MYCOBACTERIAL INTERSPERSED REPETITIVE UNIT-VARIABLE NUMBER TANDEM REPEAT GENOTYPING OF *MYCOBACTERIUM TUBERCULOSIS* ISOLATES USING LONG-READ NANOPORE SEQUENCING: A PRELIMINARY STUDY

Héctor Guzmán García[1], Pravech Ajawatanawong[2], Asmatullah Usmani[3] and Prasit Palittapongarnpim[4,5]

[1]Department of Microbiology, Faculty of Science; [2]Division of Bioinformatics and Data Management for Research, Faculty of Medicine Siriraj Hospital, Mahidol University, Bangkok, Thailand; [3]Department of Biology, Faculty of Education, Kandahar University, Loya Wayala, Kandahar, Afghanistan; [4]Pornchai Matungkasombut Centre for Microbial Genomics, Faculty of Science, Mahidol University, Bangkok; [5]National Science and Technology Development Agency, Pathum Thani Province, Thailand

Abstract. Worldwide, tuberculosis (TB) is one of the top ten causes of death. Molecular epidemiology has contributed significantly to understanding of TB transmission. One of the most significant methods is variable number tandem repeat (VNTR) typing, which determines variations of tandem repeat copy numbers, but the method is laborious. A MinION nanopore sequencer was applied to determine copy numbers of mycobacterial interspersed repetitive unit (MIRU)-VNTR loci of *Mycobacterium tuberculosis* isolates, which eases workload by barcode-multiplexing and analyzing 276 amplicons in a single experiment, and also enables detection of single nucleotide variants (SNVs). MIRU-VNTR loci (n = 24) were PCR amplified and amplicons, confirmed by agarose gel-electrophoresis, were pooled prior to nanopore nucleotide sequencing library preparation. Probabilities of correctly assigned genotypes were calculated using a four-parameter logistic regression model. Unambiguous genotypes were obtained from 245 (89%) amplicons and 14 putative SNV sites were detected distributed among eleven VNTR loci. In conclusion, long-read sequencing allowed uncomplicated genotyping of minisatellites of *M. tuberculosis* isolates from chronic pulmonary TB patients and confirmed the presence of SNVs within MIRU-VNTR loci.

Keywords: minisatellite, mycobacterial interspersed repetitive unit-variable number tandem repeat genotyping, nanopore nucleotide sequencing, tuberculosis

Correspondence: Prof Prasit Palittapongarnpim, Department of Microbiology, Faculty of Science, Mahidol University, Rama 6 Road, Bangkok 10400, Thailand
Tel: +66 (08) 1867 4202 E-mail: prasit.pal@mahidol.ac.th

INTRODUCTION

Tuberculosis (TB), caused by *Mycobacterium tuberculosis* (*Mtb*), is one of the top ten causes of death worldwide (WHO, 2019). Analysis of DNA structural variants, particularly in mycobacterial interspersed repetitive unit-variable number tandem repeat (MIRU-VNTR) loci, allows an understanding of TB epidemiology and helps to develop better control strategies (De Beer *et al*, 2012; De Beer *et al*, 2014). MIRU-VNTR genotyping relies on PCR amplification of 12-24 specific loci, followed by determination of amplicons sizes by gel-electrophoresis, which indicate the number of targets MIRU-VNTR loci (Smittipat *et al*, 2005). However, only a small number of alternative methods have been developed for determining amplicon sizes (Pang *et al*, 2011; Nikolayevskyy *et al*, 2016).

Next generation sequencing (NGS) has been applied to *Mtb* genotyping (Palittapongarnpim *et al*, 2018; Ajawatanawong *et al*, 2019). However, information regarding MIRU-VNTR copy numbers cannot be derived consistently from short reads generated by 2nd generation sequencing (Treangen and Salzberg, 2011), only 3rd generation sequencing methods allow accurate and consistent determination of VNTR copy numbers (De Coster *et al*, 2019). With its high rates of error in determination of single nucleotide variants (SNVs) by, recourse to 2nd generation sequencing method is required (Greig *et al*, 2019), but both 2nd and 3rd generation sequencings are still needed to obtain both VNTR genotyping and SNV identification.

Here, we describe a combination of classical PCR and multiplexed/barcoded Oxford nanopore sequencing technologies to determine copy number variants in MIRU-VNTR loci. The method allowed determination of amplicon sequences of 24 VNTR loci from 11 clinical *Mtb* samples in a single sequencing reaction. This could provide an alternative to conventional methods for VNTR typing or complement NGS in the study of *Mtb* genetic diversity.

MATERIALS AND METHODS

Samples collection

Mtb isolates were from chronic pulmonary TB patients registered in Chiang Rai Province, northern Thailand collected as part of a previously reported cohort study during 2003-2010, initiated by the Research Institute of Tuberculosis, Japan Anti-Tuberculosis Association (RIT/JATA), Japan and Ministry of Public Health, Thailand (Palittapongarnpim *et al*, 2018). DNA samples of *Mtb* Lineage 1 (n = 11) and H37Rv reference strain were used (Table 1).

Research protocols were approved by the Ethical Committees of Chiangrai Prachanukroh Hospital, Chiangrai Province (version 3/2556) and the Ministry of Public Health, Thailand (ref. no. CR 0032.102/15665). Prior written informed consent was obtained from all participants or parents/legal guardians.

Table 1

Mycobacterium tuberculosis isolates collected from chronic pulmonary TB patients in Chiang Rai Province, northern Thailand (2003-2010) used in the study.

Isolate ID	Oxford nanopore adapter barcode	Sequence accession number https://www.sanger.ac.uk	Sublineage classification
01	NB01	ERR718231	1.2.1.2
02	NB02	ERR718233	1.1.1.9
03	NB03	ERR718235	1.1.2.1
04	NB04	ERR718236	1.1.2.1
05	NB05	ERR718237	1.1.1.8
06	NB06	ERR718239	1.2.1.2
07	NB07	ERR718196	1.1.1.5
08	NB08	ERR718197	1.1.1.7
09	NB09	ERR718198	1.1.1.8
10	NB10	ERR718201	1.1.1.8
11	NB11	ERR718210	1.2.1.2
12 (H37Rv)	NB12	-	4.9

MIRU-VNTR PCR amplification protocol

Each MIRU-VNTR locus was amplified separately using primers shown in Supplementary Table S1. Reaction mixture (25 µl) contained 2.5 µl of 10X Taq Buffer (New England Biolabs, Ipswich, MA), 0.75 µl of 50 mM $MgCl_2$ (New England Biolabs, Ipswich, MA), 0.5 µl of 10 mM dNTPs (New England Biolabs, Ipswich, MA), 0.5 µl of 10µM of each primer (Macrogen, Seoul, South Korea), 0.125 µl of Taq DNA polymerase (New England Biolabs, Ipswich, MA), and 20 ng of genomic DNA. Thermocycling (conducted in a T100; Bio-Rad, Irvine, CA) conditions were as follows: 96°C for 10 minutes; 40 cycles at 96°C for 60 seconds, 60°C for 60 seconds and 75°C for 60 seconds; with a final step at 72°C for 7 minutes. *Mtb* H37Rv and sterile water was used as positive and negative control respectively. Amplicons were separated by 1.5% agarose gel- electrophoresis and quantified using a DS-11 Series DeNovix spectrophotometer (DeNovix, Wilmington, DE). Six random amplicons (sample ID 02, loci MIRU39 and MIRU10; sample ID 03, loci MIRU10 and MIRU24; sample ID 08, locus MIRU02, and sample ID 11, locus MIRU39) were sequenced (Macrogen, Seoul, South Korea).

MiniON sequencing protocol

MiniON sequencing sample was prepared using an SQK-LSK109 ligation sequencing and PCR barcoding kits (Oxford Nanopore Technologies Ltd, Oxford, UK). In brief, unpurified amplicons from individual *Mtb* isolates were pooled, end-repaired and dA-tailed using a NEBNext End Repair/dA-tailing kit (New England Biolabs, Ipswich, MA), then purified employing Agencourt AMPure XP beads (Beckman Coulter Inc, Brea, CA), ligated to unique dT-tailed barcode adapter sequences, NB01-NB12 (Table 1), and purified with Agencourt AMPure XP beads (Beckman Coulter Inc, Brea, CA). Equimolar barcoded DNA samples were pooled, ligated with an adapter mix (AMX; Oxford Nanopore Technologies Ltd, Oxford, UK), purified with Agencourt AMPure XP beads (Beckman Coulter Inc, Brea, CA) and the final sample (containing *Mtb* samples (n = 12); 276 amplicons) was placed in a 1D MinION R9.4 flow cell (Oxford Nanopore Technologies Ltd, Oxford, UK) and sequencing conducted for 18 hours employing a MinKNOW software version 18.05.5 (Oxford Nanopore Technologies Ltd, Oxford, UK). *Mtb* H37Rv amplicons (NB12) were used as control.

Nanopore sequencing data analysis

Raw sequence data were uploaded for "base-calling" using an EPI2ME Base Calling 1D version 2.2.8 software package (https://nanoporetech.com/nanopore-sequencing-data-analysis/). Sequences in FASTQ format were extracted from the raw FAST5 files using an Albacore version 0.8.4 software package (Oxford Nanopore Technologies Ltd, Oxford, UK). Statistical analysis of MinION sequencing data was conducted and presented using a Nanoplot version 1.26.1 software package (https://github.com/wdecoster/nanopack). Demultiplexing based on barcode adapter sequences as well as trimming of barcode adapters and chimeric reads were performed using a Porechop version 0.2.3 software package (https://github.com/rrwick/Porechop). For validating sequence identification, unique barcoded adapter sequences are required to have at least 95% identity to the reference barcode adapter sequence.

MIRU-VNTR genotyping and identification mapping methods

Reads of each VNTR locus were identified by a mapping procedure using a GraphMap 0.5.2, a software specifically designed to analyse nanopore sequencing reads, making use of new algorithms, such as gapped spaced seeds, graph mapping and longest common subsequence in k Length substrings (LCAk) (Sović *et al*, 2016). In order to remove noise generated by stutter products (defined as products formed by slipped-strand DNA synthesis at repeat regions) (Kimura *et al*, 2009), all barcoded identified reads were mapped onto a synthetic reference template FASTA containing the 24 MIRU-VNTR loci, listed according to copy numbers ranging from 0 to 15. All sequences in the reference template included the entire theoretical PCR-amplified regions of *Mtb*

H37Rv sequences using primers listed in Supplementary Table S1 and sizes of each allele are shown in Supplementary Table S2 which are available at https://doi.org/10.6084/m9.figshare.14583327.v3.

Coverage metrics procedure

Further evaluation of mapped reads was restricted to algorithms primarily based on coverage depth using a Samtools depth version 1.1 software package (Li *et al*, 2009), which creates a simple tab-separated table contain three columns, namely, reference name, position, and coverage depth. Median of read depth (MRD) (defined as median of the number of times each base in the reference amplicon sequence of each allele is covered) (Sims *et al*, 2014) and read depth uniformity (RDU) (defined as sequencing coverage between the 25th and 75th percentiles of the mapped read depth) (Trost *et al*, 2018) were computed for every possible allele of the reference template using a Plotly Chart Studio™ software package (Plotly Technologies Inc, Montréal, QC, Canada). Alleles with the highest MRD values are considered as the most likely "true" alleles.

SNVs identification procedure

A mapping data (GRAPHMAP output) method was applied to generate a table of all SNVs at all positions using SAMtools described above. Integrative Genomics Viewer (IGV) (Robinson *et al*, 2020) "consensus mode" then was used to present aligned sequence data and SNVs were identified by visual inspection. Alleles containing SNVs were pairwise-aligned to reference *Mtb* H37Rv, Beijing 2014PNGD and TCDC10 strains (GenBank accession nos. NC_000962, CP022704 and NZ_CP047164.

Statistical analysis

A 4-parameter logistic (4PL) regression model using Quest Graph™ (AAT Bioquest, Sunnyvale, CA) was employed to predict the probability that the most likely "true" alleles are indeed the "true" alleles based on MRD values. An allele is considered "true" if probability >0.95.

RESULTS

MIRU-VNTR PCR, gel-electrophoretic analysis and concentration quantification

MIRU-VNTR loci (18/24) of all 11 clinical *Mtb* isolates and reference *Mtb* H37Rv strain were successfully PCR amplified and single amplicons were observed by gel-electrophoresis in 276/288 (96%) reactions (Supplementary Figures are available at https://doi.org/10.6084/m9.figshare.14583327.v3). ETR-B, MIRU04, MIRU16, MIRU23, MIRU26, and QUB26 loci showed missing or double amplicons and were excluded from subsequent analysis. From spectrophotometric quantification of gel-excised amplicons, the highest copy number of repeats observed was ten [isolates 01 and 06 at locus Mtub21 (662 bp)] and the largest amplicon (930 bp) was obtained from isolate 06 at locus QUB26 (seven repeats) (Supplementary Fig S2 is also available at https://doi.

org/10.6084/m9.figshare.14583327.v3). Nucleotide sequencing of six random amplicons confirmed agarose gel-based amplicon size determination (data not shown).

MiniON library evaluation, sequencing and data preparation for genotyping

Total quantity of end-repaired and dA-tailed DNA, after the first purification of amplicons, was 8,783 ng and total quantity of barcoded DNA following second purification step was 2,946 ng (Supplementary Table S3 available at https://doi.org/10.6084/m9.figshare.14583327.v3). Barcoded DNA (624 ng; 52 ng per *Mtb* sample) were pooled for AMX-ligation, which, after the third purification step (420 ng), were subjected to nanopore long-read sequencing, resulting in 3,169,297 reads (average sequence length of 570 bp and average Phred Quality score of 8.02 (Supplementary Fig S3 available at https://doi.org/10.6084/m9.figshare.14583327.v3). Reads (n = 3,116,201; 98.32%) contained uniquely identified barcodes [median (range) reads of 259,683 (108,616-421,787)].

Analysis of barcoded-identified reads revealed the sequencing library comprised reads ranging in length from 4 bp to 3,376 bp, including i) truncated reads lacking complete DNA sequence at one end of the sequence, ii) chimeric reads comprising ≥2 sequences from different amplicons within the same read, and iii) stutter products attributed to slipped-strand mispairing during PCR (Supplementary Fig S4 available at https://doi.org/10.6084/m9.figshare.14583327.v3). Truncated reads and chimeric reads were discarded using Porechop version 0.2.3 software package (https://github.com/rrwick/Porechop).

MIRU-VNTR genotyping analytics

Sequencing coverage metrics were applied to discriminate between real and stutter alleles by mapping onto a synthetic reference template FASTA containing the 24 MIRU-VNTR loci, revealing a total of 1,935,503/3,116,201 (62.11%) reads able to be mapped to the reference template (Fig 1). MRD values for 271 amplicons from 24 MIRU-VNTR loci ranged from 1 to 46,993 (Table 2). Amplicons (n = 5) with MRD values = 0 were excluded from further analysis. Amplicons with relatively high MRD values were the most likely "true" alleles. As a representative, the RDU plot of MIRU02, alleles 00-10, and IGV graphic of most and least likely "true" alleles are shown in Fig 2. Examples of identification of the most likely "true" alleles for other alleles are shown in Supplementary Fig S5 available at https://doi.org/10.6084/m9.figshare.14583327.v3.

In comparison to experimental results, accuracy of prediction and MRD values are significantly correlated (p-value <0.001), with an MRD value of 84.68 corresponding to a probability of correct prediction of 0.95 (Fig 3). Based on this logistic model, it could be predicted that 245/276 (89%) amplicons were "true" alleles. The numbers of

VNTR at each of the MIRU locus for the 11 *Mtb* isolates and reference H37Rv strain (Table 3) are in agreement with results obtained by the standard PCR method, *ie* numbers of repeats ranging from 1 to 10, latter present at Mutb21 in samples 01 and 06.

Identification of VNTR loci SNVs

SNVs (n = 14) were identified in ten VNTR loci. Twelve of these SNVs sites were unique to single isolates, one SNV was present in 8 isolates and another SNV was present in 3 isolates Table 4). A SNV in MIRU16 at position 223 was identified only in three isolates, all belonging to L1.2.1.2. A SNV in MIRU02 was present eight isolates, suggesting that this SNV was common in Lineage 1. The per-base coverage reads of identified SNVs were 27,229 with a median of 5,641 reads.

DISCUSSION

Determining polymorphisms in minisatellites such as MIRU-VNTR is an important tool for *Mtb* epidemiological

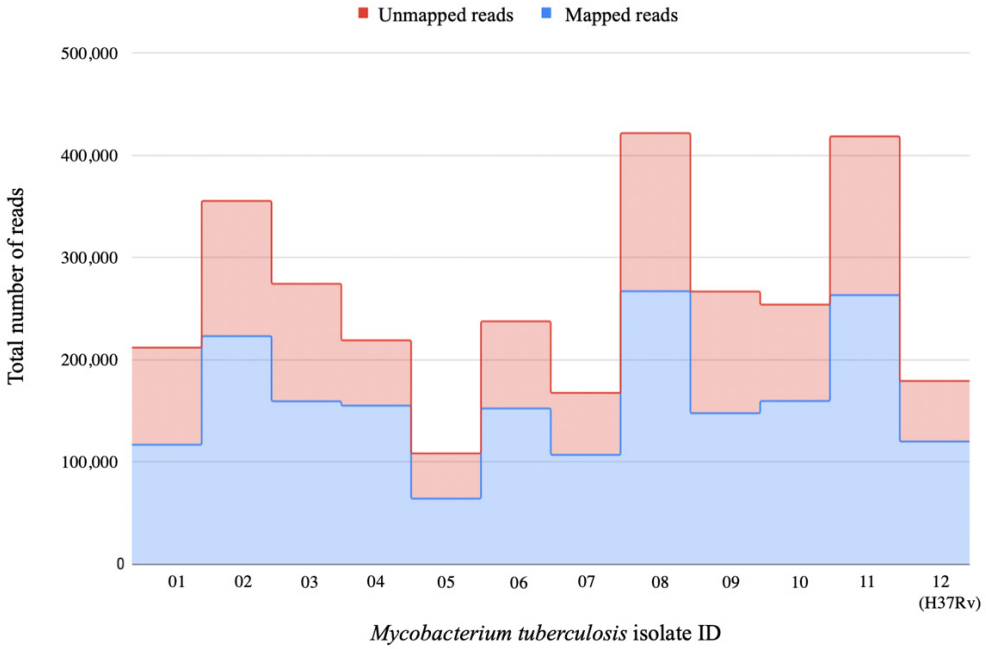

Fig 1 - Number of mapped and unmapped reads of *Mycobacterium tuberculosis* isolates collected from chronic pulmonary TB patients in Chiang Rai Province, northern Thailand (2003-2010)

Mapped and unmapped reads constitute 62.11 and 37.89% of total reads respectively.

Table 2

Median read depth (MRD) values of the most likely "true" alleles of *Mycobacterium tuberculosis* isolates collected from chronic pulmonary TB patients in Chiang Rai Province, northern Thailand (2003-2010)

Isolate ID	MIRU02	Mtub04	ETRC	MIRU04	MIRU40	MIRU10	MIRU16	Mtub21	MIRU20	QUB11b	ETRA	Mtub29	Mtub30	ETRB	MIRU23	MIRU24	MIRU26	MIRU27	Mtub34	MIRU31	Mtub39	QUB26	QUB4156	MIRU39
01	3,397	7,490	5,390	0	2,958	10	893	994	7,107	2,301	6,507	6,395	14,750	4	7,217	7,218	1,845	5,915	4,379	4,793	8,078	5	5,255	13,954
02	12,243	3,725	9,819	1	3,046	17	3,618	4,456	12,491	2,762	11,627	18,638	29,241	4	244	21,818	3,835	10,935	13,979	17,930	6,098	777	12,561	23,420
03	12,115	899	8,329	1	4,919	8	2,323	2,525	9,292	239	9,718	12,290	29,553	ND	8,649	22	3,490	5,756	7,382	6,035	3,404	233	7,370	24,999
04	4,470	15,457	4,109	43	4,929	18	ND	7,816	12,140	804	12	6,938	14,051	7	6,101	7,734	2,117	6,902	5,194	4,453	1,877	305	2,757	46,993
05	3,059	7,519	2,654	0	1,151	7	1,103	1,844	3,923	2,223	751	3,366	7,719	2	2,390	6,757	1,274	2,025	2,635	2,150	705	202	2,483	8,397
06	6,939	4,949	5,134	0	4,385	3,202	2,491	902	9,064	1,747	24,027	8,847	17,034	0	5,240	8,834	ND	7,296	7,060	8,636	8,576	554	7,292	10,290
07	5	9,152	4,079	ND	2,979	4,549	2,405	9,071	6,019	1,756	1,161	6,670	15,500	0	ND	9,465	2,797	3,831	5,979	698	905	271	4,900	14,929
08	45	31,250	9,903	ND	8,299	8,697	5,169	17,841	19,494	9,002	49	13,653	34,763	21	ND	20,291	7,161	15,595	6,208	13,324	6,032	789	9,709	29,921
09	8,986	7,896	5,070	ND	4,635	7,903	3,607	4,227	9,262	1,921	5,913	8,702	19,438	6	4	13,501	9,395	8,466	5,692	9,553	5,878	705	6,924	179
10	7,900	15,598	9,900	ND	6,552	5,537	2,629	4,301	6,943	2,559	2,093	4,284	27,607	8	3,270	16,008	3,446	2,160	3,376	8,849	7,339	ND	4,288	15,063
11	11,224	11,709	8,279	ND	7,818	5,395	2,531	1,100	11,226	4,625	5,298	13,038	4,0311	4	2,917	22,927	16,675	7,534	8,273	17,241	26,590	ND	9,193	29,477
12 (H37Rv)	6,991	20	7,333	1240	10,067	9,919	1,284	15,152	77	391	9,691	3,159	13,145	13162	3,082	383	3,931	4,672	7,504	7,965	232	104	626	5

Light blue highlight: 85 – 999; Medium blue highlight: 1,000 – 4,999; Dark blue highlight: 5,000 – 46,993

ND: not done

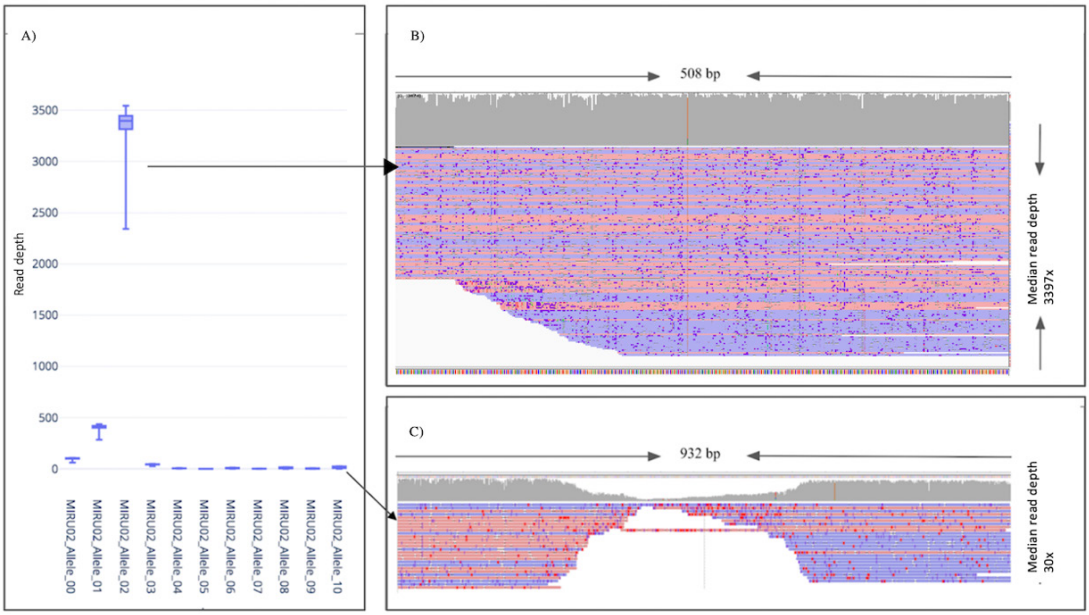

Fig 2 - Representative identification of the most and least likely "true" alleles

A: Plot of median read vs mycobacterial interspersed repetitive unit (MIRU) allele. Vertical lines (whiskers) indicate the range of per-base read depth of reference amplicon sequence; horizontal lines indicate median mapped read depth and boxes indicate read depth uniformity range.

B: Screenshot of Integrative Genomics Viewer reads of the most likely "true" allele (allele 02).

C: Screenshot of Integrative Genomics Viewer reads of a not likely "true" allele (allele 10). Blue and red lines represent forward and reverse reads orientation of the mapping respectively. Their blue and red highlight represents sequencing errors in the reads, respectively. Grey bars indicate per-base read depth accordingly.

studies, *viz* tracing spread of specific genotypes and investigating local and global genotypic population structures (Gagneux, 2017). As many VNTR loci are needed for the task, measuring copy numbers of repeat units at all loci is a laborious and time-consuming task, and, consequently, methods to simplify the extraction of copy number information from PCR products have been developed, such as duplexing PCR VNTR amplification with minimal overlap (Yasmin *et al*, 2016) or quadruplexing PCR VNTR amplification with fluorescent-labelled primers followed by fragment analysis using an automated capillary sequencer (Nikolayevskyy *et al*, 2016).

We successfully employed the recently introduced long-read nanopore multiplex sequencing method (Jain *et al*, 2016; Nguyen *et al*, 2017) to obtain copy numbers of the standard 24 MIRU-VNTR loci for genotyping 11 clinical *Mtb* isolates in a single sequencing experiment. In this preliminary study, we only tested the method on alleles that were detected by the standard PCR protocol. However, analyzing nanopore sequencing data could be challenging when the coverage of certain alleles present in the sequencing library results low.

Softwares such as GraphMap and Samtools can be performed not only for MRD-based genotyping but also for generating high quality consensus sequences. Nanopore sequencing technologies offer support tools to enhance analysis throughput, *eg* barcode

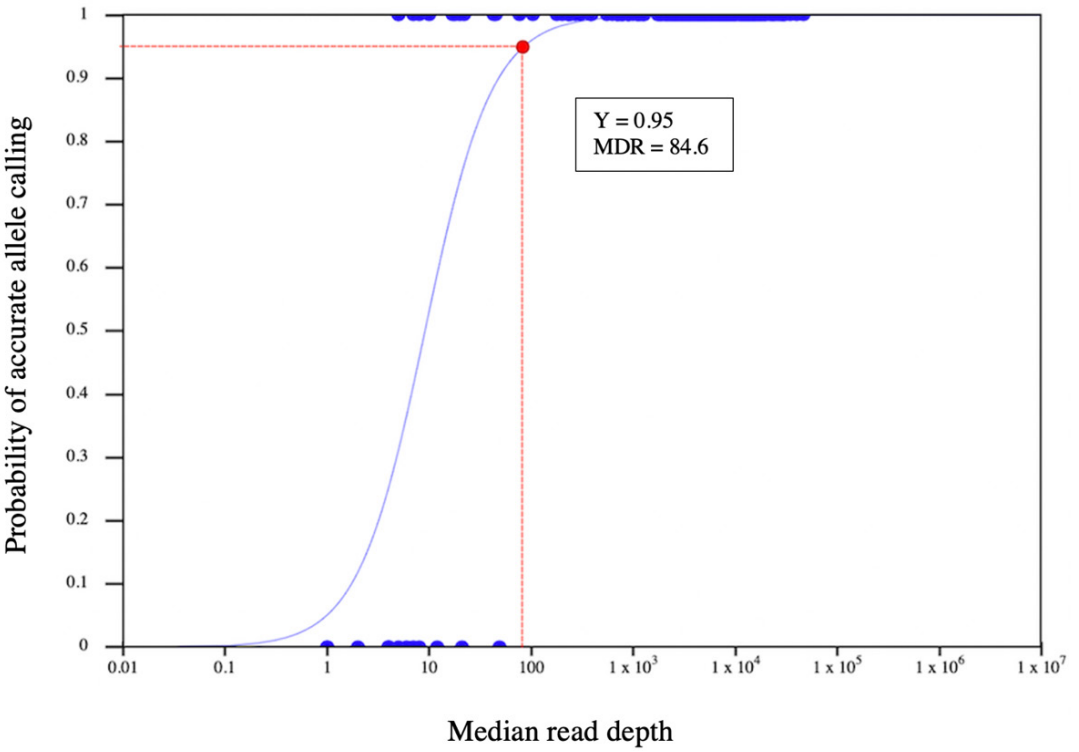

Fig 3 - Logistic regression curve fitting of correct genotyping as a function of median read depth (MRD)

Red dot represents correct allele assignment when probability of accurate read calling (Y) is ≥0.95 and MRD value ≥84.6.

Table 3

Mycobacterial interspersed repetitive unit-variable number tandem repeats at loci of *Mycobacterium tuberculosis* isolates collected from chronic pulmonary TB patients in Chiang Rai Province, northern Thailand (2003-2010)

Isolate ID	MIRU02	Mtub04	ETRC	MIRU04	MIRU40	MIRU10	MIRU16	Mtub21	MIRU20	QUB11b	ETRA	Mtub29	Mtub30	ETRB	MIRU23	MIRU24	MIRU26	MIRU27	Mtub34	MIRU31	Mtub39	QUB26	QUB4156	MIRU39
01	2	1	4		2		3	10	2	1	4	3	2		6	2	2	3	3	4	2		1	3
02	2	2	4		3		2	5	2	2	6	3	2		5	2	2	3	3	2	4	6	1	3
03	2	2	4		3		2	6	2	5	2	3	2	NA	4		2	3	4	5	5	6	1	3
04	2	2	4		3		NA	3	2	7		3	2		4	2	2	3	3	5	5	6	1	3
05	2	2	4		4		2	5	2	2	7	3	2		4	2	2	3	3	5	6	6	1	3
06	2	1	4		2	4	3	10	2	1	3	3	2		4	2	NA	3	3	4	2	7	1	3
07		2	4	NA	3	4	2	3	2	2	7	3	2		NA	2	2	3	3	6	6	6	1	3
08		2	4	NA	3	4	2	3	2	2		3	2		NA	2	2	4	3	5	4	6	1	3
09	1	2	4	NA	3	4	2	5	2	2	7	3	2			2	2	3	3	5	4	6	1	2
10	2	2	4	NA	3	4	2	5	2	2	7	3	2		4	2	2	3	3	5	4	NA	1	3
11	2	1	4	NA	2	4	3	3	2	1	4	3	2		4	2	2	3	3	4	2	NA	1	3
12 (H37Rv)	2		4	3	1	3	2	2		5	3	4	2	3	6	1	3	1	3	3	5	5	2	

Light blue highlight: 1 – 5 copies of the repeat unit: Dark blue highlight: 6 – 10 copies of the repeat unit; Purple highlight: not genotyped

NA: not applicable

Table 4

Single nucleotide variant (SNV) distribution in variable number tandem repeats (VNTRs) of *Mycobacterium tuberculosis* (Mtb) isolates collected from chronic pulmonary TB patients in Chiang Rai Province, northern Thailand (2003-2010)

VNTR locus	Number of copies of alleles	Nucleotide position	Base in H37Rv	Base in isolate	Number of isolates with SNV (n = 11)
MIRU02	2	144	G	C	1
		241	A	G	8 (n = 8)
ETR-C	4	338	C	T	1
MIRU16	3	223	G[a]	C	3
MIRU20	2	316	G	A	1
QUB11b	5	100	A	C	1
ETR-A	3	52	T	G	1
		299	T	A	1
	4	224	T[b]	A	1
MIRU23	6	235	G	C	1
		288	G	C	1
MIRU27	3	335	C	T	1
Mtub34	3	444	T	C	1
MIRU39	3	616	G	A	1

[a]Mtb Beijing2014PNGD strain; [b]Mtb TCDC10 strain

adapter systems that enable sequencing of up to 96 samples in the same sequencing experiment. An alternative method to determining VNTR copy numbers by nanopore sequencing is to directly sequence the whole genome (Bainomugisa *et al*, 2018). However, this latter approach would substantially increase the number of sequencing reactions and may still require 2nd generation sequencing to accurately identify SNVs. For only VNTR typing, a two-step PCR sequencing procedure can significantly reduce cost (Perez-Lago *et al*, 2015; Christensen *et al*, 2015). Moreover, future releases of sequencing kits that allow more multiplexing would bring the cost down further.

In conclusion, this preliminary study of long-read nanopore sequencing incorporating a multiplexing enabled ready identification of copy numbers of mycobacterial interspersed repetitive unit-variable number tandem repeat loci of *Mycobacterium tuberculosis* isolates from a small number of chronic pulmonary TB patients and allowed identification of single nucleotide variants present within such loci. A larger number of samples will be required to evaluate the feasibility and cost-benefit of this latest innovation in DNA sequencing.

ACKNOWLEDGEMENTS

The authors thank Ms Kulawadee Wonganun for her enthusiastic support and constant assistance in carrying out experiments, Mr Wuthiwat Ruangchai for providing important suggestions, continuous involvement and guidance on bioinformatics and students and staff of the Department of Microbiology, Faculty of Science, Mahidol University for their cooperation, company and valuables comments in the manuscript preparation. The research was supported in part by the Japanese SATREPS program in collaboration with the Department of Medical Sciences, Ministry of Public Health Thailand, the B-HIV Research Foundation of Thailand, the University of Tokyo and the Japan Research Institute of Tuberculosis (JATA).

CONFLICTS OF INTEREST DISCLOSURE

The authors declare no conflicts of interest.

REFERENCES

Ajawatanawong P, Yanai H, Smittipat N, *et al*. A novel Ancestral Beijing sublineage of *Mycobacterium tuberculosis* suggests the transition site to Modern Beijing sublineages. *Sci Rep* 2019; 9: 13718.

Bainomugisa A, Duarte T, Lavu E, *et al*. A complete high-quality MinION nanopore assembly of an extensively drug-resistant *Mycobacterium tuberculosis* Beijing lineage strain identifies novel variation in repetitive PE/PPE gene regions. *Microb Genom* 2018; 4: e0001488.

Christensen K D, Dukhovny D, Siebert U, Green RC. Assessing the costs and cost-effectiveness of genomic sequencing. *J Pers Med* 2015; 5: 470-86.

De Beer JL, Ködmön C, van Ingen J, Supply P, van Soolingen D, Global Network for Molecular Surveillance of Tuberculosis 2010. Second worldwide proficiency study on variable number of tandem repeats typing of *Mycobacterium tuberculosis* complex. *Int J Tuberc Lung Dis* 2014; 18: 594-600.

De Beer JL, Kremer K, Ködmön C, Supply P, van Soolingen D, Global Network for the Molecular Surveillance of Tuberculosis 2009. First worldwide proficiency study on variable-number tandem-repeat typing of *Mycobacterium tuberculosis* complex strains. *J Clin Microbiol* 2012; 50: 662-9.

De Coster W, De Rijk P, De Roeck A, *et al*. Structural variants identified by Oxford Nanopore PromethION sequencing of the human genome. *Genome Res* 2019; 29: 1178-87.

Gagneux S, editor. Strain variation in the *Mycobacterium tuberculosis* complex: its role in biology, epidemiology and control. New York City, NY: Springer International Publishing; 2017.

Greig DR, Jenkins C, Gharbia S, Dallman TJ. Comparison of single-nucleotide variants identified by Illumina and Oxford Nanopore technologies in the context of a potential outbreak of Shiga toxin-producing *Escherichia coli*. *Gigascience* 2019; 8: giz104.

Jain M, Olsen HE, Paten B, Akeson M. The Oxford Nanopore MinION: delivery of nanopore sequencing to the genomics community. *Genome Biol* 2016; 17: 239.

Kimura M, Sakamuri RM, Groathouse NA, *et al*. Rapid variable-number tandem-repeat genotyping for Mycobacterium leprae clinical specimens. *J Clin Microbiol* 2009; 47: 1757-66.

Li H, Handsaker B, Wysoke, A, *et al*. The Sequence Alignment/Map format and SAMtools. *Bioinformatics* 2009; 25: 2078-9.

Nguyen SH, Duarte T, Coin LJM, Cao MD. Real-time demultiplexing Nanopore barcoded sequencing data with npBarcode. *Bioinformatics* 2017; 33: 3988-90.

Nikolayevskyy V, Trovato A, Broda A, Borroni E, Cirillo D, Drobniewski F. MIRU-VNTR genotyping of *Mycobacterium tuberculosis* strains using QIAxcel technology: a Multicentre evaluation study. *PloS One* 2016; 11: e0149435.

Palittapongarnpim P, Ajawatanawong P, *et al*. Evidence for host-bacterial co-evolution via genome sequence analysis of 480 Thai *Mycobacterium tuberculosis* lineage 1 isolates. *Sci Rep* 2018; 8: 11597.

Pang Y, Zhou Y, Wang S, *et al*. A novel method based on high resolution melting (HRM) analysis for MIRU-VNTR genotyping of *Mycobacterium tuberculosis*. *J Microbiol Methods* 2011; 86: 291-7.

Pérez-Lago L, Martínez Lirola M, Herranz M, Comas I, Bouza E, García-de-Viedma D. Fast and low-cost decentralized surveillance of transmission of tuberculosis based on strain-specific PCRs tailored from whole genome sequencing data: a pilot study. *Clin Microbiol Infect* 2015; 21: 249.e1-9.

Robinson JT, Thorvaldsdóttir H, Turner D, Mesirov JP. igv.js: an embeddable JavaScript implementation of the Integrative Genomics Viewer (IGV), 2020 [cited 2021 Feb 10]. Available from: URL: https://www.biorxiv.org/content/10.1101/2020.05.03.075499v1.full.pdf

Sims D, Sudbery I, Ilott NE, Heger A, Ponting CP. Sequencing depth and coverage: key considerations in genomic analyses. *Nat Rev* Genet 2014; 15: 121-32.

Smittipat N, Billamas P, Palittapongarnpim M, *et al*. Polymorphism of variable-number tandem repeats at multiple loci in *Mycobacterium tuberculosis*. *J Clin Microbiol* 2005; 43: 5034-43.

Sović I, Šikić M, Wilm A, Fenlon SN, Chen S, Nagarajan N. Fast and sensitive mapping of nanopore sequencing reads with GraphMap. *Nat Commun* 2016; 7: 11307.

Treangen TJ, Salzberg SL. Repetitive DNA and next-generation sequencing: computational challenges and solutions. *Nat Rev* Genet 2011; 13: 36-46.

Trost B, Walker S, Wang Z, *et al*. A comprehensive workflow for read depth-based identification of copy-number variation from whole-genome sequence data. *Am J Hum Genet* 2018; 102: 142-55.

World Health Organization (WHO). Global tuberculosis report, 2019 [cited 2021 Feb 10]. Available from: URL: https://apps.who.int/iris/bitstream/handle/10665/329368/9789241565714-eng.pdf

Yasmin M, Le Moullec S, Siddiqui RT, De Beer J, Sola C, Refrégier G. Quick and cheap MIRU-VNTR typing of Mycobacterium tuberculosis species complex using duplex PCR. *Tuberculosis* 2016; 101: 160-3.