

1 **MinION Nanopore Sequencing of Multiple Displacement Amplified Mycobacteria DNA**
2 **Direct from Sputum**

3
4
5 Sophie George^{1,2,3¶}, Yifei Xu^{1,2¶}, Nicholas Sanderson^{1,2}, Alasdair TM Hubbard^{1,#a}, David T.
6 Griffiths^{1,2}, Marcus Morgan⁴, Louise Pankhurst^{1,3}, Sarah J. Hoosdally^{1,2}, Dona Foster^{1,2},
7 Samantha Thulborn⁵, Esther Robinson⁶, E. Grace Smith⁶, Priti Rathod⁶, A. Sarah Walker^{1,2,3},
8 Timothy E. A. Peto^{1,2,3}, Derrick W. Crook^{1,2,3}, Kate E. Dingle^{1,2*}

9
10

11 ¹ Nuffield Department of Clinical Medicine, John Radcliffe Hospital, Oxford University, UK

12
13 ² National Institute for Health Research (NIHR) Oxford Biomedical Research Centre, John
14 Radcliffe Hospital, Oxford, UK

15
16 ³ NIHR Oxford Health Protection Research Unit in Healthcare Associated Infection and
17 Antimicrobial Resistance at Oxford University in partnership with Public Health England,
18 Oxford, UK

19
20

21 ⁴ Microbiology Department, Oxford University Hospitals NHS Trust, Oxford, UK.

22 ⁵ Respiratory Medicine Unit, Nuffield Department of Medicine, John Radcliffe Hospital,
23 University of Oxford, UK

24
25

26 ⁶ PHE National Mycobacteria Reference Service - North and Central, Birmingham Public
27 Health Laboratory, UK

28 #aCurrent address: Department of Parasitology, Liverpool School of Tropical Medicine,
29 Pembroke Place, Liverpool, UK

30
31

32 ¶ These authors contributed equally to the study.

33 * Corresponding author

34 Email: kate.dingle@ndm.ox.ac.uk

35

34 **Short title:** Nanopore Sequencing of Mycobacteria DNA amplified direct from Sputum

36 **ABSTRACT**

37 Sequencing of pathogen DNA directly from clinical samples offers the possibilities of rapid
38 diagnosis, faster antimicrobial resistance prediction and enhanced outbreak investigation. The
39 approach is especially advantageous for infections caused by species which grow very slowly
40 in culture, such as *Mycobacteria tuberculosis*. Since the pathogen of interest may represent as
41 little as 0.01% of the total DNA, enrichment of the input material for target sequences by
42 specific amplification and, or depletion of non-target DNA (human, other bacteria) is
43 essential for success. Here, we investigated the potential of isothermal multiple displacement
44 amplification by Phi29 polymerase. We directed the amplification reaction towards
45 Mycobacteria DNA in sputum samples by exploiting in our oligonucleotide primer design,
46 their high GC content (approximately 65%) relative to human DNA. Amplified DNA was
47 then sequenced using the Oxford Nanopore Technology MinION. In addition, a model
48 system comprising standardised ‘mock clinical samples’ was designed. Pooled infection
49 negative human sputum samples were spiked with enumerated *Mycobacterium bovis* (BCG)
50 Pasteur strain at concentrations spanning the typical range at which *Mycobacterium*
51 *tuberculosis* is found in human sputum samples (10^6 - 10^1 BCG cells/ml). To assess the
52 amount of BCG sequence enrichment achieved, sample DNA was sequenced both before, and
53 after amplification. Reads from amplified samples, which mapped to a BCG reference
54 genome, comprised short repeated sequences - apparently transcribed multiple times from the
55 same fragment of BCG DNA. Therefore post-amplification, the samples were enriched for
56 BCG sequences relative to unamplified sequences (8,101 BCG reference mapped reads,
57 increasing to 28,617 at 10^6 BCG cells/ml sample), but BCG genome coverage declined
58 markedly (for example 89.4% to 4.1%). In summary, the use of standardised mock clinical
59 samples allowed direct comparison of data from different Mycobacteria enrichment

60 experiments and sequencing runs. However, optimal conditions for multiple displacement
61 amplification of minority Mycobacteria DNAs, remain to be identified.

62 INTRODUCTION

63 The World Health Organization (WHO) estimates that in 2016, *Mycobacterium tuberculosis*
64 complex caused 6.3 million new TB cases and 1.67 million deaths worldwide (including
65 374,000 among HIV-positive people) [1]. In addition, *Mycobacterium abscessus*, *M. avium*
66 complex, *M. kansasii*, *M. malmoense*, and *M. xenopi* are currently the most clinically
67 important of the >160 known non-tuberculous mycobacteria (NTM) species [2-4]. Correct
68 diagnosis and antimicrobial resistance determination are essential to ensure appropriate
69 treatment of Mycobacteria infections. However, when based on growth in culture, this may
70 take up to 80 days from initial presentation, increasing the risk of poor clinical outcomes and
71 failure to identify and control transmission.

72

73 Routine whole genome sequencing of Mycobacteria using Illumina MiSeq has accelerated
74 laboratory diagnosis of Mycobacteria by Public Health England (PHE) [5, 6]. Samples are
75 cultured until positive, which usually occurs within 1-2 weeks if the sample is smear positive
76 (but up to 5-6 weeks if bacterial load is low), then total DNA is extracted and sequenced
77 using the Illumina platform [5, 7]. WGS diagnostics can be completed in a median of 9 days
78 (IQR 6-10) [5]. Antimicrobial resistance predictions are based on nucleotide sequence data
79 [8], and phylogenetic analyses identify transmission events and outbreaks [9, 10]. The
80 information gained from WGS methods is typically available to the clinician within three to
81 four weeks of the sample being taken. Further savings in cost and time could potentially be
82 achieved by determining Mycobacteria genome sequences from DNA extracted directly from
83 clinical samples, thus eliminating the need for culture altogether.

84

85 Whole genome sequencing of pathogens direct from clinical samples is technically
86 challenging. Samples vary in terms of volume, numbers of human and bacterial cells and the

87 concentration of target organisms. Mycobacteria DNA, for example, can represent as little as
88 0.01% of the total DNA extracted from sputum [11]. Small scale studies employing direct
89 from sample sequencing have reported 0.002 - 0.7% sequence coverage of the *M.*
90 *tuberculosis* genome (using differential lysis and a DNA extraction kit) [12], and up to 90%
91 genome coverage with 20x depth (20/24 samples), (using the SureSelect target enrichment
92 method, Agilent, USA) [13, 14]. The study by Brown et al. [13] predicted both Mycobacteria
93 species and antibiotic susceptibility, but the cost (\$350 per sample) and duration of the
94 protocol (2 to 3 days) could prevent its use. The ideal 'direct from sample' methodology
95 would be simple, low cost and portable, to facilitate use in remote, low-income settings
96 where the burden of infection is greatest, and provision of clinical diagnostic services and
97 treatment is severely limited.

98

99 Potential advantages of adopting the Nanopore sequencing platform (Oxford Nanopore
100 Technology, ONT, Oxford, UK) include the possibility of increased read lengths [15, 16] and
101 consequent improved *de novo* assemblies, avoiding the need for a reference genome [16].
102 The accuracy of DNA sequences obtained using the Nanopore platform is constantly
103 improving; 99.9% can be achieved when Nanopolish is used to improve consensus accuracy
104 [17].

105

106 Enrichment of target pathogen sequences within total extracted DNA is a prerequisite for
107 direct-from-sample sequencing. The technique of isothermal multiple displacement
108 amplification (MDA) using Phi29 DNA polymerase [18] shows promise, since µg quantities
109 of DNA can be generated from minimal template (1-10ng) [19-21]. In the present study, we
110 investigated the possibility of biasing MDA towards Mycobacteria DNA in sputum samples,
111 prior to sequencing the DNA directly using the Oxford Nanopore Technology MinION.

112 MATERIALS AND METHODS

113 Standardised Samples for Method Development

114 A model system comprising standardised ‘mock clinical samples’ was designed. Pooled
115 infection negative human sputum samples were spiked with enumerated “Bacille de Calmette
116 et Guérin” *Mycobacterium bovis* (BCG) Pasteur strain (attenuated derivative of
117 *Mycobacterium bovis* [22]) at known concentrations. This allowed the results of different
118 experiments to be compared.

119

120 BCG Culture and Enumeration

121 BCG Mycobacteria Growth Incubator Tube (MGIT) culture conditions were optimised with a
122 two-step process facilitating the growth of single, rather than clusters or ‘flakes’ of BCG
123 cells. Firstly, freshly prepared MGIT culture tubes (Becton Dickinson, Wokingham, UK)
124 were inoculated sparsely with 10 µl BCG frozen stock. After 30 days incubation at 37°C, the
125 cultures were vortexed vigorously. ‘Flakes’ comprising large numbers of BCG cells were
126 allowed to settle for 10 minutes. Fresh MGIT tubes were prepared, with the addition of
127 Tween 80 (Acros Organics, Geel, Belgium) (0.5% final concentration) to encourage BCG
128 growth as single cells [23]. These fresh tubes were inoculated using 200 µl ‘settled’ BCG
129 culture. After 18 days incubation at 37°C, BCG cells were harvested and counted as follows.

130

131 The BCG culture was vortexed vigorously and 1 ml was removed. ‘De-clumped’ BCG cells
132 were pelleted by centrifugation for 10 minutes (13,000 rpm), then the pellet was resuspended
133 in 100 µl crystal violet stain (Pro Lab Diagnostics, Birkenhead, UK). Cells were counted
134 using a Petroff Hausser counting chamber (Hausser Scientific, Horsham, PA, USA) for
135 bacteria enumeration. After counting, the enumerated BCG stock (in MGIT culture fluid) was

136 stored at -20 °C in 1 ml aliquots until required. At this point, a 10 fold dilution series of BCG
137 cells was made in phosphate buffered saline.

138

139 Mock clinical samples were prepared by pooling ten anonymised infection negative sputum
140 samples (from asthmatic patients) ((see research ethics statement below). Pooled sputum was
141 liquefied by treatment with an equal volume of freshly prepared working strength Sputasol
142 (Oxoid, Thermo Scientific, Paisley, UK). The sputum was incubated at 37°C with occasional
143 vortexing, until liquefaction was complete. 1ml aliquots of the negative sputum samples were
144 spiked with cells from the BCG titration.

145

146 **Research Ethics Statement**

147 The protocol for this study was approved by London – Queen Square Research Ethics
148 Committee (17/LO/1420). Human samples were collected under approval of East Midlands
149 Research Ethics Committee (08/H0406/189) and all subjects gave written informed consent
150 in accordance with the Declaration of Helsinki.

151

152 **DNA Extraction directly from Mock Clinical Samples**

153 Each sample underwent a saline wash to remove extraneous human DNA. After
154 centrifugation at 13,200 rpm for 15 minutes the supernatant was discarded, then the pellet
155 was resuspended in 1 ml sterile phosphate buffered saline and centrifugation repeated. The
156 pellet was transferred in 100 µl molecular grade H₂O to a 0.5 ml plain skirted tube
157 (STARLAB, Hamburg, Germany) containing 0.8 g of aliquoted 0.1 mm silica beads (lysing
158 matrix B, MP Biomedicals, Santa Ana, USA). The mixture underwent bead beating (3x40s, 3
159 minute interval, 6.0 m/s) on a Fast Prep-24 machine (MP Biomedicals) followed by
160 centrifugation at 13,200 rpm for 10 minutes. DNA was recovered from 50 µl supernatant

161 using 1.8x volume magnetic AMPure XP beads (Beckman Coulter, High Wycombe, UK).
162 After vortexing for 20 seconds, and magnetic separation for 10 minutes, the supernatant
163 above the beads was replaced with 200µl of 80% EtOH. This was removed after 1 minute and
164 the wash step repeated, after which the beads were air dried for 10 minutes. DNA was eluted
165 in 26µl of 1 x TE buffer (pH8, Sigma Aldrich, Dorset, UK). DNA concentration, integrity
166 and fragment size were measured by Qubit Fluorometer (Rugby, UK) and TapeStation
167 (Stockport, UK) respectively.

168

169 **Multiple Displacement Amplification by Phi29 DNA polymerase**

170 DNA (1 ng/10 µl) extracted from mock clinical samples was denatured by incubation for 3
171 minutes at 96 °C, then transferred to ice, where the rest of the MDA reaction was assembled.
172 The final 20 µl reaction comprised 1x phi29 DNA polymerase reaction buffer (New England
173 BioLabs, Hitchin, UK), 0.1 mg/ml BSA, 5 mM dNTPs, 10 µM oligonucleotide primers (see
174 below) with a modified 3'-terminal endonuclease resistant phosphorothioate bond, 5 mM
175 MgCl₂, and 2 µl Phi29 DNA polymerase (20 units) (New England Biolabs). Two alternative
176 primers were tested; 'random' hexamers containing 65% GC, or 'most frequent 10mers'
177 based on the most frequent 10 bp sequence repeats identified in Mycobacteria genome (S1
178 Table). Incubation was at 30 °C for 16 hours. Amplified DNA was purified using AMPure
179 XP beads, quantitated, and 1 µg was digested to remove branched structures using 1 µl T7
180 Endonuclease I (New England BioLabs, Hitchin, UK) in 20 µl reaction volume at 37°C for 1
181 hour, followed by a second AMPure XP bead purification step.

182

183 **Oxford Nanopore Technology (ONT) Sequencing Library Preparation and Sequencing**

184 Digested DNA (700-900 ng/µl) was prepared for ONT sequencing according to the
185 manufacturer's 1D Native barcoding genomic DNA protocol using SQK-LSK108 and EXP-

186 NBD103 kits (Oxford Nanopore Technology, Oxford, UK). Each sequencing library
187 comprised seven barcoded DNA samples and was sequenced using MinION R9.4 SpotON
188 flow cells for 48 hours.

189

190 **Bioinformatics**

191 MinION reads were basecalled using Albacore v2.0.2 (Oxford Nanopore Technology,
192 Oxford, UK). We used Porechop (v0.2.2, <https://github.com/rrwick/Porechop>) to perform
193 stringent barcode demultiplexing of the sequencing data. Porechop confirms the presence of
194 the barcode sequence at both the start and end of each read; reads were acceptable only if the
195 same barcode was found at both ends, otherwise the read was discarded. This level of
196 stringency was achieved by setting the “require_two_barcodes” option in Porechop and
197 setting the threshold for the barcode score at 60. The basic statistics of the sequencing data
198 were reported using NanoPack [24]. Then, reads from each sample were mapped to the BCG
199 reference sequence (GenBank AM408590; the 16S rRNA region {1498360, 1499896} was
200 masked) using Minimap2 [25]. Integrative Genomics Viewer [26] was used to view the
201 resulting alignment profiles. The number of reads (i) mapped to the BCG reference, (ii)
202 fitting the definition of supplementary alignments (as below), and (iii) alignment length were
203 analysed using Pysam (<https://github.com/pysam-developers/pysam>). Repeated BCG-derived
204 sequences were found within the contiguous sequence of certain individual reads. These
205 reads therefore could not be aligned linearly to the BCG reference. One of the linear BCG
206 repeats within such reads was referred to as the “representative alignment” and the additional
207 repeats were referred to as “supplementary alignment(s)”. ‘Supplementary alignments’ were
208 considered to be present if the start and end of their BCG alignment positions occurred within
209 10 bp of the representative alignment, ie up to 10 bp could occur between the repeated
210 sequences. The histogram for the alignment length against the read length and the number of

211 repeats in each read was plotted by using ggplot2 implemented in R ([https://www.r-](https://www.r-project.org/)
212 [project.org/](https://www.r-project.org/)).

213 **RESULTS**

214 **Experimental Design**

215 The range of Mycobacteria cell concentrations typically found in *Mycobacterium*
216 *tuberculosis*-positive sputum [27] were represented in standardised mock clinical samples,
217 the BCG dilution series ranging from 10^6 to 10^1 BCG cells per ml liquefied sputum. Total
218 DNA was extracted from these sputum samples and a negative sputum control. The DNA
219 was sequenced on the R9.4 flow cell, to establish the proportions of BCG and other DNAs
220 present (Table 1: no amplification). This extracted DNA then formed the template for testing
221 molecular methods based on differently primed phi29 amplification reactions, aiming to
222 selectively enrich for BCG DNA within the total (Table 1: amplification).

223 **Table 1: Comparison of sequence data obtained for BCG-spiked sputum samples; with and without prior amplification.**
 224 Number of reads, mapped reads, genome coverage, supplementary alignment, and mean alignment length for un-amplified samples and phi29
 225 amplified samples (different primers) at the concentrations of BCG cells shown. Supplementary alignments occurred when the contiguous
 226 sequence of an individual read could not be aligned linearly to the reference sequence. Thus, one of the linear alignments in a repeat-containing
 227 read is referred as the “representative alignment” and the others (repeats of this sequence) are referred as “supplementary alignment(s)”. * ratio
 228 not given when the number of mapped read is less than 10.
 229

Amplification	BCG concentration (cells/ml)	Number of reads	Mapped reads	Genome coverage (%)	Supplementary alignment (ratio to mapped reads*)	Mean Alignment length
None	10 ⁶	71,029	8,068	89.4	125 (0.0)	1305
	10 ⁵	38,752	491	13.3	2 (0.0)	1276
	10 ⁴	9,543	11	0.3	0 (0.0)	1106
	10 ³	98,845	20	0.5	0 (0.0)	1047
	10 ²	96,293	2	0.2	1	2402
	10 ¹	32,743	0	0.1	0	
	0	70,929	4	0.2	0	1168
Phi29 65% GC primers	10 ⁶	144,331	28,617	4.1	12,0415 (4.2)	205
	10 ⁵	126,490	7,359	1.2	27602 (3.8)	214
	10 ⁴	58,851	84	0.1	495 (5.9)	171
	10 ³	76,870	10	0.1	28 (2.8)	210
	10 ²	105,612	0	0.0	0	
	10 ¹	92,841	1	0.0	1	162
	0	139,173	1	0.0	2	309
Phi29 MF 10mer primers	10 ⁶	201,038	31,893	4.6	54,651 (1.7)	181
	10 ⁵	112,595	7,405	1.6	14,759 (2.0)	322
	10 ⁴	148,550	2	0.0	0	342
	10 ³	148,612	4	0.0	1	116
	10 ²	186,505	1	0.0	1	145
	10 ¹	91,657	0	0.0	0	
	0	275,190	2	0.0	1	102

231 **Selective Enrichment of Mycobacteria sequences using Phi29 DNA polymerase.**

232 The high GC content of the Mycobacteria genome (for *M. tuberculosis* 65.6% GC) [28]
233 relative to most of the human genome (<50% GC for ~92% of the genome and 50-60% GC in
234 ~7% genome [29]) was exploited in our experimental design. MDA was primed using 65%
235 GC biased ‘random’ hexamers, or MF (most frequent) 10mer primers (S1 Table). An
236 unamplified DNA control was included at each sample concentration. Amplification products
237 (or control DNA) (200ng) were sequenced using the Oxford Nanopore Technologies (ONT)
238 MinION R9.4 platform (see S2 Table for summary of sample DNA concentration,
239 amplification, sequencing library preparation, and statistics of sequencing data).

240

241 For the unamplified BCG-spiked samples, the percentage of total DNA reads which mapped
242 to the BCG reference increased with increasing BCG cell concentration (Table 1). For
243 example, at 10^6 cells/ml, 11.4% of total reads mapped to the BCG reference and 89.4% of the
244 BCG genome was covered. The comparable results for Phi29 amplified (65% GC hexamer
245 primed) extracted DNA showed an increased proportion of BCG reads within the total -
246 19.8% of the total reads. However, BCG genome coverage was much lower, at 4.1% (Table 1
247 upper panel).

248

249 The BCG reference-mapped alignment profiles of samples which had undergone Phi29
250 amplification (65% GC hexamer primed) contained a large number of ‘supplementary
251 alignments’ (as defined in Materials and Methods) (Table 1). This repeat feature was virtually
252 absent in the sequence data from non-Phi29 amplified samples. The ratios between Phi29
253 amplified sequences forming representative and supplementary alignments were 4.2 (at 10^6
254 BCG cells/ml), 3.8 (10^5), 5.9 (10^4), and 2.8 (10^3), respectively (Table 1). The mean BCG

255 alignment length (about 200 nucleotides) in the 65% GC hexamer-primed Phi29 amplified
256 sequences was considerably shorter than that of the standard, unamplified sequence data.

257

258 **Explanation for biased Phi29 amplification**

259 To explain these findings, we hypothesised that Phi29 DNA Polymerase enriched the samples
260 in terms of their total BCG derived DNA content, but amplified short sequences to high
261 depths, ie the reads that mapped to the BCG reference in the Phi29 amplified data comprised
262 short repeated sequences potentially transcribed from the same fragment of DNA.

263

264 Visualization of the mapping profile revealed that BCG-like reads were split into multiple
265 small fragments which each mapped to same region of the reference genome (Fig 1). More
266 than 67% of the reads (at 10^6 BCG cells/ml) which mapped to the BCG reference comprised
267 repeats. Two, three, four, and five repeats were observed in 15.7%, 11.3%, 8.4%, and 6.3% of
268 the reads, respectively.

269

270 **Fig 1.**

271 **Analysis of reference BCG mapping profile for Phi29 amplified sequence data at 10^6**
272 **cells/ml BCG.**

273 (A) Plot of alignment length against read length.

274 (B) Histogram showing number of repeats in each read that mapped to BCG reference.

275 (C) The same histogram as (B) but focusing only on the number of repeats within the range
276 from 1 to 20 per read.

277

278 The overall GC content of the Phi29 '65% GC hexamer-primed' amplified sequences was
279 very close to the BCG average of 65.6%. The BCG sequences amplified from the 10^5 and 10^6

280 spiked samples contained GC 65.5% (within 1.2% genome coverage) and 65.7% (in 4.1%
281 genome coverage) respectively, compared to 65.6% across the whole of the genome. This
282 suggests there was no amplification bias towards BCG sequences which have GC content
283 away from the mean, and that the repeatedly amplified BCG sequences were not unusual in
284 terms of their overall GC content. The distribution of the most abundant Phi29 amplification
285 products across the sequence of the BCG reference genome also indicated that obvious
286 amplification hot spots were absent (Fig 2).

287

288 **Fig 2.**

289 **Distribution of most abundant Phi29 amplification products relative to the BCG**
290 **reference genome indicates the absence of obvious amplification hot spots.**

291 Comparison of the regions of the BCG genome amplified by Phi29 primed either using 65%
292 GC hexamers (phi) or most frequent 10mers (mf) at 10^6 and 10^5 BCG cells/ml. The top
293 10% nucleotide positions with highest depth of mapping coverage are shown.

294

295 **BCG Reads detected in the Negative Control/Low Concentration Spikes**

296 The negative control sample (negative sputum with zero BCG cells added) contained a small
297 number of reads (less than five) which mapped to the BCG reference (Table 1). This also
298 occurred in sputum samples spiked at low BCG concentrations (10^2 and 10^1 cells/ml) in both
299 the phi29 amplified and unamplified control sequence data. All samples had been sequenced
300 while ‘multiplexed’ – the addition of a barcode sequence to each sample during library
301 preparation allowed ‘de-multiplexing’ to be performed bioinformatically after sequencing.
302 Despite the fact that we implemented stringent bioinformatic barcode removal for de-
303 multiplexing, which successfully removed most of this cross-contamination, a low level

304 remained. This issue was confirmed to be bioinformatics-based, when samples were run of
305 single flow cells (not multiplexed).

DISCUSSION

Multiple Displacement Amplification of DNA by Phi29 polymerase is an attractive choice for experiments aiming to generating large quantities of DNA ($\geq\mu\text{g}$) from very small ($\leq\text{ng}$) amounts under isothermal conditions [18]. Advantages include a low error rate due to 3',5'-exonuclease 'proofreading' activity (error rate $\sim 9.5 \times 10^{-6}$), the capacity to synthesise DNA molecules $>70\text{kb}$ long and the possibility of virtually whole genome amplification [19, 30-32]. Relative to PCR-based methods, more DNA is amplified by at least an order of magnitude, and good genome coverage and reduced amplification bias of genomic DNA from human cells has been reported [33]. Long DNA fragments provide ideal input for the Nanopore MinION sequencing platform, which in turn generates long reads offering the possibility of *de novo*, rather than reference based genome assemblies.

MDA has also shown promise for the accurate and unbiased amplification of whole bacterial genomes from uncultivable, or slow growing species, and even 'single cell genomics' [34, 35]. MDA with random hexamers has been used to amplify *Xylella fastidiosa* (Gram negative plant pathogen, 52% CG content) DNA directly from approximately 1000 target cells, yielding over 4 μg of high molecular weight DNA and achieving uniform genome coverage relative to unamplified DNA [34]. Coverage of *Coxiella burnetii* (fastidious obligate intracellular pathogen, 42.5% GC) was similarly representative, as assessed by PCR [36]. Work in our laboratory aiming to sequence Mycobacteria directly from sputum samples has previously used 3% NaOH (Nac-Pac Red; Alpha-Tec Systems, Vancouver, WA, USA) to deplete sputum of non-Mycobacteria, together with a 'Molysis kit' (Molzym Life Science, Bremen, Germany) to reduce human DNA contamination [11]. An important issue arising from these pre-treatments is that for most samples, insufficient DNA remains for direct sequencing using the Nanopore MinION. Here, we aimed to investigate the possibility of

eliminating the need for such pre-treatment, while amplifying microgram quantities of DNA enriched for Mycobacteria sequences.

Our experiments employed 65% GC biased hexamers to favour amplification of the BCG genome (65% GC content) relative to the human genome (CG content <50% for ~92% of the genome and 50-60% CG for ~7% genome [29]). This achieved two to five fold enrichment for BCG sequences (Table 1) but at the expense of genome coverage (for example 89.4% genome coverage decreased to 4.1% at the 10^6 BCG spike concentration, Table 1). Post amplification, certain regions of the genome were covered at extremely high depth. The reason for the high coverage in certain regions, but not others is unknown. It may be unrelated to the mean GC content of these sequences, because this was the same within amplified sequences as the mean for the whole genome. The absence of obvious amplification hotspots conserved between experiments (Fig 2) suggests regions of high coverage may occur stochastically.

The difficulty of amplifying GC rich sequences from a complex mixture by MDA has been reported previously [37]; species with the highest GC content underwent significantly less amplification from an environmental (soil) sample compared to low GC bacteria. Yilmaz et al. [38] evaluated three different commercially available kits, including NEB Phi29 used in our study. They also observed amplification bias against high (G+C)-content templates in bacteria amplified from sludge and compost communities. Our use of 65% GC biased hexamers (also MF10mers Table 1) with the NEB Phi29 polymerase was insufficient to achieve unbiased amplification of the GC rich BCG genome. Similar bias against GC rich sequences has been observed previously [39]; MDA of DNA extracted from tumour samples reproducibly distorted gene dosage representation in the amplified DNA, reflecting the GC

content of different regions of the template. Also, a study of copy number variants within the human genome created hundreds of potentially confounding MDA artefacts that could obscure authentic copy number variants, which were reproducible and influenced by GC content [40]. There is also evidence of stochastic effects originating from the amplification of very low amounts of genomic template from a single bacterium [41] - locus representation values ranged from 0.1% to 1,211%.

The reason MDA is biased against GC rich templates is unclear, but it could reflect the higher melting temperature of GC rich DNA relative to AT rich sequences. In addition to the conditions described above, we tested reaction conditions which are known to alleviate GC-melting related issues in PCR, by effectively reducing the melting temperature of the DNA (PCR additives Q-solution and DMSO), as well as increasing the incubation temperature to 35 °C and 40 °C. A novel thermostable mutant of Phi29 polymerase, designated WGA-X (Thermofisher) has been described which amplifies DNA at 45°C, [42] and offers improved amplification of high GC content templates, but it was not commercially available. We also tested the Phi29-polymerase based Qiagen REPLI-g kit (data not shown) because it uses alkaline DNA denaturation to improve the uniformity of DNA denaturation while minimising DNA fragmentation or generation of abasic sites (relative to heat denaturation), and because it's been reported to work at 40°C [43]. This kit was also tested by Yilmaz et al. [38] and it performed best with respect to GC bias. Unfortunately, none of these modifications improved the genome coverage achieved in our study. Further experiments with shorter amplification incubation times were also performed, with the aim of potentially reducing the amplification bias, but these were unsuccessful.

The challenges of amplifying a minority, GC rich target DNA from within a complex mixture remain. Here, establishing mock clinical samples containing defined numbers of BCG cells represented a key step forward, because the data from different method development experiments could be compared. This material is proving invaluable in further work aiming to optimise ‘direct from sample’ Mycobacteria genome sequencing. In conclusion, optimal conditions under which Phi29 polymerase might be directly amplify minority GC rich templates without bias, remain to be identified.

ACKNOWLEDGEMENTS

The views expressed in this publication are those of the authors and not necessarily those of the NHS, NIHR, the Department of Health or Public Health England.

DWC and TEAP are NIHR senior investigators.

AUTHOR CONTRIBUTIONS

Conceptualization and Methodology: Sophie George, Yifei Xu, Alasdair TM Hubbard, Louise Pankhurst, Sarah J Hoosdally, Dona Foster, A. Sarah Walker, Timothy EA Peto, Derrick W Crook, Kate. E Dingle.

Resources: Samantha Thulborn, Esther Robinson, E Grace Smith, Priti Rahood, Marcus Morgan.

Investigation: Sophie George, Yifei Xu, Alasdair TM Hubbard, David T Griffiths, Marcus Morgan, Kate E Dingle.

Formal Analysis: Yifei Xu, Nicholas Sanderson, Timothy EA Peto.

Software: Yifei Xu, Nicholas Sanderson.

Visualisation: Sophie George, Yifei Xu, Timothy EA Peto, Kate E Dingle.

Writing – Original Draft Preparation: Kate E Dingle.

Writing – Review & Editing: Sophie George, Yifei Xu, Alasdair TM Hubbard, David T Griffiths, Marcus Morgan, Louise Pankhurst, Sarah J Hoosdally, Dona Foster, Samantha Thulborn, Esther Robinson, E Grace Smith, Priti Rahood, A. Sarah Walker, Timothy EA Peto, Derrick W Crook, Kate. E Dingle.

CONFLICT OF INTERESTS

The authors declare no conflicts of interest.

REFERENCES

1. World Health Organization Global Tuberculosis Report 2017. Available from: <http://apps.who.int/iris/bitstream/handle/10665/259366/9789241565516-eng.pdf;jsessionid=CD9974AE6BF74DBAA832955DAEC120E7?sequence=1>
2. Wassilew N, Hoffmann H, Andrejak C, Lange C. Pulmonary Disease Caused by Non-Tuberculous Mycobacteria. *Respiration*. 2016;91: 386-402.
3. van der Werf MJ, Ködmön C, Katalinić-Janković V, Kummik T, Soini H, Richter E, et al. Inventory study of non-tuberculous mycobacteria in the European Union. *BMC Infect Dis*. 2014;14:62.
4. van Ingen J, Hoefsloot W, Dekhuijzen PN, Boeree MJ, van Soolingen D. The changing pattern of clinical *Mycobacterium avium* isolation in the Netherlands. *Int J Tuberc Lung Dis*. 2010;14: 1176-1180.
5. Pankhurst LJ, Del Ojo Elias C, Votintseva AA, Walker TM, Cole K, Davies J, et al. Rapid, comprehensive, and affordable mycobacterial diagnosis with whole-genome sequencing: a prospective study. *Lancet Respir Med*. 2016;4: 49-58.
6. Walker TM, Cruz ALG, Peto TE, Smith EG, Esmail H, Crook DW. Tuberculosis is changing. *Lancet Infect Dis*. 2017;17: 359-361.
7. Votintseva AA, Pankhurst LJ, Anson LW, Morgan MR, Gascoyne-Binzi D, Walker TM, et al. Mycobacterial DNA extraction for whole-genome sequencing from early positive liquid (MGIT) cultures. *J Clin Microbiol*. 2015;53: 1137-1143.
8. Walker TM, Kohl TA, Omar SV, Hedge J, Del Ojo Elias C, Bradley P, et al. Whole-genome sequencing for prediction of *Mycobacterium tuberculosis* drug susceptibility and resistance: a retrospective cohort study. *Lancet Infect Dis*. 2015;15: 1193-1202.

9. Walker TM, Ip CL, Harrell RH, Evans JT, Kapatai G, Dediccoat MJ et al. Whole-genome sequencing to delineate *Mycobacterium tuberculosis* outbreaks: a retrospective observational study. *Lancet Infect Dis.* 2013;13: 137-146.
10. Walker TM, Lalor MK, Broda A, Ortega LS, Morgan M, Parker L, et al. Assessment of *Mycobacterium tuberculosis* transmission in Oxfordshire, UK, 2007–12, with whole pathogen genome sequences: an observational study. *Lancet Respir Med.* 2014;2: 285-292.
11. Votintseva AA, Bradley P, Pankhurst L, del Ojo Elias C, Loose M, Nilgiriwala K, et al. Same-day diagnostic and surveillance data for tuberculosis via whole-genome sequencing of direct respiratory samples. *J Clin Microbiol.* 2017;55: 1285-1298.
12. Doughty EL, Sergeant MJ, Adetifa I, Antonio M, Pallen MJ. Culture independent detection and characterisation of *Mycobacterium tuberculosis* and *M. africanum* in sputum samples using shotgun metagenomics on a benchtop sequencer. *PeerJ* 2014;2: e585.
13. Brown AC, Bryant JM, Einer-Jensen K, Holdstock J, Houniet DT, Chan JZ, et al. Rapid whole-genome sequencing of *Mycobacterium tuberculosis* isolates directly from clinical samples. *J Clin Microbiol.* 2015;53: 2230-2237.
14. Nimmo C, Doyle R, Burgess C, Williams R, Gorton R, McHugh TD, et al. Rapid identification of a *Mycobacterium tuberculosis* full genetic drug resistance profile through whole genome sequencing directly from sputum. *Int J Infect Dis.* 2017;62: 44-46.
15. Jain M, Koren S, Miga KH, Quick J, Rand AC, Sasani TA, et al. Nanopore sequencing and assembly of a human genome with ultra-long reads. *Nat Biotechnol.* 2018;36: 338-345.
16. Loman NJ, Quick J, Simpson JT. A complete bacterial genome assembled de novo using only nanopore sequencing data. *Nat Methods.* 2015;12: 733-735.
17. Wick RR, Judd LM, Holt KE. Comparison of Oxford Nanopore basecalling tools. 2018. Available from: <https://github.com/rrwick/Basecalling-comparison#references>.

18. Lasken RS, Egholm, M. Whole genome amplification: abundant supplies of DNA from precious samples or clinical specimens. *Trends Biotechnol.* 2003;21: 531-535.
19. Blanco L, Bernad A, Lázaro JM, Martín G, Garmendia C, Salas M. Highly efficient DNA synthesis by the phage phi 29 DNA polymerase. Symmetrical mode of DNA replication. *J Biol Chem.* 1989;264: 8935-8940.
20. Dean FB, Nelson JR, Giesler TL, Lasken RS. Rapid amplification of plasmid and phage DNA using phi29 DNA polymerase and multiply-primed rolling circle amplification. *Genome Res.* 2001;11: 1095-1099.
21. Huang L, Ma F, Chapman A, Lu S, Xie XS. Single-Cell Whole-Genome Amplification and Sequencing: Methodology and Applications. *Annu Rev Genomics Hum Genet.* 2015;16: 79-102.
22. Fine PEM, Carneiro IAM, Milstien JB, Clemens CJ. Issues Relating to the Use of BCG in Immunization Programmes. A Discussion Document. Geneva: World Health Organization. 1999. Available from:
http://apps.who.int/iris/bitstream/handle/10665/66120/WHO_V_B_99.23.pdf?sequence=1&isAllowed=y
23. Caceres N, Vilaplana C, Prats C, Marzo E, Llopis I, Valls J, et al. Evolution and role of corded cell aggregation in *Mycobacterium tuberculosis* cultures. *Tuberculosis* 2013;93: 690-698.
24. De Coster W, D'Hert S, Schultz DT, Cruts M, Van Broeckhoven C. NanoPack: visualizing and processing long-read sequencing data. *Bioinformatics (Internet).* 2018. Available from: <http://dx.doi.org/10.1093/bioinformatics/bty149>.
25. Li H. "Minimap2: pairwise alignment for nucleotide sequences." *Bioinformatics* 2018 1: 7.

26. Robinson JT, Thorvaldsdóttir H, Winckler W, Guttman M, Lander ES, Getz G, et al. Integrative genomics viewer. *Nat Biotechnol.* 2011;29: 24-26.
27. CDC (Centers for Disease Control and Prevention). Core Curriculum on Tuberculosis: What the Clinician Should Know. Chapter 4: Diagnosis of Tuberculosis Disease. Available from: <https://www.cdc.gov/tb/education/corecurr/>.
28. Cole ST, Brosch R, Parkhill J, Garnier T, Churcher C, Harris D, et al. Deciphering the biology of *Mycobacterium tuberculosis* from the complete genome sequence. *Nature* 1998;393: 537-544.
29. Vente JC, Adams M, Myers EW, Li PW, Mural RJ, Sutton GG, et al. The sequence of the human genome. *Science* 2001;291: 1304-1351.
30. Garmendia C, Bernad A, Esteban JA, Blanco L, Salas M. The bacteriophage phi 29 DNA polymerase, a proofreading enzyme. *J Biol Chem.* 1992;267: 2594-2599.
31. Esteban JA, Salas M, Blanco L. Fidelity of phi 29 DNA polymerase. Comparison between protein-primed initiation and DNA polymerization. *J Biol Chem.* 1993;268: 2719-2726.
32. Paez JG, Lin M, Beroukhir R, Lee JC, Zhao X, Richter DJ, et al. Genome coverage and sequence fidelity of phi29 polymerase-based multiple strand displacement whole genome amplification. *Nucleic Acids Res.* 2004;32 :e71.
33. Hosono S, Faruqi AF, Dean FB, Du Y, Sun Z, Wu X, et al. Unbiased whole-genome amplification directly from clinical samples. *Genome Res.* 2003;13: 954-964.
34. Detter JC, Jett JM, Lucas SM, Dalin E, Arellano AR, Wang M, et al. Isothermal strand-displacement amplification applications for high-throughput genomics. *Genomics.* 2002;80: 691-698.
35. Stepanauskas R. Single cell genomics: an individual look at microbes. *Curr Opin Microbiol.* 2012;15: 613-620.

36. Kumar S, Gangoliya SR, Berri M, Rodolakis A, Alam SI. Whole genome amplification of the obligate intracellular pathogen *Coxiella burnetii* using multiple displacement amplification. *J Microbiol Methods*. 2013;95: 368-72.
37. Direito SO, Zaura E, Little M, Ehrenfreund P, Röling WF. Systematic evaluation of bias in microbial community profiles induced by whole genome amplification. *Environ Microbiol*. 2014;16: 643-657.
38. Yilmaz S, Allgaier M, Hugenholtz P. Multiple displacement amplification compromises quantitative analysis of metagenomes. *Nat Methods* 2010;7: 943-944.
39. Bredel M, Bredel C, Juric D, Kim Y, Vogel H, Harsh GR, et al. Amplification of whole tumor genomes and gene-by-gene mapping of genomic aberrations from limited sources of fresh-frozen and paraffin-embedded DNA. *J Mol Diagn*. 2005;7: 171-182
40. Pugh TJ, Delaney AD, Farnoud N, Flibotte S, Griffith M, Li HI, et al. Impact of whole genome amplification on analysis of copy number variants. *Nucleic Acids Res*. 2008;36: e80.
41. Raghunathan A, Ferguson HR, Bornarth CJ, Song W, Driscoll M, Laske RS. Genomic DNA amplification from a single bacterium. *Appl Environ Microbiol*. 2005;71: 3342-3347.
42. Stepanauskas R, Fergusson EA, Brown J, Poulton NJ, Tupper B, Labonté JM, et al, Improved genome recovery and integrated cell-size analyses of individual uncultured microbial cells and viral particles. *Nat Commun*. 2017;8: 84.
43. Alsmadi O, Alkayal F, Monies D, Meyer BF. Specific and complete human genome amplification with improved yield achieved by phi29 DNA polymerase and a novel primer at elevated temperature. *BMC Res Notes*. 2009 24;2: 48.

Figure 1

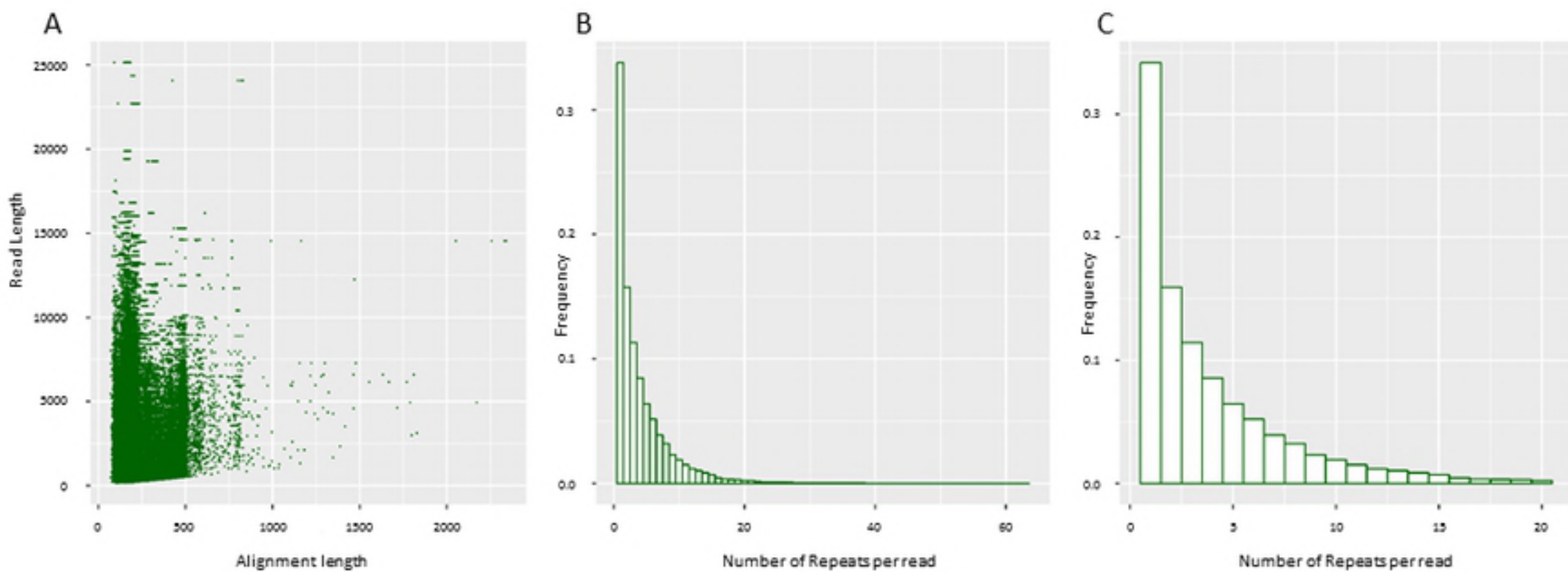


Figure 1

Figure 2

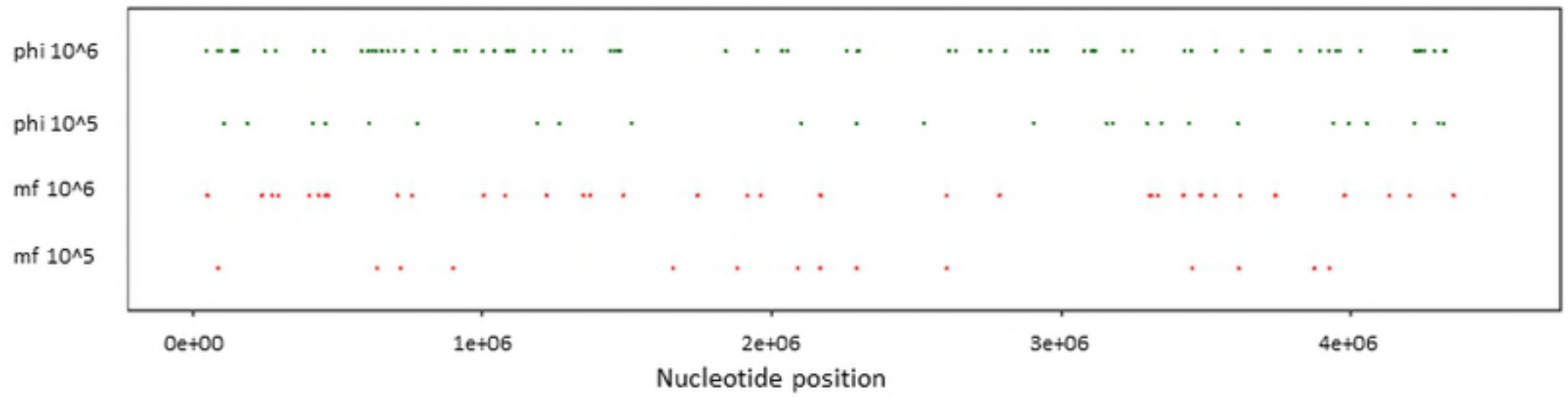


Figure 2