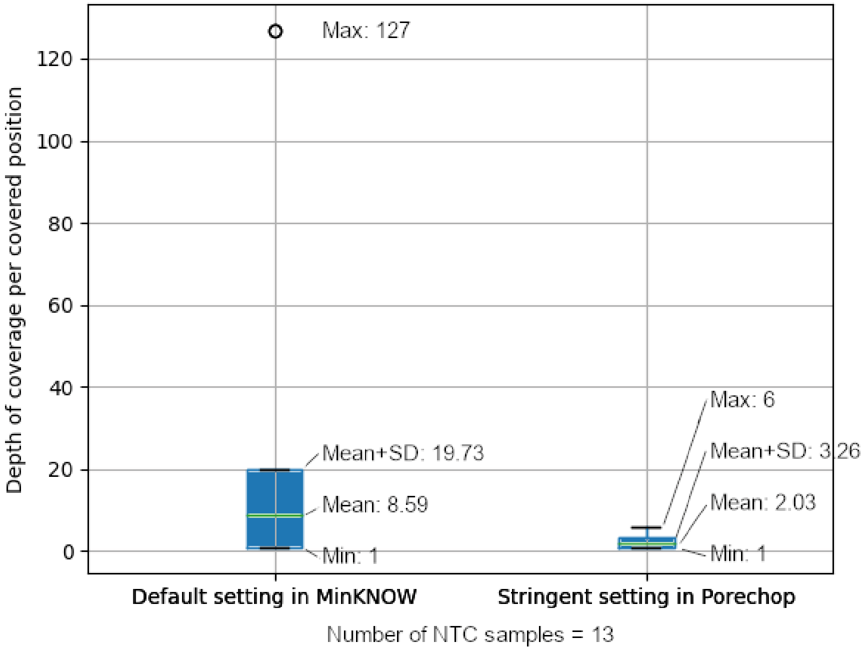
The average percentage of sequencing reads mapped to the reference genome was significantly reduced to 0.06% with a stringent demultiplexing process with Porechop (index percent identity 85%, two-barcodes, discard-middle) (P value < 0.005), and the average DP per covered position was dropped from 8.59 to 2.03 (Fig 5a).

In the sequencing run with libraries carrying the universal dual-index adapters (Supplementary Table 7), the number of sequencing reads in 83 empty positions both ranged from 2 to 105 (one empty position carried zero reads), and all these reads were successfully mapped to the H37RV reference genome. The average and the maximum depth of coverage per covered position were 1.775 and 10 respectively.

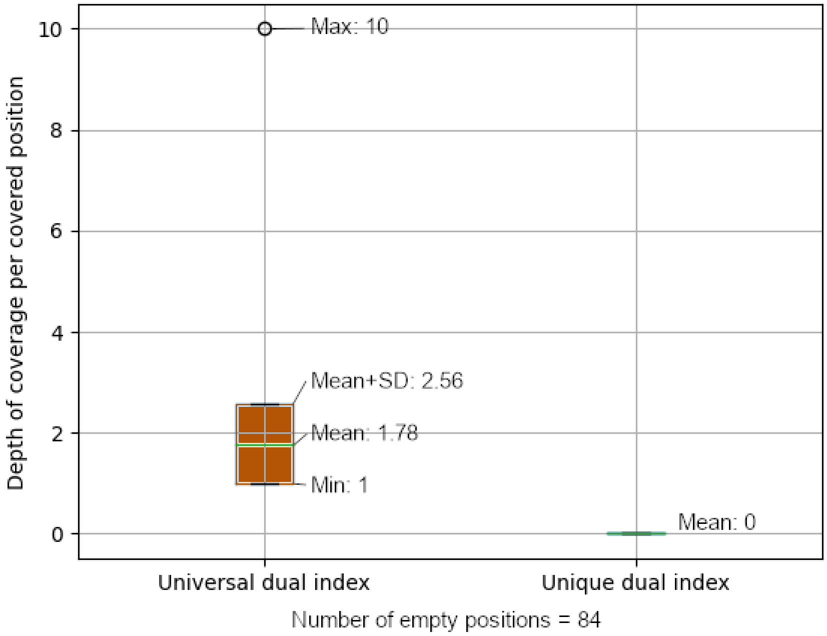
In contrast, the sequencing run with the same twelve libraries carrying UDI generated zero reads in empty positions (Fig 5b).

Fig.5a) Comparison analysis of the mapped reads contamination in NTC samples (N=13) between default parameter setting in MinKNOW demultiplexing (index percent identity 60%) and an additional stringent demultiplexing process (index percent identity 85%, two-barcodes, discard-middle) with Porechop. 5b) Comparison analysis of the mapped reads contamination in empty samples (N=84) between using universal index and UDI index.

5a)



5b)



4-1-3. Background nasal/oral flora interference

An increasing trend of noisy variants associated with the increasing adjusted *IS6110* PCR Ct values (lower MTB gDNA content) of the spike-in samples and clinical specimens was discovered (Supplementary Table 8a and 8c). We then applied a filtering strategy for the ONT and NGS datasets. The quality control filter removed sequencing reads with the readlength less than 700 and/or average base quality 7 in ONT, while it removed the sequencing reads with the readlength less than 150 and average base quality 7 in NGS. Next, applied to both ONT and NGS, the decoy filter removed mapped reads aligned to the eHOMD decoy database or human reference with AS ≥ 1000 or AS-to-read-length-ratio ≥ 1.5 in ONT and AS-to-read-length-ratio ≥ 1 in NGS. Finally, the sequencing reads were filtered out in taxon filter if there was partial alignment to non-MTB in the RefSeq database.

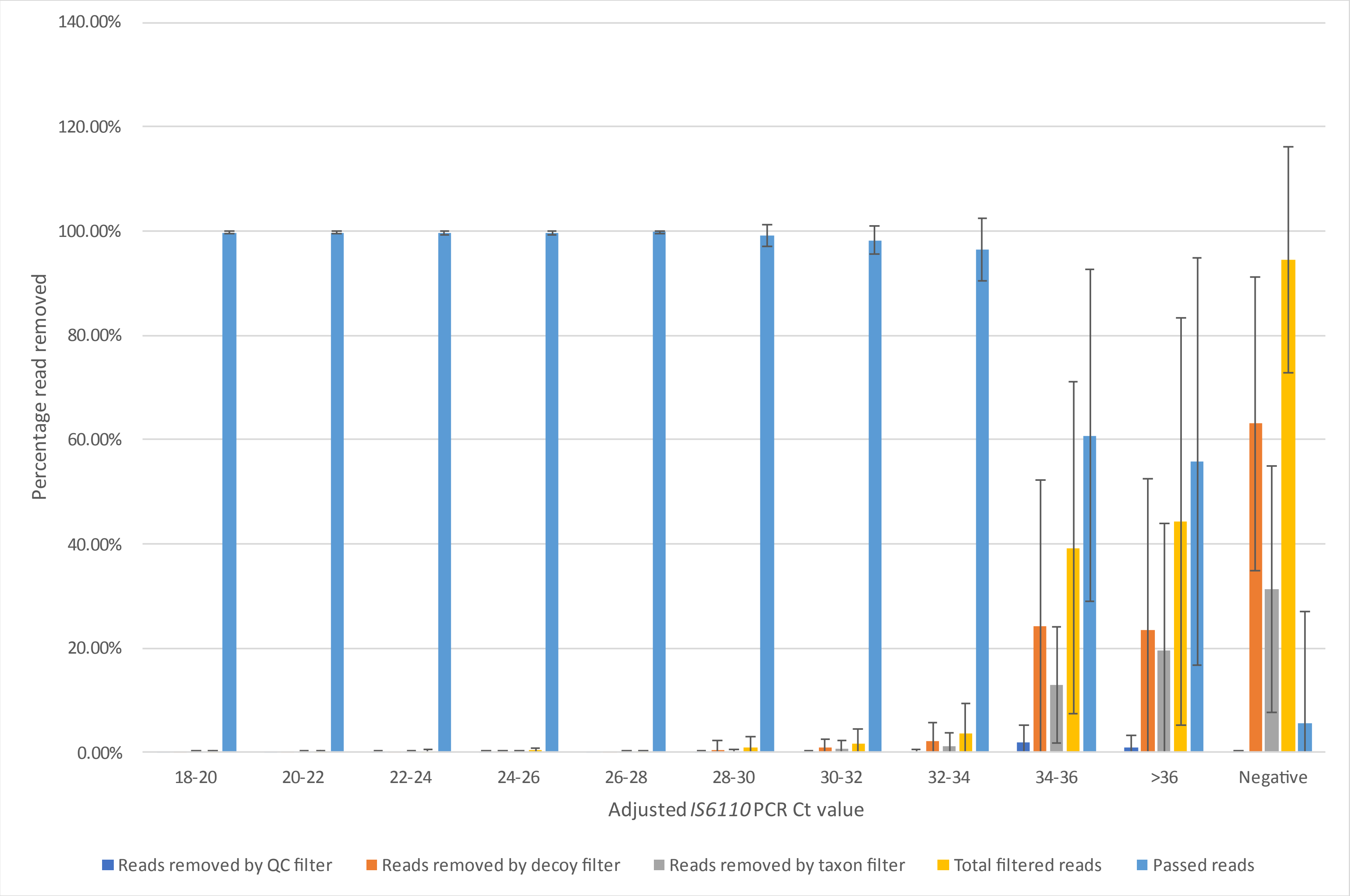
For samples with high levels of MTB gDNA (adjusted IS6110 PCR Ct value 34), the mean percentage of mapped reads removed by the filters was consistently low (10%). The percentage of removed reads dramatically increased afterwards (Fig. 6a). The mean percentage of removed reads reached 94.49% for the negative samples.

Similarly, the mean change in the number of variants was generally less than 2 for samples with the adjusted IS6110 PCR Ct value < 30, but significantly increased to tens and even hundreds afterwards (P value < 0.05) (Fig. 6b).

Similarly, in NGS dataset (Supplementary Table 8b and 8d), the decoy and taxon filters removed low percentage of mapped reads (mean < 10%) for samples with high level of MTB gDNA (i.e. low PCR Ct value). The removed reads gradually increased along with increasing ct value (Fig. 7a).

The mean percentage removal for the negative samples was 93.64%. Like ONT, the mean change in the number of variants was also less than 2 for samples with the adjusted IS6110 PCR Ct value < 30 and it considerably raised to tens and even hundreds afterwards (P value < 0.05) (Fig. 7b).

6a)


140.00%
120.00%
100.00%
80.00%
60.00%
40.00%
20.00%
0.00%
Percentage read removed
18-20
20-22
22-24
24-26
26-28
28-30
30-32
32-34
34-36
>36
Negative
Adjusted *IS6110* PCR Ct value
Reads removed by QC filter
Reads removed by decoy filter
Reads removed by taxon filter
Total filtered reads
Passed reads

6b)

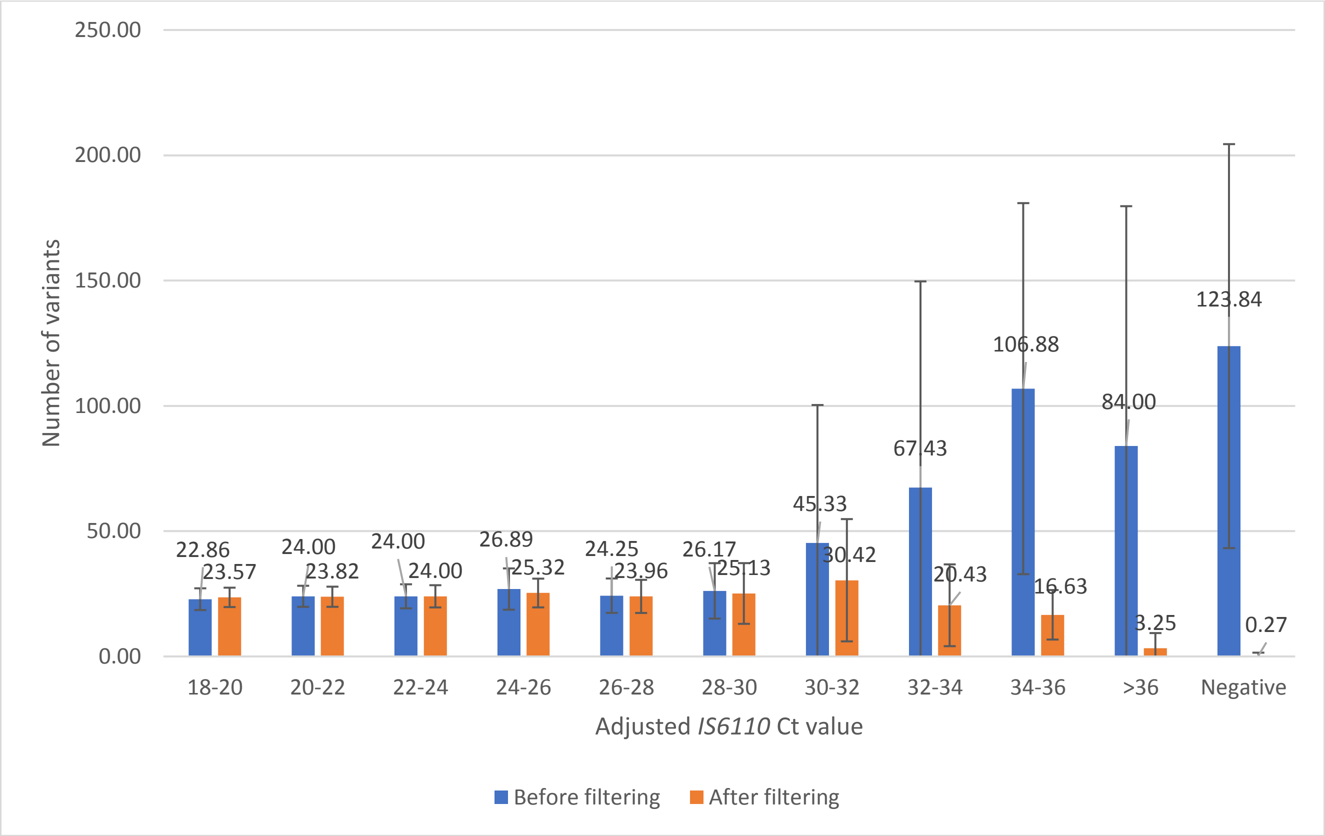
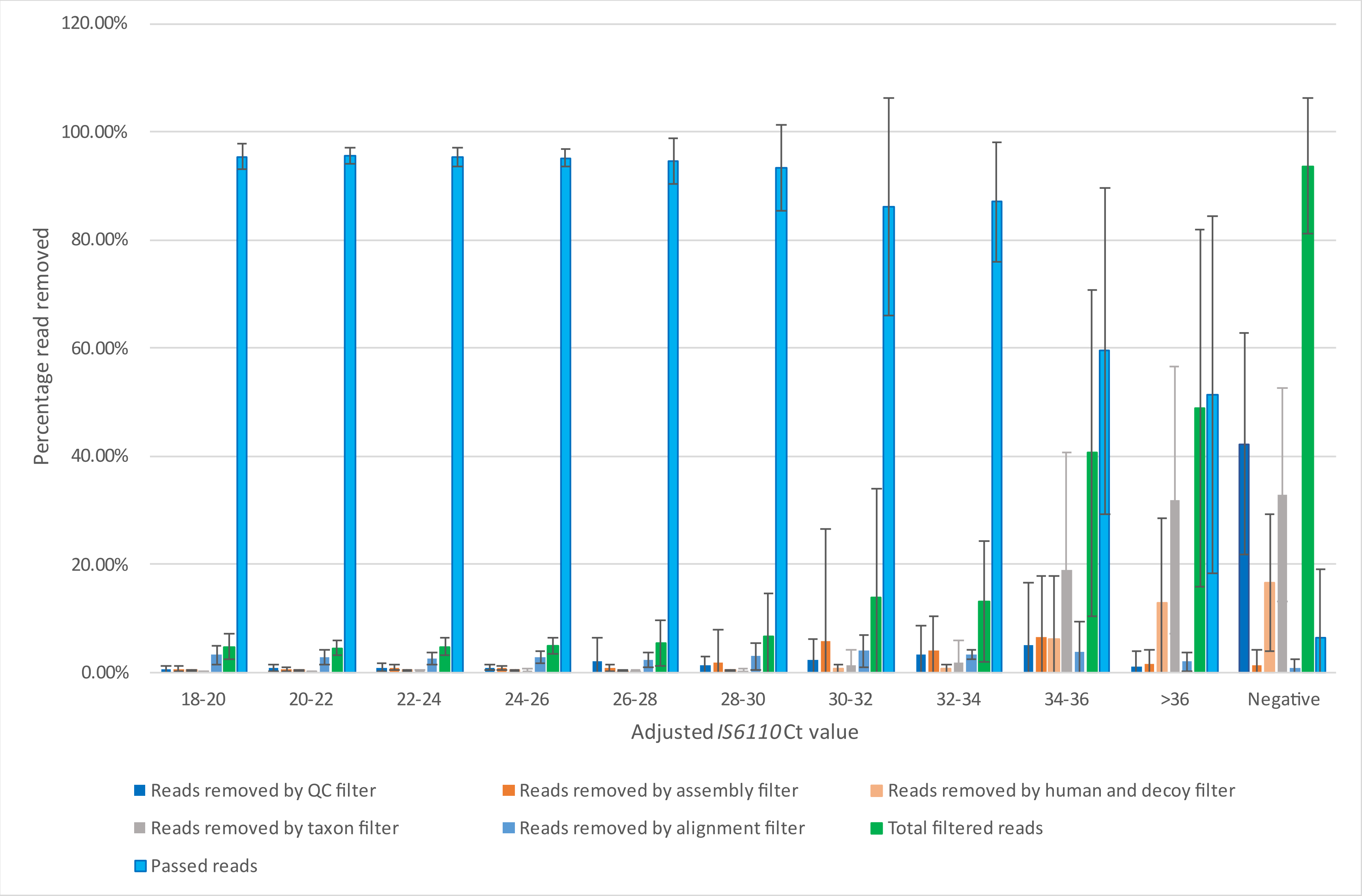


Fig. 6a) The mean percentage of removed mapped reads among 55 clinical isolate/H37RV spike-in samples. 130 clinical specimens, 37 TB-negative samples in the ONT workflow. The reads removed by quality control filter, decoy filter, taxon filter were increased dramatically when the IS6110 ct value >34 (i.e. samples with low MTB gDNA content). The error bars represented the corresponding standard deviations.

b) The mean of number of variants before and after the filtering with the decoy and taxon filters in the ONT workflow against the increasing *IS6110* Ct values (decreasing *MTB* DNA contents in the samples). The percentage of removed mapped reads and the changed number of variants (usually decreasing) was positive associated with the *IS6110* Ct values. The number above each bar represented the mean value and error bars represented the corresponding standard deviations.

7a)



120.00%
100.00%
80.00%
60.00%
40.00%
20.00%
0.00%
Percentage read removed
18-20
20-22
22-24
24-26
26-28
28-30
30-32
32-34
34-36
>36
Negative
Adjusted *IS6110* Ct value
Reads removed by QC filter
Reads removed by assembly filter
Reads removed by human and decoy filter
Reads removed by taxon filter
Reads removed by alignment filter
Total filtered reads
Passed reads

7b)

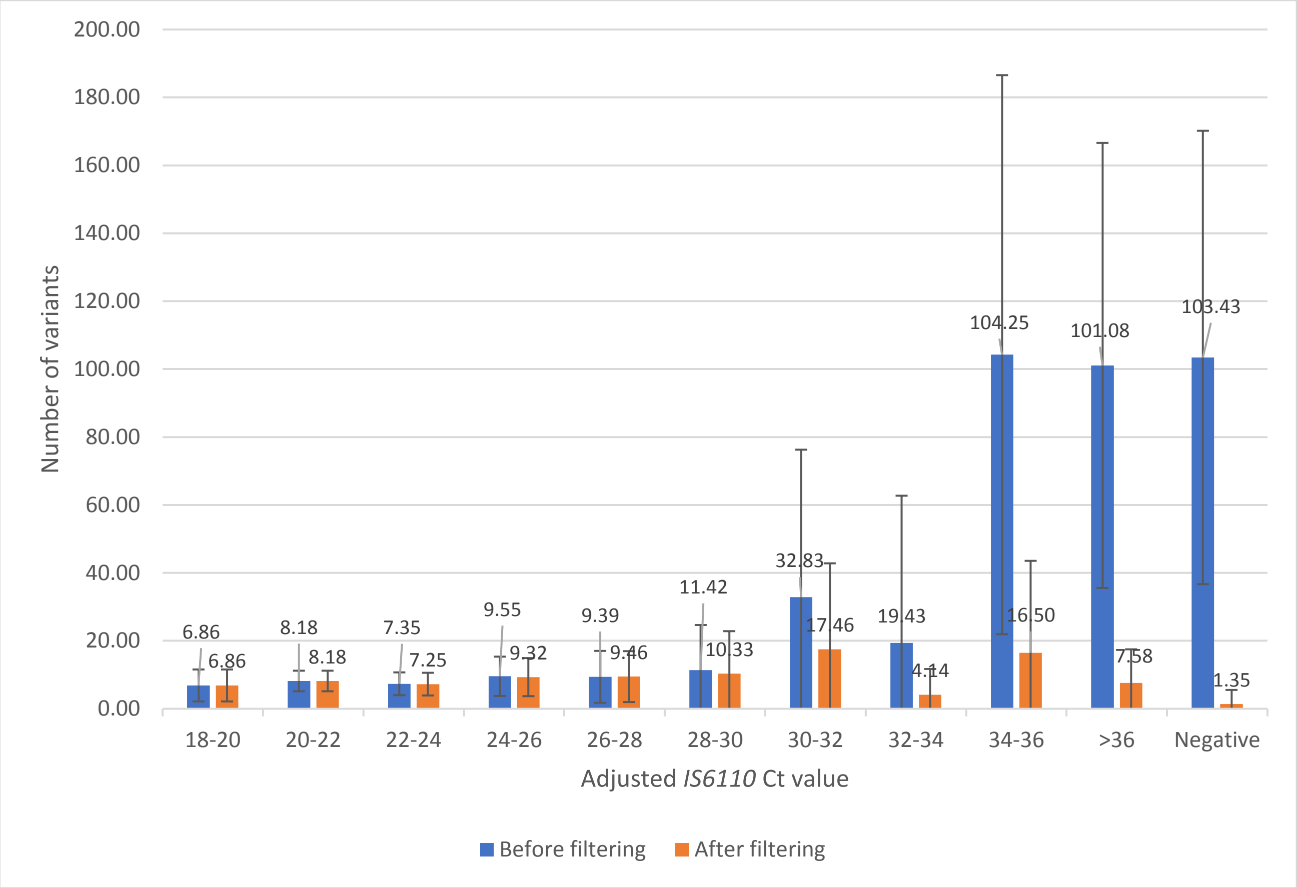


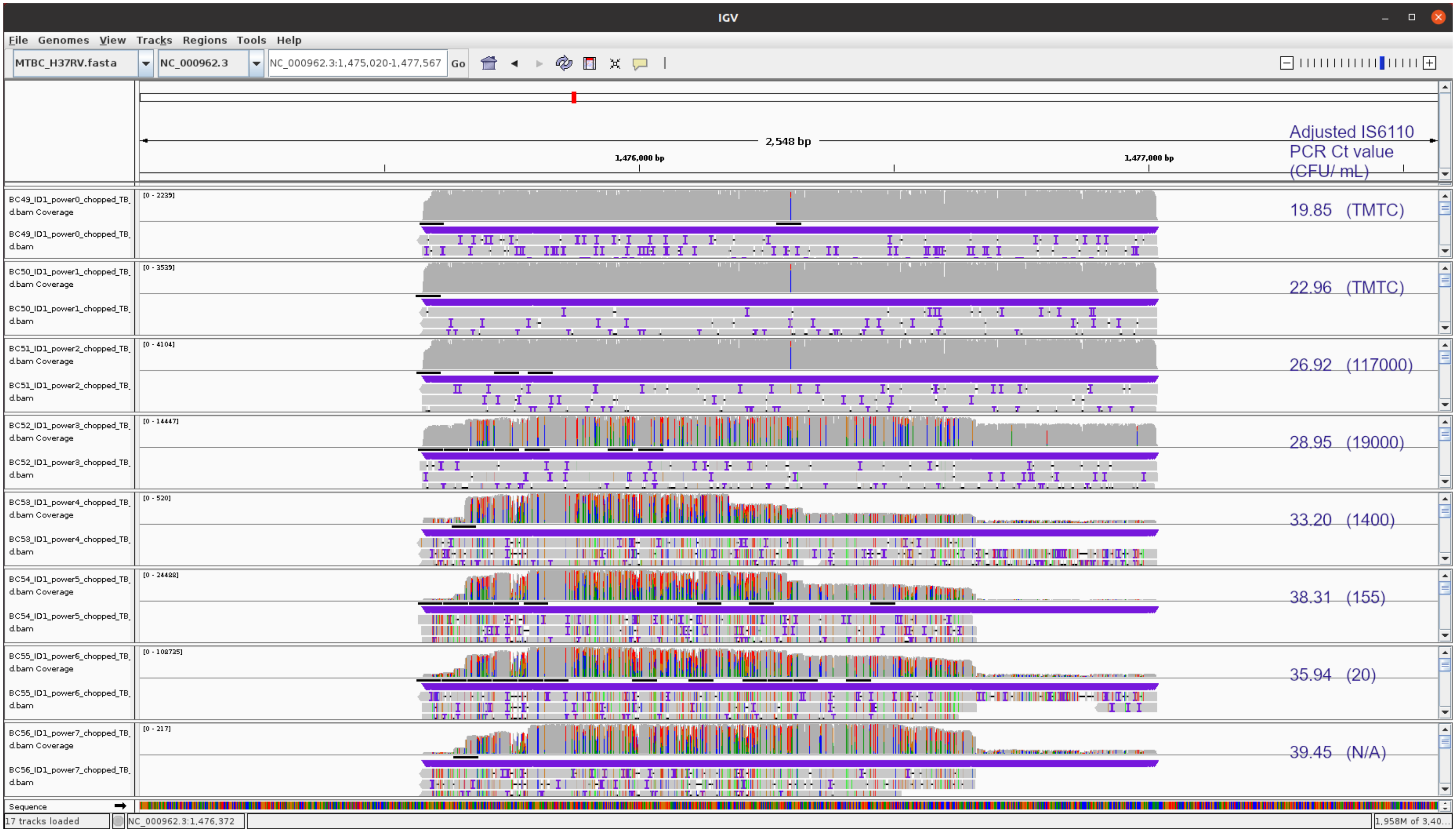
Fig. 7a) The mean percentage of removed mapped reads among 55 clinical isolate/H37RV spike-in samples. 130 clinical specimens, 37 TB-negative samples in the NGS workflow. The reads removed by filters increased gradually along with the increasing *IS6110* PCR ct value. The error bars represented the corresponding standard deviations.

7b) the mean of number of variants before and after the filtering with multiple filters in the NGS workflow against the increasing *IS6110* PCR Ct values (decreasing *MTB* DNA contents in the samples). Like ONT workflow, more false variants was removed by the filters in samples with higher IS6110 PCR ct value. The number above each bar represented the mean value and error bars represented the corresponding standard deviations.

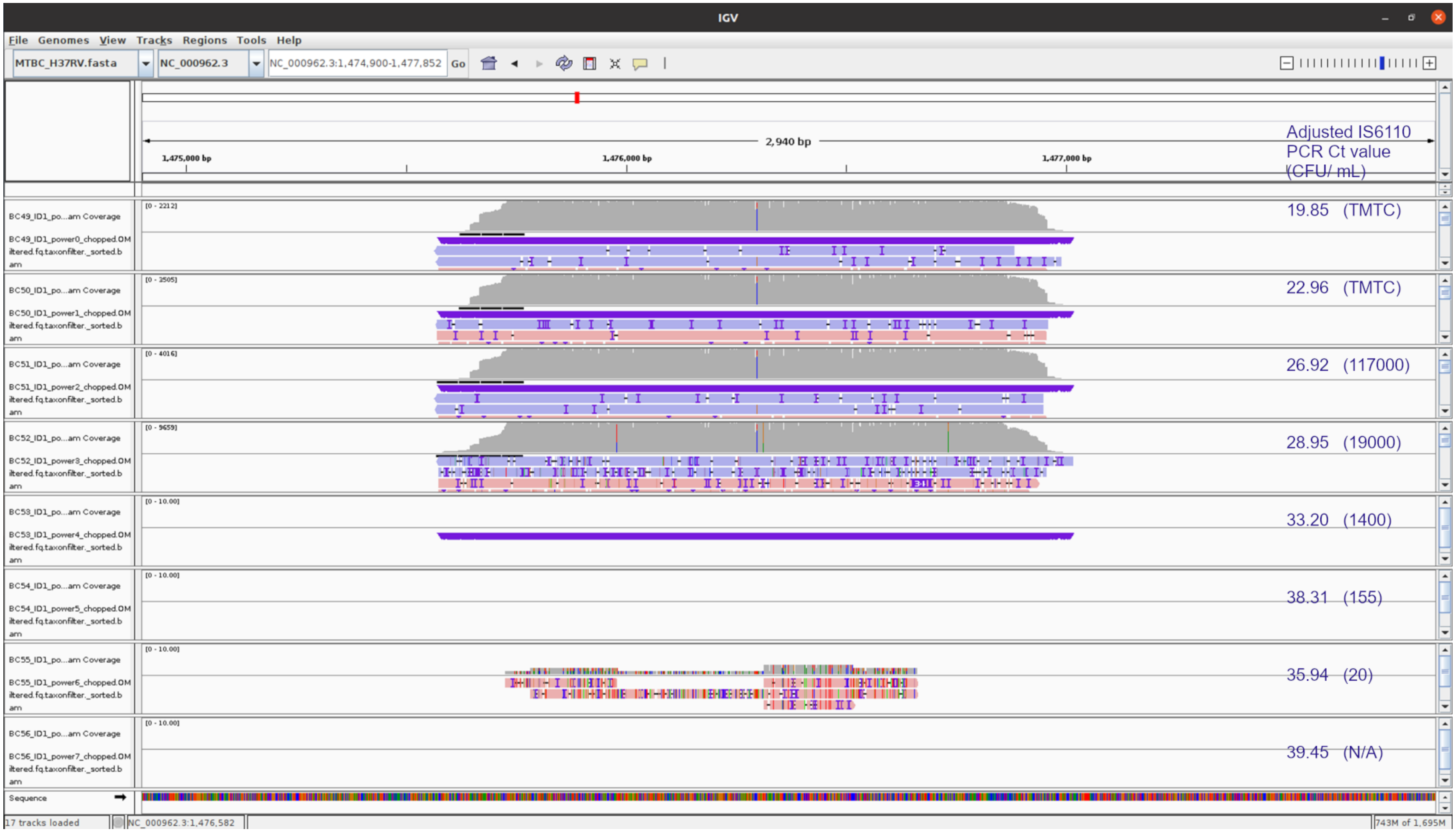
In genomic region *rrl*, without any filtering, the number of noisy variants is positively associated with the adjusted IS6110 PCR Ct value (Fig. 8a). After filtering, sequencing reads harboring noisy variants were removed, with zero coverage from the adjusted IS6110 PCR Ct value of 33.2 (Fig. 8b). Most of the filtered mapped reads in ONT workflow were removed in decoy filter (Fig. 8c, Supplementary Table 9a). Of which, the highest percentage of unique read count was genus *Haemophilus (58.98%)*, followed by *Eikenella (29.54%)*. In the taxon filter, the highest percentage of unique read count was genus *Mycolicibacterium (34.08%)*, followed by *Eikenella* (28.15%) and *Haemophilus* (26.71%). Only a very low percentage 0.14% was uniquely mapped to the *homo sapiens* reference genome in taxon filter (Fig. 8d).

Different from the ONT workflow, most of the mapped reads from NGS were filtered in the taxon filter (Fig. 8e, Supplementary Table 9b). Of which, the top dominating genus was *Mycobacteroides* (51.46%), the second was *Gemella* (8.3%). Compared with the ONT workflow, a higher percentage (26.06%) of filtered reads from NGS mapped to the *homo sapiens* reference genome was observed in taxon filter (Fig. 8f). In the decoy filter, the top genus was *Oribacterium (53.73%),* contributing more than half of the filtered reads.

8a)


IGV
File Genomes View Tracks Regions Tools Help
MTBC_H37RV.fasta
NC_000962.3
NC_000962.3:1,475,020-1,477,567
Go
2,548 bp
1,476,000 bp
1,477,000 bp
Adjusted IS6110 PCR Ct value (CFU/ mL)
BC49_ID1_power0_chopped_TB_d.bam Coverage
BC49_ID1_power0_chopped_TB_d.bam
[0 - 2239]
19.85 (TMTC)
BC50_ID1_power1_chopped_TB_d.bam Coverage
BC50_ID1_power1_chopped_TB_d.bam
[0 - 3539]
22.96 (TMTC)
BC51_ID1_power2_chopped_TB_d.bam Coverage
BC51_ID1_power2_chopped_TB_d.bam
[0 - 4104]
26.92 (117000)
BC52_ID1_power3_chopped_TB_d.bam Coverage
BC52_ID1_power3_chopped_TB_d.bam
[0 - 14447]
28.95 (19000)
BC53_ID1_power4_chopped_TB_d.bam Coverage
BC53_ID1_power4_chopped_TB_d.bam
[0 - 520]
33.20 (1400)
BC54_ID1_power5_chopped_TB_d.bam Coverage
BC54_ID1_power5_chopped_TB_d.bam
[0 - 24488]
38.31 (155)
BC55_ID1_power6_chopped_TB_d.bam Coverage
BC55_ID1_power6_chopped_TB_d.bam
[0 - 108735]
35.94 (20)
BC56_ID1_power7_chopped_TB_d.bam Coverage
BC56_ID1_power7_chopped_TB_d.bam
[0 - 217]
39.45 (N/A)
Sequence
17 tracks loaded
NC_000962.3:1,476,372
1,958M of 3,40...

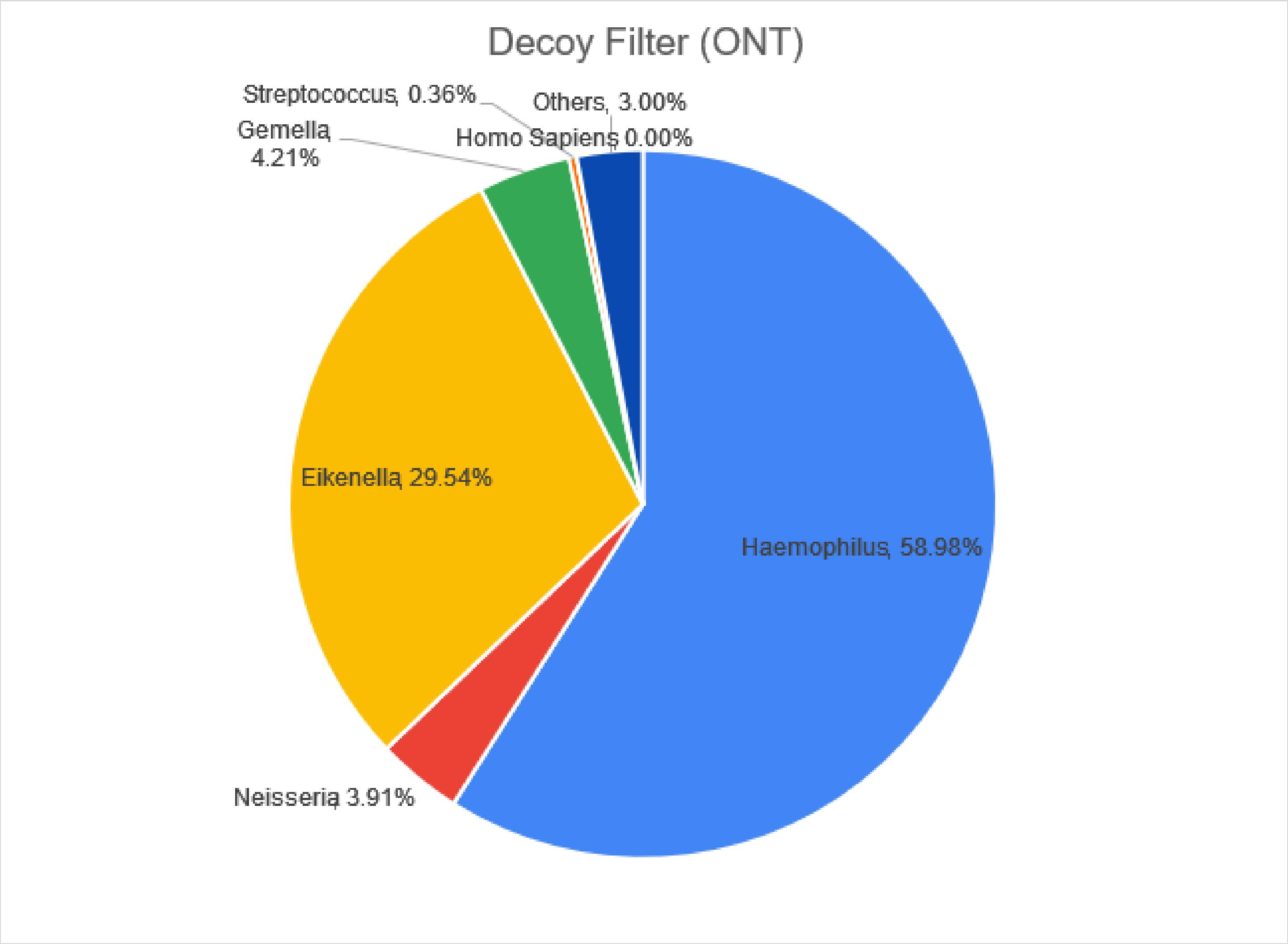
8b)


IGV
File Genomes View Tracks Regions Tools Help
Go
2,940 bp
1,475,000 bp
1,476,000 bp
1,477,000 bp
Adjusted IS6110
PCR Ct value
(CFU/ mL)
BC49_ID1_po...am Coverage
BC49_ID1_power0_chopped.OM
iltered.fq.taxonfilter_sorted.b
am
[0 - 2212]
19.85 (TMTC)
BC50_ID1_po...am Coverage
BC50_ID1_power1_chopped.OM
iltered.fq.taxonfilter_sorted.b
am
[0 - 2505]
22.96 (TMTC)
BC51_ID1_po...am Coverage
BC51_ID1_power2_chopped.OM
iltered.fq.taxonfilter_sorted.b
am
[0 - 4016]
26.92 (117000)
BC52_ID1_po...am Coverage
BC52_ID1_power3_chopped.OM
iltered.fq.taxonfilter_sorted.b
am
[0 - 9659]
28.95 (19000)
BC53_ID1_po...am Coverage
BC53_ID1_power4_chopped.OM
iltered.fq.taxonfilter_sorted.b
am
[0 - 10.00]
33.20 (1400)
BC54_ID1_po...am Coverage
BC54_ID1_power5_chopped.OM
iltered.fq.taxonfilter_sorted.b
am
[0 - 10.00]
38.31 (155)
BC55_ID1_po...am Coverage
BC55_ID1_power6_chopped.OM
iltered.fq.taxonfilter_sorted.b
am
[0 - 10.00]
35.94 (20)
BC56_ID1_po...am Coverage
BC56_ID1_power7_chopped.OM
iltered.fq.taxonfilter_sorted.b
am
[0 - 10.00]
39.45 (N/A)
Sequence

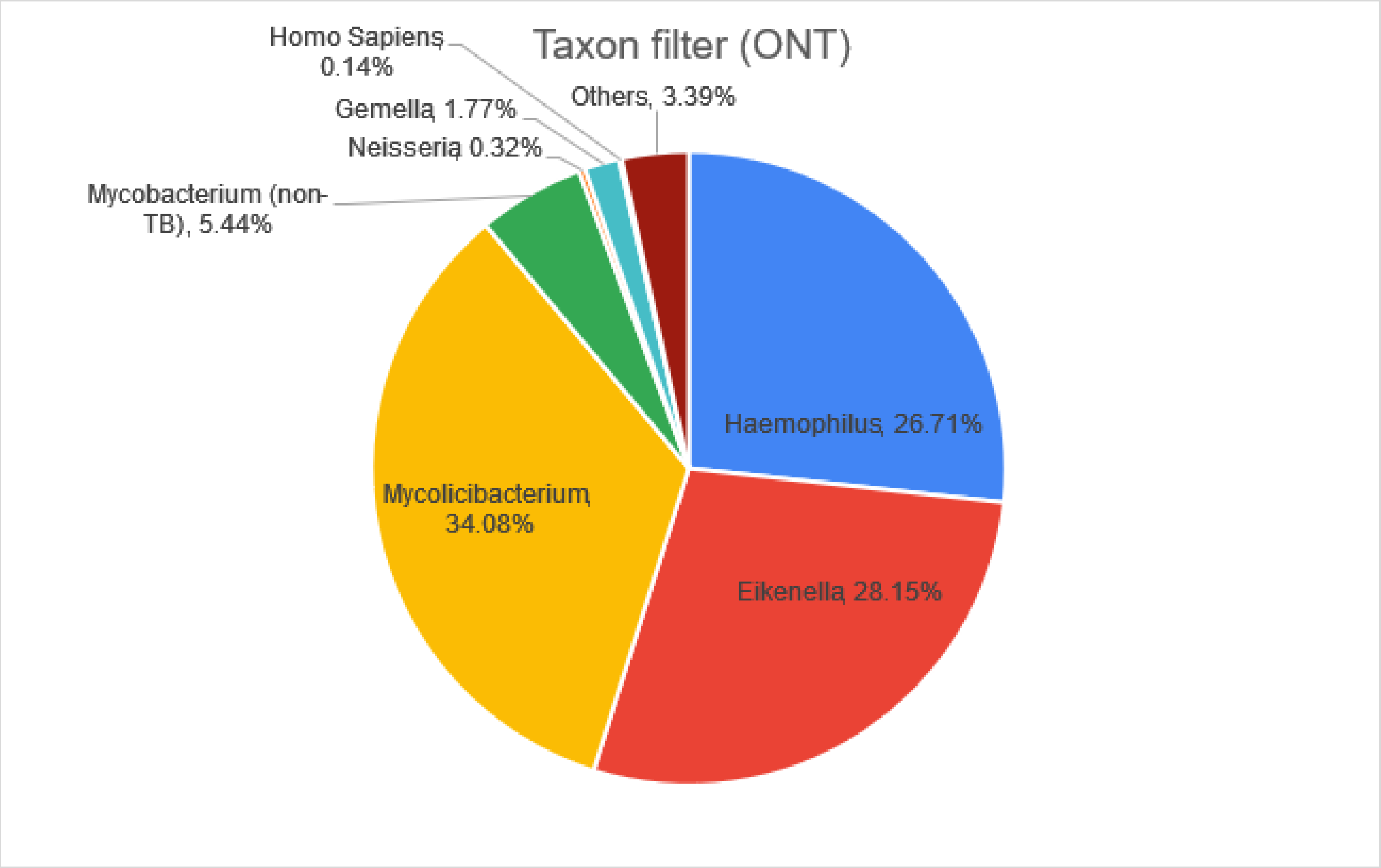
MTBC_H37RV.fasta NC_000962.3 NC_000962.3:1,474,900-1,477,852

17 tracks loaded NC_000962.3:1,476,582 743M of 1,695M

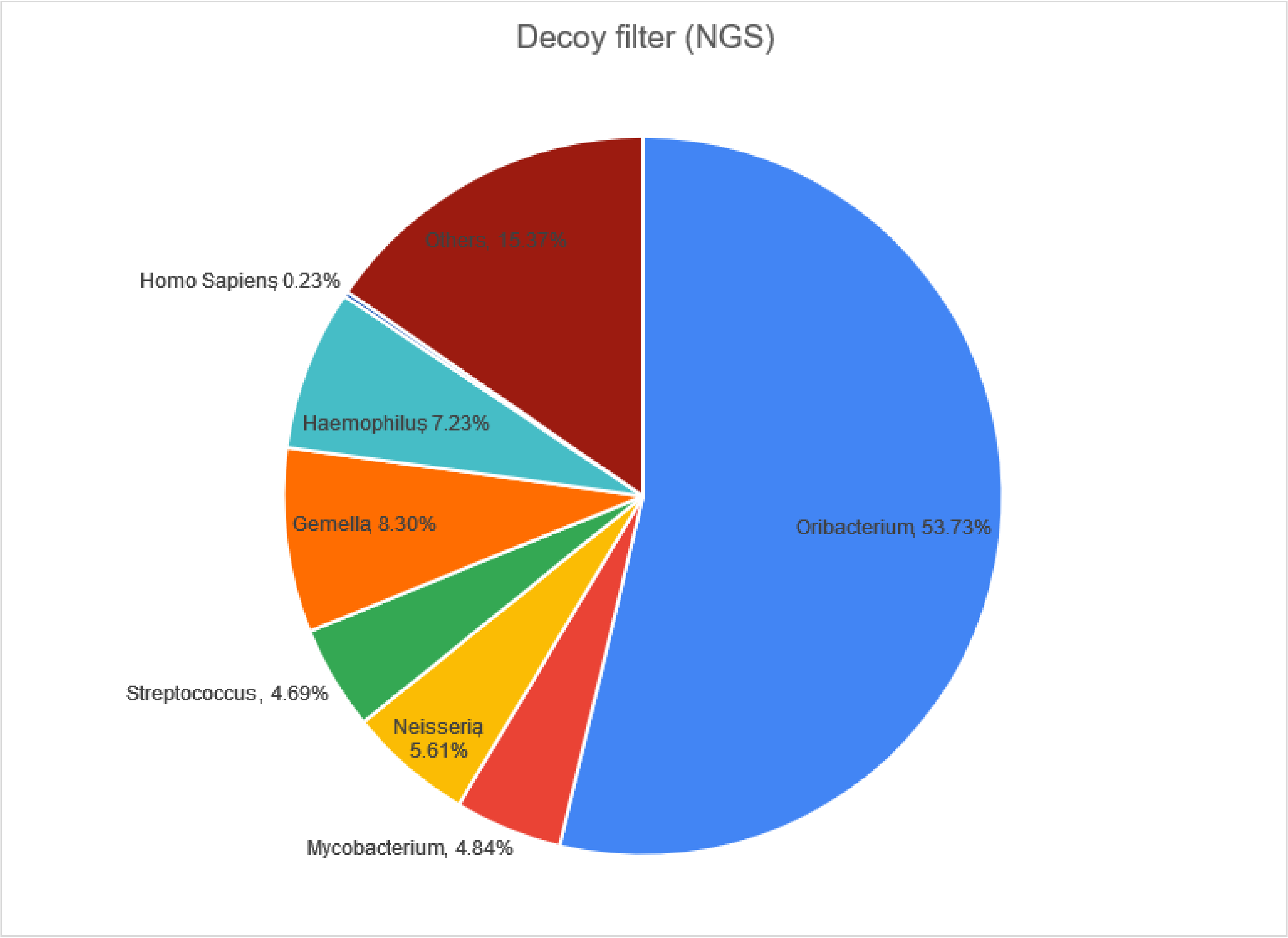
8c)


Decoy Filter (ONT)
Streptococcus, 0.36%
Others, 3.00%
Gemella, 4.21%
Homo Sapiens, 0.00%
Eikenella, 29.54%
Haemophilus, 58.98%
Neisseria, 3.91%

8d)


Taxon filter (ONT)
Homo Sapiens, 0.14%
Gemella, 1.77%
Neisseria, 0.32%
Others, 3.39%
Mycobacterium (non-TB), 5.44%
Haemophilus, 26.71%
Mycolicibacterium, 34.08%
Eikenella, 28.15%

8e)


Decoy filter (NGS)
Others, 15.37%
Homo Sapiens, 0.23%
Haemophilus, 7.23%
Gemella, 8.30%
Streptococcus , 4.69%
Neisseria, 5.61%
Mycobacterium, 4.84%
Oribacterium, 53.73%

8f)

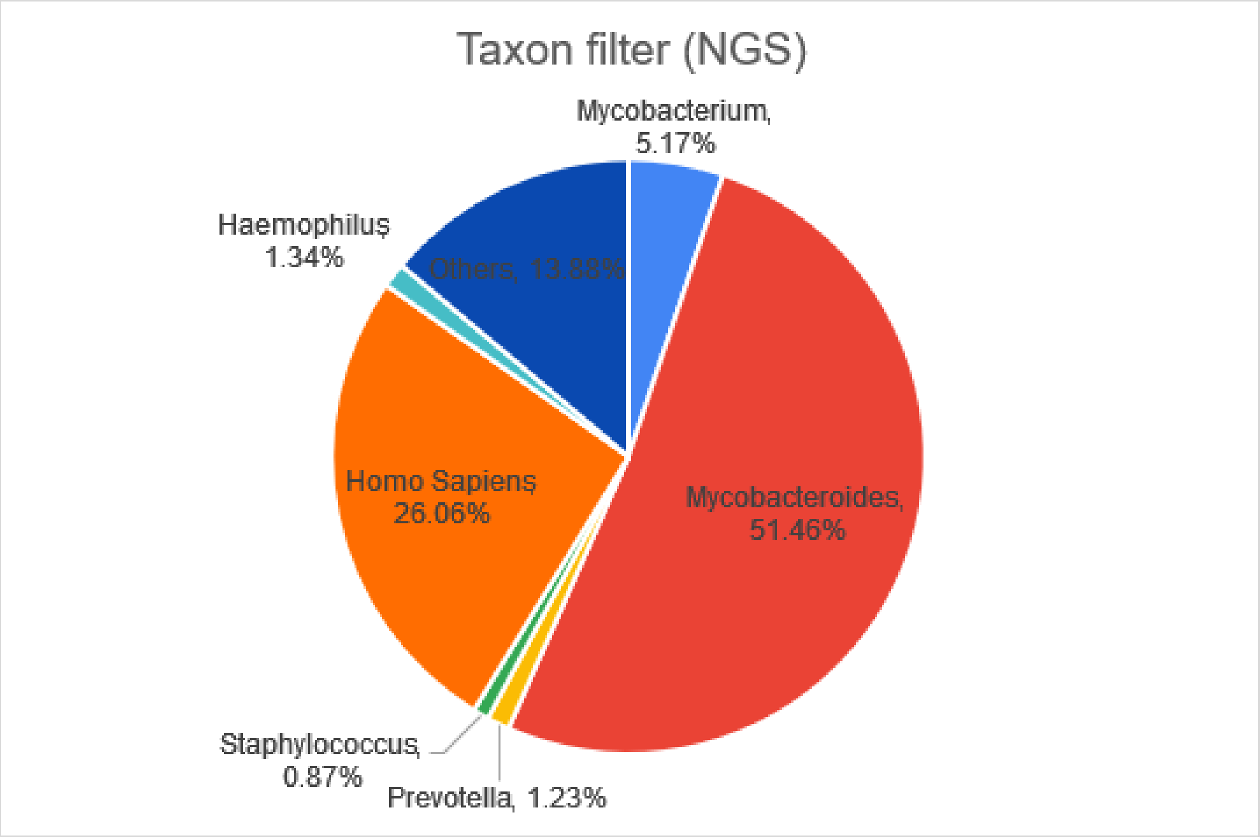


Fig 8a) The Integrative Genomics Viewer (IGV) showed the mapping of sequencing reads to the *rrl* gene of MTB reference genome NC_000962.3) for a spike-in samples containing serially diluted MTB gDNA. The number of random variants (colored) increased in the samples with lower level of MTB. Fig. 8b) The sequencing reads harboring the noisy variants at the *rrl* gene of the same sample were removed by the decoy and taxon filters.

The distribution of the nasal/oral flora species and host genomic DNA in the unique alignments filtered in 8c) decoy filter and 8d) taxon filter in ONT workflow. The distribution of the nasal/oral flora species and host genomic DNA in the unique alignments filtered in 8e) decoy filter and 8f) taxon filter in NGS workflow.

4-1-4. Limit of Detection

Linear equations for 19 different genomic regions derived from the 55 spike-in samples (validation set) were constructed with the analysis results summarized in Fig. 9a (for ONT), Fig. 9b (for NGS), and Supplementary Table 10. The area under the curve (AUC) for ONT ranged from 0.932 to 0.985, while the AUC for the NGS ranged from 0.888 to 0.979.

By calculating the adjusted *IS6110* Ct value required for having a probability of 0.9 of reaching the log average DP 1.7 (DP 50) in all targeted regions, the LOD (the lowest adjusted IS6110 Ct value) of the first-line drug panel (*katG*, *furA-KatG* intergenic regions, *mabA-inhA* promoter, *rpoB*, *ubiA*, *embB*, *rpsA*, and *pncA*) and the overall panel in ONT were 27.78 and 25.06, respectively. The LOD of the first line drug panel and the overall panel in NGS were 23.69 and 22.72 respectively.

The testing cohort, which contained 130 clinical specimens, was used to test the prediction accuracy of the logistic regression model (Supplementary Table 10c). With the probability cutoff of 0.9, the true positive rate in ONT ranged from 92.96% to 100% (mean 97.71%, SD 1.88%), but the true negative rate widely ranged between 14.46% and 50% (mean 28.62%, SD 9.72%). Similarly, the true positive rate in NGS ranged from 88.89% to 100% (mean 94.17%, SD 2.4%), but the true negative rate dropped to the range between 17.65% and 80% (mean 32.05%, SD 13.32%).

In the testing cohort, 57 samples were positive for both the acid-fast bacilli (AFB) smear and culture, whereas 73 samples were AFB negative. For those samples with positive AFB smear results, 94.74% (54/57) and 89.47% (51/57) were respectively first-line drug interpretable (k*atG,* *furA-katG* intergenic, m*abA-inhA* promoter, *inhA* structural, *rpoB, ubiA, embB. rpsA*, and *pncA*) and overall interpretable (all 19 target regions) in both ONT and NGS. For those samples with AFB smear negative results, 63.01% (46/73) and 57.73% (42/73) were first-line drug interpretable and overall interpretable, respectively.

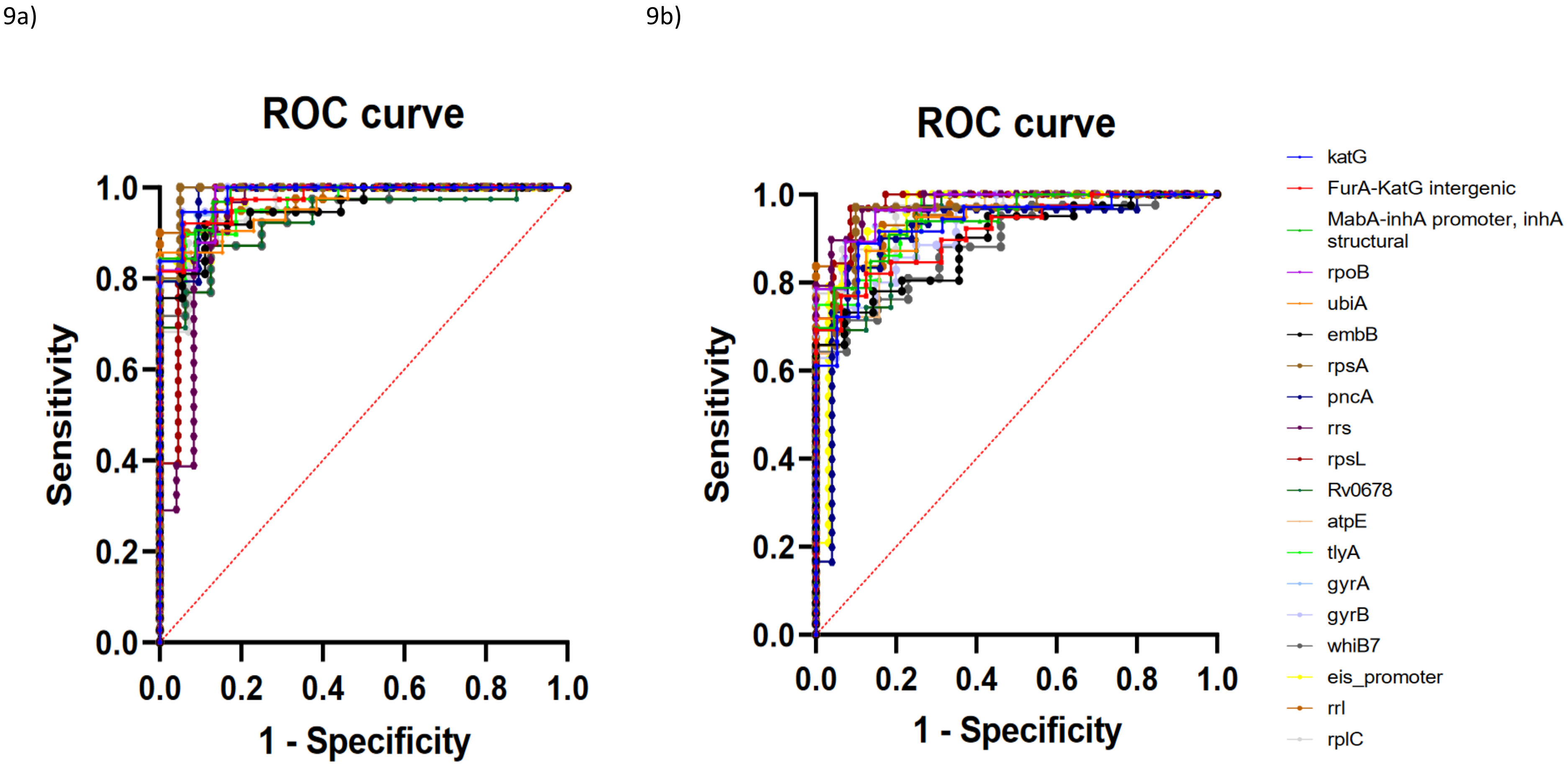


Fig. 9a) The ROC curve for evaluating the logistic regression analysis of the validation set (55 spike-in samples) that was used for determining the LOD in ONT. The Beta values were listed on Supplementary Table 10a. Fig. 9b) The ROC curve for evaluating the logistic regression analysis of the validation set (55 spike-in samples) that was used for determining the LOD in NGS. The Beta values were listed on Supplementary Table 10b. The diagonal red dash line was the reference line if there was a random relationship between the adjusted IS6110 Ct value and the DP (depth of coverage) >=50.

4-1-5. The diagnostic accuracy of the drug resistance detection workflow

For verification of performance of our new assays, nine clinical isolates, three spike-in samples (with clinical isolates WC026, WC069, and 150B) included in our previous study [95] were sequenced to reproduce the genotypic results presented in the previous study. The redesigned primer set for direct sequencing was comparable to the primer set in previous studies with the 100% concordance in genotypic drug resistance results (Supplementary Table 11). Within the adjusted IS6110 PCR Ct value of 27.78, AMR associated mutations conferring resistance to INH, RIF, EMB, PZA, STR, CAP, KAN, and AMK in the WC069 spike-in series (up to dilution power 7, equivalent to the adjusted *IS6110* Ct value of 26.11) was detected in ONT, that was consistent with NGS, except RIF in sample of dilution power 7 that was classified inconclusive because of low DP. In the WC026 spike-in series, the resistances to INH and FQ were detected in samples up to dilution power 5 (equivalent to the adjusted *IS6110* Ct value of 27.72) in both ONT and NGS. No AMR associated mutation was found in samples up to dilution power 6 (equivalent to the adjusted *IS6110* Ct value of 25.58) in the 150B spike-in series in ONT; this was consistent with NGS except for the *eis* promoter in the sample of dilution power 6 because of the low DP. As expected, pure susceptibility was also found in all the H37RV spike-in series (up to dilution power 1 in ID1, ID2, and ID4 series, respectively), though *eis promoter* and *rrs* were uncertain in samples ID1_power0 and ID2_power0, respectively, because of low DP in either ONT or NGS (Supplementary Table 12). Surprisingly, uncertain genotypic results were found in several regions in ID1_power2 because of low DP in NGS, though its adjusted IS6110 Ct value was below the cutoff. In short, the agreement of genotypic results between ONT and NGS in primer versions 1 and 2 was 100%.

With its well-known high base accuracy, NGS is used as a reference method for evaluating the genotyping performance of ONT (as summarized in Table 3 and Supplementary Table 13). In the testing cohort, 85 of the 130 clinical specimens fell below the cutoff for the adjusted IS6110 PCR Ct value of 28 (up from 27.78 in MabA-inhA promoter, inhA structural). Of these, 77 out of 85 clinical specimens were first-line drug interpretable, and 73 out of these 77 clinical specimens were overall interpretable. In the first drug interpretable group, 5 clinical specimens and 72 clinical specimens were respectively reported as resistant and susceptible in both ONT and NGS, and 8 clinical specimens were inconclusive because of low VAF or DP in either ONT or NGS. In the overall interpretable group, 11 clinical specimens and 62 clinical specimens were respectively reported as resistant and susceptible in both ONT and NGS, and 12 clinical specimens were inconclusive. On the other hand, 45 out of 130 clinical specimens were above the cutoff for the adjusted IS6110 Ct value of 28. Of these, 23 clinical specimens were first-line interpretable, and 20 of these 23 clinical specimens were overall-interpretable. In the first drug interpretable group, 3 clinical specimens and 20 clinical specimens were respectively reported as resistant and susceptible in both ONT and NGS, and 22 clinical specimens were inconclusive because of low VAF or DP in either ONT or NGS. In the overall interpretable group, 2 clinical specimens, and 18 clinical specimens were respectively reported as resistant and susceptible in both ONT and NGS, and 25 clinical specimens were inconclusive. Of note, no disagreement was found in all clinical specimens with available genotypic results in both ONT and NGS. The agreement of genotypic results between ONT and NGS was 100% in this cohort. The genotypic results of all specimens are listed in Supplementary Tables 13a and 13b.

Table 3: The agreement of the genotypic results between ONT and NGS in 130 clinical specimens (Eighty-five samples were below adjusted IS6110 Ct value 28 and 45 samples were above this adjusted IS6110 Ct value).

	Adjusted IS6110 Ct value < 28	Adjusted IS6110 Ct value > 28
Number of samples	85	45
First line drug interpretable	77	23
Agreement in resistance	5	3
Agreement in susceptible	72	20
Total number of disagreement	0	0
Number of samples inconclusive	8	22
Overall interpretable	73	20
Agreement in resistance	11	2
Agreement in susceptible	62	18
Total number of disagreement	0	0
Number of samples inconclusive	12	25

Phenotypic drug susceptibility test (pDST) results were available for 48 out of 130 clinical samples (Supplementary Table 14), while the results for the remaining samples were pending because the drug was unavailable for testing. Also, bedaquiline was not available for pDST for all samples. Overall, there were 18 true resistant results (4 in INH, 3 in RIF, 2 in EMB, 1 in PZA, 6 in STR, 1 in KAN, and 1 in AMK), 453 true susceptible results (40 in INH, 43 in RIF, 46 in EMB, 47 in PZA, 40 in STR, 47 in KAN, 47 in AMK, 47 in CAP, 48 in FQ, and 48 in LZD), 7 false resistant results (3 in INH, 1 in RIF, 2 in STR, and 1 in CAP), and 1 false susceptible result in RIF. Sample 23953 was inconclusive for INH as there was no coverage in the MabA-inhA promoter or inhA structural region (Table 4). The true positive (resistant) rate was 0.94 (18/(18+1)), while the true negative (susceptible) rate was 0.984 (453/(453+7)). The precision and the recall were 0.72 (18/(18+7)) and 0.94 (18/(18+1)) respectively, and so the F1 score was 0.815. Among the false-positive (resistant) results, most of the false results were found in sample IDs 19395, 19396, and 21065R. Mutation 1673425 (C>T) at the *MabA-inhA* promoter was found in all these samples. Mutation 781687 (A>G) at rpsL was found in sample IDs 19395 and 19396. Mutation 1473246 (A>G) at *rrs* was found in sample ID 21065R. Mutation 761161 (T>C) at *rpoB* was found in sample ID 21729. Additionally, a false susceptible result for RIF was found in sample ID 21065R (Table 5).

Table 4) The agreement between the genotypic results and the pDST results of 48 clinical specimens.

	Overall										
	First-line drugs				Second-line drugs						
	Isoniazid	Rifampicin	Ethambutol	Pyrazinamide	Streptomycin	Kanamycin	Amikacin	Capreomycin	Fluoroquinolone	Bedaquiline	Linezolid
True resistant	4	3	2	1	6	1	1	0	0	N/A	0
True susceptible	40	43	46	47	40	47	47	47	48	N/A	48
False resistant	3	1	0	0	2	0	0	1	0	N/A	0
False susceptible	0	1	0	0	0	0	0	0	0	N/A	0
N/A (VF<0.2 and/or DP<20)	1	0	0	0	0	0	0	0	0	N/A	2
N/A (pDST is not available)	0	0	0	0	0	0	0	0	0	N/A	0

Table 5) The summary of false resistant and susceptible results.

Drug	Sample ID	Agreement	Gene	Associated mutations (DP, VAF), amino acid change
Isoniazid (INH)	19395	False resistant	*MabA-inhA* promoter, *inhA* structural	1673425 (C>T), (0.97, 2895), C-15T
	19396	False resistant	*MabA-inhA* promoter, *inhA* structural	1673425 (C>T), (0.97, 2796), C-15T
	21065R	False resistant	*MabA-inhA* promoter, *inhA* structural	1673425 (C>T), C-15T
Rifampicin	21729	False resistant	*rpoB*	761161 (T>C), (0.92, 5206), Leu533Pro
	21065R	False susceptible	*rpoB*	
Streptomycin	19395	False resistant	*rpsL*	781687 (A>G), (0.9, 1938), Lys43Arg
	19396	False resistant	*rpsL*	781687 (A>G), (0.9, 3068), Lys43Arg
Capreomycin	21065R	False resistant	*rrs*	1473246 (A>G), (0.98, 3513), A1401G

4-1-6. Time to report and operation cost

In the ONT workflow, genomic DNA extraction and post-DNA-extraction cleanup, and the multiplex PCR with its post-PCR cleanup, were held on the first working day. Then, the barcoding PCR and library preparation were held in the next morning, followed by the commencement of the sequencing run in the afternoon on the second working day. By assuming 24 samples per sequencing batch, the sequencing run lasted around 16 hours, which allowed the subsequent data analysis on the third working day. The AMR report was available the same working day. The total time to complete the clinical report for ONT was three working days.

In the NGS workflow, like the ONT workflow, genomic DNA extraction, post-DNA-extraction cleanup, multiplex PCR, and post-PCR cleanup were conducted on the first working day. Library preparation and library quality control (QC) were done on the second working day. The library was sequenced with MiSeq Nano V2 (250 x 2 cycles) on the third working day. After the 28-hour sequencing, assuming 48 samples per sequencing batch, the subsequent analysis was performed in the morning of the fourth working day, and so the AMR report was available in the afternoon. The total time to clinical report for NGS workflow was 4 working days (Fig. 10).

The cost per sample was USD 45.47 for ONT (assuming 24 samples per sequencing batch) and USD 59.72 for NGS (assuming 48 samples per sequencing batch).

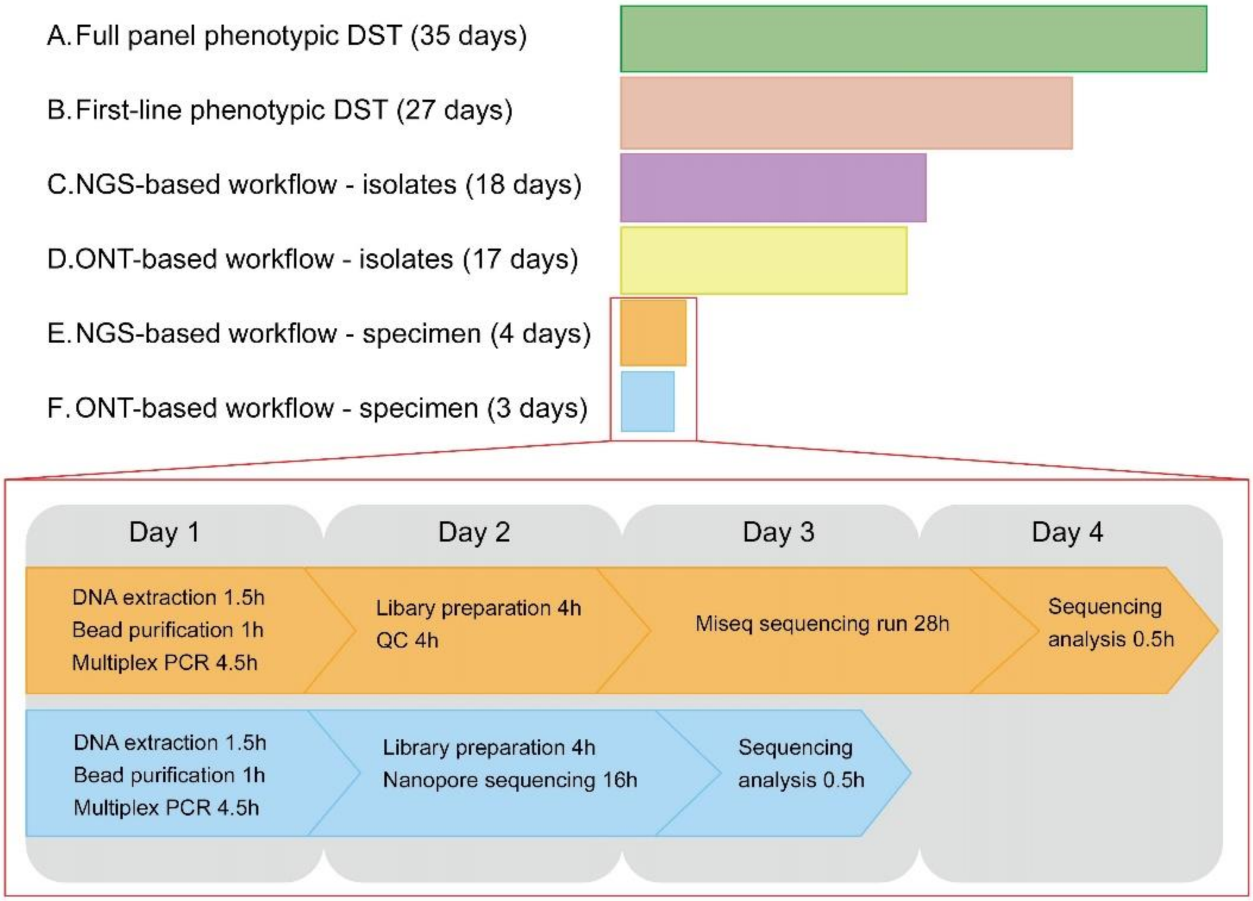


Fig 10) The breakdown of time to report for ONT and NGS workflow for direct AMR detection in *MTB*.

4-2. Targeted sequencing workflow for direct ARV resistance detection in HIV in plasma samples

4-2-1 Assay optimization

To facilitate the long-range targeted genome amplification, two commercial master mix claimed for long-range PCR (Platinum SuperFi II PCR Master Mix, and LongAmp Taq 2X Master Mix) were tested with six plasma samples (KB2056, KB2061, KB2064, KB2065, KB2066, and KB2067). The reaction setup and the cycling condition were list in Table 6. Though the amplicon concentration from targeted genomic amplification with LongAmp was higher than those with SuperFi II, no clear band with the expected amplicon length could be observed in gel electrophoresis, while clear bands with the expected amplicon length could be observed in those reactions with SuperFi II (Fig. 11). This suggested that SuperFi II outperformed LongAmp in long-range amplification in this workflow. This also suggested the DNA concentration measurement after post-amplification cleanup did not have any reference value to true amplicon concentration in the sample.

Table 6) The reaction setup and the cycling conditions with SuperFi II PCR Master Mix and LongAmp Taq 2X Master Mix.

LongAmp Taq 2X Master Mix

Component	**Volume (ul)**	**Final concentration**
LongAmp Taq 2X Master Mix	12.5	1X
10 uM Forward primer	1	0.4 uM
10 uM Reverse primer	1	0.4 uM
Template DNA	8	
Nuclease-free water	2.5	
Total volume	**25**	

Cycling condition

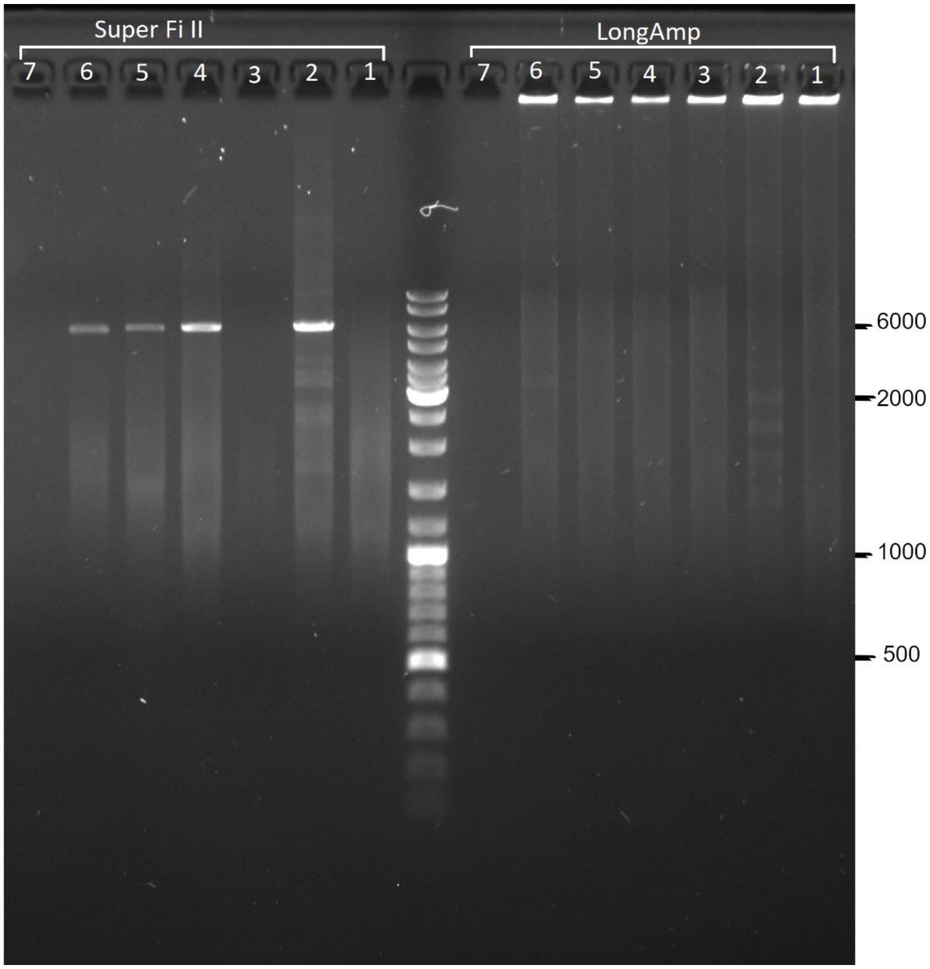
Cycle Step	**Temp**	**Time**
Heat denaturation	94°C	3 min
35 cycles	94°C	30 s
	60	30 s
	65°C	5 min
Final extension	65°C	10 min
	4°C	hold

SuperFi II PCR Master Mix

Component	Volume (uL)	Final concentration
2X Platinum SuperFi II PCR Master Mix	25	1X
10 uM Forward primer	2.5	0.5 uM
10 uM Reverse primer	2.5	0.5 uM
Template cDNA (directly from reverse transcription reaction)	20	
Total volume	50	

Cycling condition

Cycle step	Temperature	Time	
Heat activation	98°C	2 minutes	
Denaturation	98°C	10 seconds	35 cycles
Annealing	60°C	10 seconds	
Extension	72°C	4 minutes 30 seconds	
Final extension	72°C	5 minutes	
Hold	4°C	∞	



Tube	Sample ID (ezDNase-treated)	Viral load (copies/ uL)
1	KB2056	4.97
2	KB2061	2745.36
3	KB2064	57.14
4	KB2065	927.34
5	KB2066	307.97
6	KB2067	329.59
7	NTC	N/A

Fig. 11) Gel electrophoresis of the amplicons after long-range targeted genomic region (~6kb). The key size markers on the ladder were listed on the right side of the figure. The sample IDs and the viral load were listed in the above table.

On the other hand, the effect of human DNA removal with ezDNase on amplification efficiency was also studied. The same six samples were taken to long-range targeted genome amplification with or without pretreatment with ezDNase. The band intensity for those amplicons with pretreatment of ezDNase was generally higher than those without pretreatment of ezDNase. This concluded that human DNA removal with ezDNase was recommended prior to targeted genome amplification (Fig. 12).

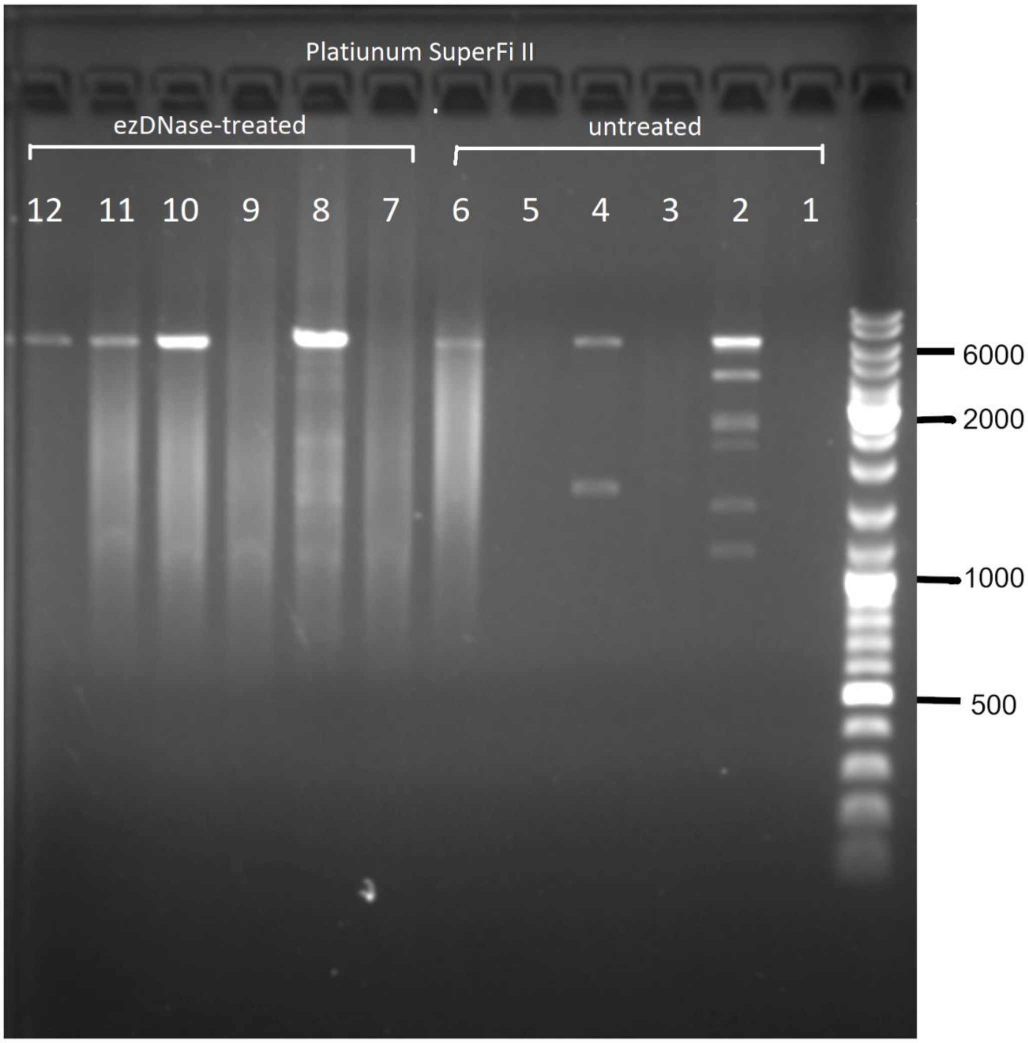


Fig. 12) Gel electrophoresis of PCR products of the same six samples after long-range targeted genome amplification with or without human gDNA removal with ezDNase.

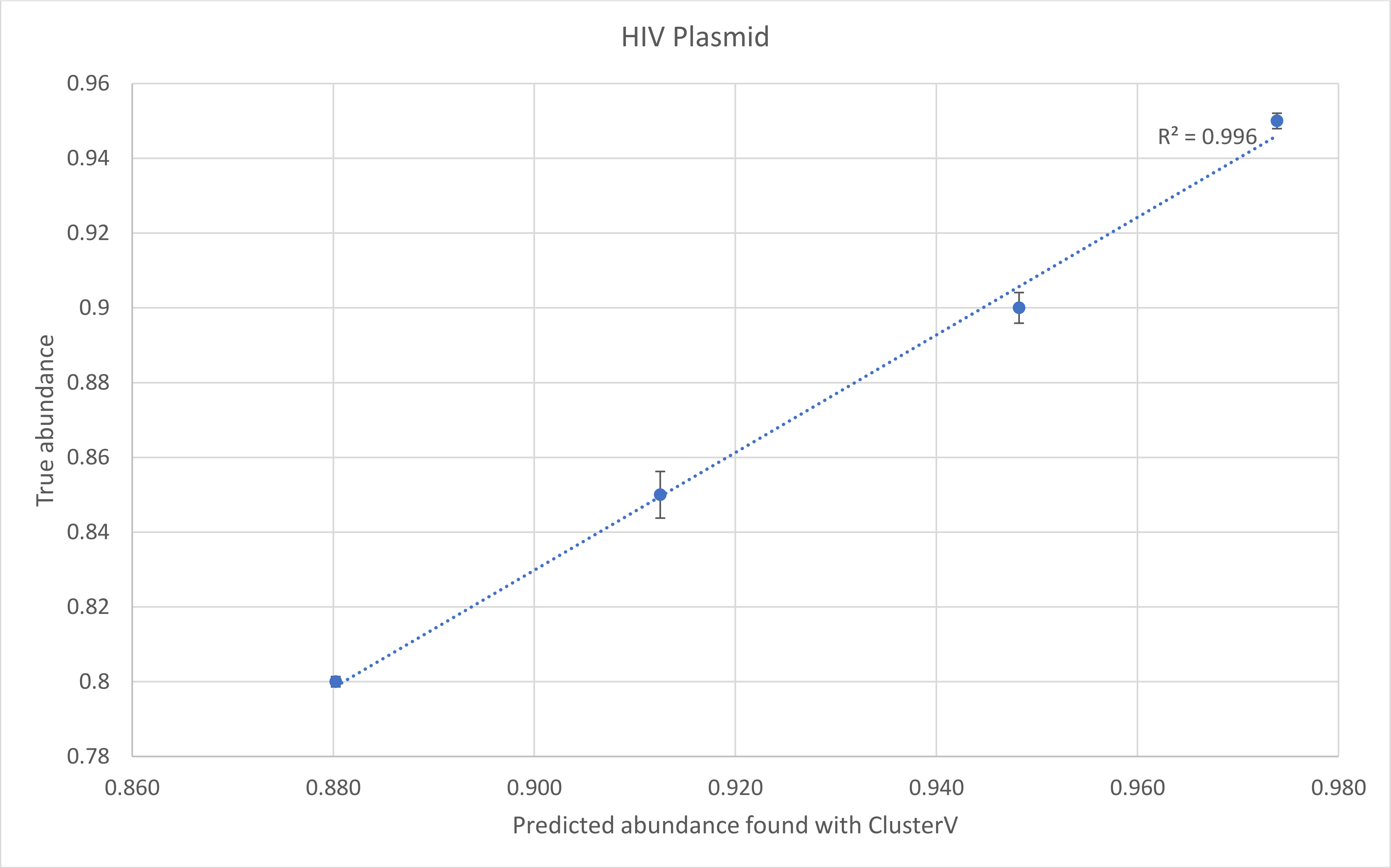
4-2-2. Validation of clustering performance in the ONT workflow

To test the clustering performance of ClusterV in the ONT workflow, samples with a unique mutation pattern were used for generating the *in-silico* simulation data set and gradient series in a known mixing ratio. In principle, the predicted abundance (mixing ratio) of different samples determined by ClusterV should be consistent with the true abundance (the corresponding mixing ratio) in the data set or the gradient series. The in-silico simulation dataset was prepared by mixing the HIV plasmid and two clinical samples (Sample ID: KB2061 and KB2979) in various combinations (10:10:80, 33:33:33, 80:20:0, 50:50:0, and 5:95:0). Meanwhile, 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.

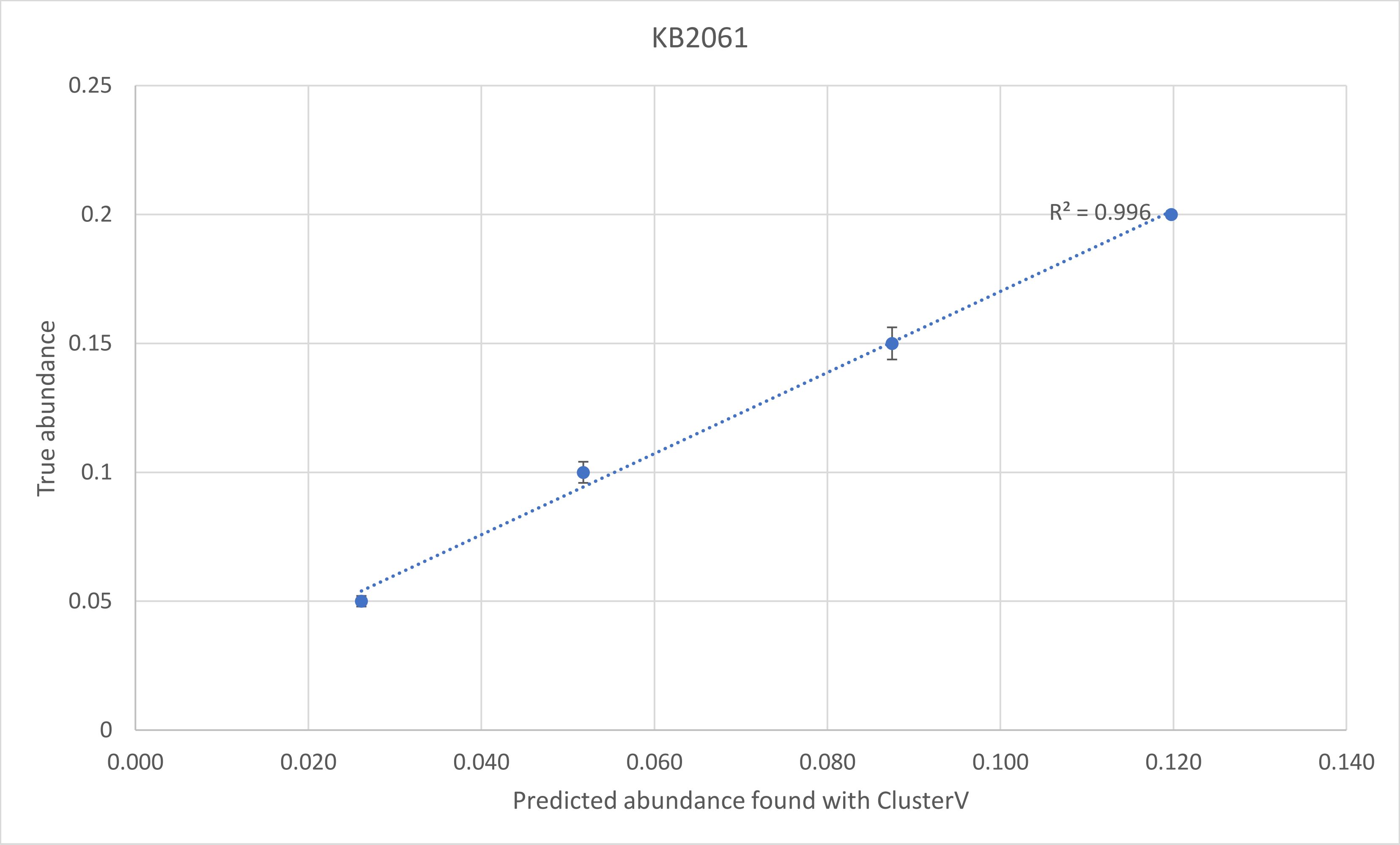
In short, only one quasispecies was found in the individual three samples (HIV plasmid, KB2061, and KB2979) used in the gradient series and *in silico* simulation data set with a median VAF > 0.9, confirming their qualification in evaluating the clustering performance.

After the clustering with ClusterV, the number of identified quasispecies matched to the actual number of quasispecies in both gradient series and *in silico* simulation data set. Also, the predicted abundance found with ClusterV was in a linear relationship with the true abundance. The R squares for the HIV plasmid and KB2061 gradient series were both 0.996 (Figs. 13a and 13b), while the R square for the *in silico* simulation data set was 0.9939 (Fig. 13c). In general, the median VAF for variants found in each quasispecies was 0.89 or above (Supplementary Table 15).

13a)


HIV Plasmid
R² = 0.996
True abundance
0.96
0.94
0.92
0.9
0.88
0.86
0.84
0.82
0.8
0.78
0.860
0.880
0.900
0.920
0.940
0.960
0.980
Predicted abundance found with ClusterV

13b)


KB2061
True abundance
0.25
0.2
0.15
0.1
0.05
0
0.000
0.020
0.040
0.060
0.080
0.100
0.120
0.140
Predicted abundance found with ClusterV
R² = 0.996

13c)

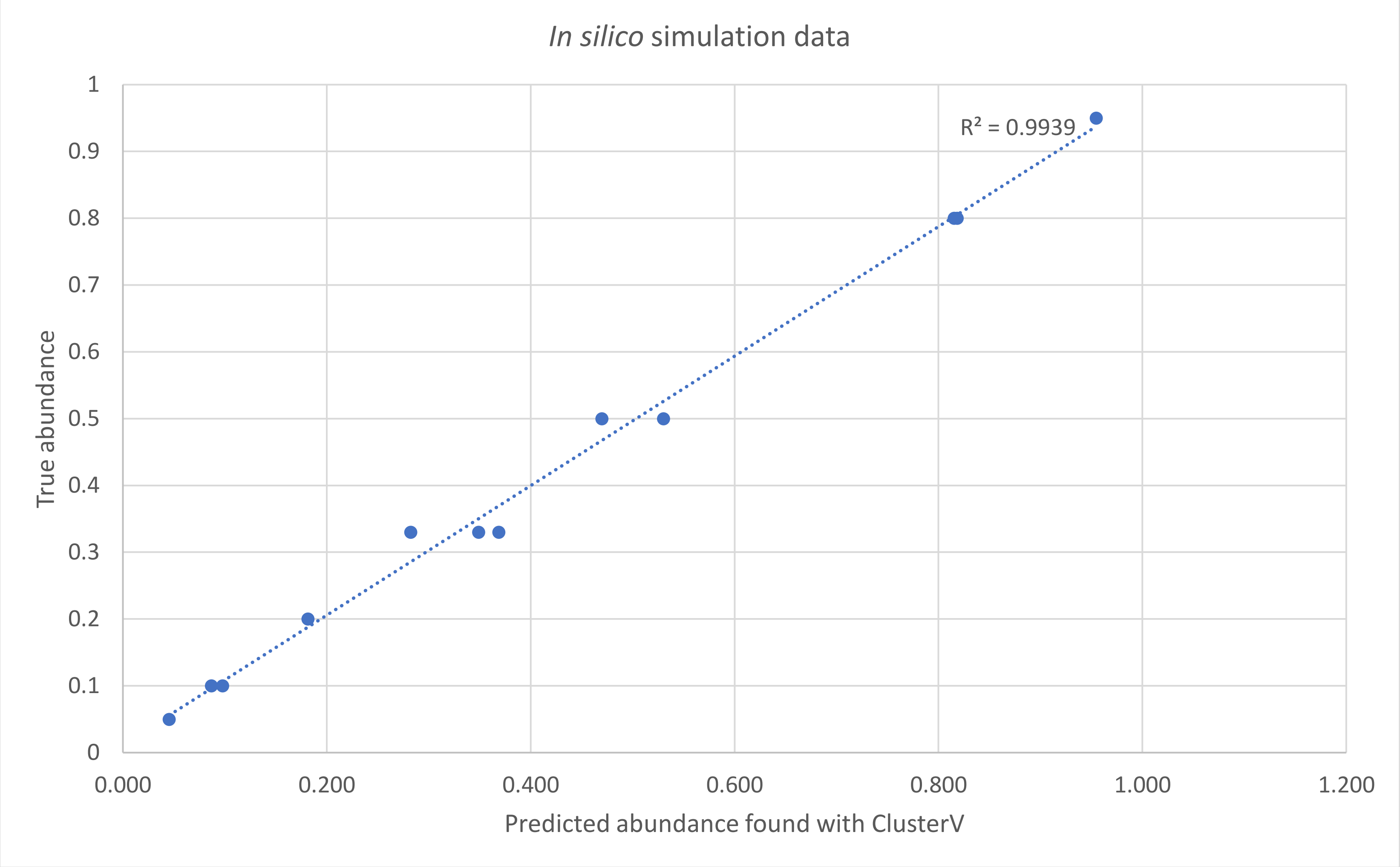


Fig. 13) The comparison between the true abundance and the predicted abundance with ClusterV in a) HIV plasmid, b) KB2061 in gradient study, and c) *In silico* simulation data.

4-2-3. The variant calling and diagnostic performance of the targeting ONT sequencing workflow

Fifty-nine samples out of 77 plasma samples were successfully sequenced with the ONT workflow. The Sanger sequencing results were available for all these samples, covering PR and RT (covering amino acid positions 1–402), and the Sanger sequencing results for INT were only available in 16 samples. On the other hand, NGS results were available in 43 samples but not in 16 samples because of insufficient sample volume.

After clustering with ClusterV, 28 out of 59 samples (47.45%) consisted of only one subtype, 10 samples (16.94%) consisted of two subtypes (16.94%), 6 samples (10.17%) consisted of three subtypes, and the remaining 19 samples (25.42%) consisted of more than three subtypes. Of which, subtype B and CRF01_AE were dominant (accounting for 74.19%), followed by CRF07_BC (11.29%). Three samples, KB2974, KB2980, and KB2998, carried a mixture of subtypes (Supplementary Table 16a).

A total of 4,104 amino acid mutations were found in 59 samples with ONT (Fig. 14, Supplementary Table 16b). Of which, 2,200 mutations (53.6%) were concordant with Sanger, 271 mutations (6.6%) were discordant, 1,506 (36.7%) mutations could not be validated with Sanger because of the unavailable Sanger sequencing results, and 127 (3.1%) mutations were classified as unknown amino acid mutations (Fig. 14a).

Of the 1,506 mutations found in ONT that could not be validated with Sanger and those 271 discordant mutations, 1153 (28%) and 222 (5%) mutations were respectively concordant and discordant with NGS, and 49 (1%) discordant mutations could not be validated with NGS. The remaining 353 mutations could not be validated without available NGS data. On the other hand, 135 amino acid mutations were uniquely found in Sanger but not in the ONT workflow (Fig. 14b). Of which, only 27 mutations were classified as true mutations because they were concordant with NGS, while 62 mutations were false mutations because of their discordance, and 46 mutations could not be validated because of unavailable NGS results.

As a result, 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. By correlating the log overall VAF with the number of true and false amino acid mutations, an ROC curve with AUC 0.903 was constructed. The overall VAF of reaching the true mutation rate of 0.9 of getting true amino acid mutations was 0.4, with the true mutation rate of 0.9072 and the false mutation rate of 0.1712 (Fig. 15, Supplementary Table 17).

Thirty AVR-resistance associated mutations were found in 22 samples (Fig. 14c, Supplementary Table 18a). Twenty-six out of thirty-three mutations were considered true mutations because of their concordance with Sanger. Six out of thirty-three mutations were discordant with Sanger, including four false mutations because of their discordance with NGS, one true mutation with their concordance with NGS, and one true mutation with no corresponding Sanger results but concordance with NGS. One mutation was uncertain because of unavailable NGS results. On the other hand, 3 AVR resistance associated mutations in mixed alleles were uniquely found in Sanger (Supplementary Table 18b). One was considered a true mutation because it was concordant with NGS, while the other two were false mutations due to their discordance with NGS.

As a result, the precision and the recall of the diagnostic performance were respectively 0.875 (28/32) and 0.965 (28/(28+1)), and so the F1 score was 0.918. With the cutoff of 0.4 from the

ROC analysis, a total of seven mutations (three were respectively concordant and discordant with Sanger, and one could not be validated with Sanger) with an overall VAF below the cutoff were excluded (Fig. 14d). The precision and the recall of the diagnostic performance were respectively 0.96 (24/25) and 0.96 (24/(24+1)), and so the F1 score increased to 0.96.

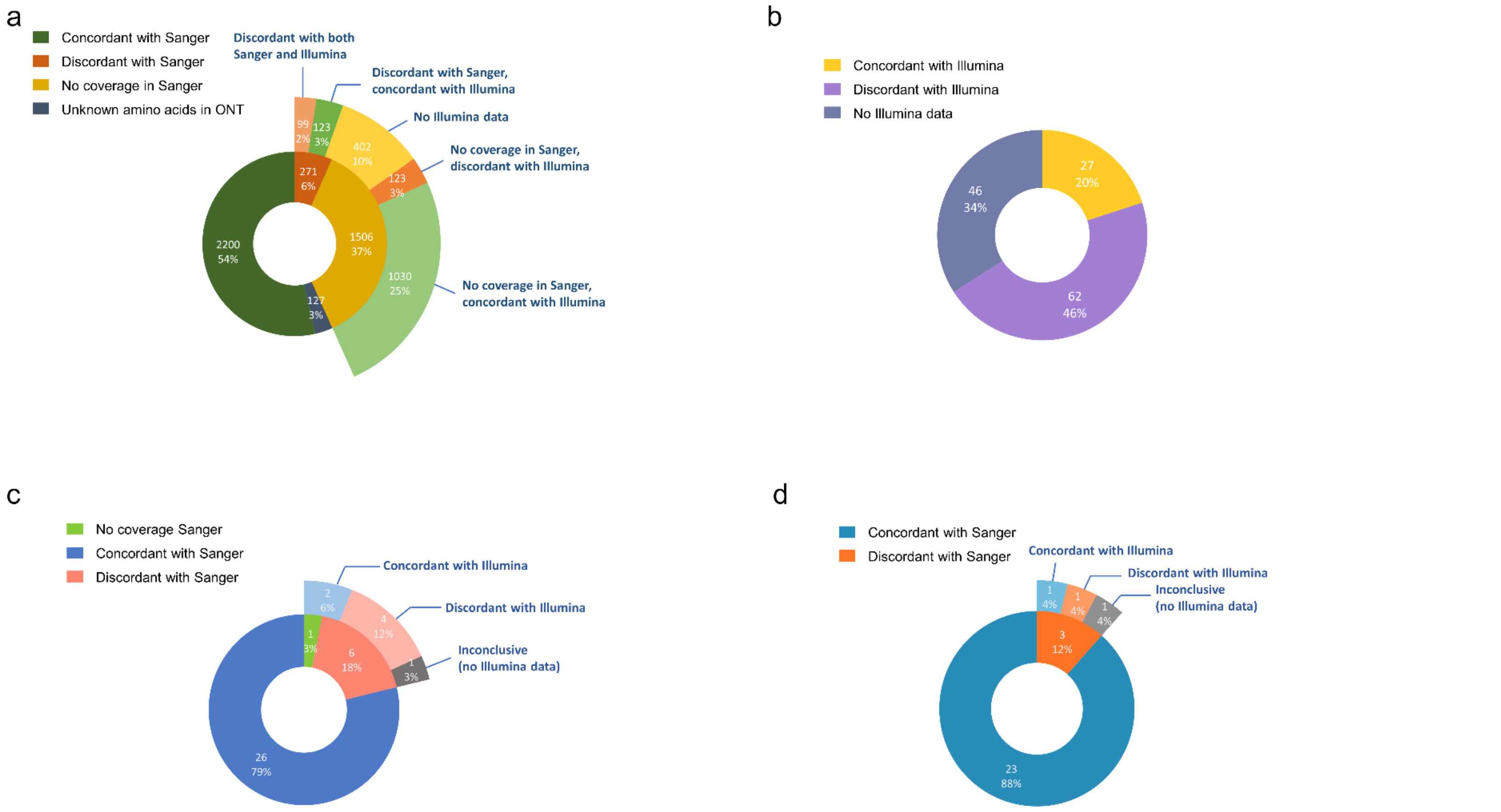


Fig. 14a) The distribution of the amino acid mutations (n=4,104) found in ONT with the concordance or discordance with Sanger or NGS. 14b) The agreement of unique amino acid mutations (n=135) detected in Sanger with the NGS. 14c) The distribution of AVR ssociated amino acid mutations (n=33) found in ONT with the concordance or discordance with Sanger or NGS before the application of the overall VAF cutoff 0.4, and 14d) after cutoff.

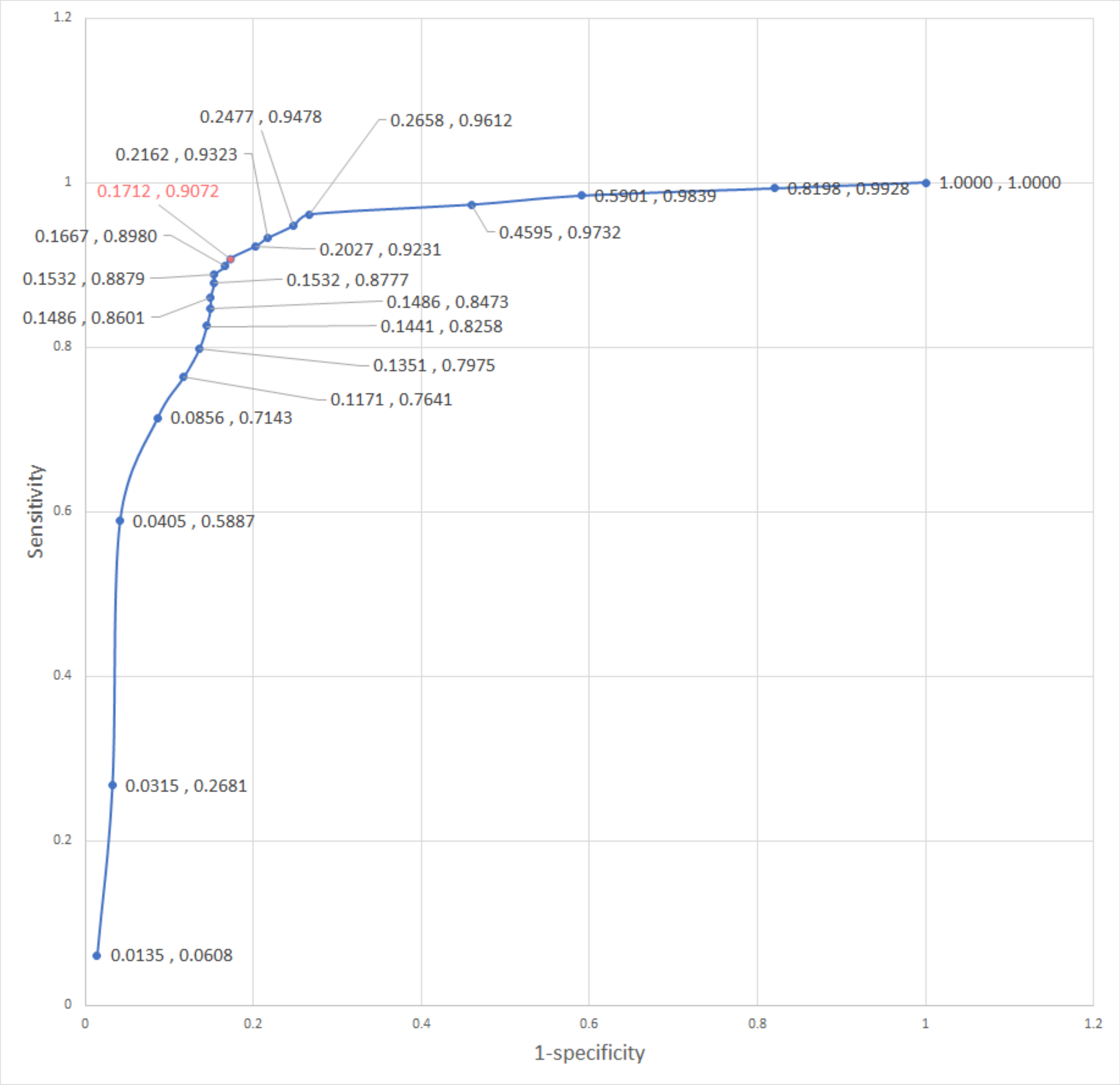


Fig. 15) The ROC analysis of variant calling performance in the ONT workflow. Each data point on the curve displayed sensitivity and false positive rate. The overall VAF cutoff 0.4 was highlighted in red.

4-2-4. Detection of the mutations in the quasispecies

With the aid of hierarchical clustering in ONT workflow, different ARV resistance-associated mutations profiles were found between quasispecies in a sample. For example, in the sample KB0270, mutation G73S in the PR gene conferring potential low to low PI resistance (ATV, FPV, IDV, SQV) and mutations M41L, S68G, and M184V in the RT gene conferring potentially low to high NRTI resistance (ABC, DDI, NFV, FTC, and LMV) were commonly shared with the quasispecies KB2070_1 and KB0270_2 with a total abundance of 100%, but mutation T215F was found only in the quasispecies KB0270_2 (abundance of 32.27%), conferring additional low-intermediate resistance levels of NRTI (ABC, AZT, D4T, DDI, and TDF) (Fig. 16a - c).

In the other example sample KB2987, mutation V106I in RT conferring potentially low NNRTI resistance (DOR, ETR, NVP, and NPV) was found in four out of seven quasispecies (KB2987_2, KB2987_3, KB2987_4, KB2987_5, and KB2987_7) with the total abundance 63.2%, while a low overall AF (0.063) mutation G190E conferring intermediate-high resistance of DOR, EFV, NVP, RPV, and ETR was exclusively found in KB2987_6 (abundance 0.067) that was confirmed with NGS (Fig. 16 d-f).

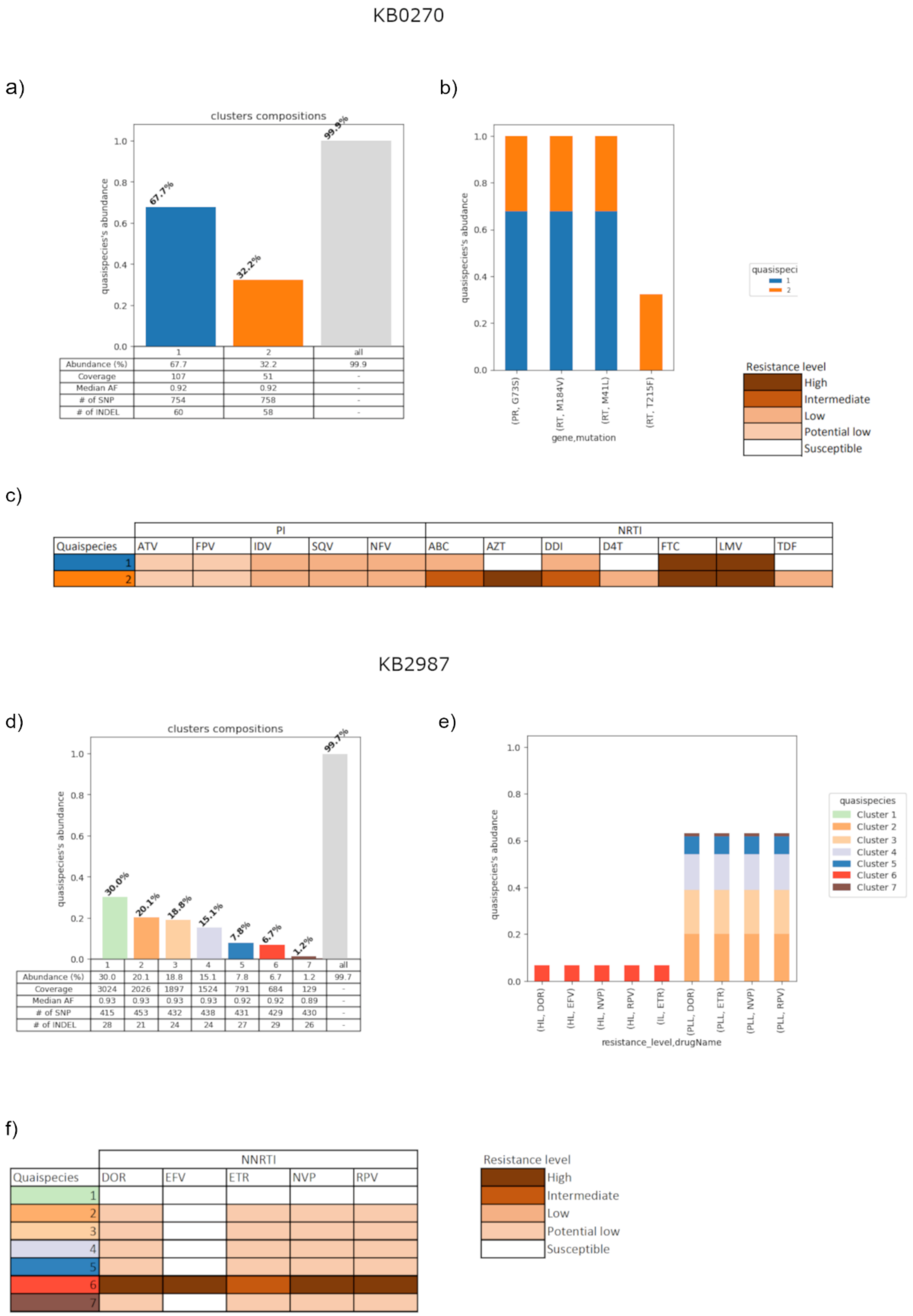


Fig. 16) Illustration of different mutation patterns of HIV quasispecies in samples KB0270 and KB2987. 16a) The example in KB0270 with its different abundance, 16b) resistance patterns, and 16c) resistance levels in different quasispecies. 16d) The example in KB2987 with its different abundance, 16e) resistance patterns, and 16f) resistance levels found in different quasispecies.

4-2-5. Limit of detection

Two linear equations from logistic regression models (Fig. 17) were constructed for two DP cutoffs, the minimum 50 and the suggested 1500, for bioinformatic analysis. The meaning of DP>=50 was depth of coverage enough for variant calling in high confidence for one quasispecies with using Clair-ensemble embedded in ClusterV. Higher DP allowed the detection of low abundance quasispecies with higher confidence, hence DP >= 1500 was recommended to detect quasispecies in the abundance as low as 0.03. The AUCs for the cutoffs of DP 50 and DP 1500 were 0.853 and 0.8549, respectively. By calculating the viral load (copies/uL) required for having a probability of 0.9 of reaching DP 50 and DP 1500, the LOD for DP 50 was 303.9 copies/uL and 1930.6 copies/uL, respectively.

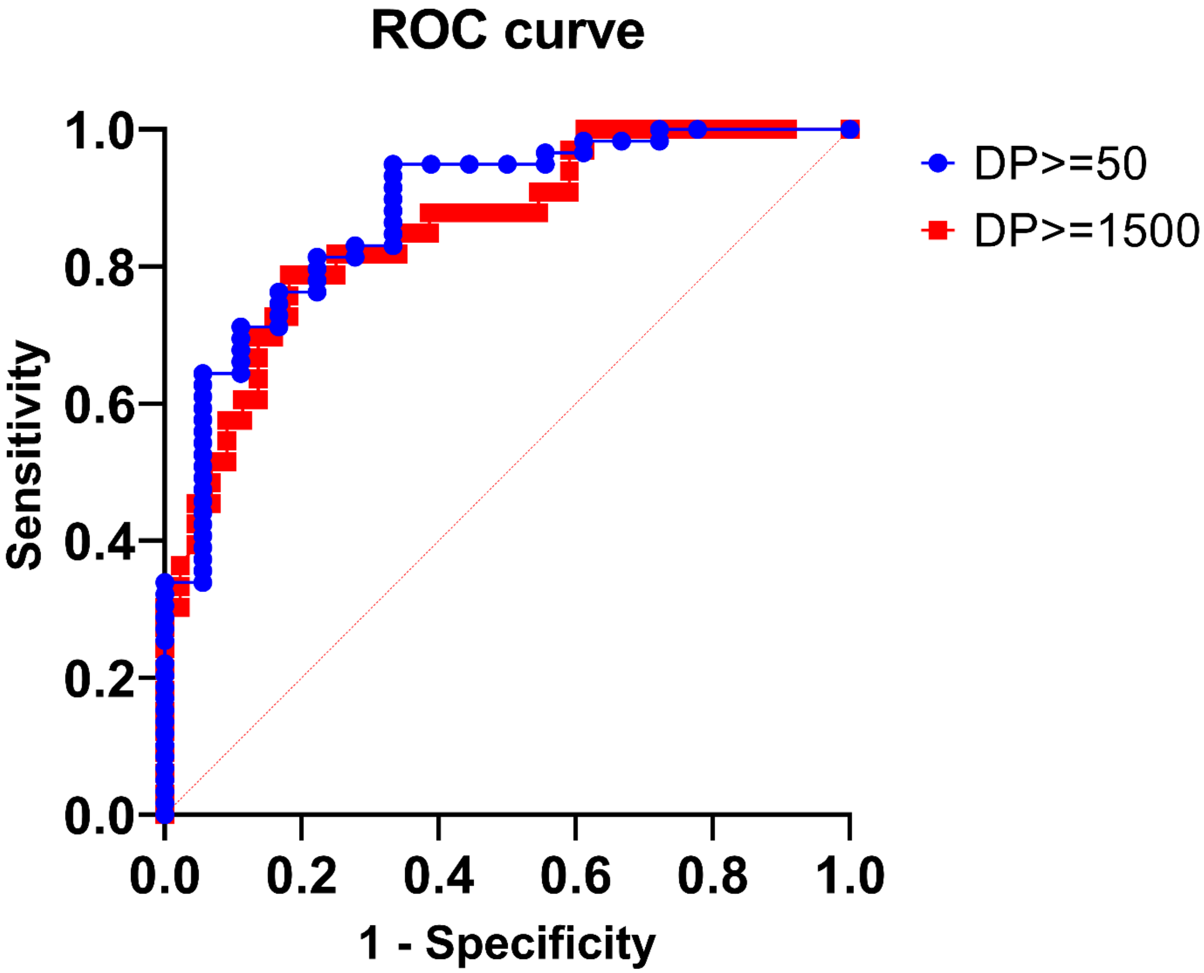


Fig. 17) The ROC curve for evaluating the logistic regression analysis of 77 plasma samples that was used for determining the LOD in HIV ONT workflow. The Beta and the AUC values were listed on Supplementary Table 19a. The diagonal red dash line was the reference line if there was a random relationship between the adjusted IS6110 Ct value and the DP (depth of coverage) cutoffs.

4-2-6. Time to report and operation cost

For a batch of 12 plasma samples, the first working day was assigned for viral RNA extraction (5 hours) and RT-PCR (3.5 hours). On the second working day, the amplicons were purified and were then taken to library preparation (3 hours), followed by the commencement of nanopore sequencing. After the 48-hour sequencing, it should typically require 2 hours (based on a computer with two 12-core Intel Xeon Silver 4116 processors with 126 GB of RAM running in 10 threads) for downstream bioinformatic analysis and clinical report generation. In short, the time to report was four working days (Fig. 18). It would take around 8 hours for a computer with 32 GB of RAM and a CPU of Intel(R) Xeon(R) E5-2678 v3 or equivalent, with a clock speed of 2.5 GHz or equivalent. The cost per sample was USD 120.93 (assuming 12 samples per sequencing batch) and USD 88.02 (assuming 24 samples per sequencing batch) (Supplementary Table 20).

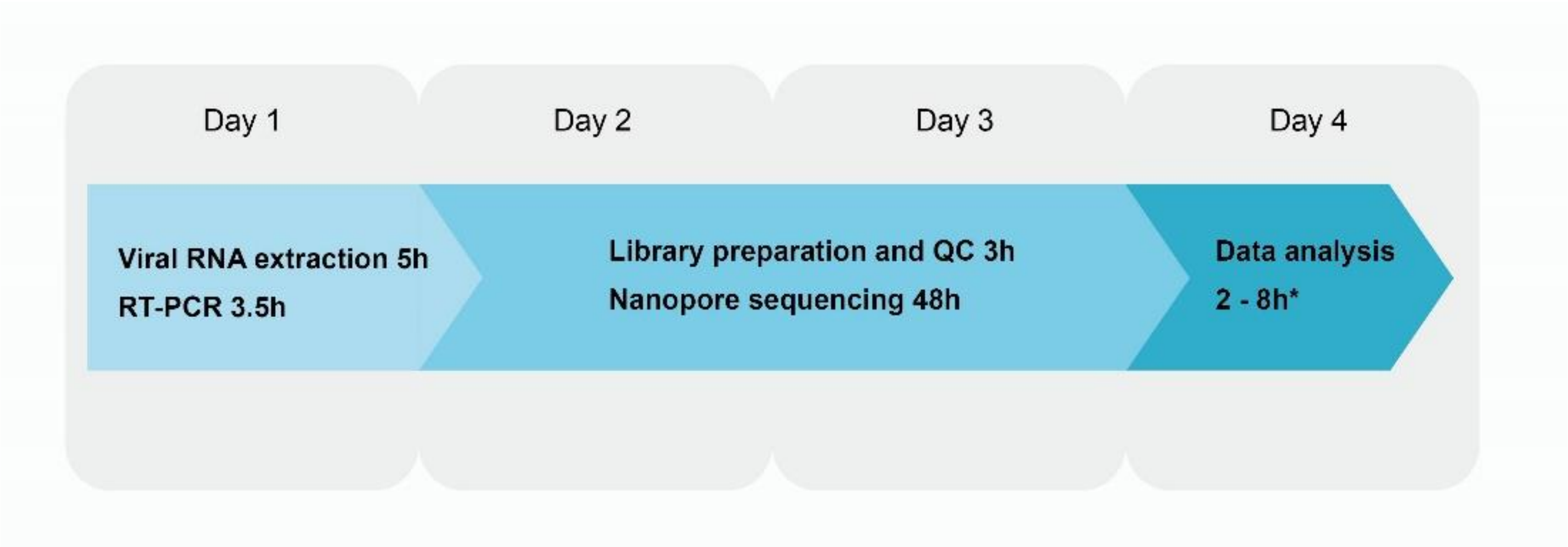


Fig 18) The breakdown of time to report in ONT workflow for direct AVR resistance detection in HIV.

5. Discussion

5-1. Target sequencing workflow for direct AMR detection in MTB in sputum samples

A targeted sequencing panel and a bioinformatic analysis workflow were developed, covering 19 common antibiotic resistance genes associated with well-known first-line, second-line, and anti-XDR-TB antibiotics. Compared with our previously published workflow for clinical isolates, this workflow was developed for direct specimen sequencing, which further reduced the time to report from two weeks to around four workdays by skipping the 14-day MTB culture (Fig. 10). The comprehensive drug resistance information and the saved time would benefit the patient in the choice of antibiotics during medical prescription, especially for those carrying drug-resistant MTB or intolerant to certain antibiotics in complicated cases. It also helps in disease control by minimizing the spread of the disease. Another advantage was the storage space savings from sequencing data. The sequencing data size of amplicon sequencing should be much lower than WGS, assuming the same average DP of covered target regions was reached.

Though the advantages of a direct sequencing workflow can be seen, a few challenges should be taken into account for the transition from the working principle to on-site clinical diagnosis. In this study, index misassignment was reported in both nanopore sequencing and NGS workflows. The major reasons were the index hopping with dual-index adapters in NGS, which was consistent with other studies [123], and the lenient default demultiplexing setting in MinKNOW in ONT. Index misalignment may cause cross contamination between the patient samples, especially when a sample with a low Ct value (high MTB gDNA content) masks another sample with a high Ct value (low MTB gDNA content). The ‘leaked’ sequencing reads can significantly contribute to the depth of coverage and can lead to false variant call. To improve the situation, the use of unique dual index adapters (UDI) in NGS and the stringent settings in Porechop and MinKNOW for demultiplexing in ONT were recommended.

Another challenge is the possible nasal or oral interference with the genotypic results. Unlike pure clinical isolates, clinical specimens are a mixture of human cells, nasal or oral flora, and the MTB itself. In this study, an increasing number of noisy variants was associated with an increasing adjusted *IS6110* PCR Ct value. The power of the decoy strategy successfully removed a portion of reads and several variants that were positively associated with the adjusted *IS6110* PCR Ct value. The change in the number of variants was not necessarily negative, and it could be positive in some cases, meaning that some variants were recovered after filtering. The removed mapped reads in the decoy filter and the taxon filter revealed a diversity of bacterial species in normal oral and nasal flora that shared similar genomic sequences with MTB. This confirmed the presence of nasal or oral interference with the genotypic results. The risk of interference was the masking of the original MTB genotypic results with the variants from the nasal and oral flora, especially in the case of a genomic location with low or zero DP from MTB read sequences. So far, none of the random variants in our database hit the AMR resistance mutations in our drug resistance database, but the hitting chance is a concern.

To overcome this challenge, the decoy strategy consisting of multiple filters was introduced. Other than traditional filtering on readlength, base quality score, mapping score, DP, and more, reads mapped to MTB H37RV were additionally mapped to the reference genome in human (GRCh38) and eHOMD in the decoy filter, followed by the RefSeq in the taxon filter. These filters significantly removed the reads of interference in samples with a high adjusted *IS6110* Ct value and even in negative samples. Of note, there were still mapped reads to the H37RV after filtering in negative samples. Usually, these reads mapped to a few target regions should be screened out with our LOD setting. Of course, there was still room for improvement, such as the regular updating of the databases included in the filtering process. Theoretically, the principle of the decoy strategy can be applied to direct sequencing of the patient samples for drug resistance detection in other infectious diseases.

The DP of a genomic position was an important parameter for valid variant calling. However, the DP may vary because of the variation of the TB-gDNA content in neighboring samples in the same sequencing batch and the rough estimation of actual library concentration. In this data set, the average DP of the target regions within an adjusted *IS6110* PCR Ct value interval (for example, 23 - 25) mostly did not follow the normal distribution. To determine the LOD of this workflow, a binary standard was set based on a sample with known adjusted *IS6110* PCR Ct values that reached the log average DP cutoff in all targeted regions. Also, by considering the DP variance within a region, a conservative higher log average DP cutoff of 1.7 (equivalent to 50) was set. The LODs of the first-line drug panel and the overall panel in ONT were 27.78 and 25.06, respectively.

The LODs of the first-line drug panel and the overall panel in NGS were 23.69 and 22.72, respectively.

Comparing the primers in our previous studies, the modified primer set reduced the enzyme bias, favoring the amplification of shorter target regions. The primer set specifically amplified target regions in both clinical isolates and spike-in samples as well as clinical specimens within the LOD. The variants called out in clinical isolates in the previous studies could also be called out in the primer validation test for both ONT and NGS. The genotypic results of ONT were completely concordant with NGS in all these sample types. This demonstrated that the performance of this primer set in specific amplification and variant calling was robust in clinical isolates and clinical specimens.

Notably, the agreement between ONT and NGS was 100% in the testing cohort (especially for the samples below the cutoff), meaning that the high variant calling accuracy of both sequencing technologies for AMR detection in clinical specimens. From the partial pDST results of 48 clinical samples, the F1 score was only 0.815, which was mainly due to the false-positive results from Sample IDs 19395, 19396, and 21065R. A well-known mutation, C-15T, at the *MabA-inhA* promoter in these three samples did not confer INH resistance, and the mutation Lys43Arg at *rpsL* in Sample IDs 19395 and 19396 did not confer STR resistance. Also, the mutation Leu533Pro at rpoB in Sample ID 21729 did not confer RIF resistance, and the mutation A1401G at rrs in Sample ID 21065R did not confer CAP resistance. On the other hand, there was a false susceptible result of RIF in Sample ID 21065R. One possible reason for the false resistance results was the slow growth rate of the bacteria and the low-level AMR caused the false culture-negative results given a fixed incubation time. Also, there might be unknown mutations and mechanisms outside of these regions reversing the resistance or susceptibility status. Partial pDST results were available for further concordance studies for these samples, but the high agreement between the genotypic results from ONT and NGS proved the working principle for clinical specimens. With the success of the targeted sequencing workflow, the time to report for clinical isolates can be reduced from more than two weeks to a few working days. This quickly provides the clinicians with the AMR profiles for the decision of an anti-TB regimen for the patients and avoids unnecessary waste of time on inappropriate use of antibiotics. Also, it favors resource and sample management by reducing the accumulation of culture wares during the long incubation period.

5-2. Targeted sequencing workflow for direct ARV resistance detection in HIV in plasma samples

Benefiting from the launch of long-sequencing technology and high-fidelity polymerase for long cDNA amplification, the ONT workflow was able to provide a detailed ARV resistance profile in each quasispecies of a sample. This workflow was proven to demonstrate high agreement for high VAF amino acid mutations with Sanger sequencing. Also, it showed the association of the amino acid mutations with different quasispecies after clustering.

Variant calling in both Sanger sequencing and NGS can provide genotype information for a sample with mixed quasispecies, but it cannot provide the linkage of the amino acid mutations to the quasispecies. The first advantage of long sequencing and clustering is the ability to associate amino acid mutations with different quasispecies. In the above highlighted examples such as KB0270 and KB2987, and other samples such as KB1895 and KB2974 (Supplementary Figures S1 and S2), AVR resistance-associated mutations were exclusively found in some quasispecies in a sample. The KB2987 was even a special example of one quasispecies in low abundance carrying an intermediate-high level of resistance to five NNRTIs, while other quasispecies in high abundance carried potential low level NNRTI resistance at that moment. Another special example KB2974 contained a mixture of quasispecies in different subtypes (Subtype B and Subtype CRF01_AE). An accessory DRM V179D in RT was found in only Subtype B quasispecies (KB2974_1, KB2974_3, KB2974_4, KB2974_5, and KB2974_6), while an accessory DRM S68G in RT was found in only Subtype CRF01_AE quasispecies (KB2974_2 and KB2974_8). To the best of my knowledge, coinfection with different HIV subtypes with different resistance profiles is rare, but this special example suggested it could happen. However, only Subtype CRF01_AE was reported in Sanger sequencing. This implied the potential for progression to ARV resistance, possibly caused by the co-infection of two HIV strains that individually carried different resistance profiles. This case requires clinical attention to confirm the development of resistance. This also implied the long-read sequencing technology together with the hierarchical clustering strategy could assist to reveal the association of the detailed drug resistance profile to different quasispecies levels that might be missed with Sanger sequencing.

Theoretically, the association of ARV resistance to the quasispecies in a sample can display the detailed ARV resistance profile, including how the ARV resistance was distributed between different quasispecies (Fig. 19). In a scenario where more than one ARV is found in a sample, with only the VAF as the indicator for determining the ARV resistance, the detailed drug resistance profile cannot be identified as there may be different combinations of DRMs distributed in quasispecies. In the case of a quasispecies carrying all the DRMs, the regimen covering those ARVs will no longer suppress the viral load and cause the virological rebound. However, in the case of the resistance to those ARVs, which are separately distributed to different quasispecies, the prescription of those ARVs may still be able to suppress the HIV activities in the life cycle, and so the patient will not miss the opportunity to receive this regimen. As a result, the power of long-read sequencing and hierarchical clustering grants more genetic evidence for highly active antiretroviral therapy (HAART) treatment.

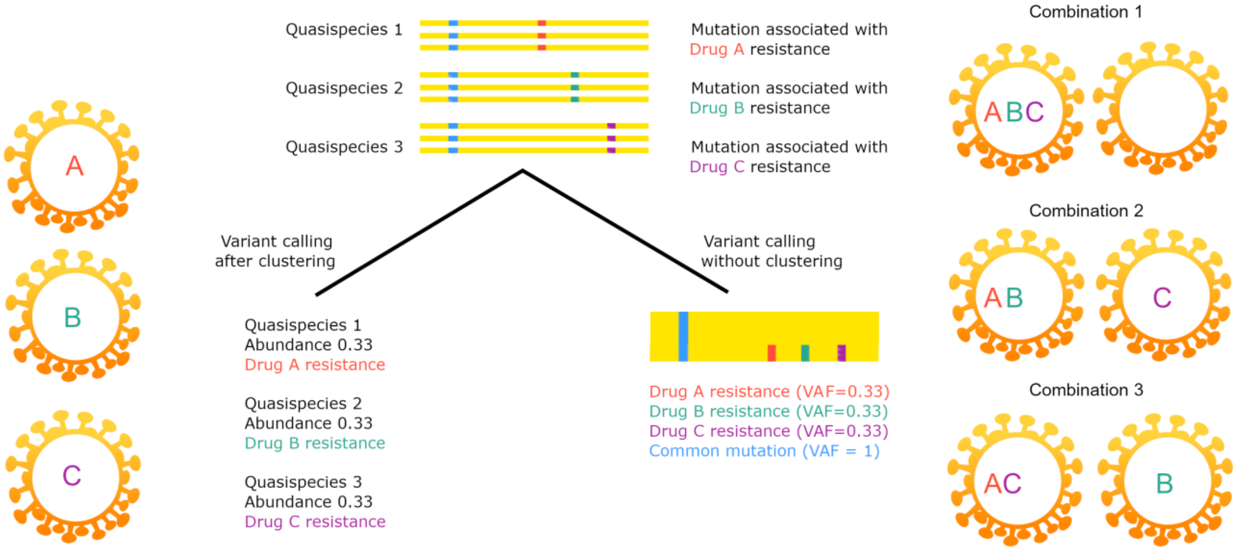
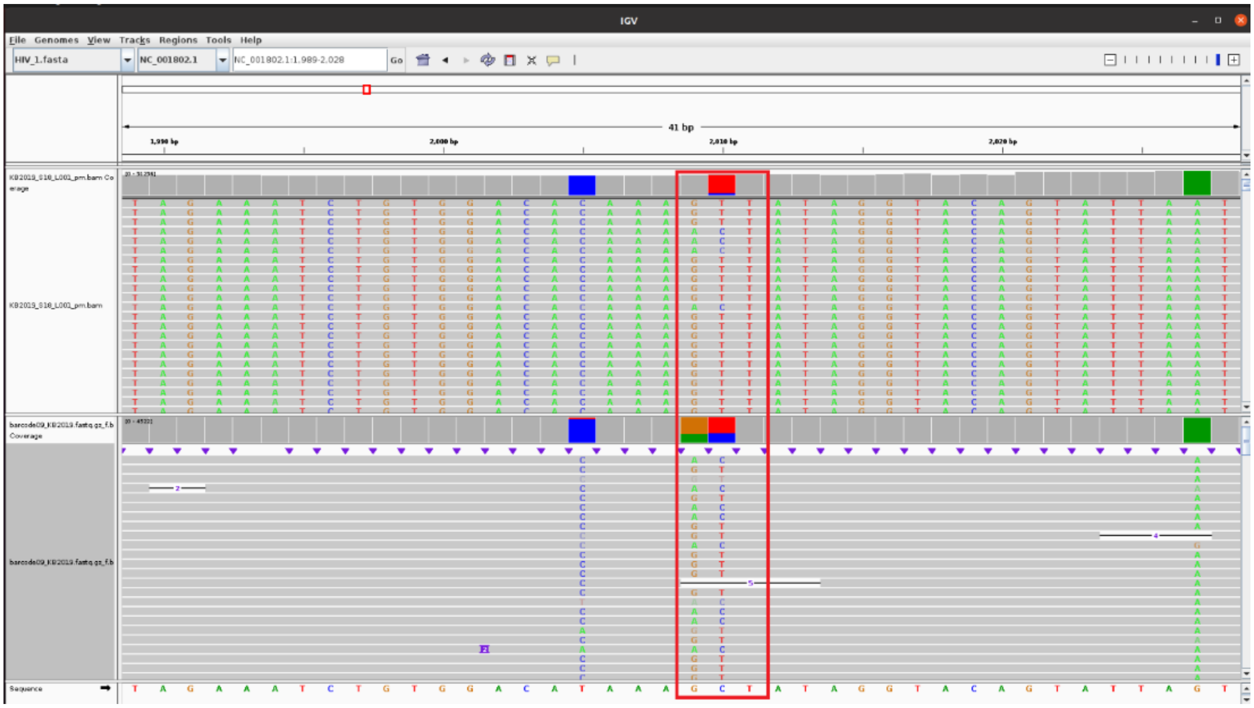


Fig. 19) The explanation of how the linkage of ARV resistance to different HIV quasispecies revealed the clearer ARV resistance profile for medical regimen design. With long-read sequencing and hierarchical clustering, detailed drug resistance profile in quasispecies could be revealed (left). Without hierarchical clustering, different combinations of the linkage of DRMs to different quasispecies were possible.

The second benefit of using hierarchical clustering prior to variant calling is to avoid the misinterpreted variants in a genetic code containing two or three mutations at mixed alleles. Hierarchical clustering can group the reads with the unique genomic mutation patterns before the final variant calls in different quasispecies. For example, in Sanger sequencing, mutation A71I in PR was interpreted in sample KB2019 but the genetic codes ACT (encoding theronine T) and GTT (encoding valine V) were interpreted in both ONT and NGS instead (Fig. 20a). The analysis in Sanger sequencing could not originate the genomic mutations from different sequencing reads. Another example of mutation V111M in RT was found in sample KB2987 (Fig. 20b), but only ATA (encoding isoleucine I) and GTG (encoding valine V) were reported in ONT and NGS.

20a)


IGV
File Genomes View Tracks Regions Tools Help
HIV_1.fasta
NC_001802.1
NC_001802.1:1,989-2,028
Go
41 bp
1,990 bp
2,000 bp
2,010 bp
2,020 bp
KB2019_S18_L001_pm.bam Coverage
KB2019_S18_L001_pm.bam
barcode09_KB2019.fastq.gz_f.b Coverage
barcode09_KB2019.fastq.gz_f.b
Sequence
T A G A A A T C T G T G G A C A T A A A G C T A T A G G T A C A G T A T T A G T

20b)

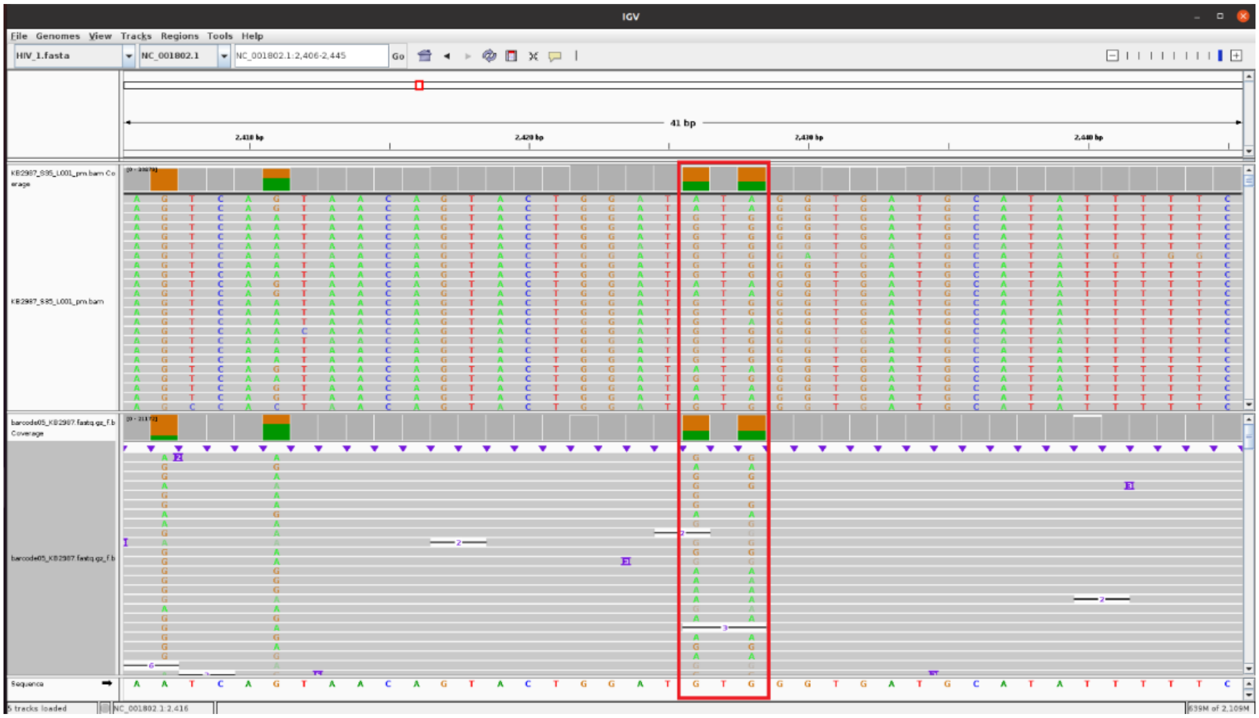

IGV
File Genomes View Tracks Regions Tools Help
HIV_1.fasta
NC_001802.1
NC_001802.1:2,406-2,445
Go
41 bp
2,410 bp
2,420 bp
2,430 bp
2,440 bp
KB2987_S85_L001_pm.bam Coverage
KB2987_S85_L001_pm.bam
barcode05_KB2987.fastq.gz_f.b Coverage
barcode05_KB2987.fastq.gz_f.b
Sequence
A A T C A G T A A C A G T A C T G G A T G T G G T G A T G C A T A T T T T T C
5 tracks loaded
NC_001802.1:2,416

Fig. 20a) Two examples a) KB2019 and b) KB2987) demonstrated how the amino acid mutations were misinterpreted in Sanger, comparing with the genetic codes found in ONT and NGS (in red square).

The third benefit of hierarchical clustering is subtype identification in mixed quasispecies. For example, after clustering, sample KB2974 was found to contain a mixture of subtype B and subtype CRF01_AE, whereas only subtype CRF01_AE was reported in Sanger. Another example, Sample KB2980, was a mixture of subtypes CRF07_BC and CRF01_AE, but only CRF07_BC was identified in Sanger. In another example, Sample KB2998, two subtypes were detected: subtype CRF07_BC and subtype B+C, which were respectively identified in the top and third ranks in Sanger. This means the consensus generated from Sanger sequencing may not truly reveal all the subtypes in a sample.

Hierarchical clustering can be resources-intensive, to lower the computation burden, the clustering strategy employed in this study was based on multiple rounds of variant selection with near VAF peaks rather than a distance matrix between the sequences. With the recommended computer specifications (two 12-core Intel Xeon Silver 4116 processors with 126 GB of RAM), the analysis time is around 2 hours. The analysis time is around 8 hours if a lower computer specification (32 GB of RAM and an Intel(R) Xeon(R) CPU E5-2678 v3 or equivalent) is used. This design favors the setup of this workflow in clinical centers that require short analysis times and low computer specifications.

In the study of abundance prediction in gradient series and an *in-silico* simulation data set, a direct linear relationship was demonstrated between the predicted abundance from the ONT workflow and the corresponding true abundance, though they were generally slightly lower than the true abundance. This meant the ONT workflow was able to detect the true minor variants in the samples. However, from the ROC analysis of 59 plasma samples, the false positive rate was higher than 0.2 when the overall VAF was below 0.35, whereas the cutoff of 0.4 could balance the high sensitivity (~0.91) and low false positive rate (~0.17).

By applying this cutoff to ARV resistance detection, the F1 score was increased from 0.918 to 0.96, as three out of four false mutations with an overall VAF below this cutoff were filtered out. Of note, two mutations (K70R in RT in KB2016 and N348I in RT in KB2971) with overall VAF above this cutoff were discordant with Sanger. The mutation N348I in KB2971 was inconclusive as it could not be validated without an available NGS result, but the mutation K70R in KB2016 was considered false because it was discordant with NGS. The reason was still unclear, but one possible reason was the low viral load of these two samples (104.79 copies/ uL and 98.79 copies/ uL in KB2016 and KB2971, respectively), which possibly led to occasional amplification of only one of the quasispecies and dominated the overall VAF. On the other hand, there were true mutations in the marginal overall VAF (<0.4); these mutations could be reported as possibly true and could be followed.

There were limitations in this study. Firstly, no longitudinal samples were available for the same patient. It was hard to trace the temporal change in the abundance of the ARV resistance quasispecies. Secondly, though NGS was employed to validate the clustering performance, VF variation might be caused by the random fragmentation of defective genomes [124]. Unlike the

ONT workflow, these short amplicons could not be eliminated with the filtering parameters (such as long read length) in the downstream bioinformatic analysis. Thirdly, the clustering performance was validated with the gradient series and *in-silico* simulation data set, but it can also be validated with other long-sequencing technologies such as Pacific Biosciences (PacBio). The limited read length in Sanger sequencing and NGS could not be the reference for comparison in this case. Lastly, the F1 score based on only 22 ARV-resistant samples might be vulnerable to any discordance; more samples would be included for re-evaluating the overall ARV resistance detection in the future.

5-3. Lessons from the target sequencing workflow development

5-3-1. The adoption of sequencing technologies to drug resistance detection in clinical laboratories

Unlike in basic research, the daily sample size for clinical tests can be considerably larger as the tests are used to serve hundreds or even thousands of patients. Time and resource management are always the top priorities. Practically, there are a few conditions to be considered for the adoption: 1) The number of samples included in a sequencing batch (usually related to the required coverage or DP for downstream analysis) reflects the cost per sample. The lower cost per sample benefits the usage of the sequencing for clinical applications. Furthermore, the high flexibility of the sample number in a sequencing batch reduces the time required to begin a sequencing run. 2) The data generated by sequencing per sample can be a burden for data storage. Compact data size per sample no doubt favors the long-term storage of accumulated numbers of samples. 3) A shorter time to report allows for quick decisions about the regimens and favors resource management. 4) Simple handling steps usually minimize human error in sequencing operations.

5-3-2. Direct whole genome sequencing versus target sequencing

Direct whole genome sequencing (WGS) is one approach for drug resistance detection of MTB in sputum samples. With the aid of rapid library preparation, this approach can reduce the hand-on time. However, with the high content of human genomic DNA (>80%, with a mapping quality score >=50) (Supplementary Table 21) and less than 1% for MTB gDNA, a larger sequencing capacity is required for providing sufficient coverage for subsequent analysis, with the drawback of increased reagent cost. In ONT, a new sequencing model called adaptive sequencing can help to retain the MTB sequencing reads once a portion of the reads has been successfully mapped to the reference genome and eject the non-TB ones. The time to reach sufficient coverage is still under evaluation, as the MTB gDNA contents may highly vary between the sputum samples. An alternative is human gDNA depletion during gDNA extraction. Though around 80% of human gDNA can be removed prior to library preparation [125], with the presence of nasal and oral flora, the extent of improvement in MTB sequencing yield is still unknown. Also, not all the genomic regions are associated with AMR. The coverage of the whole genome no doubt occupies a large portion of sequencing capacity, unless the genomic information is necessary for further exploration of underlying mechanisms or phylogenetics for public health control.

Target enrichment with amplicon sequencing is still a good option to increase the sequencing data yield of interest. Other than allowing a high depth of coverage for variant calling analysis, the smaller data size per sample supports a larger number of samples per sequencing batch, which lowers the data storage and cost per sample.

5-3-3. Contamination - the potential source of false results

Avoiding contamination is always important in clinical molecular biology, as nucleic acid amplification and sequencing are highly sensitive to trace amounts of nucleic acid, including those from contamination. In good clinical practice, molecular biology reagents should be prepared separately from the site where the clinical sample was inputted, and all the processes should be carried out in a clean environment. More contamination sources were discovered in this study, which should be prioritized in research and development for direct sequencing of clinical samples. In a clinical sample containing background nasal or oral flora, sequencing reads from non-MTB species may interfere with variant calling results, resulting in false susceptibility or resistance results. The presence of interference is related to a decrease in MTB content in the samples. Another source of contamination confirmed in this study is index misassignment, meaning that the sequencing reads of one sample leak to another sample in the same sequencing batch. The choice of the traditional dual index in NGS and the lenient demultiplexing setting in ONT can cause this problem. Interestingly, false results may be observed when a sample with high MTB content contaminates another sample with low MTB content.

Several studies done by other research teams worldwide successfully prove the working principle of direct sequencing of respiratory samples for drug resistance detection. To our knowledge, they rarely highlight the above-mentioned potential problems. Other than exploring the sequencing power in this clinical application, a guideline for the research and development of the sequencing panel or the standard operating procedure is necessary for avoiding the contamination problem.

5-3-4. The database – the key element in sequencing workflow development

Databases containing drug resistance-associated amino acids or genomic markers are usually used for checking the presence of those mutations in the samples. Many well-known key markers are already involved in databases for MTB and HIV. These markers are validated in basic research worldwide. However, phenotypic resistance involves complicated cellular pathway networks, and even mutations compensate for the loss of fitness caused by the key drug resistance-associated mutations in the same gene. More proteins in MTB are being discovered that may interact with AMK, revealing more unknown mechanisms of resistance to this antibiotic [126]. Even some key mutations do not always indicate AMR; for example, A1401G in the *rrs* gene does not always confer CAP resistance in MTB [127]. Also, the workflows were designed for detecting the AMR-associated genomic markers, but AMR caused by overexpression of some genes is undetectable. For example, overexpression of gene Rv2170 may lead to INH resistance in wildtype MTB by acetylation and hydrolysis of INH into isonicotinic acid and acetylhydrazine [128]. Slow growth status may trigger downregulation of the expression of *katG* [129]. This implies more studies are required for determining the causative relationship between the genetic markers and the associated AMR, especially given the discordance between the genotypic results and the phenotypic results.

5-3-5. The room for improvement in the workflow development

In the development of target sequencing workflow specifically for *MTB*, the DNA contents in the sputum samples were quantified with quantitative PCR targeting IS6110 insertion elements and AFB Smear results. As GeneExpert was also one common clinical test for AMR detection, expansion of the testing cohort covering the samples with GeneExpert MTB assay results may provide the users more references if the samples meet the LOD of this workflow. On the other hand, with the ongoing exploration of AMR genetic markers, regular updating of the database, such as mutations in gene panD, to improve diagnostic accuracy is necessary. Finally, the agreement between the genotypic results and the phenotypic results could not be completely determined as only partial pDST results were available at this moment. Discordance with pDST results would help reveal more novel underlying mechanisms of resistance.

In the development of target sequencing workflow for HIV, retrospective longitudinal studies can be used for tracking the change in ARV resistance profile with a patient group. If a new AVR resistance-associated mutation is detected in a sample, it may become more significant over time unless there is a change in the regimen.

6. Conclusion

Two target sequencing workflows for direct AMR detection in MTB in sputum and direct ARV resistance detection in HIV in plasma were successfully developed. The 100% agreement between ONT and NGS in AMR detection in MTB and the high F1 score of 0.96 with a threshold from ROC analysis in AVR resistance detection in HIV proved the working principle of these workflows. Other than the high diagnostic performance, this study demonstrated how long sequencing could be applied for associating the amino acid mutations with different HIV quasispecies in a sample that could provide a clearer ARV resistance profile. The success of the development implies the time to report for these two workflows can be only a few working days, which allows for a quick decision on the appropriate regimen for the patients and favors resource and sample management. However, from the journey in the workflow development, other than the bright side of the sequencing technologies, the challenges for adoption of these technologies for direct detection of clinical specimens were revealed, including index misassignment and nasal/oral flora interference in sputum that could lead to false results, together with the suggestion in choosing an index set and the employment of the decoy strategy, respectively, for minimizing the effects of these problems. The findings in this study can be a good reference for future direct sequencing developments in infectious diseases.

7. References

1. Lao, H.Y., et al., *The Clinical Utility of Two High-Throughput 16S rRNA Gene Sequencing Workflows for Taxonomic Assignment of Unidentifiable Bacterial Pathogens in Matrix-Assisted Laser Desorption Ionization-Time of Flight Mass Spectrometry.* J Clin Microbiol, 2022. **60**(1): p. e0176921.

2. Gu, W., et al., *Rapid pathogen detection by metagenomic next-generation sequencing of infected body fluids.* Nat Med, 2021. **27**(1): p. 115-124.

3. WHO, *Global Tuberculosis Report 2021*. 2021.

4. IGO, W.H.O.C.B.-N.-S., *WHO Consolidated Guidelines on Tuberculosis, Module 4: Treatment - Drug-Resistant Tuberculosis Treatment* 2020.

5. Conradie, F., et al., *Treatment of Highly Drug-Resistant Pulmonary Tuberculosis.* N Engl J Med, 2020. **382**(10): p. 893-902.

6. Jankute, M., et al., *Assembly of the Mycobacterial Cell Wall.* Annu Rev Microbiol, 2015. **69**: p. 405-23.

7. Torres, J.N., et al., *Novel katG mutations causing isoniazid resistance in clinical M. tuberculosis isolates.* Emerg Microbes Infect, 2015. **4**(7): p. e42.

8. Yu, S., et al., *Reduced affinity for Isoniazid in the S315T mutant of Mycobacterium tuberculosis KatG is a key factor in antibiotic resistance.* J Biol Chem, 2003. **278**(17): p. 14769-75.

9. Unissa, A.N., et al., *Overview on mechanisms of isoniazid action and resistance in Mycobacterium tuberculosis.* Infect Genet Evol, 2016. **45**: p. 474-492.

10. Lempens, P., et al., *Isoniazid resistance levels of Mycobacterium tuberculosis can largely be predicted by high-confidence resistance-conferring mutations.* Sci Rep, 2018. **8**(1): p. 3246.

11. Ando, H., et al., *Downregulation of katG expression is associated with isoniazid resistance in Mycobacterium tuberculosis.* Mol Microbiol, 2011. **79**(6): p. 1615-28.

12. Campbell, E.A., et al., *Structural mechanism for rifampicin inhibition of bacterial rna polymerase.* Cell, 2001. **104**(6): p. 901-12.

13. Lin, Y.H., et al., *Resistance profiles and rpoB gene mutations of Mycobacterium tuberculosis isolates in Taiwan.* J Microbiol Immunol Infect, 2013. **46**(4): p. 266-70.

14. Zeng, M.C., Q.J. Jia, and L.M. Tang, *rpoB gene mutations in rifampin-resistant Mycobacterium tuberculosis isolates from rural areas of Zhejiang, China.* J Int Med Res, 2021. **49**(3): p. 300060521997596.

15. Rando-Segura, A., et al., *Molecular characterization of rpoB gene mutations in isolates from tuberculosis patients in Cubal, Republic of Angola.* BMC Infect Dis, 2021. **21**(1): p. 1056.

16. Shubladze, N., N. Tadumadze, and N. Bablishvili, *Molecular patterns of multidrug resistance of Mycobacterium tuberculosis in Georgia.* Int J Mycobacteriol, 2013. **2**(2): p. 73-78.

17. Rifat, D., et al., *In vitro and in vivo fitness costs associated with Mycobacterium tuberculosis RpoB mutation H526D.* Future Microbiol, 2017. **12**(9): p. 753-765.

18. Brandis, G. and D. Hughes, *Genetic characterization of compensatory evolution in strains carrying rpoB Ser531Leu, the rifampicin resistance mutation most frequently found in clinical isolates.* J Antimicrob Chemother, 2013. **68**(11): p. 2493-7.

19. Xu, Z., et al., *Transcriptional Approach for Decoding the Mechanism of rpoC Compensatory Mutations for the Fitness Cost in Rifampicin-Resistant Mycobacterium tuberculosis.* Front Microbiol, 2018. **9**: p. 2895.

20. Zhang, L., et al., *Structures of cell wall arabinosyltransferases with the anti-tuberculosis drug ethambutol.* Science, 2020. **368**(6496): p. 1211-1219.

21. Zhu, C., et al., *Molecular mechanism of the synergistic activity of ethambutol and isoniazid against Mycobacterium tuberculosis.* J Biol Chem, 2018. **293**(43): p. 16741-16750.

22. Starks, A.M., et al., *Mutations at embB codon 306 are an important molecular indicator of ethambutol resistance in Mycobacterium tuberculosis.* Antimicrob Agents Chemother, 2009. **53**(3): p. 1061-6.

23. Safi, H., et al., *Allelic exchange and mutant selection demonstrate that common clinical embCAB gene mutations only modestly increase resistance to ethambutol in Mycobacterium tuberculosis.* Antimicrob Agents Chemother, 2010. **54**(1): p. 103-8.

24. Park, Y.K., et al., *Correlation of the phenotypic ethambutol susceptibility of Mycobacterium tuberculosis with embB gene mutations in Korea.* J Med Microbiol, 2012. **61**(Pt 4): p. 529-534.

25. Bwalya, P., et al., *Characterization of embB mutations involved in ethambutol resistance in multi-drug resistant Mycobacterium tuberculosis isolates in Zambia.* Tuberculosis (Edinb), 2022. **133**: p. 102184.

26. Li, W., *Bringing Bioactive Compounds into Membranes: The UbiA Superfamily of Intramembrane Aromatic Prenyltransferases.* Trends Biochem Sci, 2016. **41**(4): p. 356-370.

27. Safi, H., et al., *Evolution of high-level ethambutol-resistant tuberculosis through interacting mutations in decaprenylphosphoryl-beta-D-arabinose biosynthetic and utilization pathway genes.* Nat Genet, 2013. **45**(10): p. 1190-7.

28. Gopal, P., et al., *Pharmacological and Molecular Mechanisms Behind the Sterilizing Activity of Pyrazinamide.* Trends Pharmacol Sci, 2019. **40**(12): p. 930-940.

29. Rajendran, V. and R. Sethumadhavan, *Drug resistance mechanism of PncA in Mycobacterium tuberculosis.* J Biomol Struct Dyn, 2014. **32**(2): p. 209-21.

30. Petrella, S., et al., *Crystal structure of the pyrazinamidase of Mycobacterium tuberculosis: insights into natural and acquired resistance to pyrazinamide.* PLoS One, 2011. **6**(1): p. e15785.

31. Demirci, H., et al., *A structural basis for streptomycin-induced misreading of the genetic code.* Nat Commun, 2013. **4**: p. 1355.

32. Carter, A.P., et al., *Functional insights from the structure of the 30S ribosomal subunit and its interactions with antibiotics.* Nature, 2000. **407**(6802): p. 340-8.

33. Wang, Y., et al., *The roles of rpsL, rrs, and gidB mutations in predicting streptomycin-resistant drugs used on clinical Mycobacterium tuberculosis isolates from Hebei Province, China.* Int J Clin Exp Pathol, 2019. **12**(7): p. 2713-2721.

34. Khosravi, A.D., et al., *Frequency of rrs and rpsL mutations in streptomycin-resistant Mycobacterium tuberculosis isolates from Iranian patients.* J Glob Antimicrob Resist, 2017. **9**: p. 51-56.

35. Vocat, A., et al., *Bioluminescence for assessing drug potency against nonreplicating Mycobacterium tuberculosis.* Antimicrob Agents Chemother, 2015. **59**(7): p. 4012-9.

36. Sarathy, J.P., G. Gruber, and T. Dick, *Re-Understanding the Mechanisms of Action of the Anti-Mycobacterial Drug Bedaquiline.* Antibiotics (Basel), 2019. **8**(4).

37. Preiss, L., et al., *Structure of the mycobacterial ATP synthase Fo rotor ring in complex with the anti-TB drug bedaquiline.* Sci Adv, 2015. **1**(4): p. e1500106.

38. Degiacomi, G., et al., *In vitro Study of Bedaquiline Resistance in Mycobacterium tuberculosis Multi-Drug Resistant Clinical Isolates.* Front Microbiol, 2020. **11**: p. 559469.

39. Omar, S.V., et al., *Bedaquiline-Resistant Tuberculosis Associated with Rv0678 Mutations.* N Engl J Med, 2022. **386**(1): p. 93-94.

40. He, W., et al., *Prevalence of Mycobacterium tuberculosis resistant to bedaquiline and delamanid in China.* J Glob Antimicrob Resist, 2021. **26**: p. 241-248.

41. Villellas, C., et al., *Unexpected high prevalence of resistance-associated Rv0678 variants in MDR-TB patients without documented prior use of clofazimine or bedaquiline.* J Antimicrob Chemother, 2017. **72**(3): p. 684-690.

42. Hooper, D.C. and G.A. Jacoby, *Topoisomerase Inhibitors: Fluoroquinolone Mechanisms of Action and Resistance.* Cold Spring Harb Perspect Med, 2016. **6**(9).

43. Aldred, K.J., et al., *Fluoroquinolone interactions with Mycobacterium tuberculosis gyrase: Enhancing drug activity against wild-type and resistant gyrase.* Proc Natl Acad Sci U S A, 2016. **113**(7): p. E839-46.

44. Pantel, A., et al., *Extending the definition of the GyrB quinolone resistance-determining region in Mycobacterium tuberculosis DNA gyrase for assessing fluoroquinolone resistance in M. tuberculosis.* Antimicrob Agents Chemother, 2012. **56**(4): p. 1990-6.

45. Von Groll, A., et al., *Fluoroquinolone resistance in Mycobacterium tuberculosis and mutations in gyrA and gyrB.* Antimicrob Agents Chemother, 2009. **53**(10): p. 4498-500.

46. Wang, Z., et al., *Characterization of Fluoroquinolone-Resistant and Multidrug-Resistant Mycobacterium tuberculosis Isolates Using Whole-Genome Sequencing in Tianjin, China.* Infect Drug Resist, 2022. **15**: p. 1793-1803.

47. Mujuni, D., et al., *Molecular characterisation of second-line drug resistance among drug resistant tuberculosis patients tested in Uganda: a two and a half-year's review.* BMC Infect Dis, 2022. **22**(1): p. 363.

48. Lin, Y., et al., *The antituberculosis antibiotic capreomycin inhibits protein synthesis by disrupting interaction between ribosomal proteins L12 and L10.* Antimicrob Agents Chemother, 2014. **58**(4): p. 2038-44.

49. Vianna, J.F., et al., *Binding energies of the drugs capreomycin and streptomycin in complex with tuberculosis bacterial ribosome subunits.* Phys Chem Chem Phys, 2019. **21**(35): p. 19192-19200.

50. Laughlin, Z.T. and G.L. Conn, *Tuberactinomycin antibiotics: Biosynthesis, anti-mycobacterial action, and mechanisms of resistance.* Front Microbiol, 2022. **13**: p. 961921.

51. Maus, C.E., B.B. Plikaytis, and T.M. Shinnick, *Molecular analysis of cross-resistance to capreomycin, kanamycin, amikacin, and viomycin in Mycobacterium tuberculosis.* Antimicrob Agents Chemother, 2005. **49**(8): p. 3192-7.

52. Maus, C.E., B.B. Plikaytis, and T.M. Shinnick, *Mutation of tlyA confers capreomycin resistance in Mycobacterium tuberculosis.* Antimicrob Agents Chemother, 2005. **49**(2): p. 571-7.

53. Sowajassatakul, A., et al., *Molecular characterization of amikacin, kanamycin and capreomycin resistance in M/XDR-TB strains isolated in Thailand.* BMC Microbiol, 2014. **14**: p. 165.

54. Campbell, P.J., et al., *Molecular detection of mutations associated with first- and second-line drug resistance compared with conventional drug susceptibility testing of Mycobacterium tuberculosis.* Antimicrob Agents Chemother, 2011. **55**(5): p. 2032-41.

55. Georghiou, S.B., et al., *Frequency and Distribution of Tuberculosis Resistance-Associated Mutations between Mumbai, Moldova, and Eastern Cape.* Antimicrob Agents Chemother, 2016. **60**(7): p. 3994-4004.

56. Rana, V., et al., *Molecular Epidemiology and Polymorphism Analysis in Drug-Resistant Genes in M. tuberculosis Clinical Isolates from Western and Northern India.* Infect Drug Resist, 2022. **15**: p. 1717-1732.

57. Ramirez, M.S. and M.E. Tolmasky, *Amikacin: Uses, Resistance, and Prospects for Inhibition.* Molecules, 2017. **22**(12).

58. Chen, W., et al., *Unusual regioversatility of acetyltransferase Eis, a cause of drug resistance in XDR-TB.* Proc Natl Acad Sci U S A, 2011. **108**(24): p. 9804-8.

59. Zaunbrecher, M.A., et al., *Overexpression of the chromosomally encoded aminoglycoside acetyltransferase eis confers kanamycin resistance in Mycobacterium tuberculosis.* Proc Natl Acad Sci U S A, 2009. **106**(47): p. 20004-9.

60. Reeves, A.Z., et al., *Aminoglycoside cross-resistance in Mycobacterium tuberculosis due to mutations in the 5' untranslated region of whiB7.* Antimicrob Agents Chemother, 2013. **57**(4): p. 1857-65.

61. Islam, M.M., et al., *Prevalence and molecular characterization of amikacin resistance among Mycobacterium tuberculosis clinical isolates from southern China.* J Glob Antimicrob Resist, 2020. **22**: p. 290-295.

62. Long, K.S. and B. Vester, *Resistance to linezolid caused by modifications at its binding site on the ribosome.* Antimicrob Agents Chemother, 2012. **56**(2): p. 603-12.

63. Beckert, P., et al., *rplC T460C identified as a dominant mutation in linezolid-resistant Mycobacterium tuberculosis strains.* Antimicrob Agents Chemother, 2012. **56**(5): p. 2743-5.

64. Makafe, G.G., et al., *Role of the Cys154Arg Substitution in Ribosomal Protein L3 in Oxazolidinone Resistance in Mycobacterium tuberculosis.* Antimicrob Agents Chemother, 2016. **60**(5): p. 3202-6.

65. Zimenkov, D.V., et al., *Examination of bedaquiline- and linezolid-resistant Mycobacterium tuberculosis isolates from the Moscow region.* J Antimicrob Chemother, 2017. **72**(7): p. 1901-1906.

66. Du, J., et al., *Low Rate of Acquired Linezolid Resistance in Multidrug-Resistant Tuberculosis Treated With Bedaquiline-Linezolid Combination.* Front Microbiol, 2021. **12**: p. 655653.

67. Wasserman, S., et al., *Linezolid resistance in patients with drug-resistant TB and treatment failure in South Africa.* J Antimicrob Chemother, 2019. **74**(8): p. 2377-2384.

68. Dreyer, V., et al., *High fluoroquinolone resistance proportions among multidrug-resistant tuberculosis driven by dominant L2 Mycobacterium tuberculosis clones in the Mumbai Metropolitan Region.* Genome Med, 2022. **14**(1): p. 95.

69. Holec, A.D., et al., *Nucleotide Reverse Transcriptase Inhibitors: A Thorough Review, Present Status and Future Perspective as HIV Therapeutics.* Curr HIV Res, 2017. **15**(6): p. 411-421.

70. Diallo, K., M. Gotte, and M.A. Wainberg, *Molecular impact of the M184V mutation in human immunodeficiency virus type 1 reverse transcriptase.* Antimicrob Agents Chemother, 2003. **47**(11): p. 3377-83.

71. Sarafianos, S.G., et al., *Structure and function of HIV-1 reverse transcriptase: molecular mechanisms of polymerization and inhibition.* J Mol Biol, 2009. **385**(3): p. 693-713.

72. Das, K., S.E. Martinez, and E. Arnold, *Structural Insights into HIV Reverse Transcriptase Mutations Q151M and Q151M Complex That Confer Multinucleoside Drug Resistance.* Antimicrob Agents Chemother, 2017. **61**(6).

73. Acosta-Hoyos, A.J. and W.A. Scott, *The Role of Nucleotide Excision by Reverse Transcriptase in HIV Drug Resistance.* Viruses, 2010. **2**(2): p. 372-394.

74. Patel, P.H. and H. Zulfiqar, *Reverse Transcriptase Inhibitors*, in *StatPearls*. 2022: Treasure Island (FL).

75. Cilento, M.E., K.A. Kirby, and S.G. Sarafianos, *Avoiding Drug Resistance in HIV Reverse Transcriptase.* Chem Rev, 2021. **121**(6): p. 3271-3296.

76. Tie, Y., et al., *Atomic resolution crystal structures of HIV-1 protease and mutants V82A and I84V with saquinavir.* Proteins, 2007. **67**(1): p. 232-42.

77. Hayashi, H., et al., *Dimerization of HIV-1 protease occurs through two steps relating to the mechanism of protease dimerization inhibition by darunavir.* Proc Natl Acad Sci U S A, 2014. **111**(33): p. 12234-9.

78. Ghosh, A.K., H.L. Osswald, and G. Prato, *Recent Progress in the Development of HIV-1 Protease Inhibitors for the Treatment of HIV/AIDS.* J Med Chem, 2016. **59**(11): p. 5172-208.

79. Wang, R.G., H.X. Zhang, and Q.C. Zheng, *Revealing the binding and drug resistance mechanism of amprenavir, indinavir, ritonavir, and nelfinavir complexed with HIV-1 protease due to double mutations G48T/L89M by molecular dynamics simulations and free energy analyses.* Phys Chem Chem Phys, 2020. **22**(8): p. 4464-4480.

80. Maertens, G.N., A.N. Engelman, and P. Cherepanov, *Structure and function of retroviral integrase.* Nat Rev Microbiol, 2022. **20**(1): p. 20-34.

81. Jozwik, I.K., D.O. Passos, and D. Lyumkis, *Structural Biology of HIV Integrase Strand Transfer Inhibitors.* Trends Pharmacol Sci, 2020. **41**(9): p. 611-626.

82. Hare, S., et al., *Structural and functional analyses of the second-generation integrase strand transfer inhibitor dolutegravir (S/GSK1349572).* Mol Pharmacol, 2011. **80**(4): p. 565-72.

83. Delelis, O., et al., *Impact of Y143 HIV-1 integrase mutations on resistance to raltegravir in vitro and in vivo.* Antimicrob Agents Chemother, 2010. **54**(1): p. 491-501.

84. Abram, M.E., et al., *Impact of primary elvitegravir resistance-associated mutations in HIV-1 integrase on drug susceptibility and viral replication fitness.* Antimicrob Agents Chemother, 2013. **57**(6): p. 2654-63.

85. Quashie, P.K., et al., *Differential effects of the G118R, H51Y, and E138K resistance substitutions in different subtypes of HIV integrase.* J Virol, 2015. **89**(6): p. 3163-75.

86. Quashie, P.K., et al., *Characterization of the R263K mutation in HIV-1 integrase that confers low-level resistance to the second-generation integrase strand transfer inhibitor dolutegravir.* J Virol, 2012. **86**(5): p. 2696-705.

87. Mukhatayeva, A., et al., *Antiretroviral therapy resistance mutations among HIV infected people in Kazakhstan.* Sci Rep, 2022. **12**(1): p. 17195.

88. Pang, X., et al., *HIV drug resistance and HIV transmission risk factors among newly diagnosed individuals in Southwest China.* BMC Infect Dis, 2021. **21**(1): p. 160.

89. Watera, C., et al., *HIV drug resistance among adults initiating antiretroviral therapy in Uganda.* J Antimicrob Chemother, 2021. **76**(9): p. 2407-2414.

90. McClung, R.P., et al., *Transmitted Drug Resistance Among Human Immunodeficiency Virus (HIV)-1 Diagnoses in the United States, 2014-2018.* Clin Infect Dis, 2022. **74**(6): p. 1055-1062.

91. Northrop, A.J. and L.W. Pomeroy, *Forecasting Prevalence of HIV-1 Integrase Strand Transfer Inhibitor (INSTI) Drug Resistance: A Modeling Study.* J Acquir Immune Defic Syndr, 2020. **83**(1): p. 65-71.

92. Votintseva, A.A., et al., *Same-Day Diagnostic and Surveillance Data for Tuberculosis via Whole-Genome Sequencing of Direct Respiratory Samples.* J Clin Microbiol, 2017. **55**(5): p. 1285-1298.

93. Doyle, R.M., et al., *Direct Whole-Genome Sequencing of Sputum Accurately Identifies Drug-Resistant Mycobacterium tuberculosis Faster than MGIT Culture Sequencing.* J Clin Microbiol, 2018. **56**(8).

94. Doughty, E.L., et al., *Culture-independent detection and characterisation of Mycobacterium tuberculosis and M. africanum in sputum samples using shotgun metagenomics on a benchtop sequencer.* PeerJ, 2014. **2**: p. e585.

95. Tafess, K., et al., *Targeted-Sequencing Workflows for Comprehensive Drug Resistance Profiling of Mycobacterium tuberculosis Cultures Using Two Commercial Sequencing Platforms: Comparison of Analytical and Diagnostic Performance, Turnaround Time, and Cost.* Clin Chem, 2020. **66**(6): p. 809-820.

96. Cabibbe, A.M., et al., *Application of Targeted Next-Generation Sequencing Assay on a Portable Sequencing Platform for Culture-Free Detection of Drug-Resistant Tuberculosis from Clinical Samples.* J Clin Microbiol, 2020. **58**(10).

97. Kambli, P., et al., *Targeted next generation sequencing directly from sputum for comprehensive genetic information on drug resistant Mycobacterium tuberculosis.* Tuberculosis (Edinb), 2021. **127**: p. 102051.

98. Feuerriegel, S., et al., *Rapid genomic first- and second-line drug resistance prediction from clinical Mycobacterium tuberculosis specimens using Deeplex-MycTB.* Eur Respir J, 2021. **57**(1).

99. Leung, K.S., et al., *Clinical utility of target amplicon sequencing test for rapid diagnosis of drug-resistant Mycobacterium tuberculosis from respiratory specimens.* Front Microbiol, 2022. **13**: p. 974428.

100. Biswas, K., et al., *The nasal microbiota in health and disease: variation within and between subjects.* Front Microbiol, 2015. **9**: p. 134.

101. Metzner, K.J., 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**(2): p. 239-47.

102. Hedskog, C., et al., *Dynamics of HIV-1 quasispecies during antiviral treatment dissected using ultra-deep pyrosequencing.* PLoS One, 2010. **5**(7): p. e11345.

103. Kapoor, A., et al., *Multiple independent origins of a protease inhibitor resistance mutation in salvage therapy patients.* Retrovirology, 2008. **5**: p. 7.

104. Kyeyune, F., et al., *Low-Frequency Drug Resistance in HIV-Infected Ugandans on Antiretroviral Treatment Is Associated with Regimen Failure.* Antimicrob Agents Chemother, 2016. **60**(6): p. 3380-97.

105. Boltz, V.F., et al., *Linked dual-class HIV resistance mutations are associated with treatment failure.* JCI Insight, 2019. **4**(19).

106. Arias, A., et al., *Sanger and Next Generation Sequencing Approaches to Evaluate HIV-1 Virus in Blood Compartments.* Int J Environ Res Public Health, 2018. **15**(8).

107. Manyana, S., et al., *HIV-1 Drug Resistance Genotyping in Resource Limited Settings: Current and Future Perspectives in Sequencing Technologies.* Viruses, 2021. **13**(6).

108. WHO/HIVRESNET, *HIV Drug resistance laboratory operational framework*. 2017.

109. Leung, K.S., et al., *Diagnostic evaluation of an in-house developed single-tube, duplex, nested IS6110 real-time PCR assay for rapid pulmonary tuberculosis diagnosis.* Tuberculosis (Edinb), 2018. **112**: p. 120-125.

110. Leung, C.M., et al., *MegaPath: sensitive and rapid pathogen detection using metagenomic NGS data.* BMC Genomics, 2020. **21**(Suppl 6): p. 500.

111. Bushnell, B., J. Rood, and E. Singer, *BBMerge - Accurate paired shotgun read merging via overlap.* PLoS One, 2017. **12**(10): p. e0185056.

112. Li, H. and R. Durbin, *Fast and accurate short read alignment with Burrows-Wheeler transform.* Bioinformatics, 2009. **25**(14): p. 1754-60.

113. Li, D., et al., *MEGAHIT v1.0: A fast and scalable metagenome assembler driven by advanced methodologies and community practices.* Methods, 2016. **102**: p. 3-11.

114. BD, V.d.A.G.O.C., *Genomics in the Cloud: Using Docker, GATK, and WDL in Terra (1st Edition)*. 2020: O'Reilly Media.

115. Lui, W.W., et al., *MegaPath-Nano: Accurate Compositional Analysis and Drug-level Antimicrobial Resistance Detection Software for Oxford Nanopore Long-read Metagenomics.* 2020 Ieee International Conference on Bioinformatics and Biomedicine, 2020: p. 329-336.

116. Li, H., *Minimap2: pairwise alignment for nucleotide sequences.* Bioinformatics, 2018. **34**(18): p. 3094-3100.

117. Leung, A.W., et al., *ECNano: A cost-effective workflow for target enrichment sequencing and accurate variant calling on 4800 clinically significant genes using a single MinION flowcell.* BMC Med Genomics, 2022. **15**(1): p. 43.

118. Shafer, R.W., *Rationale and uses of a public HIV drug-resistance database.* J Infect Dis, 2006. **194 Suppl 1**(Suppl 1): p. S51-8.

119. Dreyer, V., et al., *Detection of low-frequency resistance-mediating SNPs in next-generation sequencing data of Mycobacterium tuberculosis complex strains with binoSNP.* Sci Rep, 2020. **10**(1): p. 7874.

120. Spencer, D.H., et al., *Performance of common analysis methods for detecting low-frequency single nucleotide variants in targeted next-generation sequence data.* J Mol Diagn, 2014. **16**(1): p. 75-88.

121. Van Poelvoorde, L.A.E., et al., *Strategy and Performance Evaluation of Low-Frequency Variant Calling for SARS-CoV-2 Using Targeted Deep Illumina Sequencing.* Front Microbiol, 2021. **12**: p. 747458.

122. Lee, E.R., et al., *Performance comparison of next generation sequencing analysis pipelines for HIV-1 drug resistance testing.* Sci Rep, 2020. **10**(1): p. 1634.

123. MacConaill, L.E., et al., *Unique, dual-indexed sequencing adapters with UMIs effectively eliminate index cross-talk and significantly improve sensitivity of massively parallel sequencing.* BMC Genomics, 2018. **19**(1): p. 30.

124. Kuniholm, J., C. Coote, and A.J. Henderson, *Defective HIV-1 genomes and their potential impact on HIV pathogenesis.* Retrovirology, 2022. **19**(1): p. 13.

125. Hasan, M.R., et al., *Depletion of Human DNA in Spiked Clinical Specimens for Improvement of Sensitivity of Pathogen Detection by Next-Generation Sequencing.* J Clin Microbiol, 2016. **54**(4): p. 919-27.

126. Sharma, D., et al., *Cytosolic Proteome Profiling of Aminoglycosides Resistant Mycobacterium tuberculosis Clinical Isolates Using MALDI-TOF/MS.* Front Microbiol, 2016. **7**: p. 1816.

127. Reeves, A.Z., et al., *Disparities in capreomycin resistance levels associated with the rrs A1401G mutation in clinical isolates of Mycobacterium tuberculosis.* Antimicrob Agents Chemother, 2015. **59**(1): p. 444-9.

128. Arun, K.B., et al., *Acetylation of Isoniazid Is a Novel Mechanism of Isoniazid Resistance in Mycobacterium tuberculosis.* Antimicrob Agents Chemother, 2020. **65**(1).

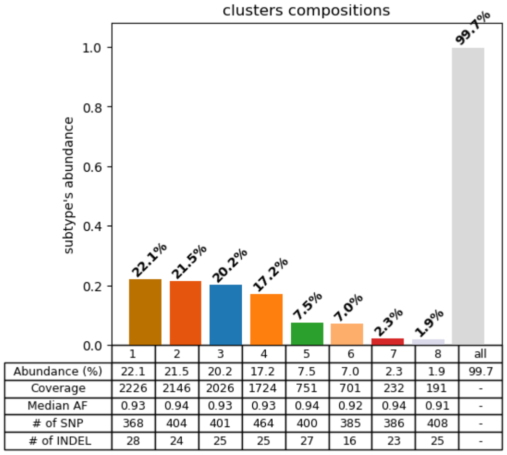
129. Niki, M., et al., *A novel mechanism of growth phase-dependent tolerance to isoniazid in mycobacteria.* J Biol Chem, 2012. **287**(33): p. 27743-52.

Supplementary Figures

S1) The example in KB1895 with its different abundance, 12b) resistance patterns, and 12c) resistance levels in different quasispecies.

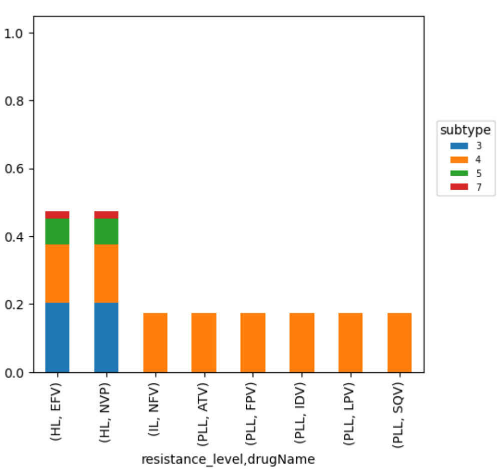
KB1895

S1a)

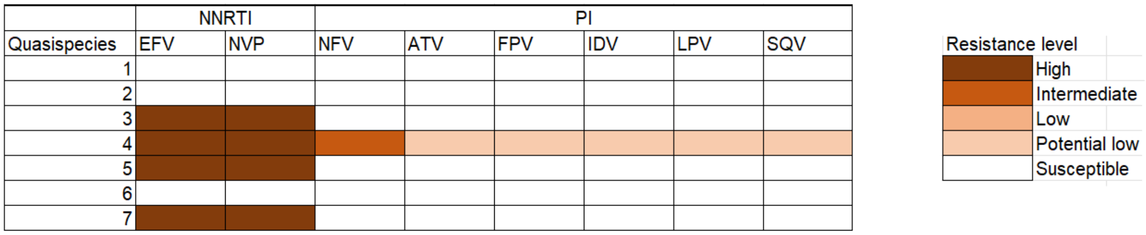


	1	2	3	4	5	6	7	8	all
Abundance (%)	22.1	21.5	20.2	17.2	7.5	7.0	2.3	1.9	99.7
Coverage	2226	2146	2026	1724	751	701	232	191	-
Median AF	0.93	0.94	0.93	0.93	0.94	0.92	0.94	0.91	-
# of SNP	368	404	401	464	400	385	386	408	-
# of INDEL	28	24	25	25	27	16	23	25	-

S1b)



S1c)



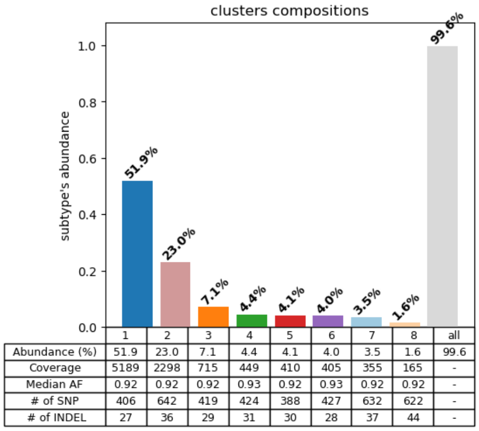
	NNRTI		PI					
Quasispecies	EFV	NVP	NFV	ATV	FPV	IDV	LPV	SQV
1								
2								
3	🟫	🟫						
4	🟫	🟫	🟧	░	░	░	░	░
5	🟫	🟫						
6								
7	🟫	🟫						

Resistance level	
🟫	High
🟧	Intermediate
▨	Low
░	Potential low
□	Susceptible

S2) The example in KB2974 with its different abundance, 12b) resistance patterns, and 12c) resistance levels in different quasispecies.

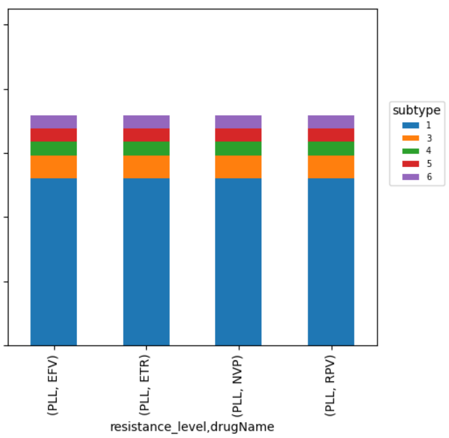
KB2974

S2a)

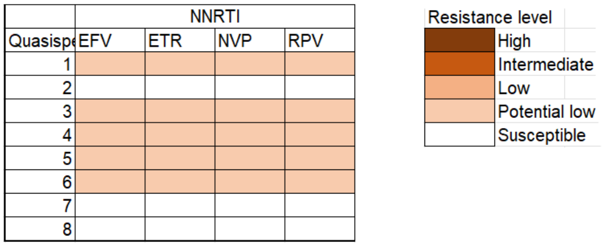


	1	2	3	4	5	6	7	8	all
Abundance (%)	51.9	23.0	7.1	4.4	4.1	4.0	3.5	1.6	99.6
Coverage	5189	2298	715	449	410	405	355	165	-
Median AF	0.92	0.92	0.92	0.93	0.92	0.93	0.92	0.92	-
# of SNP	406	642	419	424	388	427	632	622	-
# of INDEL	27	36	29	31	30	28	37	44	-

S2b)



S2c)



	NNRTI			
Quasispe	EFV	ETR	NVP	RPV
1	░	░	░	░
2				
3	░	░	░	░
4	░	░	░	░
5	░	░	░	░
6	░	░	░	░
7				
8				

Resistance level	
█	High
▓	Intermediate
▒	Low
░	Potential low
□	Susceptible