

1 **Enrichment of long DNA fragments from mixed samples for Nanopore sequencing**

2 **Eckert SE¹, Chan JZ-M¹, Houniet D¹, the PATHSEEK consortium², Breuer J³, Speight G¹**

3

4 Author affiliations: 1) Oxford Gene Technology, Begbroke Science Park, Begbroke Hill, Woodstock
5 Road, Begbroke, Oxfordshire OX5 1PF UK; 2) a list of participants can be found in the
6 Acknowledgement; 3) UCL Division of Infection & Immunity, Cruciform Building, Gower Street,
7 London WC1E 6BT UK

8 Correspondence to: Graham.Speight@ogt.com

9

10 **Abstract**

11 Whole-genome sequencing of pathogenic organisms directly from clinical samples combines
12 detection and genotyping in one step. This can speed up diagnosis, especially for slow-growing
13 organisms like *Mycobacterium tuberculosis* (*Mtb*), which need considerable time to grow in
14 subculture, and can provide vital information for effective personalised treatment. Within the
15 PATHSEEK project, we have developed a bait-capture approach to selectively enrich DNA/RNA from
16 specific bacterial and viral pathogens present in clinical samples. Here, we present a variation of the
17 method that allows enrichment of large fragments of target DNA for sequencing on an Oxford
18 Nanopore MinION™ sequencer. We enriched and sequenced cDNA from Influenza A (FluA), genomic
19 DNA (gDNA) from human cytomegalovirus (CMV) and from two strains of *Mtb*, and present an
20 evaluation of the method together with analysis of the sequencing results from a MinION™ and an
21 Illumina MiSeq sequencer. While unenriched FluA and CMV samples had no reads matching the
22 target organism due to the high background of DNA from host cell lines, enriched samples had 56.7%
23 and 90.9% on-target reads respectively for the best quality Nanopore reads.

24 Introduction

25 Amidst increased occurrence of extensive or totally drug-resistant pathogens and antibiotic
26 overuse¹, high-throughput sequencing methods, particularly those which can be readily applied in a
27 clinical setting, are being used to aid and refine diagnosis in a timely fashion². Data from whole-
28 genome sequencing provides a wealth of information such as identification of resistance markers
29 carried by the infecting agent(s), allowing for rapid, targeted and personalised treatment. However,
30 DNA extracted from clinical samples consists of a mixture of low amounts of pathogen genetic
31 material and overwhelming quantities of human and commensal DNA. Previous studies have shown
32 that it is possible to bypass the traditional culture-based diagnosis and obtain informative sequence
33 data from metagenomic samples, but the throughput is low and the method prohibitively costly³.
34 The EU-funded FP7 PATHSEEK project (<http://www.pathseek.eu/>) has developed a disruptive
35 diagnostic platform to sequence bacterial and viral pathogens directly from clinical samples^{4, 5}. This
36 enrichment approach employs custom baits to capture genomic material from the target pathogens,
37 thereby removing the human and commensal DNA, and allowing greater throughput of samples on
38 Illumina sequencers. However, as this method is optimised for short-read sequencers such as the
39 Illumina MiSeq and the Ion PGM, it is very difficult to resolve highly repetitive regions such as those
40 found in cytomegalovirus⁶, or provide evidence of recombination events such as those seen in
41 *Chlamydia trachomatis*. Members of the PATHSEEK consortium joined the Oxford Nanopore
42 Technologies (ONT, Oxford, UK) MinION™ access program to assess the suitability of its long-read
43 platform, the MinION™, for the targeted enrichment method.

44 Here, we present an adaption to the PATHSEEK method – enrichment of DNA fragments of between
45 1 and 15 kb for sequencing on long-range platforms. We compare sequence data from unenriched
46 and enriched cultured FluA and CMV samples, run on the MinION™ and Illumina (San Diego, CA,
47 USA) MiSeq platforms. We also mixed cultured *Mtb* gDNA from two different strains with human
48 DNA to assess the efficiency of enrichment by hybridisation for longer bacterial DNA fragments. Long

49 genomic fragments were readily purified from a background of the cell line used for producing the
50 viruses, or, in case of *Mtb*, admixed human DNA.

51

52

53

54 **Materials and Methods**

55 **Samples**

56 Mycobacterial gDNA from strains H37Rv and the extensively drug-resistant clinical strain *MtbC* were
57 kind gifts from A. Brown, L. J Schreuder, T Parish (Barts and The London School of Medicine and
58 Dentistry, Queen Mary University of London, London, UK), P. Butcher and J. Dhillon⁷ (Institute of
59 Infection and Immunity, St. George's Hospital, University of London, UK). To simulate clinical
60 samples, *Mtb* DNA was mixed with human gDNA (Male, #G1471, Promega, Madison WI, USA), to
61 10% (450 ng human DNA, 50 ng *Mtb* DNA) or 90% (450 ng *Mtb*, 50 ng human) prior to processing.

62 RNA from Influenza strain A A/PR/8/34 #0111041v H1N1, grown in MDCK Cocker Spaniel kidney cell
63 line, was obtained from the Public Health England Culture Collection (Porton Down, UK), and reverse
64 transcribed with NEBNext RNA First Strand Synthesis Module #E7525 and NEBNext mRNA Second
65 Strand Synthesis Module #E6111 (New England Biolabs, Hitchin, UK) according to the manufacturer's
66 instructions. The sample was cleaned up with DNA Clean & Concentrator columns (#D4013, Zymo
67 Research, Irvine CA, USA).

68 DNA from CMV strain Merlin grown in fibroblast cell culture was subjected to PreCR (#M0309, New
69 England Biolabs, Ipswich MA, USA) enzymatic repair according to the manufacturer's
70 recommendations after shearing. It was cleaned up with 100 µl (1 volume) Agencourt AMPure XP

71 beads (#A63880, Beckman Coulter High Wycombe, UK), according to the manufacturer's
72 instructions, and eluted in 25 μ l H₂O.

73

74 **Sample preparation and long-fragment hybridisation**

75 CMV and *Mtb* Samples (500 ng) were diluted in TE to an end volume of 80 μ l, and sheared in Covaris
76 g-TUBEs (#520079, Covaris, Woburn MA, USA) with two passages at 7200 rpm/4200 g for 1 min in a
77 desktop centrifuge (#5242, Eppendorf, Hamburg, Germany). FluA samples were not sheared as the
78 cDNA fragments, derived from eight segments of negative-sense RNA of approximately 700 – 2300
79 nucleotides (nt), were size-compatible with Nanopore sequencing.

80 Concentrations and fragment sizes were determined with a Qubit fluorometer (dsDNA BR Assay Kit
81 #Q32850, Life Technologies Ltd, Paisley, UK), and Agilent Tape Station (Genomic DNA ScreenTape
82 #5067-5365 and Genomic DNA Reagents #5067-5366, Agilent, Santa Clara CA, USA) according to
83 manufacturers' instructions.

84 PATHSEEK custom baits for the target organisms FluA, CMV and *Mtb* were designed using an in-
85 house Perl script developed by the PATHSEEK consortium, using a database of 4968 H1N1 and 2966
86 H3N2 FluA genomes, 115 partial and complete CMV genomes and the *Mtb* strain H37Rv reference
87 genome (AL123456.3) respectively. Sheared gDNA (CMV, *Mtb*) and cDNA (FluA) samples (500 ng)
88 were hybridised and captured using 2 μ l FluA and CMV baits, or 5 μ l for *Mtb* baits per reaction.
89 Hybridisation and washing were performed using a modified version of the SureSelect^{XT} Target
90 Enrichment for Illumina Paired-End Multiplexed Sequencing as described previously^{4,5}. The workflow
91 used in this study is outlined in Figure 1. Samples (30 μ l) were then heated to 95°C for 3 min, and
92 cooled to 35°C (ramp: 4°C/min) to release the target fragments from the baits and streptavidin
93 beads.

94 Hybridised samples were split into two aliquots and half was used for Nanopore library preparation
95 with ONT kit versions SQK-MAP003 for *Mtb* H37Rv and SQK-MAP004 for CMV, FluA, *MtbC*. The
96 remainder was used to generate Illumina-compatible libraries, using another modified version of the
97 SureSelect^{XT} Target Enrichment for Illumina Paired-End Multiplexed Sequencing (Figure 1).

98

99 **Nanopore library preparation, sequencing and analysis**

100 Following renaturation, end repair and dA-tailing were performed with enzymes from the
101 SureSelect^{XT} kit (#5500-0075, Agilent) as specified by the manufacturer. AMPure XP-purified, dA-
102 tailed samples were ligated for 15 min at room temperature in 50 µl volume with 20 µl PCR adapters
103 (ONT, sequencing kits SQK-MAP003 and SQK-MAP004), 5 µl 10x T4 DNA ligase reaction buffer and 2
104 µl T4 DNA ligase (Agilent). They were cleaned up with 90 µl AMPure XP beads and eluted in 50 µl
105 H₂O. This ligated DNA (48 µl) was amplified by long-range PCR with 2 µl PCR primers (ONT) and 50 µl
106 Long Amp Taq 2x Master mix (#M0287, NEB) with the following program: 95°C 3 min, 15-18 cycles of
107 [95°C 15 sec, 62°C 15 sec, 65°C 10 min], 65°C 20 min, 4°C hold. PCR reactions were cleaned up with
108 100 µl AMPure XP beads and eluted in 50 µl H₂O. Concentrations and fragment sizes were measured
109 with a Qubit fluorometer and Agilent Tape Station as before.

110 A second round of end repair and dA-tailing was performed on 1 µg of enriched, amplified PCR
111 product using SureSelect^{XT} reagents as described above, but without purification after dA-tailing.
112 Instead, leader/hairpin ligation and sample clean-up were performed according to the ONT protocol
113 for kit SQK-MAP003 (for strain *Mtb* H37Rv only) or SQK-MAP004, in protein LoBind tubes (1.5 ml,
114 #0030108116, Eppendorf). In detail, 10 µl adapter mix and 2 µl hairpin adapter (ONT) were
115 incubated for 10 min at room temperature with 30 µl dA-tailed sample and 50 µl 2x blunt/TA ligase
116 master mix (#M0367, NEB) in 100 µl volume. SQK-MAP003 libraries were cleaned up with AMPure
117 XP beads, SQK-MAP004 with Dynabeads for His-Tag isolation and pulldown (#10103D, Life

118 Technologies) according to the respective ONT protocols. Libraries were eluted from the beads at
119 room temperature for 10 min with 25 µl elution buffer (ONT). Library concentrations were typically
120 2-10 ng/µl, as assessed by Qubit fluorometer.

121 Before each MinION™ run, flowcells were quality-tested with the script MAP_Platform_QC
122 (MinKnow software version 0.46.2.8 to 0.49.2.9), then loaded with 6 µl library and 4 µl fuel mix in
123 140 µl EP buffer (ONT), and run with script MAP_48Hr_Sequencing_Run, typically for 24h. Reads
124 were analysed by the Metrichor 2D basecalling (versions 2.19 to 2.29) cloud-based platform, and the
125 resulting fast5 files ("pass" and "fail" quality) converted to fasta format with Poretools⁸. BLASR⁹ and
126 LAST¹⁰ were used to align reads to the pathogen reference sequences (human CMV herpes virus
127 HHV-5 GU179001.1, *Mtb* strain H37Rv NC_018143.2, and Influenza strain H1N1, A/Puerto
128 Rico/8/1934). Files were further tested with both aligners against background human
129 (Human_g1k_v37, www.1000genomes.org) or dog (Ensembl CanFam3.1 GCA_000002285.2;
130 NC_006583.3) sequences, and the ONT adapters used for PCR.

131 The FluA control sample that did not undergo hybridisation (75 ng) was PCR-amplified to 500 ng as
132 described previously, then prepared as recommended in the ONT Genomic DNA sequencing protocol
133 SQK-MAP004. For the non-hybridised CMV sample, 500 ng was used directly for Nanopore library
134 preparation (SQK-MAP004).

135

136 **Illumina library preparation from long, hybridisation-enriched fragments**

137 After the long-fragment hybridisation, samples of *Mtb* H37Rv, *MtbC*, CMV and FluA were sheared
138 with a Covaris AFA instrument (Covaris, Woburn MA, USA) to 200 nt fragment size and converted
139 into Illumina-compatible libraries using a modified SureSelect^{XT} protocol (Figure 1) and Agilent
140 reagents as before. Briefly, samples were end-repaired, dA-tailed, adapters ligated, and DNA
141 amplified (6 cycles) as described in the protocol. Following sample purification, the PCR products

142 were re-amplified using post-capture indexed PCR primers for a further 15 cycles. The resulting
143 indexed libraries were quantified by Qubit and Agilent Tape Station as before, and pooled.
144 Sequencing was performed on an Illumina MiSeq machine with paired-end 600V3 kits (#MS-102-
145 3003) with automatic adapter trimming. Results from the Illumina MiSeq runs were analysed with
146 the Picard pipeline (<http://broadinstitute.github.io/picard/>).

147

148

149

150 **Results**

151 **Comparison of Nanopore library size and read length**

152 CMV and *Mtb* gDNA was sheared to fragments of 10-15 kb using Covaris g-TUBEs. FluA cDNA was
153 used without previous shearing for both enriched and non-enriched Nanopore libraries. Table 1
154 shows the median fragment length of the input DNA after shearing, as determined on an Agilent
155 Tape Station. Interestingly, the distribution of the unsheared FluA cDNA on the Tape Station showed
156 distinct peaks of 120 nt, 300 nt, 500 nt, 800 nt, 1-3 kb, with fragments up to 15 kb; these correspond
157 to transcripts from the eight FluA genomic segments NC_002016 to NC_002023, and cell line DNA.
158 PCR-amplified samples had wide ranges of sizes both within and between individual reactions, with
159 PCR products about half the size of the original DNA used for hybridisation. The Nanopore reads
160 (raw data in the European Nucleotide Archive, Study PRJEB12651) were similarly variable in length,
161 as indicated by the size of the standard deviations in Table 1. “Pass” quality (both strands read while
162 passing through the nanopore, resulting in higher confidence) reads were, on average, 1.2 to 4.2x
163 the length of “fail” quality reads. PCR-amplified samples were considerably smaller than the original
164 DNA fragments. Sequenced reads were shorter still, but with a wide range, reflected in high

165 standard deviations in Table 1. Reads from non-hybridised samples were longer than enriched
166 samples, either due to fragment damage during the hybridisation and wash processes, or during
167 PCR, which preferentially amplifies shorter fragments.

168

169 **Table 1:** Average DNA fragment sizes. The table shows input, post-PCR library size, and average
170 Nanopore read length (“pass” and “fail” quality) with standard deviations (SD) of the samples used in
171 this study. The non-hybridised CMV sample was not PCR-amplified as enough material was available
172 to proceed directly to sequencing. *FluA samples were not sheared.

Samples/input sizes	Sheared (nt)	PCR-amplified (nt)	Sequenced pass reads (nt/nt SD)	Sequenced fail reads (nt/nt SD)
FluA non-hybridised	120, 300, 500, 800, 1000, 3000+*	99-4000	1598/1191	805/946
FluA	120, 300, 500, 800, 1000, 3000+*	370-4000	773/683	533/733
CMV non-hybridised	1960, 49000	-	3176/2291	487/1203
CMV	12500	1587, 5640	1528/975	1083/1099
<i>Mtb</i> H37Rv	13800	2000-7000	2402/1865	757/1855
<i>Mtb</i> C	15000	1500	759/355	596/713

173

174

175 Comparison of BLASR and LAST aligners

176 We used BLASR⁹ and LAST¹⁰ (with the settings used in Quick et al.¹¹) for the alignment of Nanopore
177 reads to their respective references (pathogen and human/dog cell line). We found that BLASR
178 alignment resulted in fewer reads, with higher identity to the reference strains, and lower standard
179 deviation. In contrast, the LAST aligner produced more reads aligning to the reference, with lower

180 identity and higher standard deviation. Table 2 shows statistics for the similarities to the target
181 references obtained with the two aligners.

182

183 **Table 2:** Average similarity and length (with standard deviations, SD) of Nanopore reads aligned to
184 the pathogen targets using BLASR and LAST.

Sample	BLASR alignment		LAST alignment	
	Similarity of reads to target (%/SD %)	Length of reads (nt/SD nt)	Similarity of reads to target (%/SD %)	Length of reads (nt/SD nt)
FluA hybridised	80/6.3	185/109	74.9/6.9	344/150
CMV hybridised	79.6/6.4	940/835	72.1/7.9	1576/1102
<i>MtbC</i> hybridised	80.3/6	305/224	71.8/8.1	628/958
<i>Mtb</i> H37Rv hybridised	77.9/5.7	1182/1056	71.9/8.3	1744/1255

185

186 Most “pass” quality reads aligned to either the target organism or the respective cell line, whereas
187 most “fail” quality reads did not match to either target, cell line, or sequences in the PubMed
188 Nucleotide database (November 2015). A small number of FluA (85), CMV (315) and *MtbC* (10)
189 Nanopore reads matched to both target pathogen and cell line, predominantly in the output of the
190 LAST aligner (Table 3). However, when compared to the Nucleotide database
191 (<http://blast.ncbi.nlm.nih.gov/Blast.cgi>, November 2015), these reads only aligned to either
192 reference and not both, indicating suboptimal performance of these aligners for Nanopore reads.

193

194 **Table 3:** Percentages of reads aligned to target pathogen and cell line/human DNA in the samples
195 prepared for this study.

Sample	Target pathogen			Cell line/human DNA		
	% of total reads aligned	% "pass" reads aligned	% "fail" reads aligned	% of total reads aligned	% "pass" reads aligned	% "fail" reads aligned
FluA non-hybridised	0.0	0.0	0.0	27.1	75.2	23.7
FluA hybridised	10.9	57.2	9.5	9.3	49.4	8.2
CMV non-hybridised	1.2	5.9	1.0	22.5	100.0	19.3
CMV hybridised	45.5	98.7	35.0	3.6	6.0	3.1
<i>MtbC_10</i> hybridised	0.8	5.9	0.8	8.2	23.5	8.2
<i>MtbC_90</i> hybridised	4.4	88.1	3.9	6.7	17.3	6.6
<i>Mtb</i> H37Rv_10 hybridised	7.3	32.8	3.9	11.4	46.6	6.6
<i>Mtb</i> H37Rv_90 hybridised	3.4	5.9	1.7	10.5	12.6	9.2

196

197 **Comparison of enriched and non-enriched Nanopore libraries**

198 Amplification and subsequent sequencing of the long DNA fragments demonstrates the success of
 199 the hybridisation and library preparation for both Nanopore and Illumina sequencing. Analysis of the
 200 42261 reads obtained from a non-enriched, PCR-amplified FluA cDNA library run on the Nanopore
 201 MinION™ found 75.2% pass and 23.7% fail reads aligned to the MDCK dog cell line used for
 202 cultivation of the virus, whilst only one read aligned to the FluA reference H1N1. After hybridisation
 203 and amplification, 57.2% of pass and 9.5% of fail reads (34211 reads in total) from a Nanopore run
 204 could be aligned to FluA. This amounts to a total FluA coverage of 71.2x. Though there was generally
 205 good coverage across the eight FluA segments, Figure 2 shows a wide variation of number of reads
 206 per fragment, as well as distinct peaks of coverage of FluA cDNA. Similarly, the frequency of cell line
 207 reads dropped to 49.4% (pass) and 8.2% (fail) (Table 3).

208

209 The un-enriched CMV library (432 unique reads in total) produced five reads (1.2%) matching the
210 CMV reference HHV-5, while 97 reads (22.5%) matched the human_g1k_v37 reference. After
211 hybridisation of the DNA with the CMV-specific bait set, we obtained 37589 reads from three runs,
212 with almost all (98.7%) pass reads and 35% of fail reads aligning to the CMV reference. This amounts
213 to a coverage of 89.9x of the CMV genome. Figure 3 shows the output of these reads aligned to the
214 reference.

215

216 A comparison of the consensus sequence generated from the enriched CMV reads aligned to the
217 CMV HHV-5 reference using the genomic similarity search tool YASS¹⁴ found the former had 99.4%
218 similarity to the reference (233854 of 235230 nucleotide residues). The conflicting/mismatch
219 residues are mostly gaps in the Nanopore consensus sequence at 46364-46433 (proteins UL34 and
220 UL35), 147820-147830 (helicase-primase subunit UL102), 194363-194698 and 195851-95977. The
221 last two regions of difference coincide with inverted repeat regions⁶ (194344 to 195667, 195090 to
222 197626). A number of mismatches to the reference HHV-5 were identified upstream of base 1270;
223 these were due to low coverage of this region by Nanopore reads. We found regions with low (<5x)
224 coverage had a high number of mismatches compared to the reference, but areas of greater
225 coverage matched near-perfectly.

226

227 For *Mtb* strain H37Rv, we obtained 2028 unique pass and 9961 unique fail reads (0.077x coverage),
228 for the strain *MtbC*, 202 pass and 46711 fail reads (0.182x). Distribution of reads from both *Mtb*
229 strains aligning to the H37Rv genome is relatively even (Figure 4). Localized high-coverage areas with
230 multiple reads in both strains were found at a number of genomic positions encoding transposases,
231 e.g. 887488-887429, 890363-889044, 1539822-1538580, 2640242-2635594, 3547252-3544391,
232 3789669-3788312 of strain H37RV (NC_018143.2). The areas with increased coverage can also be

233 observed in Illumina-generated datasets (Figure 4), presumably due to the redundancy of the
234 sequence, which could result in localised increased aligning of reads.

235

236 **Sequencing of enriched long fragments on the Illumina MiSeq**

237 To assess the success of the long fragment hybridisation, Illumina libraries were generated from half
238 of the hybridised material, and sequenced on a MiSeq (Table 4). The coverage profile from the
239 enriched FluA and *Mtb* samples produced similar distribution, and reads from both the Nanopore
240 and Illumina sequencers showed preferential enrichment of the same regions (Figure 4 shows *Mtb*).
241 In contrast, aligned Nanopore CMV reads generated from the long fragment-enriched sample had a
242 slightly different coverage profile to their Illumina counterpart. The longer Nanopore reads formed
243 wide peaks, whilst mapped Illumina reads clustered in narrow stacks with deep troughs (Figure 5).
244 This could be due to the positioning of baits, which could preferentially enrich short fragments that
245 overlap well with them. Some gaps are visible in the Nanopore coverage, these are bridged by
246 relatively few reads and do not always correspond to gaps in the Illumina coverage.

247

248 **Table 4:** Statistics of coverage generated by Picard (<http://broadinstitute.github.io/picard/>) in
249 Illumina MiSeq runs of independent hybridisations of long fragments.

Sample	Fragment hybridisation	Number of reads aligned to target pathogen	% reads aligned to target pathogen	Mean depth of pathogen coverage	% target bases covered at 10x
FluA	Long	2664967	59	2089	97
CMV	Long	1765332	94	957	96
10% <i>Mtb</i> H37Rv (1)	Long	6906339	84	295	99

10% Mtb H37Rv (2)	Long	1071332	62	50	99
10% Mtb H37Rv (3)	Long	8193188	96	315	99
10% Mtb H37Rv (4)	Long	2942898	56	100	98
90% Mtb H37Rv (5)	Long	2258868	93	99	99
90% Mtb H37Rv (6)	Long	6978112	97	297	99
90% Mtb H37Rv (7)	Long	9926437	87	342	99
90% Mtb H37Rv (8)	Long	15382452	96	521	98
Mtb H37Rv 9)	Short	3982148	99	169	99
90% MtbC (1)	Long	689141	18	24	85
MtbC (2)	Short	2980023	99	115	97

250

251

252

253 Discussion

254 This study explores the enrichment of long fragments using baits designed and used in previous
255 studies^{4,5} for whole-genome sequencing of specific pathogens from clinical or mixed samples. These,
256 instead of the 200 nt strands conventionally used for Illumina library construction, give a wider
257 range of possibilities for the deconvolution of repeat regions, detection of translocations and larger
258 indels, or mate-pair libraries. Pathogens with small genomes, in our case cDNA from FluA, could be
259 sequenced without previous shearing, potentially preserving structural information and avoiding
260 assembly problems. We show that the enriched long fragments can be used for libraries for both
261 Illumina sequencers (after re-shearing), and the third-generation long-range sequencer Oxford

262 Nanopore MinION™. The portable, relatively inexpensive, and low-footprint MinION™ sequencers
263 have been used in settings where conventional Illumina sequencing would be difficult¹⁵. Nanopore
264 sequencing has previously been used to detect structural variants in large genomes¹⁶; our method
265 could be employed as a non-amplicon-based alternative for this application. As the enrichment
266 approach is platform-agnostic, it could also be used to generate libraries compatible with the PacBio
267 RS II and Sequel systems.

268 Previous work¹⁷ has shown that detection of moderate to high titres of pathogen DNA (chikungunya
269 virus, Ebola and Hepatitis C virus) from human blood samples is possible using Nanopore
270 sequencing. However, this direct sequencing approach cannot take full advantage of the capabilities
271 of sequencing for strain typing and variant identification, especially in bacterial infections, due to
272 their larger genomes, and/or low titre samples. In our Nanopore sequencing runs with un-enriched
273 FluA or CMV DNA, we detected very low numbers of reads from the pathogen compared to those
274 from the host cell line culture. In contrast, sequencing data from enriched DNA produced good
275 coverage of the FluA and CMV genomes and partial coverage of the *Mtb* genome. High and even
276 coverage of the target pathogen genome¹⁸, gained through targeted enrichment, can compensate
277 for the errors seen in Nanopore sequencing, and facilitate the detection of potential resistance to
278 antibiotics and identification of mixed infections and minor variants.

279 We found Illumina sequencing results from material derived from long fragment hybridisation (Table
280 4) showing good and even coverage, similar to the standard PATHSEEK method of short-fragment
281 hybridisation^{4, 5}. This, together with the fact that high-molecular weight post-enrichment fragments
282 could be readily ligated and amplified, highlights the suitability for long fragments for hybridisation
283 and its downstream applications.

284 We analysed our Nanopore datasets with two different aligners - BLASR produced fewer reads with
285 higher identity to the target pathogen and lower standard deviation, LAST identified more reads
286 with lower identity and higher standard deviation, similar to the results of Kilianski et al.¹⁹. The

287 difference of up to 40% in sequences identified as matches to the reference between the two
288 aligners, highlights that neither works optimally for aligning Nanopore reads to their reference. A
289 small number of reads reported as containing both cell line and pathogen matches, mainly from the
290 output of LAST, could not be confirmed. As reported elsewhere^{17, 19}, a large percentage of Nanopore
291 reads (mainly the “fail” quality) could not be aligned to either the target pathogen, or human (*Mtb*),
292 the human cell line (CMV) or the dog cell line (FluA), respectively, and show no similarities when
293 compared to the NCBI Nucleotide Database (November 2015).

294

295 The persisting, and, in some cases, increasing pressure of pathogenic viruses and bacteria on human
296 and animal health underlines the need for fast, accurate and up-to-date diagnosis. Next-generation
297 sequencing can enable clinicians to identify both pathogen species and genotype in one step,
298 significantly aiding individual treatment with respect to drug resistance²⁰ and monitoring of
299 outbreaks²¹. The increasing interest in whole-genome sequencing for diagnosis has led to the drive
300 to take it out of the research environment and into the clinic to reduce the time between taking a
301 specimen, identifying the pathogen and obtaining actionable genomic data. Nanopore sequencing,
302 coupled with data streaming and real-time analysis, has the possibility to bring sequencing closer to
303 patient, or to be used in settings without easy access to an Illumina machine. The Nanopore
304 sequencing platform, coupled with our long-fragment enrichment method, can be applied to a range
305 of pathogens and clinical samples. However, our experiments were performed with a 16-hour
306 hybridisation reaction, a time-consuming and rate-limiting step. In the future, this has the potential
307 to be shortened to four hours by using a different hybridisation protocol. Though not tested in this
308 study, addition of molecular bar-codes as outlined in the “Sequencing using the PCR Barcoding Kit”
309 ONT protocol, will allow for several clinical samples of enriched viral DNA to be run simultaneously
310 on one MinION™ flowcell. This, coupled with increasing accuracy of the MinION™ reads, will also

311 reduce the coverage necessary for strain and variant identification, making this method suitable for
312 diagnostic purposes in the future.

313

314

315

316 Acknowledgements

317 We would like to thank Dietrich Lueersen, David Blaney and Dan Swan for their help in the analysis
318 of the data; Richard Milne at Department for Virology, UCL Medical School (Royal Free Campus,
319 Rowland Hill Street, London, UK) for the kind gift of CMV Merlin strain, Amanda Brown, Lise J
320 Schreuder and Tanya Parish at the Barts and The London School of Medicine and Dentistry (Queen
321 Mary University of London, London, UK), for strain *Mtb* H37Rv, Philip Butcher and Jasvir Dhillon at St.
322 George's Hospital (University of London, London, UK) for the generous donation of strain *Mtb* C. Past
323 and present members of the PATHSEEK consortium are: Judith Breuer, Rachel Williams, Mette
324 Theilgaard Christiansen, Josie Bryant, Sofia Morfopoulou, Helena Tutill, Erika Yara-Romero, Charlotte
325 Williams, Dan Depledge (UCL), Martin Schutten, Saskia Smits, Georges M.G.M. Verjans, Freek B. van
326 Loenen, Anne van der Linden, Albert Osterhaus (Erasmus MC); Katja Einer-Jensen, Martin Ludvigsen,
327 Roald Forsberg (CLC Bio); James Clough, Graham Speight, Jacqueline Chan, Jolyon Holdstock, Sabine
328 Eckert, Mike McAndrew, Amanda Brown (OGT).

329

330

331 Competing interests

332 S.E.E is a Nanopore shareholder.

333 References

- 334 1. Carlet J. The world alliance against antibiotic resistance: consensus for a declaration. Clin
335 Infect Dis. 2015; 60(12): 1837-1841.
- 336 2. Wlodarska M, Johnston JC, Gardy JL, Tang P. A microbiological revolution meets an ancient
337 disease: improving the management of tuberculosis with genomics. Clin Microbiol Rev. 2015; 28(2):
338 523-539.
- 339 3. Doughty EL, Sergeant MJ, Adetifa I, Antonio M, Pallen MJ. Culture-independent detection
340 and characterisation of *Mycobacterium tuberculosis* and *M. africanum* in sputum samples using
341 shotgun metagenomics on a benchtop sequencer. PeerJ. 2014; 2: e585.
- 342 4. Christiansen MT, Brown AC, Kundu S, Tutill HJ, Williams R, Brown JR, et al. Whole-genome
343 enrichment and sequencing of *Chlamydia trachomatis* directly from clinical samples. BMC Infect Dis.
344 2014; 14: 591.
- 345 5. Brown AC, Bryant JM, Einer-Jensen K, Holdstock J, Houniet DT, Chan JZ, et al. Rapid Whole-
346 Genome Sequencing of *Mycobacterium tuberculosis* Isolates Directly from Clinical Samples. J Clin
347 Microbiol. 2015; 53(7): 2230-2237.
- 348 6. Masse MJ, Karlin S, Schachtel GA, Mocarski ES. Human cytomegalovirus origin of DNA
349 replication (oriLyt) resides within a highly complex repetitive region. Proc Natl Acad Sci U S A. 1992;
350 89(12): 5246-5250.
- 351 7. Witney AA, Gould KA, Arnold A, Coleman D, Delgado R, Dhillon J, et al. Clinical application of
352 whole-genome sequencing to inform treatment for multidrug-resistant tuberculosis cases. J Clin
353 Microbiol. 2015; 53: 1473–1483.

- 354 8. Loman NJ, Quinlan AR. Poretools: a toolkit for analyzing Nanopore sequence data.
355 *Bioinformatics* 2014; 23: 3399-3401.
- 356 9. Chaisson MJ, Tesler G. Mapping single molecule sequencing reads using basic local
357 alignment with successive refinement (BLASR): application and theory. *BMC Bioinformatics* 2013; 13:
358 238.
- 359 10. Kielbasa SM, Wan R, Sato K, Horton P, Frith MC. Adaptive seeds tame genomic sequence
360 comparison. *Genome Research* 2011; 21(3): 487–493.
- 361 11. Quick J, Quinlan AR, and Loman NJ. A reference bacterial genome dataset generated on the
362 MinION™ portable single-molecule nanopore sequencer. *GigaScience* 2014; 3: 22.
- 363 12. Robinson JT, Thorvaldsdóttir H, Winckler W, Guttman M, Lander ES, Getz G, et al. Integrative
364 Genomics Viewer. *Nature Biotechnology* 2011; 29: 24–26.
- 365 13. Thorvaldsdóttir H, Robinson JT, Mesirov JP. Integrative Genomics Viewer (IGV): high-
366 performance genomics data visualization and exploration. *Briefings in Bioinformatics* 2013; 14: 178-
367 192.
- 368 14. Noe L, Kucherov G. YASS: enhancing the sensitivity of DNA similarity search. *Nucleic Acids*
369 *Res.* 2005; 33(2): W540-W543.
- 370 15. Quick J, Loman NJ, Duraffour S, Simpson JT, Severi E, et al. Real-time, portable genome
371 sequencing for Ebola surveillance. *Nature* 2016; 530(7589):228-32.
- 372 16. Norris AL, Workman RE, Fan Y, Eshleman JR, Timp W. Nanopore sequencing detects
373 structural variants in cancer. *Cancer Biol Ther.* 2016; 17(3):246-53.
- 374 17. Greninger AL, Naccache SN, Federman S, Yu G, Mbala P, Bres V, et al. Rapid metagenomic
375 identification of viral pathogens in clinical samples by real-time nanopore sequencing analysis.
376 *Genome Med.* 2015; 7: 99.

- 377 18. Judge K, Harris SR, Reuter S, Parkhill J, Peacock SJ. Early insights into the potential of the
378 Oxford Nanopore MinION for the detection of antimicrobial resistance genes. *J Antimicrob*
379 *Chemother.* 2015; 70(10): 2775-2778.
- 380 19. Andy Kilianski A, Haas JL, Corriveau EJ, Liem AT, Willis KL, Kadavy DR, et al. Bacterial and viral
381 identification and differentiation by amplicon sequencing on the MinION nanopore sequencer.
382 *GigaScience* 2015; 4: 12.
- 383 20. Köser CU, Ellington MJ, Peacock SJ. Whole-genome sequencing to control antimicrobial
384 resistance. *Trends Genet.* 2014; 30(9): 401–407.
- 385 21. Chin CS, Sorenson J, Harris JB, Robins WP, Charles RC, Jean-Charles RR, et al. The Origin of
386 the Haitian Cholera Outbreak Strain. *N Engl J Med.* 2011; 364(1): 33–42.
- 387

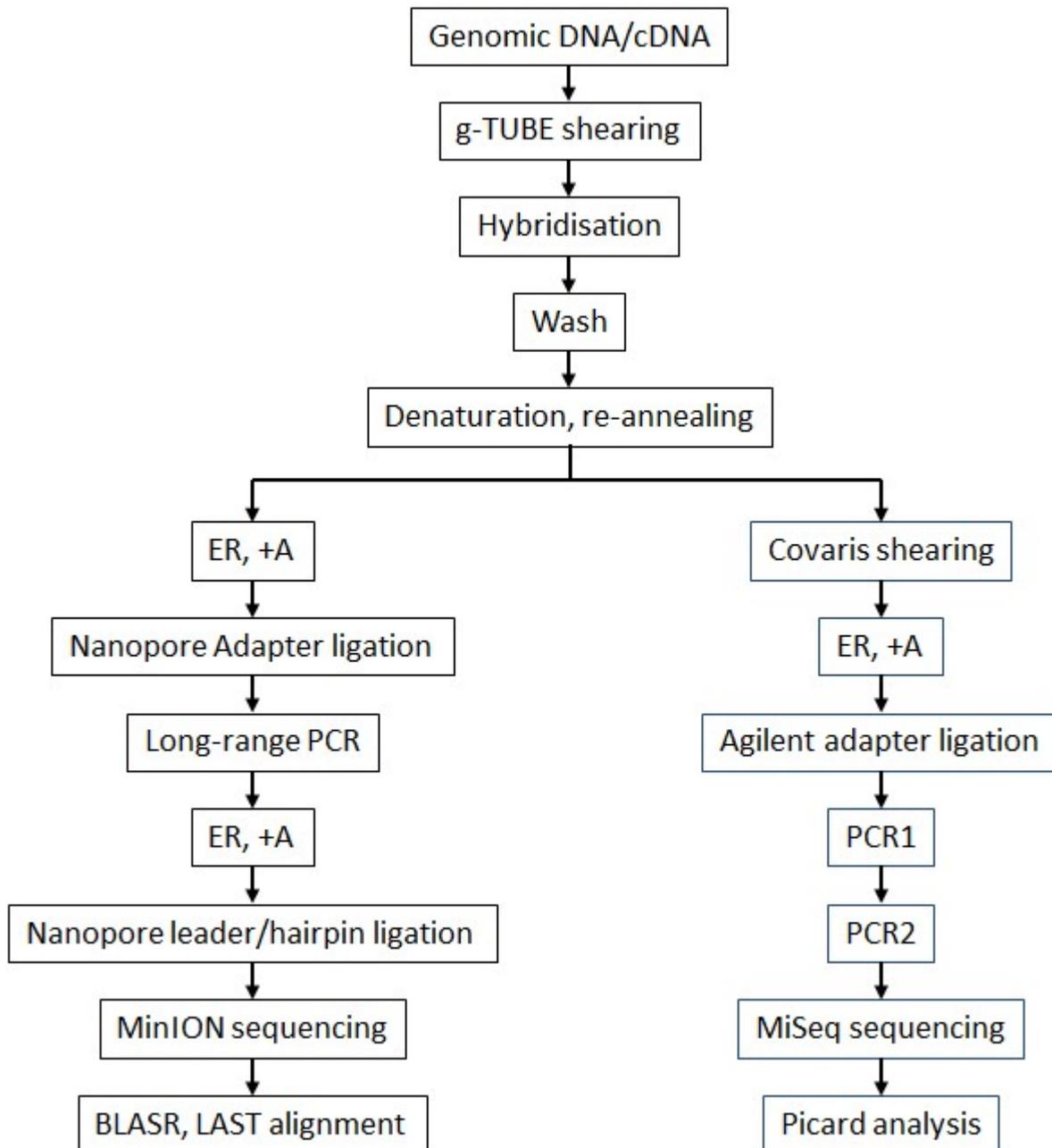
388 **Figures**

389

390 **Figure 1:** Workflow for hybridisation and sequencing of long-fragment-enriched pathogen DNA. ER:

391 end repair of fragments, +A: dA-tailing.

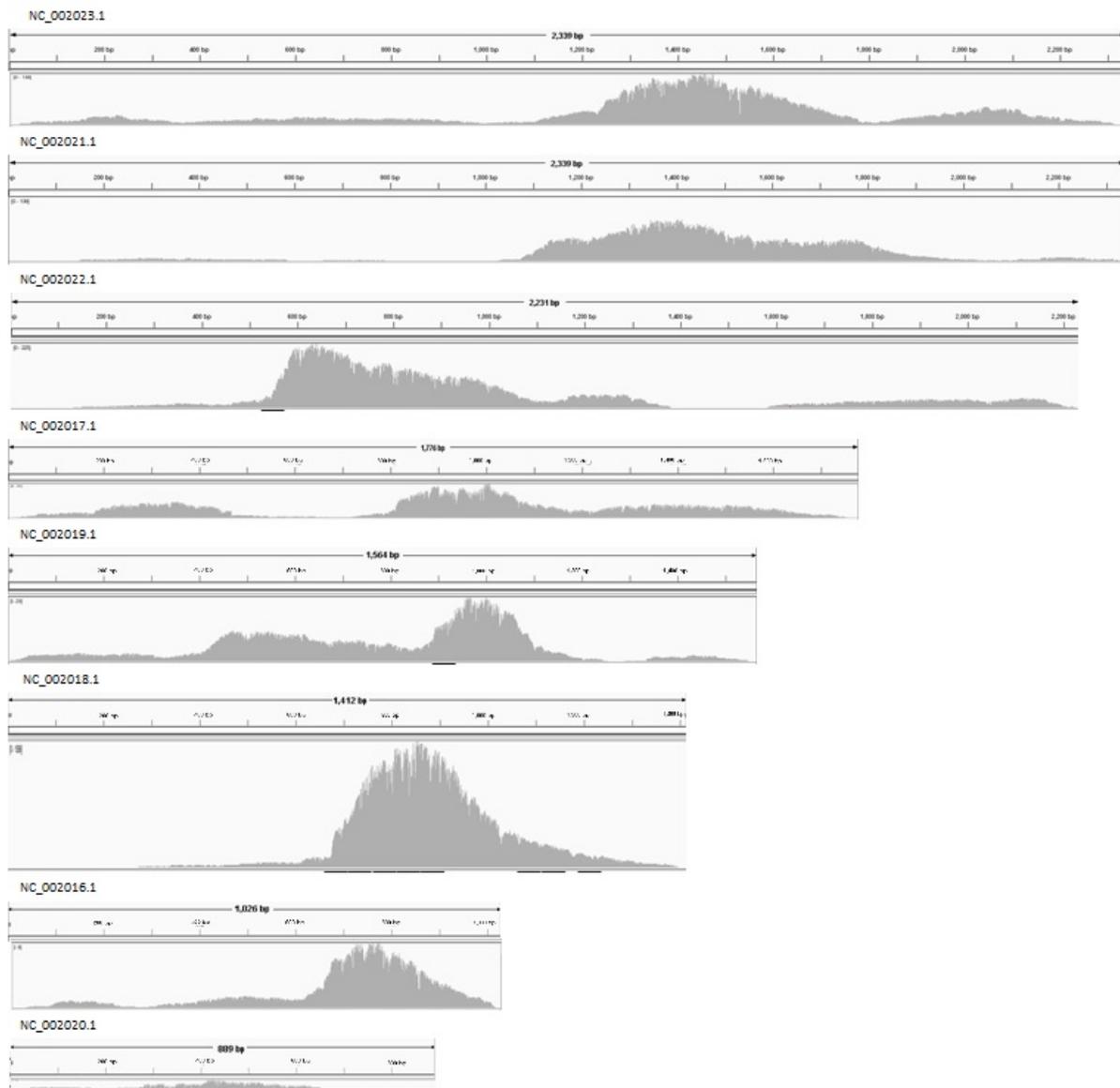
392



393

394 **Figure 2: Nanopore reads from enriched FluA cDNA, aligned to reference FluA H1N1 with BLASR,**
395 **visualized in the Integrated Genome Viewer^{12, 13} (IGV).** Maximum coverage results for the fragments
396 are: 139 (NC_002023.1), 139 (NC_002021.1), 225 (NC_002022.1), 51 (NC_002017.1), 219
397 (NC_002019.1), 1589 (NC_002018.1), 185 (NC_002016.1), 16 (NC_002020.1).

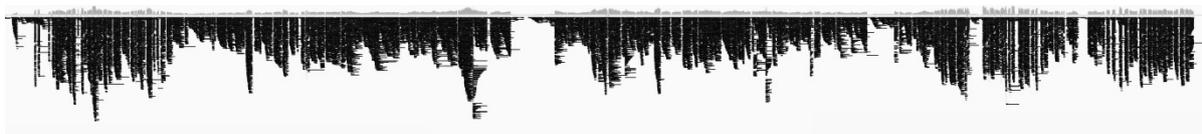
398



399

400

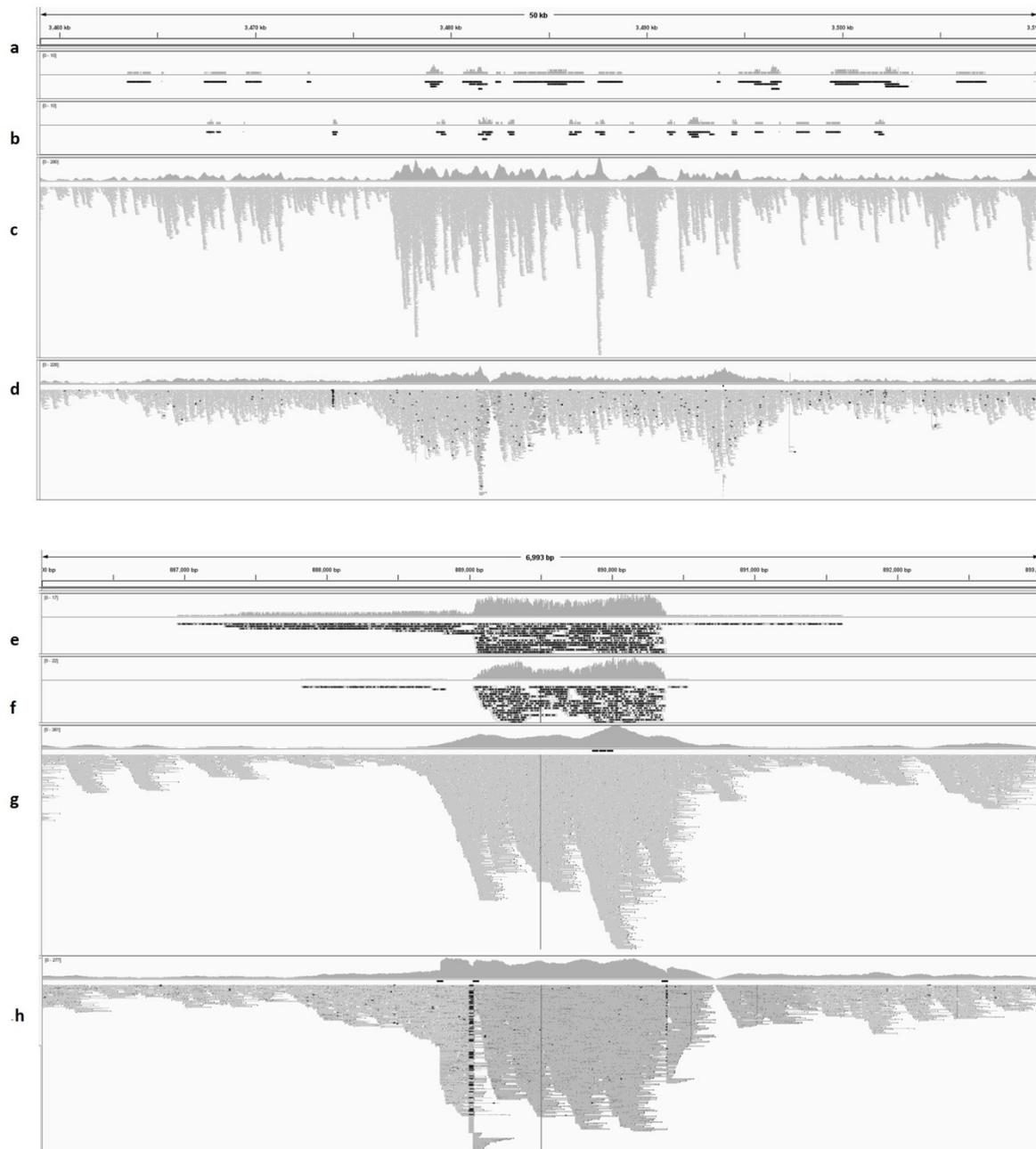
401 **Figure 3: Coverage of the CMV strain HHV-5 genome with Nanopore reads, aligned with BLASR and**
402 **visualized with IGV. The maximum coverage for this plot is 291.**



403

404

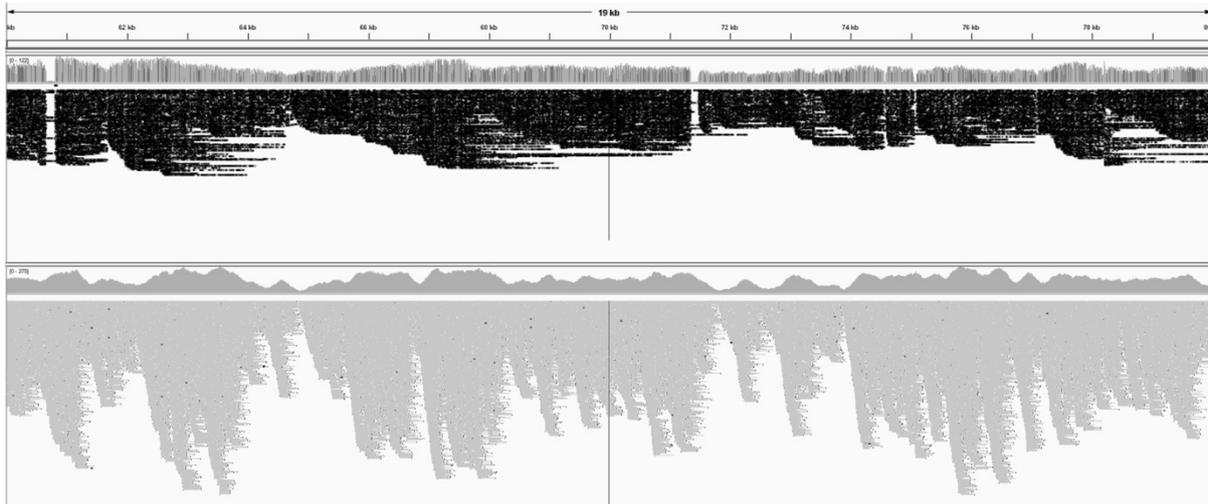
405 **Figure 4:** Alignments of Nanopore reads of libraries from DNA of strains *Mtb* H37Rv (a), *Mtb*C (b) and
406 Illumina-sequenced long-fragment-hybridised *Mtb* H37Rv (c) and *Mtb*C (d), in position 3,459,000-
407 3,510,000 of the genome of strain H37Rv show low but even coverage by Nanopore reads (a-d).
408 Panels e-h show the same samples in the same order, in a region of high coverage (886,000-
409 893,000), visualised in IGV. Regions with increased coverage in the Nanopore reads correspond to
410 higher read depth in the Illumina coverage.



411
412

413 **Figure 5:** Alignment of Nanopore reads (top) and Illumina reads (bottom) to reference human CMV
414 herpes virus HHV-5 GU179001.1, positions 60,000-80,000, visualised in IGV. The longer Nanopore
415 reads cover the reference more evenly but show some gaps that are well-covered by Illumina reads.

416



417