

1 **Comparison of single nucleotide variants identified by Illumina and Oxford Nanopore technologies**
2 **in the context of a potential outbreak of Shiga Toxin Producing *Escherichia coli*.**

3

4

5 David R Greig, Claire Jenkins, Saheer Gharbia & Timothy J Dallman*.

6

7 National Infection Service, Public Health England, London, NW9 5EQ.

8 *Corresponding author.

9

10

11 Author details:

12 David R Greig:

13 Email – David.Greig@phe.gov.uk

14

15 Claire Jenkins:

16 Email - Claire.Jenkins1@phe.gov.uk

17

18 Saheer Gharbia

19 Email – Saheer.Gharbia@phe.gov.uk

20

21 Timothy J Dallman:

22 Email - Tim.Dallman@phe.gov.uk

23

24

25 Keywords –Oxford Nanopore, Illumina, Variant calling, STEC, outbreak.

26

27

28 **Abstract**

29 **Background**

30 We aimed to compare Illumina and Oxford Nanopore Technology (ONT) sequencing data from the
31 two isolates of STEC O157:H7 to determine whether concordant single nucleotide variants were
32 identified and whether inference of relatedness was consistent with the two technologies.

33 **Results**

34 For the Illumina workflow, the time from DNA extraction to availability of results, was approximately
35 40 hours in comparison to the ONT workflow where serotyping, Shiga toxin subtyping variant
36 identification were available within seven hours. After optimisation of the ONT variant filtering, on
37 average 95% of the discrepant positions between the technologies were accounted for by
38 methylated positions found in the described 5-Methylcytosine motif sequences, CC(A/T)GG. Of the
39 few discrepant variants (6 and 7 difference for the two isolates) identified by the two technologies, it
40 is likely that both methodologies contain false calls.

41 **Conclusions**

42 Despite these discrepancies, Illumina and ONT sequences from the same case were placed on the
43 same phylogenetic location against a dense reference database of STEC O157:H7 genomes
44 sequenced using the Illumina workflow. Robust SNP typing using MinION-based variant calling is
45 possible and we provide evidence that the two technologies can be used interchangeably to type
46 STEC O157:H7 in a public health setting.

47

48

49

50

51 **Background**

52 Shiga toxin producing *Escherichia coli* (STEC) O157:H7 is a zoonotic, foodborne pathogen defined by
53 the presence of phage-encoded Shiga toxin genes (*stx*) [1]. Disease symptoms range from mild
54 through to severe bloody diarrhoea, often accompanied by fever, abdominal cramps and vomiting
55 [2]. The infection can progress to Haemolytic Uremic Syndrome (HUS), characterized by kidney
56 failure and/or cardiac and neurological complications [3,4]. Transmission from an animal reservoir,
57 mainly ruminants, occurs by direct contact with animals or their environment, or by the
58 consumption of contaminated food products with reported vehicles including beef and lamb meat,
59 dairy products, raw vegetables and salad [2,4].

60

61 STEC O157:H7 belongs to multi-locus sequence type clonal complex (CC) 11, with all but a small
62 number of variants belonging to sequence type ST11. CC11 comprises three main lineages (I, II and
63 I/II) and seven sub-lineages (Ia, Ib, Ic, IIa, IIb, IIc and I/II) [5]. There are two types of Shiga toxin, Stx1
64 and Stx2. Stx1 has four subtypes (1a-1d) and Stx2 has seven subtypes (2a-2g). Subtypes 1a, 2a, 2c,
65 and rarely 2d, are found in STEC O157:H7. Strains harbouring *stx2a* are significantly associated with
66 cases that develop HUS [2,6]. As well as harbouring *stx* encoding prophage, STEC O157:H7 has an
67 additional prophage repertoire accounting for at least 20% of the chromosome.

68

69 The implementation of whole genome sequencing (WGS) data for typing STEC has improved the
70 detection and management of outbreaks of foodborne disease [6]. Single nucleotide polymorphism
71 (SNP) typing offers an unprecedented level of strain discrimination and can be used to quantify the
72 genetic relatedness between groups of genomes. In general, for clonal bacteria, the fewer
73 polymorphisms identified between pairs of strains, the less time since divergence from a common
74 ancestor and therefore the increased likelihood that they are from the same source population.
75 Therefore, it is paramount that variant detection for typing is accurate, highly specific and
76 concentrated on positions of neutral evolution to ensure the correct interpretation of the sequence
77 data within the epidemiological context of an outbreak. It has been previously shown that different
78 bioinformatics analysis approaches for variant identification exhibit detection variability [7,8]. It is
79 therefore important that within a particular analysis, workflow parameters to filter identified
80 variants to achieve optimum sensitivity and specificity are appropriately optimised.

81

82 Short read sequencing platforms, such as those provided by Illumina, have been adopted by public
83 health agencies for infectious disease surveillance worldwide [9] and have proved to be a robust and
84 accurate method for quantifying relatedness between bacterial genomes. High-throughput Illumina

85 sequencing although cost effective, often requires batch processing of hundreds of microbial isolates
86 to achieve cost savings and therefore this approach offers less flexibility for urgent, small scale
87 sequencing often required during public health emergencies [10]. In contrast, Oxford Nanopore
88 Technologies (ONT) offers a range of rapid real-time sequencing platforms from the portable
89 MinION to the higher throughput GridION and PromethION models, although at this time lower read
90 accuracy compared to Illumina data suggests accurate variant calling maybe problematic.

91

92 In September 2017, Public Health England (PHE) was notified of two cases of HUS in two children
93 admitted to the same hospital on the same night. STEC O157:H7 was isolated from the faecal
94 specimens of both cases. In order to rapidly determine whether or not the cases were part of a
95 related phylogenetic cluster and therefore likely to be epidemiologically linked to each other, or to
96 any other cases in the PHE database, we sequenced both isolates using the MinION platform and
97 integrated the ONT sequencing data with a dense reference database of Illumina sequences. We
98 aimed to compare Illumina and ONT sequencing data from the two isolates to assess the utility of
99 the ONT method for urgent, small scale sequencing, and to determine whether the same single
100 nucleotide variants were identified and whether inference of relatedness was consistent with the
101 two technologies.

102

103 **Data description**

104 Paired-end FASTQ files were generated from the Illumina HiSeq 2500 for both samples (cases). Raw
105 long-read data (FAST5) was generated from the MinION and basecalled using Albacore (FASTQ) in
106 real-time. Both technologies derived FASTQ reads were trimmed and filtered (Trimmomatic,
107 Porechop, Filtlong) before being aligned (BWA, Minimap2) to a reference genome (NC_002695.1).
108 Variant positions were called using GATK before being imported into SnapperDB. Full processing
109 details can be found within the methods section.

110

111 **Results**

112 *Comparison of typing results generated by Illumina and ONT workflows*

113 To consider the potential benefits of real-time sequencing to enhance opportunities for early
114 outbreak detection, the timelines from DNA extraction to result generation for Illumina and ONT
115 workflows were evaluated (Figure 1) and the relationship between yield, time and genome coverage
116 plotted (Figure 2). For the ONT workflow, the time from DNA extraction to completion of the
117 sequencing run was 28 hours. A total yield of 0.45 Gbases for the isolate from Case A and 0.59
118 Gbases for the isolate from Case B was achieved which corresponds to an equivalent coverage of the

119 Sakai O157 STEC reference genome (5.4Mb) of 81.29X and 108.30X for isolate A and B respectively.
120 The average PHRED quality score for all reads in Case A was 9.87 and Case B was 9.47, which is
121 approximately 1 error every 10 bases. Base-calling and analysis was performed in real-time and
122 serotyping, Shiga toxin subtyping and variant identification were available within six hours and
123 twenty minutes of the 24-hour sequencing run. With respect to the Illumina sequencing workflow,
124 the time from DNA extraction to availability of results, assuming there were no breaks in the
125 process, was just under 40 hours (Figure 1).

126

127 The species identification, serotype, MLST profile and Shiga toxin subtype results generated by both
128 Illumina and ONT workflows were concordant with both isolates identified as *Escherichia coli*
129 O157:H7 ST11 (12,12,8,12,15,2,2), *stx2a* and *stx2c*. During the ONT sequencing run, the bacterial
130 species was unambiguously identified in less than one minute for both cases (Figure 1). Additionally,
131 using Krocus, a confirmed MLST was generated for Case A at 1:54 hours and Case B at 10:39 hours
132 into the sequencing run. This was the point at which the last read required to generate a consensus
133 on the MLST was base-called. By 93 minutes for Case A and 41 min for Case B, it was possible to
134 determine the *E. coli* O157:H7 serotype, and *stx2a* and *stx2c* were detected at 58 and 24 minutes
135 into the sequencing run for Case A and Case B, respectively.

136

137 *Optimisation of ONT variant calling*

138 To compare Illumina and ONT sequences within a standardised framework it was necessary to
139 optimise the parameters for variant filtering within GATK2 to compensate for the lower read
140 accuracy observed in the ONT data. Using Case B for the optimisation, base calls in the ONT data
141 were classified as true positives (variant base detected by both methods), false positives (variant
142 base in ONT, reference base in Illumina), true negatives (reference base in Illumina and ONT) or false
143 negatives (variant base in Illumina, reference base in ONT). To disregard areas of the genome that
144 the ONT reads could map to (and therefore identify variants) but were ambiguously mapped with
145 Illumina reads, pre-filtering was performed by masking regions annotated as phage in the reference
146 genome and those that could not be accurately self-mapped with simulated reference Illumina
147 FASTQ reads. Figure 3 plots the precision (the proportion of true positives with respect to all
148 positives calls) against the recall/sensitivity (the proportion of true positives identified with respect
149 to all true positives) for an array of consensus ratio cut-offs for each of the masking strategies.
150 Similar areas 'under the curve' were achieved for the different masking strategies with slightly
151 higher precision at lower recall achieved with 'self-masking' (AUC – 0.71) and slightly higher recall at
152 lower precision with explicit masking of the Sakai prophage (AUC – 0.75). The absence of a masking

153 strategy markedly affects the precision of variant calling with ONT data, in comparison of Illumina as
 154 a gold standard (AUC – 0.30). To identify the optimum consensus cut-off for filtering ONT variants
 155 processed through GATK the F1 score was calculated at each consensus cut-off. A consensus cut-off
 156 of 0.8 maximised the precision and recall (Figure 4) irrespective of the filtering methods.

157

158 *Investigation of the discrepant variants identified between the Illumina and ONT data*

159 After optimised quality and prophage filtering there were 266 and 101 base positions for Cases A
 160 and B respectively that were discordant between the ONT and Illumina sequencing data. The
 161 majority of discrepancies were where the ONT data identified a variant not identified in the Illumina
 162 data (261/266 (98.12%) and 95/101 (94.06%) discrepant base positions for Cases A and B
 163 respectively). In contrast the Illumina data identified 5 (1.88%) discrepant base positions as variants
 164 for Case A and 6 (5.94%) for case B (Table 1) not identified by the ONT data.

165

Variants and reason for omission.	Case A		Case B	
	Illumina VCF	ONT VCF	Illumina VCF	ONT VCF
Total # of variants against the reference genome post quality filtering.	2076		1424	
# of variants with masked due to location in phage	708		531	
# of discrepant variants called between case A and B alone.	266		101	
# of variants in each VCF.	5	261	6	95
# of variants with methylated positions masked.	0	260	0	94
Final variants.	5	1	6	1

166

167 **Table 1** – Table showing the breakdown of the total number of variants of each technology against
 168 the reference genome, followed by the numbers of masked variants within prophage or methylated
 169 positions.

170

171 For both cases the most common discrepant variant were adenines classified as guanines in the ONT
 172 data with respect to the Illumina data (and reference), accounting for 68.05% (181/266) for Case A
 173 and 72.28% (73/101) for Case B. The second most common discrepancy was thymine being
 174 classified as cytosine in the ONT data accounting for 29.70% (79/266) in Case A and 22.80% (21/101)
 175 in Case B (Table 1). Of the transitions described above, 97.74% (Case A) and 93.07% (Case B)
 176 occurred when the variant was between two homopolymeric regions of multiple cytosines and
 177 guanines (Figure 5). These homopolymeric regions were similar to described DNA cytosine
 178 methylase (Dcm) binding sequences [11]. Nanopolish was subsequently used to identify likely Dcm,
 179 5' – cytosine – phosphate – guanine – 3' (CpG) and DNA adenine methyltransferase (Dam)
 180 methylation sites in the ONT sequencing data and confirmed 260/266 (97.74%) and 94/101 (93.07%)

181 discrepant variants in the ONT data were classed as methylated for Cases A and B respectively. All of
 182 which were determined to be Dcm methylation for both cases.

183

184 Once the methylated positions were masked from the analysis, there were a total of 6 (5 discrepant
 185 variants in Illumina and 1 ONT) and 7 (6 discrepant variants in Illumina and 1 ONT) discrepant SNPs
 186 between the ONT and Illumina data, for Cases A and B respectively (Table 2 & 3). Four discrepant
 187 Illumina variants are shared by both Case A and Case B. One shared variant was found in a non-
 188 coding region, another shared variant was found in *rhcC* encoding an RHS (rearrangement hotspot)
 189 protein defined by the presence of extended repeat regions. Two further shared variants were
 190 found in *dadX*, an alanine racemase gene. *dadX* is a paralogue of *alr*, also annotated as an alanine
 191 racemase in the *Sakai* reference genome with significant nucleotide similarity (>75% nucleotide
 192 identity). Both intra and inter gene repeats are known to be regions of potential false positives calls
 193 with Illumina data due to miss-mapping.

194

SNP	Position	BASE in Ref	BASE in Illum	Depth in Illum	BASE in ONT	Depth in ONT	Variant	Locus tag	Annotation
1	270,595	C	A	46	C	141	A	ECs0237	rhcC
2	379,516	A	G	114	A	100	G	NON CODING	N
3	1,681,338	C	G	59	C	61	G	ECs1685	alanine racemase 2
4	1,681,339	G	C	57	G	61	C	ECs1685	alanine racemase 2
5	2,636,513	T	C	91	T	69	C	ECs2674	hypothetical protein
6	4,709,195	A	A	86	G	82	G	ECs4673	membrane-bound ATP synthase epsilon-subunit AtpC

195

196 **Table 2** – Table showing the final discrepant SNPs between the Illumina data and ONT data for case
 197 A. Also shown is the base as it is in the reference, the Illumina called base and read depth at that
 198 position and the same for the ONT data. Finally, also included is the locus tag relative to the
 199 reference genome and the gene annotation.

200

SNP	Position	BASE in Ref	BASE in Illum	Depth in Illum	BASE in ONT	Depth in ONT	Variant	Locus tag	Annotation
1	270,595	C	A	19	C	207	A	ECs0237	rhcC
2	379,516	A	G	52	A	124	G	NON CODING	N
3	1,681,338	C	G	44	C	86	G	ECs1685	alanine racemase 2
4	1,681,339	G	C	41	G	86	C	ECs1685	alanine racemase 2
5	2,033,176	T	G	34	T	85	G	ECs2049	hypothetical protein
6	2,731,621	A	C	52	A	73	C	NON CODING	N
7	4,901,209	A	A	49	G	102	G	ECs4834	superoxide dismutase Soda

201

202 **Table 3** – Table showing the final discrepant SNPs between the Illumina data and ONT data for case
203 B. Also shown is the base as it is in the reference, the Illumina called base and read depth at that
204 position and the same for the ONT data. Finally, also included is the locus tag relative to the
205 reference genome and the gene annotation.

206

207 *Phylogenetic Analysis*

208 Using the optimised variant calling parameters both strains clustered phylogenetically in lineage Ic
209 within a dense reference database of STEC O157:H7 genomes (n=4475). However, the genomes
210 were located in distinct sub-clades (Figure 6). It was, therefore, unlikely that the isolates originated
211 from the same source, and it was concluded that Cases A and B were not epidemiologically linked.
212 Following phylogenetic analysis of the Illumina SNP typing data (Figure 6), Case A was designated a
213 sporadic case. However, Case B clustered with a concurrent outbreak, already under investigation,
214 comprising three additional cases. The Illumina sequence linked to Case B was zero SNPs different
215 from the other three cases in the cluster, whereas the ONT sequence was 7 SNPs different, when
216 excluding the methylated positions (Table 3). Based on the ONT sequencing data alone, this
217 discrepancy would have led to uncertainty as to whether or not the Case B was linked to the
218 outbreak.

219

220 *Assembly Profile*

221 The ONT-only assembly resolved to five contigs (5.73 mb) for Case A and four contigs (5.60 mb) for
222 Case B (Supplementary Table 1). In Case A, the five contigs were determined to be a single
223 chromosomal contig, a single plasmid contig (pO157) and the three prophage duplications. In Case B,
224 the four contigs were determined to be a single chromosomal contig with two plasmids (one being
225 the pO157). For Case A the assembly resolved to 25 contigs (5.51mb) with a hybrid assembly and
226 668 contigs (5.45 mb) with an Illumina only assembly. Case B resolved to 34 contigs (5.49 mb) with a
227 hybrid assembly and 575 contigs (5.42 mb) with an Illumina only assembly.

228

229 Alignment of the assemblies (Supplementary Figures 1 and 2) revealed several locations within the
230 ONT-only assembly that there were absent in the hybrid and Illumina-only assemblies. In Case A,
231 there were 8 regions only present within the ONT-only chromosome assembly, of which 7 are
232 related to prophage regions (Supplementary Figure 1). In case B, there were 10 chromosomal
233 regions in the ONT-only assembly that did not align to the other assemblies. All 10 regions were
234 associated with prophage regions (Supplementary Figures 2).

235

236 **Discussion**

237 In this study, the two isolates sequenced using ONT were unambiguously identified as STEC O157:H7
238 ST 11 *stx2a/stx2c* in less than 15 hours and it was possible to distinguish the genetic relatedness
239 between the isolates within 377 minutes. The WGS turn-around time from DNA extraction and
240 library preparation, to sequencing and analysis via the Illumina workflow at PHE, is three to six days.
241 Although this turnaround time is rapid for a service utilising batch processing on the HiSeq
242 platforms, the sequencing approach using the MinION, whereby individual samples or small
243 barcoded batches are loaded and results generated and analysed in real-time, has the potential to
244 be faster and more flexible. This approach is therefore ideal for urgent, small scale sequencing,
245 often required during public health emergencies. In this scenario, analysis of the ONT data provided
246 evidence that the two cases were not epidemiologically linked and, although efforts were made to
247 determine the potential source of the infection for both cases through the National Enhanced STEC
248 Surveillance System [2], an outbreak investigation was not initiated.

249
250 A current limitation of MinION sequencing is its lower read accuracy when compared to short-read
251 technologies [12,13,14,15,16]. This accuracy has improved as the technology has matured but still
252 falls short of the 99% accuracy offered by short-read platforms [15]. There are a number of factors at
253 play that contribute to the low signal to noise ratio currently inherent in the nanopore data including
254 structural similarity of nucleotides, simultaneous influence of multiple nucleotides on the signal, the
255 non-uniform speed at which nucleotides pass through the pore and the fact that the signal does not
256 change within homopolymers [15]. Despite the current limitations of the technology, when mapped
257 to reference sequences in an established database of Illumina sequences, the ONT and Illumina
258 workflow placed the sequences from the same case on the same branch in a dense reference
259 database of STEC O157:H7 genomes sequenced using the Illumina workflow.

260
261 Although analysis of the Illumina and ONT sequencing data placed the sequences on the same
262 branch on the phylogeny, there were SNP discrepancies between the sequences generated by the
263 two different workflows, even after optimisation of the parameters. The vast majority of the
264 discrepant SNPs (261/266 – 98.12% and 95/101 – 94.06 % for Cases A and B respectively) were
265 attributed to variants identified in the ONT data and not the Illumina data. The majority of
266 discrepancies (97.74% in Case A and 93.07% in Case B) were found in sequences that are the same as
267 the known 5-Methylcytosine motif sequences, CC(A/T)GG [11,17] in the ONT data. Following a
268 search of the ONT discrepant SNPs for CpG, Dam and Dcm methylation using Nanopolish, the

269 majority (97.74% and 93.07% for case A and B respectively) of the ONT discrepant SNPs were
270 identified in Dcm methylated regions.

271

272 As Nanopolish is detecting these methylated positions with the use of the raw FAST5 data, it is
273 suggested that these particular discrepancies appear during the basecalling process. Albacore
274 handles most methylation well across the three methylation models searched for by Nanopolish, for
275 example only 94 out of 13,504 methylated positions were considered incorrect by base calling for
276 Case B. However, for mapping based-SNP typing, this level of error in base calling means that it is
277 not possible to accurately determine the number of SNPs, thus potentially obscuring the true
278 phylogenetic relationship between isolates of STEC O157:H7.

279

280 The optimisation of variant filtering was performed using the Illumina data as a gold standard.
281 However, it is possible that the alignment of the Illumina data might have generated false SNPs
282 based on reads mapping to ambiguous regions of the genome, whereas the long reads obtained
283 using the ONT workflow is able to resolve these ambiguous regions and call variants, or not, at these
284 positions correctly. As the Illumina data was used as the gold standard, in this scenario SNPs
285 produced in the Illumina data would have been classed incorrectly as false negatives in the ONT
286 data. Discrepant variants identified in the Illumina data were attributed mainly to potentially false
287 mapping of Illumina reads to homologous regions of the reference genome, variants which were
288 misidentified at the same position independently in Case A and Case B. Furthermore, comparison of
289 assemblies generated by ONT reads, Illumina reads and a hybrid approach highlights the extra
290 genetic content accessible to ONT assemblies where variation can be quantified.

291

292 In this study an ONT sequencing workflow was used to rapidly rule out an epidemiological link
293 between two children admitted to the same hospital on the same day with symptoms of HUS. The
294 isolates of STEC O157:H7 from each child mapped to different clades within the same STEC O157:H7
295 lineage (1c). We provide further evidence that SNP typing using MinION-based variant calling is
296 possible when the coverage of the variation is high [15]. The error rate exhibited by ONT sequencing
297 workflows continue to improve due to developments in the pore design, the library preparation
298 methods, innovations in base-calling algorithms and the introduction of post-sequencing correction
299 tools, such as Nanopolish [15,21]. Currently, both short and long read technologies are used for
300 public health surveillance, and there is a need to integrate the outputs so that all the data can be
301 analysed in the same way. Recently, Rang et al [15] reiterated how the scientific community can
302 make valuable contributions to improving ONT read accuracy by systematically comparing

303 computational strategies as highlighted in this study and elsewhere [22]. On-going up-dates to the
304 chemistry and software tools will facilitate the robust detection of SNPs enabling ONT to compete
305 with short read platforms, ultimately enabling the two technologies to be used interchangeably in
306 clinical and public health settings.

307

308 **Methods**

309 *DNA extraction, Library preparation and Illumina Sequencing*

310 Genomic DNA was extracted from two strains of STEC O157 isolates from two HUS cases admitted to
311 the same hospital on the same night. Using a Qiagen Qiasymphony (Qiagen, Hilden, Germany) to
312 manufactures instructions, genomic DNA extracted and quantified using a Qubit and the BR dsDNA
313 Assay Kit (Thermofisher Scientific, Waltham, USA) to manufactures instructions. The sequencing
314 library was prepared by fragmenting and tagging the purified gDNA using the Nextera XT DNA
315 Sample Preparation Kits (Illumina, Cambridge, UK) to manufactures instructions. The prepared
316 library was loaded onto an Illumina HiSeq 2500 (Illumina, Cambridge, UK) at PHE and sequencing
317 performed in rapid run mode yielding paired-end 100bp reads.

318

319 *Processing and analysis of Illumina sequence data*

320 FASTQ reads were processed using Trimmomatic v0.27 [23] to remove bases with a PHRED score of
321 less than 30 from the leading and trailing ends, with reads less than 50 bp after quality trimming
322 discarded. A *k*-mer approach (<https://github.com/phe-bioinformatics/kmerid>) was used to confirm
323 the species of the samples. Sequence type (ST) assignment was performed using MOST v1.0
324 described by [24]. *In silico* serotyping was performed by using GeneFinder, an inhouse PHE
325 programme (Doumith, unpublished) which uses Bowtie v2.2.5 [25] and Samtools v0.1.18 [26] to
326 align FASTQ reads to a multifasta containing the target genes (including *wzx*, *wzy* and *fliC*). *Stx* sub-
327 typing was performed as described in [27]. Illumina FASTQ reads were mapped to the Sakai STEC
328 O157 reference genome (NC_002695.1) using BWA MEM v0.7.13 [28]. Variant positions identified by
329 GATK v2.6.5 UnifiedGenotyper [29] that passed the following parameters; >90% consensus,
330 minimum read depth of 10, Mapping Quality (MQ) \geq 30. Any variants called at positions that were
331 within the known prophages in Sakai were masked from further analyses. The remaining variants
332 were imported into SnapperDB v0.2.5 [30].

333

334 *DNA extraction, Library preparation and Nanopore Sequencing*

335 Genomic DNA was extracted and purified using the Promega Wizard Genomic DNA Purification Kit
336 (Promega, Madison, USA) with minor alterations including doubled incubation times, no vigorous

337 mixing steps (performed by inversion) and elution into 50µl of double processed nuclease free water
338 (Sigma-Aldrich, St. Louis, USA). DNA was quantified using a Qubit and the HS (High sensitivity) dsDNA
339 Assay Kit (Thermofisher Scientific, Waltham, USA) to manufactures instructions. Library preparation
340 was performed using the Rapid Barcoding Kit - SQK-RBK001 (Oxford Nanopore Technologies, Oxford,
341 UK) with each sample's gDNA being barcoded by transposase based tagmentation and pooled as per
342 manufactures instructions. The prepared library was loaded on a FLO-MIN106 R9.4 flow cell (Oxford
343 Nanopore Technologies, Oxford, UK) and sequenced using the MinION for 24 hours.

344

345 *Processing and analysis of Nanopore sequence data*

346 Raw FAST5 files were basecalled and de-multiplexed in real-time, as reads were being generated,
347 using Albacore v2.1 (Oxford Nanopore Technologies) into FASTQ files. Run metrics were generated
348 using Nanoplot v1.8.1 using default parameters [31]. Reads were processed through Porechop v0.2.1
349 using default parameters (Wick. Unpublished) [32] to remove any barcodes and adapters used in
350 SQK-RBK001. Samples were speciated using Kraken v0.10.4 [33]. A MLST was assigned using Krocus
351 with the following parameters --kmer 15, --min_block_size 300 and --margin 500 [34]. *Stx* sub-typing
352 and serotyping was determined by aligning the basecalled reads using minimap2 v2.2 [35] and
353 Samtools v1.1 [26] to a multifasta containing the *Stx* and serotype encoding genes.

354

355 For reference based variant calling FASTQ reads were mapped to the Sakai STEC O157 reference
356 genome (NC_002695.1) using minimap2 v2.2 [35]. VCFs were produced using GATK v2.6.5
357 UnifiedGenotyper [29]. Any variants called at positions that were within the known prophages in
358 Sakai were masked from further analyses. To determine the optimum consensus cut-off for ONT
359 variant detection the VCF was filtered with sequentially decreasing ad-ratio values at 0.1 intervals.
360 Using the Illumina variant calls as the gold standard, F1 scores (the weighted average of precision
361 and recall) were calculated to determine the optimal ad-ratio for processing ONT data through
362 GATK.

363

364 *Comparison of Illumina and Nanopore discrepant SNPs*

365 Nanopolish [21] was also used to detect methylation across the ONT data to compare to the
366 discrepant positions. This was performed using the call-methylation function searching for three
367 types of methylation including, the DNA adenine methyltransferase (Dam), DNA cytosine methylase
368 (Dcm) and 5' – cytosine – phosphate – guanine – 3' (CpG) models. The discrepant SNPs between the
369 Illumina and ONT for both Case A and Case B were manually visualised in Tablet v1.17.08.17 [36] in

370 order to elucidate the reason for the discrepancy. Discordant SNPs being within a homopolymeric
371 region were also quantified.

372

373 *Generation of phylogenetic trees*

374 Filtered VCF files for each of the Illumina and ONT sequencing data for each sample, were
375 incorporated, into SnapperDB v0.2.5 [30] containing variant calls from 4471 other STEC CC11
376 genomes generated through routine surveillance by Public Health England. SnapperDB v0.2.5 [30]
377 was used to generate a whole genome alignment of the 4475 genomes (including both datasets for
378 the selected strains for this study). Both methylated positions and prophage positions were masked
379 from the alignment. The alignment was processed through Gubbins V2.0.0 [37] to account for
380 recombination events. A maximum likelihood tree was then constructed using RAxML V8.1.17 [38].

381

382 *Assembly of ONT data*

383 Trimmed ONT FASTQ files were assembled using Canu v1.6 [39]. Polishing of the assemblies was
384 performed using Nanopolish v0.10.2 [21] using both the trimmed ONT FASTQs and FAST5s for each
385 respective sample accounting for methylation using the --methylation-aware option set to dcm.
386 Assemblies were reoriented to start at the *dnaA* gene (NC_000913) from *E. coli* K12, using the
387 fixstart parameter in circulator v1.5.5 [40].

388

389 *Hybrid assemblies*

390 Trimmed ONT FASTQ files were assembled using Unicycler v0.4.2 [41] with the following parameters
391 min_fasta_length=1000, mode=normal and -1 and -2 for the incorporation of each sample's
392 equivalent Illumina FASTQ. Pilon v1.23 [42] was used to correct the assembly using the Illumina
393 reads.

394

395 *Assembly of Illumina data*

396 Illumina reads were assembled using SPAdes v3.13.0 [43] with the careful parameter activated and
397 with kmer lengths of 21, 33, 55, 65, 77, 83 and 91.

398

399 *Annotation*

400 Prokka v1.13 [44] with the species set to *E. coli* was used to annotate the final assemblies.

401 Mauve snapshot_2015-02-25 (1) [45] using the "move contig" function was used to align each

402 assembly to the ONT reference as they had the least number of contigs.

403

404 **Availability of supporting data**

405 The FASTQ files for the paired read Illumina sequence data can be found on the NCBI (National
406 Center for Biotechnology Information) Sequence Read Archive (SRA); Case A accession: SRR7184397,
407 Case B accession - SRR6052929. The ONT FASTQ files, Case A accession – SRR7477814, Case B
408 accession - SRR7477813. All files can be found under BioProject - PRJNA315192.

409

410 **Abbreviations**

411 AUC: Area under curve; BWA: Burrows-Wheeler aligner; CC: Clonal complex; Dam: DNA adenine
412 methyltransferase; Dcm: DNA cytosine methylase; GATK: Genome analysis toolkit; HUS: Haemolytic
413 Uremic Syndrome; MLST: Multi-locus sequence type; NCBI: National Center for Biotechnology
414 Information; ONT: Oxford Nanopore Technologies; PHE: Public Health England; SNP: Single
415 nucleotide polymorphism; SRA: Sequence read archive; STEC: Shiga toxin-producing *Escherichia coli*;
416 VCF: Variant call format; WGS: Whole genome sequencing.

417

418 **Author contributions**

419 CJ and TJD conceptualised the project. CJ and DRG performed DNA extractions. DRG performed
420 library preparation and Nanopore sequencing. TJD and DRG processed Illumina sequence data. DRG
421 processed all ONT data. TJD performed ONT optimisation. DRG performed methylation analysis. TJD
422 and DRG performed Illumina and ONT data comparison. CJ wrote the original draft. DRG, CJ, SG and
423 TJD performed manuscript editing.

424

425 **Competing interests**

426 This project was part funded by Oxford Nanopore Technologies.

427

428 **Acknowledgements**

429 We would like to thank Oxford Nanopore Technologies for funding this research. In particular we
430 would like to thank Leila Luheshi and Divya Mirrington for their support and scientific assistance.

431 We would also like to thank the frontline NHS Laboratories for submitting the samples used in this
432 study to the Gastrointestinal Bacteria Reference Unit at Public Health England.

433 We would like to acknowledge Dr Andrew Page of the Quadram Institute for critically reviewing the
434 manuscript.

435

436 **References**

- 437 1. Croxen MA, Law RJ, Scholz R, Keeney KM, Wlodarska M, Finlay BB. Recent advances in
438 understanding enteric pathogenic *Escherichia coli*. Clin Microbiol Rev. 2013. 26(4):822-80.
439 doi: 10.1128/CMR.00022-13.
- 440 2. Byrne L, Jenkins C, Launders N, Elson R, Adak GK. The epidemiology, microbiology and
441 clinical impact of Shiga toxin-producing *Escherichia coli* in England, 2009-2012. Epidemiol
442 Infect. 2015. 143(16):3475-87. doi: 10.1017/S0950268815000746.
- 443 3. Launders N, Byrne L, Jenkins C, Harker K, Charlett A, Adak GK. Disease severity of Shiga toxin-
444 producing *E. coli* O157 and factors influencing the development of typical haemolytic
445 uraemic syndrome: a retrospective cohort study, 2009-2012. BMJ Open. 2016. 6(1):e009933.
446 doi: 10.1136/bmjopen-2015-009933.
- 447 4. Heiman KE, Mody RK, Johnson SD, Griffin PM, Gould LH. *Escherichia coli* O157 Outbreaks in
448 the United States, 2003-2012. Emerg Infect Dis. 2015. 21(8):1293-1301. doi:
449 10.3201/eid2108.141364.
- 450 5. Dallman TJ, Ashton PM, Byrne L, Perry NT, Petrovska L, Ellis R, Allison L, Hanson M, Holmes
451 A, Gunn GJ, Chase-Topping ME, Woolhouse ME, Grant KA, Gally DL, Wain J, Jenkins C.
452 Applying phylogenomics to understand the emergence of Shiga-toxin-producing *Escherichia*
453 *coli* O157:H7 strains causing severe human disease in the UK. Microb Genom. 2015.
454 1(3):e000029. doi: 10.1099/mgen.0.000029.
- 455 6. Dallman TJ, Byrne L, Ashton PM, Cowley LA, Perry NT, Adak G, Petrovska L, Ellis RJ, Elson R,
456 Underwood A, Green J, Hanage WP, Jenkins C, Grant K, Wain J. Whole-genome sequencing
457 for national surveillance of Shiga toxin-producing *Escherichia coli* O157. Clin Infect Dis. 2015.
458 61(3):305-12. doi: 10.1093/cid/civ318.
- 459 7. Olson ND, Lund SP, Colman RE, Foster JT, Sahl JW, Schupp JM, Keim P, Morrow JB, Salit ML,
460 Zook JM. Best practices for evaluating single nucleotide variant calling methods for microbial
461 genomics. Front Genet. 2015. 6(235): doi: 10.3389/fgene.2015.00235.
- 462 8. Ruffalo M, Koytürk M, Ray S, LaFramboise T. Accurate estimation of short read mapping
463 quality for next-generation genome sequencing. Bioinformatics. 2012. 28(18):i349-i355. doi:
464 10.1093/bioinformatics/bts408.
- 465 9. Timme RE, Rand H, Sanchez Leon M, Hoffmann M, Strain E, Allard M, Roberson D, Baugher
466 JD. GenomeTrakr proficiency testing for foodborne pathogen surveillance: an exercise from
467 2015. Microb Genom. 2018. 4(7). doi: 10.1099/mgen.0.000185.
- 468 10. Quick J, Loman NJ, Duraffour S, Simpson JT, Severi E, Cowley L, Bore JA, Koundouno R, Dudas
469 G, Mikhail A, Ouédraogo N, Afrough B, Bah A, Baum JH, Becker-Ziaja B, Boettcher JP, Cabeza-
470 Cabrerizo M, Camino-Sanchez A, Carter LL, Doerrbecker J, Enkirch T, Dorival IGG, Hetzelt N,
471 Hinzmann J, Holm T, Kafetzopoulou LE, Koropogui M, Kosgey A, Kuisma E, Logue CH,
472 Mazzarelli A, Meisel S, Mertens M, Michel J, Ngabo D, Nitzsche K, Pallash E, Patrono LV,
473 Portmann J, Repits JG, Rickett NY, Sachse A, Singethan K, Vitoriano I, Yemanaberhan RL,
474 Zekeng EG, Trina R, Bello A, Sall AA, Faye O, Faye O, Magassouba N, Williams CV, Amburgey
475 V, Winona L, Davis E, Gerlach J, Washington F, Monteil V, Jourdain M, Bererd M, Camara A,
476 Somlare H, Camara A, Gerard M, Bado G, Baillet B, Delaune D, Nebie KY, Diarra A, Savane Y,
477 Pallawo RB, Gutierrez GJ, Milhano N, Roger I, Williams CJ, Yattara F, Lewandowski K, Taylor J,
478 Rachwal P, Turner D, Pollakis G, Hiscox JA, Matthews DA, O'Shea MK, Johnston AM, Wilson
479 D, Hutley E, Smit E, Di Caro A, Woelfel R, Stoecker K, Fleischmann E, Gabriel M, Weller SA,
480 Koivogui L, Diallo B, Keita S, Rambaut A, Formenty P, Gunther S, Carroll MW. Real-time
481 portable genome sequencing for Ebola surveillance. Nature. 2016. 530(7589):228-32 doi:
482 10.1038/nature16996.

- 483 11. Gomez-Eichelmann MC, Levy-Mustri A, Ramirez-Santos J. Presence of 5-methylcytosine in
484 CC(A/T)GG sequences (Dcm methylation) in DNAs from different bacteria. *J Bacteriol.* 1991.
485 173(23):7692-4.
- 486 12. Mikheyev AS, Tin MM. A first look at the Oxford Nanopore MinION sequencer. *Mol Ecol*
487 *Resour.* 2014. 14(6):1097-102. doi: 10.1111/1755-0998.12324.
- 488 13. Laver T, Harrison J, O'Neill PA, Moore K, Farbos A, Paszkiewicz K, Studholme DJ. Assessing
489 the performance of the Oxford Nanopore Technologies MinION. *Biomol Detect Quantif.*
490 2015. 1-8.
- 491 14. Magi A, Semeraro R, Mingrino A, Giusti B, D'Aurizio R. Nanopore sequencing data analysis:
492 state of the art, applications and challenges. *Brief Bioinform.* 2017. doi: 10.1093/bib/bbx062.
493 Epub ahead of print.
- 494 15. Rang FJ, Kloosterman WP, de Ridder J. From squiggle to basepair: computational approaches
495 for improving nanopore sequencing read accuracy. *Genome Biol.* 2018. 19(1):90. doi:
496 10.1186/s13059-018-1462-9.
- 497 16. Senol Cali D, Kim JS, Ghose S, Alkan C, Mutlu O. Nanopore sequencing technology and tools
498 for genome assembly: computational analysis of the current state, bottlenecks and future
499 directions. *Brief Bioinform.* 2018. doi: 10.1093/bib/bby017. Epub ahead of print.
- 500 17. Marinus MG. DNA methylation in *Escherichia coli*. *Annu Rev Genet.* 1987. 21:113-31.
- 501 18. Jain M, Koren S, Miga KH, Quick J, Rand AC, Sasani TA, Tyson JR, Beggs AD, Dilthey AT, Fiddes
502 IT, Malla S, Marriott H, Nieto T, O'Grady J, Olsen HE, Pedersen BS, Rhie A, Richardson
503 H, Quinlan AR, Snutch TP, Tee L, Paten B, Phillippy AM, Simpson JT, Loman NJ, Loose M.
504 Nanopore sequencing and assembly of a human genome with ultra-long reads. *Nat*
505 *Biotechnol.* 2018. 36(4):338-345. doi: 10.1038/nbt.4060.
- 506 19. Ebler J, Haukness M, Pesout T, Marshall T, Paten, B. Haplotype-aware genotyping from noisy
507 long reads. *bioRxiv.* 2018. 293944. <https://doi.org/10.1101/293944>.
- 508 20. Sarkozy P, Jobbágy Á, Antal P. Calling homopolymer stretches from raw Nanopore reads by
509 analyzing k-mer dwell times. In: Eskola H., Väisänen O., Viik J., Hyttinen J. (eds) *EMBECC &*
510 *NBC 2017. EMBEC 2017, NBC 2017. IFMBE Proceedings, 2018. vol 65.* Springer, Singapore.
- 511 21. Loman NJ, Quick J, Simpson JT. A complete bacterial genome assembled de novo using only
512 nanopore sequencing data. *Nat Methods.* 2015. 12(8):733-5. doi: 10.1038/nmeth.3444.
- 513 22. Wick RR, Judd LM, Gorrie CL, Holt KE. Completing bacterial genome assemblies with
514 multiplex MinION sequencing. *Microb Genom.* 2017. 3(10):e000132. doi:
515 10.1099/mgen.0.000132.
- 516 23. Bolger AM, Lohse M, Usadel B. Trimmomatic: A flexible trimmer for Illumina Sequence
517 Data. *Bioinformatics.* 2014. 30(15):2114-20. doi: 10.1093/bioinformatics/btu170.
- 518 24. Tewolde R, Dallman T, Schaefer U, Sheppard CL, Ashton P, Pichon B, Ellington M, Swift C,
519 Green J, Underwood A. MOST: a modified MLST typing tool based on short read sequencing.
520 *PeerJ.* 2016. 4:e2308. doi: 10.7717/peerj.2308.
- 521 25. Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2. *Nat Methods.* 2012.
522 9(4):357-9. doi: 10.1038/nmeth.1923.
- 523 26. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R,
524 1000 Genome Project Data Processing Subgroup. The Sequence alignment/map (SAM)
525 format and SAMtools. *Bioinformatics.* 2009. 25(16):2078-9. doi:
526 10.1093/bioinformatics/btp352.

- 527 27. Ashton PM, Perry N, Ellis R, Petrovska L, Wain J, Grant KA, Jenkins C, Dallman TJ. Insight into
528 Shiga toxin gene encoded by *Escherichia coli* O157 from whole genome sequencing. PeerJ.
529 2015. 17. doi: 10.7717/peerj.739.
- 530 28. Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler Transform.
531 Bioinformatics. 2009. 25(14):1754-60. doi: 10.1093/bioinformatics/btp324.
- 532 29. McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytzky A, Garimella K, Altshuler
533 D, Gabriel S, Daly M, DePristo MA. The Genome Analysis Toolkit: a MapReduce frame- work
534 for analyzing next-generation DNA sequencing data. Genome Res. 2010. 20(9):1297-303.
535 doi: 10.1101/gr.107524.110.
- 536 30. Dallman T Ashton P, Schafer U, Jironkin A, Painset A, Shaaban S, Hartman H, Myers R,
537 Underwood A, Jenkins C, Grant K. SnapperDB: A database solution for routine sequencing
538 analysis of bacterial isolates. Bioinformatics. 2018. doi: 10.1093/bioinformatics/bty212.
- 539 31. De Coster W, D'Hert S, Schultz DT, Cruts M, Van Broeckhoven C. NanoPack: visualizing and
540 processing long-read sequencing data. Bioinformatics. 2018. 34(15):2666-9. doi:
541 10.1093/bioinformatics/bty149.
- 542 32. Wick R. Unpublished. <https://github.com/rrwick/Porechop>.
- 543 33. Wood DE, Salzberg SL. Kraken: ultrafast metagenomic sequence classification using exact
544 alignments. Genome Biol. 2014. 15(3):R46. doi: 10.1186/gb-2014-15-3-r46.
- 545 34. Page A, Keane J. Rapid multi-locus sequence typing direct from uncorrected long reads using
546 Krocus. PeerJ. 2018. 6:e5233 doi: 10.7717/peerj.5233.
- 547 35. Li H. Minimap2: fast pairwise alignment for long nucleotide sequences. Bioinformatics. 2018.
548 doi: 10.1093/bioinformatics/bty191. Epub ahead of print.
- 549 36. Milne I Stephen G, Bayer M, Cock PJ, Pritchard L, Cardle L, Shaw PD, Marshall D. Using Tablet
550 for visual exploration of second-generation sequencing data. Briefings in Bioinformatics.
551 2013. 14(2):193-202. doi: 10.1093/bib/bbs012.
- 552 37. Croucher NJ, Page AJ, Connor TR, Delaney AJ, Keane JA, Bentley SD, Parkhill J, Harris SR.
553 Rapid phylogenetic analysis of large samples of recombinant bacterial whole genome
554 sequences using Gubbins. Nucleic Acids Res. 2014. 43(3):e15. doi: 10.1093/nar/gku1196.
- 555 38. Stamatakis A. 2014. RAxML Version 8: A tool for Phylogenetic Analysis and Post-Analysis of
556 Large Phylogenies. Bioinformatics. 2014. 30(9):1312-13. doi:
557 10.1093/bioinformatics/btu033.
- 558 39. Koren S, Walenz BP, Berlin K, Miller JR, Phillippy AM. 2017. Canu: scalable and accurate long-
559 read assembly via adaptive k-mer weighting and repeat separation. Genome Res. 27(5):722-
560 36. doi: 10.1101/gr.215087.116.
- 561 40. Hunt M, Silva ND, Otto TD, Parkhill J, Keane JA, Harris SR. 2015. Circlator: automated
562 circularization of genome assemblies using long sequencing reads. Genome Biol. 16(294):1-
563 10. doi: 10.1186/s13059-015-0849-0.
- 564 41. Wick RR, Judd LM, Gorrie CL, Holt KE. 2017. Unicycler: Resolving bacterial genome
565 assemblies from short and long sequencing reads. PLoS Comput Biol. 13(6):e1005595. doi:
566 10.1371/journal.pcbi.1005595.
- 567 42. Walker BJ, Abeel T, Shea T, Priest M, Abouelliel A, Sakthikumar S, Cuomo CA, Zeng Q,
568 Wortman J, Young SK, Earl AM. 2014. Pilon: an integrated tool for comprehensive microbial
569 variant detection and genome assembly improvement. PLOS One. 9(11):e112963. doi:
570 10.1371/journal.pone.0112963.

- 571 43. Bankevich A, Nurk S, Antipov D, Gurevich AA, Dvorkin M, Kulikov AS, Lesin VM, Nikolenko SI,
572 Pham S, Prjibelski AD, Pyshkin AV, Sirotkin AV, Vyahhi N, Tesler G, Alekseyev MA, Pevzner
573 PA. 2012. SPAdes: A new genome assembly algorithm and its applications to single cell
574 sequencing. *J Comput Biol.* 19(5):455-77. doi: 10.1089/cmb.2012.0021.
- 575 44. Seemann T. 2014. Prokka: rapid prokaryotic genome annotation. *Bioinformatics.*
576 30(14):2068-9. doi: 10.1093/bioinformatics/btu153.
- 577 45. Darling ACE, Mau B, Blattner FR, Perna NT. 2004. Mauve: Multiple alignment of conserved
578 genomic sequence with rearrangements. *Genome Res.* 14(7):1394-403. doi:
579 10.1101/gr.2289704.
- 580
581
582
583

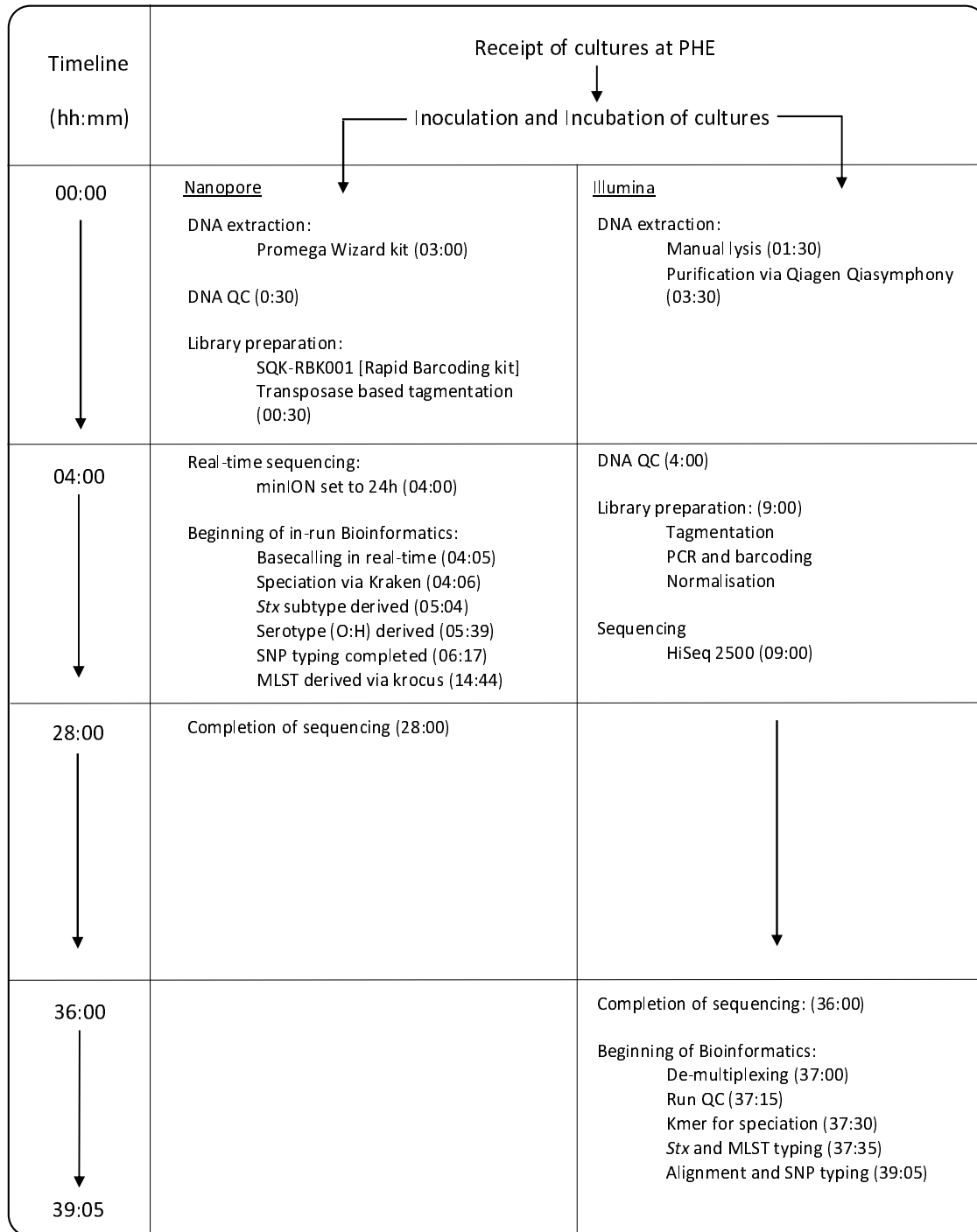


Figure 1 – Figure showing comparative timeline from beginning DNA extraction to results generation for Oxford Nanopore and Illumina technologies. Times shown the completion of the labelled event relative to the start of the assay (hh:mm).

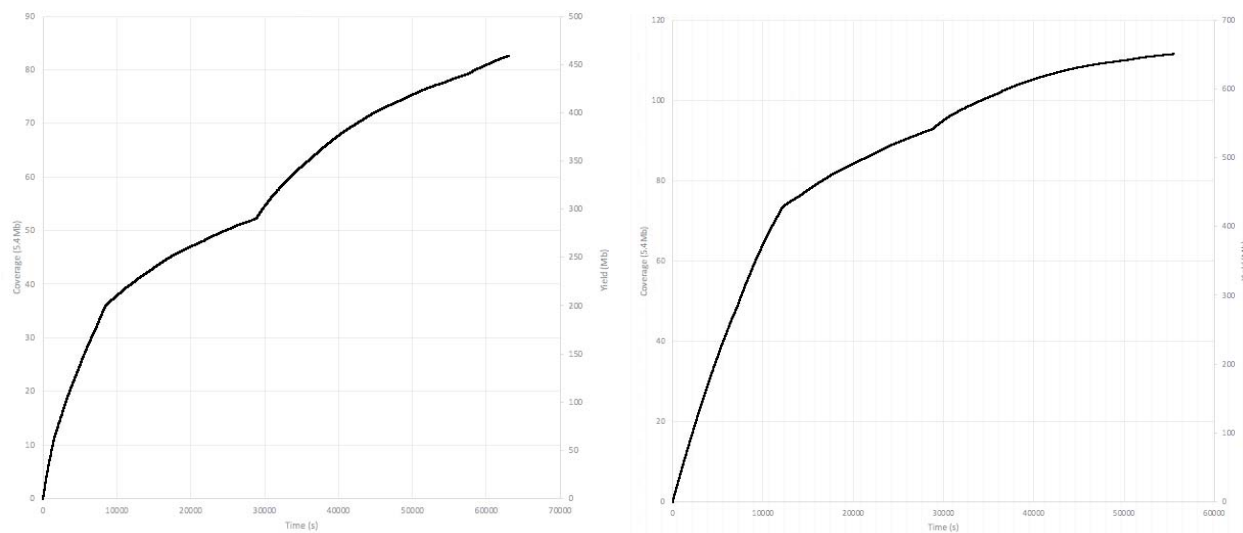


Figure 2 – Two time/yield/coverage graphs showing production of reads in real-time and the associated cumulative mapping coverage. Case A is the graph on the left and Case B is on the right.

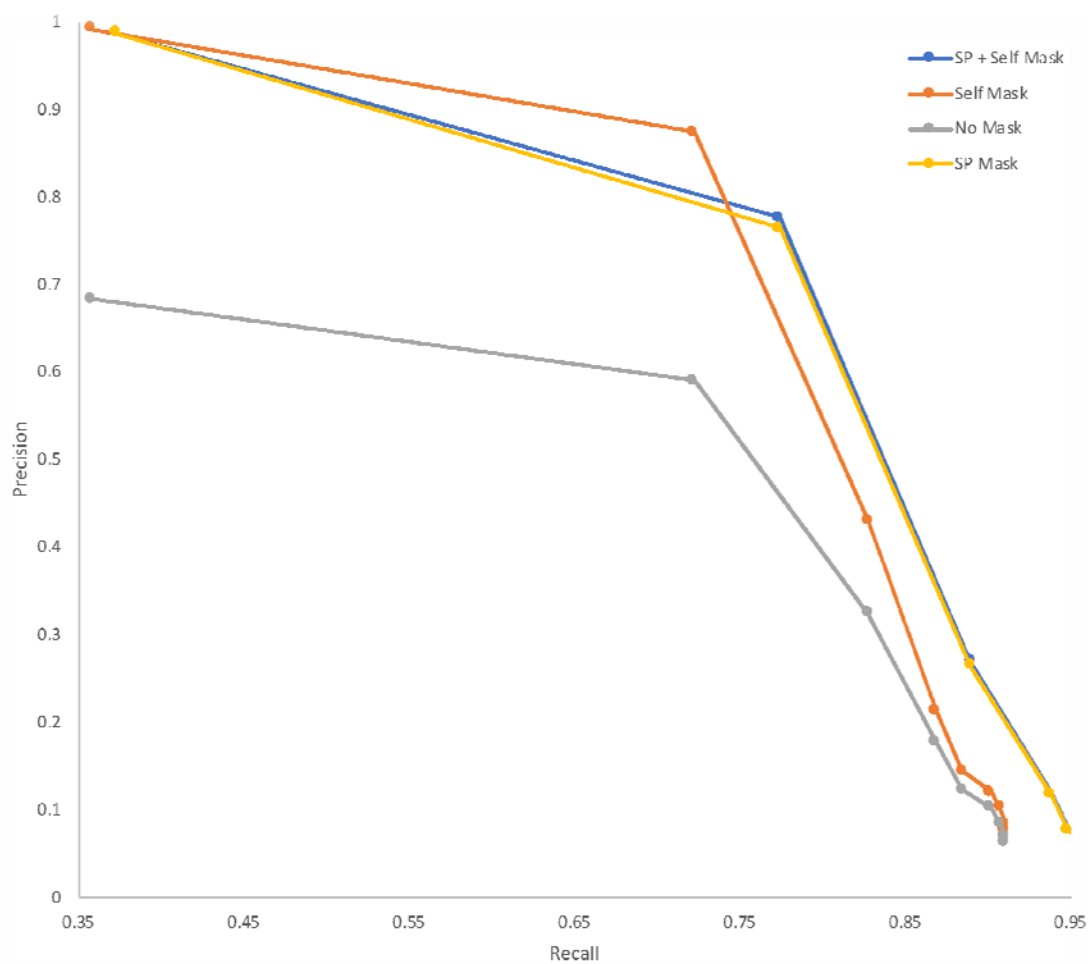


Figure 3 – Precision Vs Recall of variant calling for an array of consensus ratio cut-offs and pre-masking strategies including masking positions annotated as ‘Sakai phage’ (‘SP’) and positions that are ambiguously self-mapped (‘Self’) with simulated Illumina FASTQs from the reference genome. Performed on case B.

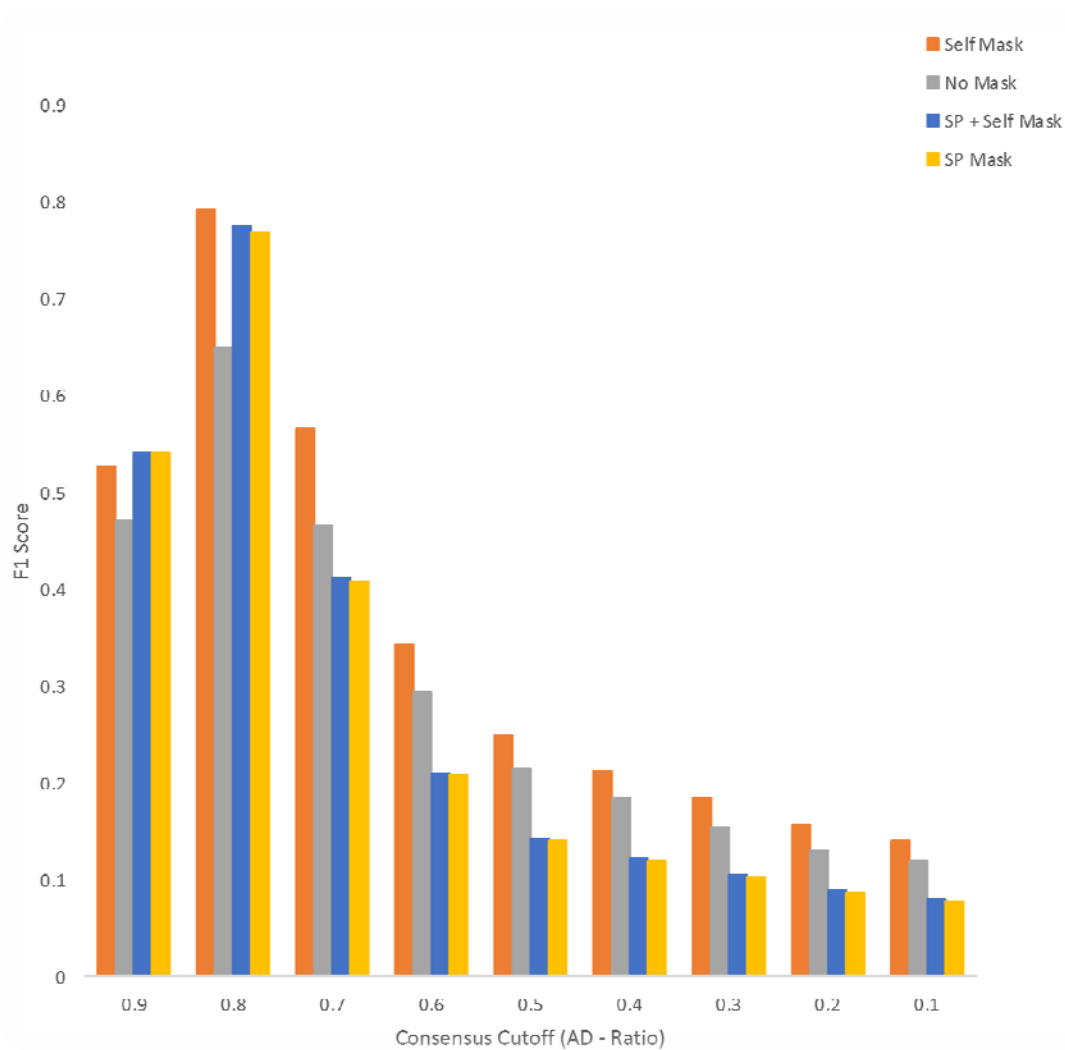


Figure 4 – F1 Score for an array of consensus ratio cut-offs and pre-masking strategies including masking positions annotated as ‘Sakai phage’ (‘SP’) and positions that are ambiguously self-mapped (‘Self’) with simulated Illumina FASTQs from the reference genome.

	Position					Case	
	-2	-1	Variant	+1	+2	A	B
Reference	C	C	A	G	G	A > G Transition (n=181)	69.62% (n=73)
Alignment	C	C	G	=	G		
Reference	C	C	T	G	G	T > C Transition (n=79)	30.38% (n=21)
Alignment	C	C	C	G	G		
Total						100% (n=260)	100% (n=94)

Figure 5 – Figure showing the two most common discrepancies in the ONT optimised GATK VCFs and a breakdown of the relative proportions of these transitions compared to the total number of

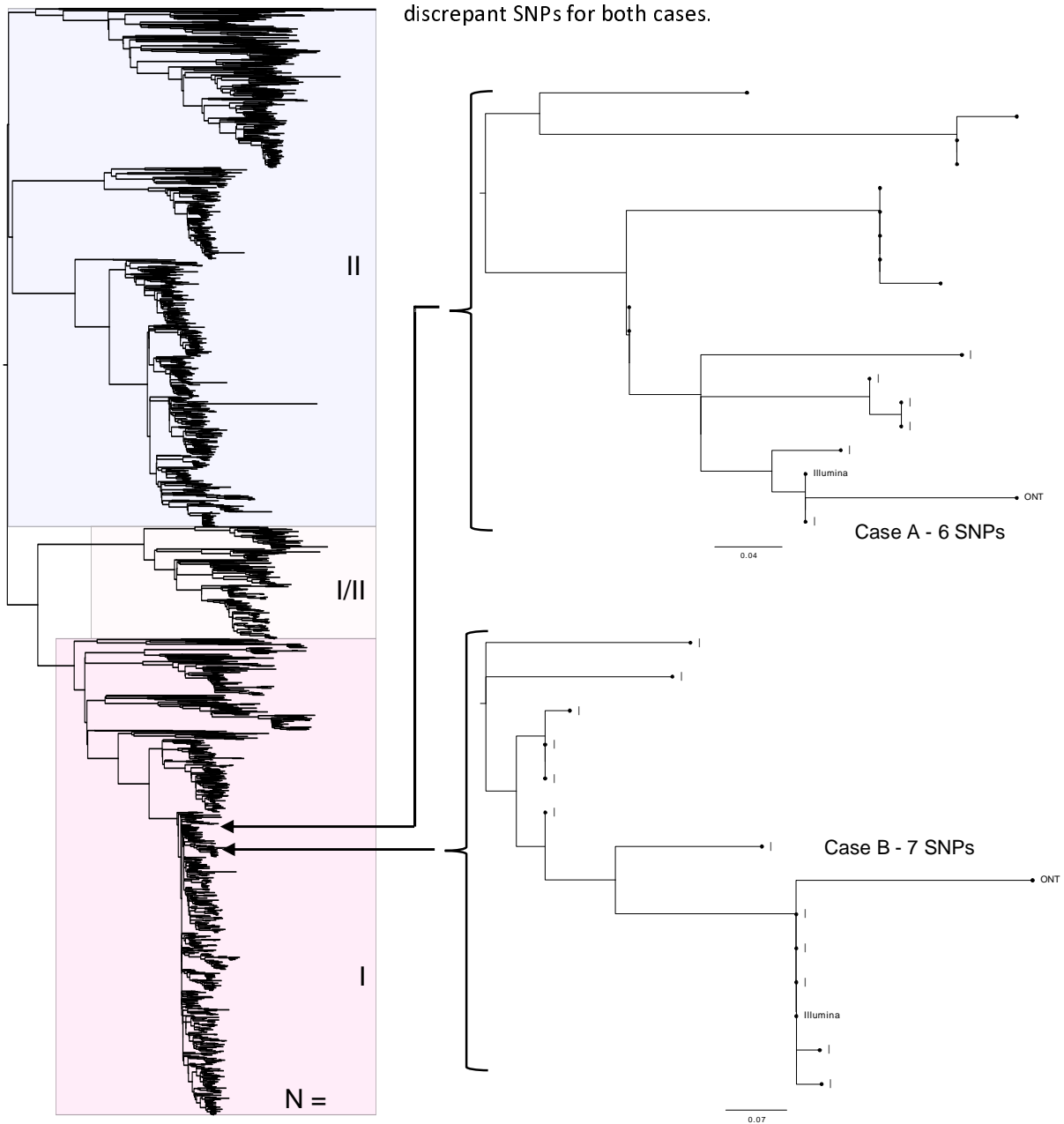


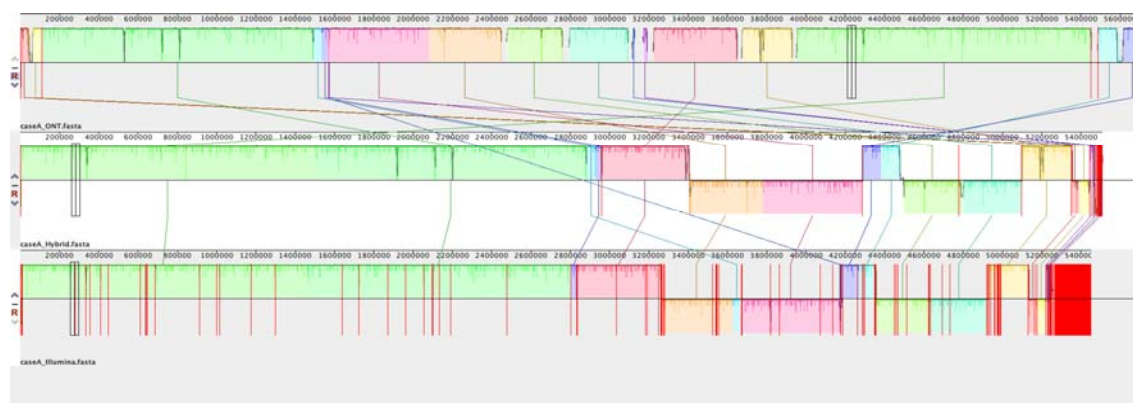
Figure 6 – Maximum likelihood tree, of a “soft core” alignment of 4475 genomes showing the tree lineages (I, I/II and II) of STEC (Clonal Complex 11). Also showing where Oxford Nanopore and Illumina sequencing data is placed within the tree for each of the two cases. All methylated positions and prophage regions have been masked.

Supplementary Tables

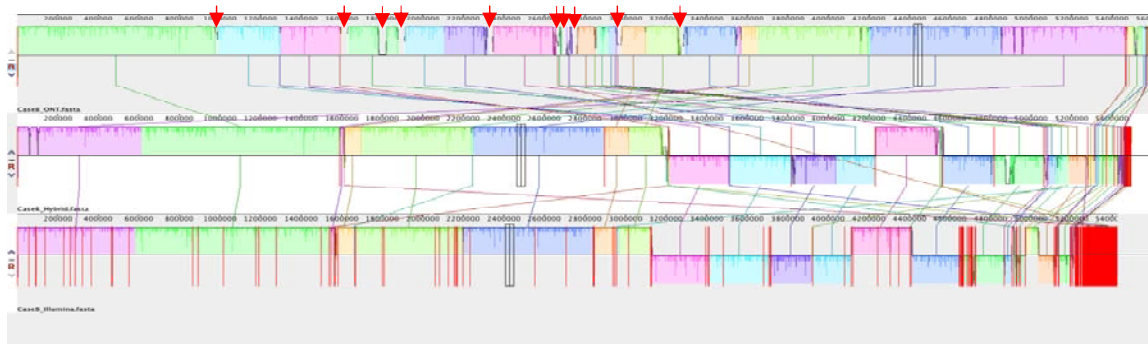
Case	# of contigs in ONT-only assembly (size bp)	# of contigs in hybrid assembly (size bp)	# of contigs in Illumina-only assembly (size bp)
A	5 (5,725,666 bp)	25 (5,506,670 bp)	668 (5,449,735 bp)
B	4 (5,620,611 bp)	34 (5,491,608 bp)	575 (5,424,436 bp)

Table 1 – Table showing the number of contigs generated and size of assembly for each assembly method for both cases.

Supplementary Figures



Supplementary figure 1 – MaVUE alignment showing regions of similarity between the ONT-only, hybrid and Illumina-only assemblies (order descending) for Case A. Also showing the chromosomal regions in the ONT-only assembly that did not match the other assemblies (red arrows).



Supplementary figure 2 – Mauve alignment showing regions of similarity between the ONT-only, hybrid and Illumina-only assemblies (order descending) for Case B. Also showing the chromosomal regions in the ONT-only assembly that did not match the other assemblies (red arrows).