

1 Rapid, multiplexed, whole genome and plasmid sequencing of foodborne pathogens using long-read
2 nanopore technology

3

4 Tonya L. Taylor¹, Jeremy D. Volkening², Eric DeJesus³, Mustafa Simmons³, Kiril M. Dimitrov¹, Glenn E.
5 Tillman³, David L. Suarez¹ and Claudio L. Afonso^{1*}

6

7 1. Exotic and Emerging Avian Viral Diseases Research Unit, Southeast Poultry Research Laboratory, US
8 National Poultry Research Center, Agricultural Research Service, USDA, 934 College Station Road,
9 Athens, GA, USA.

10 2. BASE₂BIO, Oshkosh, WI, USA.

11 3. Office of Public Health Services, Food Safety and Inspection Services, USDA, 950 College Station
12 Road, Athens, GA, USA

13

14 *Corresponding author: **Claudio L. Afonso**. Phone: +1 (706) 546 – 3642. Email:

15 Claudio.Afonso@ars.usda.gov. Mailing address: USDA ARS, Southeast Poultry Research Laboratory, 934

16 College Station Rd., Athens, GA 30605. Fax: (706) 546-3161.

17

18 **Abstract**

19 United States public health agencies are focusing on next-generation sequencing (NGS) to
20 quickly identify and characterize foodborne pathogens. Here, the MinION nanopore, long-read
21 sequencer was used to simultaneously sequence the entire chromosome and plasmids of *Salmonella*
22 *enterica subsp. enterica* serovar Bareilly and *Escherichia coli* O157:H7. A rapid, random sequencing
23 approach, coupled with *de novo* genome assembly within a customized data analysis workflow, that can
24 resolve highly-repetitive genomic regions, was developed. In sequencing runs, as short as four hours,
25 using nanopore data alone, full-length genomes were obtained with an average identity of 99.87% for
26 *Salmonella* Bareilly and 99.89% for *E. coli* in comparison to the respective MiSeq references. These long-
27 read assemblies provided information on serotype, virulence factors, and antimicrobial resistance genes.
28 Using a custom-developed, SNP-selection workflow, the potential of the nanopore-only assemblies
29 (after only 30 minutes of sequencing) for rapid phylogenetic inference, with identical topology
30 compared to the published dataset, was demonstrated. To achieve maximum quality assemblies, the
31 developed bioinformatics workflow employed additional polishing steps to correct the systematic errors
32 produced by the nanopore-only assemblies. Nanopore sequencing provided a shorter (10 hours library
33 preparation and sequencing) turnaround time compared to other NGS technologies.

34

35 **Keywords:** MinION, complete genome sequencing, plasmids, foodborne pathogen, long-reads, assembly
36 workflow, tracing.

37 Introduction

38 United States public health agencies routinely perform surveillance on microbial foodborne
39 pathogens, and in the U.S. alone each year, approximately 1 in 6 individuals are sickened by foodborne
40 illnesses, resulting in approximately 3,000 deaths¹. During outbreak responses, identification of the
41 source is instrumental for the fast removal of the contaminated items from public circulation. However,
42 specific characterization of foodborne pathogens during these surveillance programs in food production
43 and distribution is important, as it allows for early warnings and possible removal of the contaminated
44 food product(s) before the development of an outbreak¹.

45 To that end, U.S. public health agencies have employed next-generation sequencing (NGS) using
46 short-read sequencing technology in surveillance activities and outbreak response². In addition to
47 utilizing whole genome sequencing (WGS) for pathogen identification, more detailed information on the
48 pathogen such as virulence, antimicrobial resistance, serotype, and inference of possible links between
49 the sources of contamination is obtained³. WGS has provided faster identification of pathogens from
50 contaminated sources of outbreaks, reduced the number of illnesses and deaths due to the foodborne
51 infections, and decreased the number of isolates needed to link the illness to the source of
52 contamination^{4,5}.

53 Although WGS is now a routine procedure in epidemiologic investigation and surveillance of
54 foodborne pathogens, short-read, sequencing technology present challenges such as resolving repetitive
55 regions, leading to incomplete *de novo* assemblies (severe fragmentation)⁶⁻⁸. These gaps can lead to the
56 inability to determine genome organization or architecture, which can be important in determining if
57 genes are co-regulated or co-transmissible in the case of genes associated with mobile elements⁹. Even
58 though the short-reads are accurate, closed whole genome assemblies are now commonly accomplished
59 using a combination of both short-read (for base accuracy) and long-read sequencing technologies (for
60 structural accuracy)¹⁰⁻¹².

61 Long-read sequencing, enabled by single-molecule real-time (SMRT) sequencing technology that
62 has been utilized since 2004, can produce reads averaging 11kb in length, which facilitates the
63 completion of bacterial genome assemblies that are either lacking in sequencing depth at certain
64 repetitive areas of the genome or have areas that are missing reads completely using short-read
65 technology¹³. The long-reads span across these large repetitive regions¹⁴⁻¹⁶ and can provide unbiased
66 coverage of regions sequenced poorly with other technologies due to G/C content or other
67 characteristics^{13,17}. However, there is still a need for an approach that generates inexpensive, long-read
68 data in a short turnaround time to be utilized for both rapid detection of an organism, complete
69 sequencing of bacterial chromosomes and plasmids, and complementation to other sequencing
70 technologies used in both outbreak investigations and foodborne pathogen surveillance.

71 Oxford Nanopore has developed technology to fit this role in the form of the MinION nanopore-
72 based sequencer that produces long, single-molecule reads. This novel sequencer is pocket-size (10cm x
73 2cm x 3.3cm) and powered directly by a USB port from a laptop computer¹⁸. It is portable, field-
74 deployable, inexpensive, and can provide sequencing of both DNA and RNA in real time. Nanopore
75 sequencing still produces systematic errors, and for that reason, it has previously only been used as a
76 complement of short read sequencing. The use of nanopore-only assemblies for full-length genome
77 sequencing, genome structure determination, antimicrobial resistance gene identification, and
78 phylogenetic analysis would be a significant advance that would allow the universal low-cost access of
79 information necessary for critical management decisions. Since the release of the MinION platform,
80 bioinformatic tools have been steadily evolving, with the goal of using nanopore data to assemble
81 accurate, whole, bacterial genomes independent of any other sequencing technology¹⁹.

82 In this study, utilizing the nanopore technology, we aimed to simultaneously sequence and
83 assemble complete genomes of two pathogenic bacterial strains that can cause human illness
84 worldwide, *Salmonella enterica subsp. enterica* serovar Bareilly and *Escherichia coli* O157:H7. Using a

85 custom, reproducible bioinformatics workflow that employs publicly-available tools, the circularized
86 bacterial genomes and associated plasmids of both strains were assembled and polished with a final
87 error rate of only 0.1 %. This study shows that long-read nanopore sequencing can be used as a low-cost
88 method to generate closed assemblies of microbial foodborne pathogen genomes and associated
89 plasmids and demonstrates that the data is of sufficient quality for phylogenetic classification of
90 *Salmonella* isolates. These closed assemblies provide information on genome organization and can
91 complement existing characterization data from other technologies such as short-read sequencing.

92

93 **Materials and Methods**

94 *Bacterial cultures and DNA extraction*

95 The *Salmonella* Bareilly isolate (CFSAN000189) was isolated from raw shrimp in India (Biosample
96 SAMN04364135), and the *E.coli* O157:H7 isolate (FSIS11705876) was isolated from domestic, raw, ground
97 beef collected by the U.S. Department of Agriculture Food Safety and Inspection Services (USDA-FSIS) as
98 part of routine sampling of a U.S. establishment (Biosample SAMN08167607). Both bacterial isolates were
99 grown on sheep blood agar (SBA) for 24 hours at 35 °C. Total DNA from each isolate was extracted using
100 the DNeasy Blood and Tissue Kit (Qiagen, USA) following manufacturer's instructions. DNA concentrations
101 throughout the experiment were determined by using the Qubit® dsDNA HS Assay Kit on a Qubit®
102 fluorometer 3.0 (Thermo Fisher Scientific, USA).

103

104 *Library Preparation and MinION Sequencing*

105 The 1D gDNA long read selection protocol was used with the SQK-LSK108 kit (Oxford Nanopore
106 Technologies (ONT), UK) to prepare MinION-compatible libraries. The DNA shearing step was eliminated
107 from the protocol with the aim of selecting for very long reads. Approximately, 2µg of *E. coli* DNA and 2
108 µg of *Salmonella* DNA in a total of 100 µL each were added to the NEBNext® Ultra™ II End Repair/dA-

109 Tailing module (New England Biolabs (NEB), USA) for end repair and dA-Tailing, following manufacturer's
110 instructions, and purified using Agencourt AMPure XP beads (Beckman Coulter, USA). Each purified,
111 end-prepped DNA product was barcoded using a separate barcode from the 1D Native barcoding kit
112 (EXP-NBD103, ONT, UK) and following the 1D Native barcoding genomic DNA protocol. The samples
113 were then bead-purified (Beckman Coulter, USA), and equimolar amounts of each barcoded sample
114 were pooled together for a final quantity of 700ng. Adapters were ligated to the pooled sample using
115 Blunt/TA ligase (NEB, USA) following the 1D gDNA long read selection protocol. The MinION device was
116 used to sequence the created library on a new FLO-MIN106 R9.4 flow cell^{20,21}. The standard 48 hr 1D
117 sequencing protocol was initiated using the MinKNOW software (ONT, UK). Average quality and
118 coverage of the raw sequencing data were determined using CG-pipeline²².

119

120 *MiSeq Sequencing and Quality control*

121 To verify the newly developed approach used in this study, libraries for short-read WGS of the
122 *Salmonella* Bareilly and *E. coli* isolates were prepared using the Nextera XT kit (Illumina, USA) according
123 to the manufacturer's protocol. The libraries were loaded separately into a single flow cell of the 300
124 and 500 cycle MiSeq Reagent Kits v2 for *Salmonella* Bareilly and *E. coli*, respectively, and paired-end
125 sequencing (2×150 bp for *Salmonella* Bareilly and 2×250 bp for *E. coli*) was performed on the MiSeq
126 instrument (Illumina, USA). The produced raw data were analyzed using SPAdes version 3.71²³. Average
127 quality and coverage of the raw sequencing data were determined using CG-pipeline²².

128

129 *MinION Bacterial Bioinformatics Workflow for Whole Genome Assembly*

130 To analyze the MinION sequencing data, a customized workflow was developed. For subsequent
131 time analysis, the data was also analyzed at intervals from the start of the sequencing. Reads were
132 basecalled using Albacore (ONT v 2.0.2b) and subsampled for assembly using Filtlong (v.0.2.0)²⁴ to a

133 target depth of 75X. Fitlong subsampling is not random but keeps the longest and highest quality reads
134 from the input, which targets maximum sequencing depth (total bases). Read quality was weighted
135 more heavily than length ('mean_q_weight 5), as testing showed this was necessary to retain sufficient
136 coverage of small plasmids. The filtered reads were assembled using the Unicycler pipeline (v.0.4.7)²⁵.
137 This pipeline utilizes a minimap2/miniasm/racon iterative approach to assemble long-read-only data.
138 Since Unicycler sometimes fails to detect valid end overlaps, assemblies were circularized using a
139 custom script based on minimus2 (available in the workflow source repository)²⁶. Circular contigs were
140 rotated to start at a fixed position based on the reference. The consensus sequences were subjected to
141 two rounds of polishing using Nanopolish (v.0.10.2)²⁷, for which the full run (subject to time-based sub-
142 setting but prior to FitLong subsampling) was used, and Benchmarking Universal Single-Copy Orthologs
143 (BUSCO v.3.0.2)²⁸ was used to evaluate the completeness of coding sequences and degree of gene
144 fragmentation in the polished assemblies. To evaluate assembly accuracy, two procedures were used.
145 For the *Salmonella* Bareilly isolate, which has previously been sequenced and published²⁹, DNAdiff
146 (MUMMER v.3.23)³⁰ was used to evaluate both base-level and structural accuracy in the MinION
147 assembly compared to the published reference. For the *E.coli* isolate, lacking a published reference,
148 Illumina MiSeq reads were mapped to the assembly using BWA (v0.7.17), and LoFreq (v.2.1.3.1)³¹ was
149 used to call single nucleotide polymorphisms (SNPs) and small indels, from which the assembly accuracy
150 was calculated. Utilizing the short-read data, Pilon (v1.2.2)³² was used to error-correct small errors ('--fix
151 bases') in the assemblies using existing short-read data from the same isolates prior to GenBank
152 submission (SRA accession SRR498276 for *Salmonella* Bareilly; SRA accession SRR6373397 for *E. coli*
153 O157:H7).

154

155 *MinION Annotation*

156 The polished-MinION assemblies after 4 hours of sequencing were initially annotated using the
157 “Annotate From” tool within Geneious 11.1.5 and the published *Salmonella* Bareilly strain CFSAN000189
158 (GenBank Accession NC_021844) and *E. coli* O157:H7 strain 9234 (GenBank Accession CP017446)
159 sequences as references. ResFinder v.3.1 was used to locate any antimicrobial resistance genes and any
160 point mutations that would result in antimicrobial resistance³³. Additionally, to confirm the 4-hour
161 assembly annotation, the pilon-corrected, final genome sequences were submitted to GenBank to be
162 processed through the NCBI Prokaryotic Genomic Annotation Pipeline (PGAP) before being released.

163

164 *Phylogenetic Analysis*

165 Twenty-three *Salmonella* reference datasets were downloaded (Supplementary Table 4) and
166 used in tracing a foodborne outbreak in the U.S that were previously published^{29,34}. The eight sub-
167 sampled (15 mins to 1500 mins) unpolished *S. Bareilly* assemblies from this experiment were used to
168 generate simulated Illumina datasets using ART (150 × 2, 50X coverage, MiSeq platform, 300 bp mean
169 fragment length, 50 bp standard deviation)³⁵. All datasets were analyzed with a SNP-calling pipeline
170 using strain CFSAN000212 as a reference. Briefly, reads were optionally trimmed using Trim Galore
171 (Illumina datasets), aligned to the reference using BWA-MEM³⁶, called using LoFreq³¹, and filtered using
172 local scripts according to specific criteria. For Illumina datasets, the VCF files were filtered by removing
173 indels as well as any SNP with an alternate allele frequency of < 90%. Sites meeting one or more of the
174 following criteria were flagged as suspect, and these loci were ignored during matrix generation: i) sites
175 within 3 bp of a homopolymeric stretch of 4 bp or more; ii) sites occurring in a variant cluster (multiple
176 variants within 2 bp of each other; iii) sites within 10 bp of a dam or dcm methylation motif; and iv) sites
177 with observed A->G or T->C transition mutations. The remaining SNPs (64) were used to create a matrix
178 of variable sites for phylogenetic reconstruction. PhyML v3.0 was used to generate a maximum
179 likelihood SNP tree using the HYK85 model with 500 bootstrap iterations³⁷. In this analysis, the Illumina

180 result from the CFSAN000189 strain was replaced with the MinION-only assembly from sequencing the
181 same strain in this study to ensure that our data would not be topologically attracted by the reference
182 sequence and would cluster correctly without it. This tree was compared against the standard reference
183 tree (utilizing the 23 reference strains) provided by Timme et al., 2017³⁴.

184

185 *Availability of workflows, tools and code*

186 The full NextFlow workflow, Conda environment configuration, and other associated code used
187 in the analyses are publicly-available on GitHub (https://github.com/jvolkening/minion_bacterial).

188

189 **Results**

190 *Analysis of MinION and MiSeq Raw Data*

191 Before subsampling of the reads, the raw MinION sequencing data was used to estimate the
192 mean depth for *Salmonella* Bareilly and *E. coli*, respectively. A total of 2.8 billion bases from 333,298
193 *Salmonella* Bareilly reads, with an average read length of 8638 nucleotides (nt), yielded a mean depth of
194 599X. For *E. coli*, a total of 3.8 billion bases from 429,909 reads with an average read length of 8979 nt
195 were sequenced, and the mean depth was calculated to be 692X (Table 1). The shortest MinION read
196 was 85 nt, which was from the *E. coli* isolate, while the longest read was from *Salmonella* Bareilly and
197 was 129,119 nt. Both sets of MinION data had a mean read quality score above the standard ($Q \geq 10$),
198 which indicates superior quality.

199 Illumina MiSeq data was also analyzed using the same bioinformatic tool. The MiSeq raw data
200 had a depth of 57X for *Salmonella* Bareilly and 111X for *E. coli*. This sequencing technology produced
201 288 million bases from 1,930,511 *Salmonella* Bareilly reads, with an average read length of 150 nt. For *E.*
202 *coli*, a total of 556 million bases from 2,291,825 reads were sequenced that had an average read length
203 of 243 nt (Table 1). The minimum read length from both sets of bacterial sequences was 35 nt, while the

204 longest was 151 nt for *Salmonella* Bareilly and 251 nt for *E. coli*; the MiSeq mean read quality was above
205 the Q30 benchmark, signaling exceptional data.

206

207 *Assembly of MinION sequencing data*

208 The raw MinION data for both isolates were subsampled on the basis of cumulative run time in
209 order to simulate the effect of run length on final assembly quality. Subsets of reads generated in the
210 first 15, 30, 60, 120, 240, 480, and 960 minutes (mins), in addition to the full run length, were analyzed
211 (Tables 2 and 3). Four hours (240 mins) was determined as the shortest run time sufficient to assemble
212 circular sequences from all chromosomes and plasmids from both isolates and represented a point after
213 which longer run times resulted in remarkably diminishing gains in final accuracy (Fig S1). Data at each
214 of the other run time subsets is available in Tables 2, 3, S2 and S3; however, the following analyses
215 herein refer to the data collected in the first 240 mins of sequencing.

216 The MinION sequencing data was assembled using a custom Nextflow³⁸ workflow that utilized
217 publicly-available tools. Fitlong length- and quality-based subsampling to a 75X target depth resulted in
218 28,492 reads for the *Salmonella* Bareilly isolate, which were assembled into two circular contigs, the
219 chromosome and plasmid, with an average nucleotide identity of 99.87% and 100% coverage compared
220 to the reference genome (Table 2). For the *E. coli* isolate, 19,589 subsampled reads produced two
221 circular contigs, the chromosome and plasmid, with an average nucleotide identity of 99.89% compared
222 to the available MiSeq data of the same bacterium (Table 3). The *Salmonella* Bareilly genome assembled
223 into one chromosomal contig of 4,724,389 bp and one plasmid of 81,761 bp (Table 2). The *E. coli*
224 O157:H7 genome assembled into one chromosomal contig of 5,482,542 bp and one plasmid of 94,503
225 bp (Table 3).

226 The final genome assemblies utilized two rounds of polishing using Nanopolish, which
227 represented, by far, the most time-consuming and resource-intensive portion of the analysis workflow.

228 However, it also increased the overall accuracy (Fig 1a) due to a decrease in both SNPs (Fig 1b) and
229 chromosomal insertions or deletions (Fig 1c). The largest gains in accuracy were achieved from the first
230 round of polishing, while much less but still noticeable improvement was achieved with the second
231 round, particularly when examining completeness of genome annotation as measured by BUSCO.
232 However, further rounds (>2) of polishing did not significantly impact the overall assembly (Fig 1). To
233 demonstrate these data, relative time, central processing units (CPU), and memory consumption for
234 each step of the workflow can be found in supplementary Table S1.

235 For the *Salmonella* Bareilly assembly at 4 hours, the rate of single nucleotide polymorphisms
236 (SNPs) per kilobase (kb) decreased from 2.41 to 0.42 after one round of polishing and to 0.26 after two
237 rounds of polishing. At the same time point, the insertions or deletions (indels) per kb decreased from
238 3.91 to 1.14 after one round of polishing and to 1.03 after two rounds of polishing (Supplemental Table
239 S2). For the *E. coli* assembly at the same time point, the SNPs per kb decreased from 2.20 to 0.37 after
240 only one round of polishing and to 0.2 after two rounds of polishing. The indels per kb also decreased
241 from 3.86 to 1 to 0.89 (Supplemental Table S3). Additionally, the BUSCO tool was used to further analyze
242 the polished data to determine the completeness of the gene content based on quality and length of
243 alignment. The “BUSCO completeness” (fraction of expected gene complement with full-length reading
244 frames) value of both bacterial assemblies and the rounds of polishing were directly related, increasing
245 from 21 and 23% for the *Salmonella* and *E. coli* assemblies, respectively, with no polishing to 65 and 69%
246 after two rounds of polishing; the BUSCO fragmented (decreased length alignment of genes) and BUSCO
247 missing (no significant matches) values decreased correspondingly (Supplemental Table S2 and S3).

248

249 *MinION Assembly Annotation*

250 Both 4-hour, MinION-only assemblies, after two rounds of polishing with Nanopolish, were
251 initially annotated using Geneious and the most closely related, published, annotated genomes for each

252 bacterial species. The annotations were confirmed using the PGAP annotations of the final, corrected
253 assemblies. Since the *Salmonella* Bareilly genome was already completed and closed, we confirmed that
254 the genome annotation of the sequence produced by MinION was structurally identical to the
255 annotation from the Illumina/PacBio sequence already published, for example, but not limited to, the
256 two major serotyping antigens located on the chromosome: the flagellin FliC CDS and the O-antigen
257 polymerase.

258 The presence of major virulence factors in the *E. coli* MinION-only assembly were identified, as
259 well as genes that would cause possible antimicrobial resistance (Fig 2 and 3). The locus of enterocyte
260 effacement (LEE), one of the major virulence factors of enterohemorrhagic *E. coli*^{39,40} that includes the
261 gene intimin for adhesion and the type III secretion system, was annotated between positions 4,603,699
262 and 4,636,299 in this MinION-only assembly (Fig 2). Additionally, the genes expressing the Shiga toxins
263 (Stx), responsible for causing host cell damage^{39,41}, were annotated from position 3,181,004 to
264 3,181,963 for Stx subunit A and from position 3,180,723 to 3,180,992 for Stx2 subunit B (Fig 2). The
265 multidrug resistance gene Mdf(A), which encodes a membrane protein that confers resistance to a
266 multitude of clinically important drugs, including macrolides, lincosamides, and streptogramin B⁴², was
267 also identified at position 1,012,477 to 1,013,709. No other genes or point mutations that would confer
268 antimicrobial resistance were detected. Not only was the full-length chromosome of this *E. coli* O157:H7
269 isolate sequenced using MinION, but also the full-length pO157 (Fig 3). This plasmid also encodes *E. coli*
270 O157-specific virulence factors³⁹, such as hemolysin (*ehx*) identified at position 16,584 to 19,578,
271 catalase-peroxidase (*katP*) at position 76,704 to 78,356, and the type II secretion system (T2SS) at
272 position 64,056 to 85,694.

273

274 *Additional Polishing of the MinION assemblies with MiSeq Data*

275 This paper is primarily focused on optimizing sequencing run times and costs for food safety
276 applications using nanopore-only approaches. However, for submission of final sequences to GenBank,
277 the most accurate assemblies attainable were used. To this end, for both samples, assemblies produced
278 using the full run length were utilized and further error-corrected using Pilon, together with available
279 MiSeq data. Pilon utilizes the low error rate of Illumina reads mapped to the draft assembly to
280 drastically improve the local accuracy of the final sequence. The error rate decreased for both samples
281 after Pilon polishing, with accuracy rates of 99.99% and 100%, and BUSCO completeness rates of 99.7%
282 and 99.99% for *Salmonella* and *E.coli*, respectively. There were also reductions in SNPs per kb to 0.002
283 and 0.001 and indels per kb to 0.008 and 0.002 for *Salmonella* and *E.coli*, respectively. The assembled,
284 polished, and short-read error-corrected data from the full 25-hour run were the final assemblies
285 annotated and submitted to GenBank (Accession numbers CP034177- CP034178 and CP035545-
286 CP035546 with Bioproject PRJNA498670).

287

288 *Phylogenetic inference (SNP tree)*

289 The constructed maximum likelihood SNPs trees are presented in Figure 4. The tree provided
290 with the reference *Salmonella* datasets used for phylogenetic pipeline validation for foodborne
291 pathogen surveillance³⁴ is depicted in Figure 4A. To demonstrate the potential of the MinION-only
292 sequencing for phylogenetic inference, the raw data for strain CFSAN000189 sequenced in this study
293 was replaced with the data from our un-polished, MinION-only assemblies, and the tree was rebuilt
294 from raw data using our pipeline (Figure 4B). For simplicity, only the 30 mins, 240 mins and 1500 mins
295 timepoints were used for the reconstruction. The comparison between the trees built with the
296 reference datasets and the tree utilizing the MinION-only data for the CSAN000189 strain
297 demonstrates complete topological congruence between both trees. The results using all eight time
298 points showed identical topology except for the 15-minute time-point (data not shown).

299

300 **Discussion**

301 In this study, we demonstrate that long-read, nanopore sequencing technology can be used as a
302 single tool to sequence full length bacterial chromosomes and plasmids. Utilizing publicly-available tools
303 in a customized workflow, the MinION-only data produced assembled sequences with as little as 0.1%
304 error rate, which is 0.4% and 3.1% lower than previous reports^{9,19}. The tools used in our customized
305 bioinformatic workflow are publicly-available^{24,25}. The workflow was optimized and tailored to the
306 specific data being analyzed, and the additionally used local code is also made available on Github.

307 Using MinION sequencing alone, two completely closed contigs, one chromosome and one
308 plasmid, for each pathogen were assembled. This capability and the low cost make the MinION highly
309 accessible as both a primary sequencing platform, as well as a secondary platform to complement
310 laboratories' existing sequencing infrastructure. The initial investment required for the MinION is
311 drastically lower than other sequencing technologies (currently, a starter pack is \$1000, which includes
312 the instrument, two flow cells, usually \$500 each, and the respective wash, sequencing, and library
313 loading kit). Additionally, each flow cell can be used for multiple runs, and samples can be multiplexed
314 together per run to reduce the cost^{20,43}. Based on the results of barcoding and simultaneous sequencing
315 of two whole bacterial genomes and plasmids shown here, we estimate that six bacterial samples could
316 be multiplexed together to further decrease cost and sequenced in approximately 16 hours to obtain
317 complete genomic data with high accuracy.

318 The effects of increased sequencing run lengths, different criteria and weights to subsample
319 data for assembly, and increased rounds of polishing were examined for their effect on the final
320 assembly completeness and accuracy. It was observed that the nanopore reads were long enough on
321 average, and that over-aggressive length-based filtering resulted in reduced representation. Such
322 extensive subsampling would result in less complete assembly of small plasmids, which can contain

323 virulence factors of great interest for diagnostic and food safety purposes. It therefore proved critical to
324 evaluate filtering and subsampling criteria to take full advantage of the technology.

325 Our results suggest that at least one round of polishing with Nanopolish is needed to achieve
326 acceptable accuracy, and a second round provides additional improvement if the near-doubling of the
327 analysis time is warranted. The data in Table S1 is provided when only one core is utilized, but due to the
328 availability of high-performance computers, the analysis time for two rounds of polishing can decrease
329 to 6 hours using 124 cores, for example. In MinION-only assemblies, it is known that putative
330 pseudogenes caused by systematic indel errors (often near homopolymeric tracts^{19,44}), leading to
331 reading frame shifts can be an issue, as evident from the “BUSCO fragmented” column in Supplementary
332 tables S2 and S3. Even after polishing, this value was observed to be greater than 20% of expected
333 coding genes, which must be taken into consideration during annotation. However, the polished
334 assemblies, with only 0.1% error, still reveal serotype and important genes responsible for the virulence,
335 metabolism, defense, and pathogenesis of the bacterium.

336 In outbreak situations, a rapid turn-around time is necessary. Therefore, polymerase chain
337 reaction (PCR), real-time PCR assays, and other rapid diagnostic assays are still employed. However,
338 WGS is far more powerful and informative and has become routine in use as it can be coupled with
339 proper bioinformatics analysis to provide complete genome sequences in a couple of days². With the
340 MinION platform and sufficient computational resources (which can be cloud-based and thus widely
341 available), basecalled sequence data can potentially be analyzed in near-real-time as it comes off the
342 machine⁴⁵. Therefore, the MinION can be used for rapid diagnostics as initial sequencing data from pure
343 cultures can be provided in approximately 9 to 10 hours⁴⁶. The complete MinION data can be further
344 analyzed and polished after the entire sequencing run to obtain accurate, whole genomes that provide
345 detailed data on subtyping, virulence genes, antimicrobial resistance genes, and other genetic
346 characteristics. Same-day detection of antimicrobial resistance genes with 99.75% accuracy (with

347 polishing) after enriching for plasmid DNA and MinION sequencing has been recently demonstrated⁴⁷.
348 Here, we also demonstrate the potential of the un-polished, MinION-only results for rapid phylogenetic
349 inference with identical topology as compared to the reference published dataset after only 30 mins of
350 sequencing, with consistent results at all other analyzed time points. The application of the technology
351 for epidemiological tracing during real field outbreak remains to be tested and verified utilizing more
352 samples.

353 In conclusion, this low-cost, rapid, random-priming nanopore sequencing approach, coupled
354 with our customized workflow, provides sufficient data where complete genomes, including plasmids,
355 can be assembled into a single contiguous sequence with 99.89% accuracy. These data provided both
356 gene identification and genomic organization without the need for additional sequencing tools to close
357 gaps that are required by other sequencing methods. We were able to successfully sequence complete
358 bacterial genomes with the lowest error rate reported-to-date using a single sequencing method and
359 to demonstrate the potential of these results for epidemiological inference and outbreak tracing. As
360 the nanopore chemistry and bioinformatics continue to evolve, this method is promising in providing
361 a sufficient amount of accurate data to complement the current sequencing methods by resolving
362 repetitive regions of the genome, which will be instrumental in increasing the number of available
363 complete genome assemblies.

364

365 **Acknowledgments**

366 The authors would like to thank Drs. Ruth Timme and James Pettengill with the FDA Center for Food
367 Safety and Applied Nutrition for the *Salmonella* Bareilly isolate used in this manuscript. This work was
368 supported by U.S. Department of Agriculture, ARS CRIS Project 6612-32000-072.

369 The mention of trade names or commercial products in this publication is solely for the purpose of
370 providing specific information and does not imply recommendation or endorsement by the U.S.
371 Department of Agriculture. The USDA is an equal opportunity provider and employer.

372

373 **Author Contributions:**

374 Conceptualization – T. Taylor, E. DeJesus, G. Tillman, C. Afonso.; Methodology - T. Taylor and E. DeJesus.;
375 Software - J. Volkening and M. Simmons.; Formal Analysis - T. Taylor, J. Volkening, M. Simmons, K.
376 Dimitrov; Resources – G. Tillman, D. Suarez, C. Afonso.; Original Draft Preparation - T. Taylor; Review &
377 Editing - T. Taylor, J. Volkening, E. DeJesus, M. Simmons, K. Dimitrov, G. Tillman, D. Suarez, C. Afonso;
378 Supervision, Project Administration, Funding Acquisition - D. Suarez, C. Afonso.

379

380 **Conflicts of Interest**

381 The authors declare no conflict of interest.

382

383 **Data availability**

384 The final assemblies generated during the current study are available in GenBank (Accession
385 CP034177- CP034178 and CP035545-CP035546). The raw data generated during the current study are
386 available from the corresponding author on reasonable request.

387

388
389

References

- 390 1 Scallan, E., Griffin, P. M., Angulo, F. J., Tauxe, R. V. & Hoekstra, R. M. Foodborne illness acquired
391 in the United States--unspecified agents. *Emerg Infect Dis* **17**, 16-22,
392 doi:10.3201/eid1701.091101p2 (2011).
- 393 2 Sekse, C. *et al.* High Throughput Sequencing for Detection of Foodborne Pathogens. *Front*
394 *Microbiol* **8**, 2029, doi:10.3389/fmicb.2017.02029 (2017).
- 395 3 Forbes, J. D., Knox, N. C., Ronholm, J., Pagotto, F. & Reimer, A. Metagenomics: The Next Culture-
396 Independent Game Changer. *Front Microbiol* **8**, 1069, doi:10.3389/fmicb.2017.01069 (2017).
- 397 4 Struelens, M. J., Palm, D. & Takkinen, J. Enterohaggens, Shiga toxin-producing Escherichia
398 coli O104:H4 outbreak: new microbiological findings boost coordinated investigations by
399 European public health laboratories. *Euro Surveill* **16** (2011).
- 400 5 Dallman, T. J. *et al.* The utility and public health implications of PCR and whole genome
401 sequencing for the detection and investigation of an outbreak of Shiga toxin-producing
402 Escherichia coli serogroup O26:H11. *Epidemiology and Infection* **143**, 1672-1680,
403 doi:10.1017/S0950268814002696 (2015).
- 404 6 van Dijk, E. L., Jaszczyszyn, Y. & Thermes, C. Library preparation methods for next-generation
405 sequencing: tone down the bias. *Exp Cell Res* **322**, 12-20, doi:10.1016/j.yexcr.2014.01.008
406 (2014).
- 407 7 Chain, P. S. *et al.* Genomics. Genome project standards in a new era of sequencing. *Science* **326**,
408 236-237, doi:10.1126/science.1180614 (2009).
- 409 8 Nagarajan, N. & Pop, M. Sequence assembly demystified. *Nat Rev Genet* **14**, 157-167,
410 doi:10.1038/nrg3367 (2013).
- 411 9 Greig, D. R., Dallman, T. J., Hopkins, K. L. & Jenkins, C. MinION nanopore sequencing identifies
412 the position and structure of bacterial antibiotic resistance determinants in a multidrug-
413 resistant strain of enterohaggens Escherichia coli. *Microb Genom*,
414 doi:10.1099/mgen.0.000213 (2018).
- 415 10 Margos, G. *et al.* Lost in plasmids: next generation sequencing and the complex genome of the
416 tick-borne pathogen *Borrelia burgdorferi*. *BMC Genomics* **18**, 422, doi:10.1186/s12864-017-
417 3804-5 (2017).
- 418 11 Orlek, A. *et al.* Plasmid Classification in an Era of Whole-Genome Sequencing: Application in
419 Studies of Antibiotic Resistance Epidemiology. *Front Microbiol* **8**, 182,
420 doi:10.3389/fmicb.2017.00182 (2017).
- 421 12 Gonzalez-Escalona, N., Yao, K. & Hoffmann, M. Closed Genome Sequence of Salmonella enterica
422 Serovar Richmond Strain CFSAN000191, Obtained with Nanopore Sequencing. *Microbiol Resour*
423 *Announc* **7**, doi:10.1128/MRA.01472-18 (2018).
- 424 13 Koren, S. *et al.* Reducing assembly complexity of microbial genomes with single-molecule
425 sequencing. *Genome biology* **14**, R101, doi:10.1186/gb-2013-14-9-r101 (2013).
- 426 14 Utturkar, S. M. *et al.* Evaluation and validation of de novo and hybrid assembly techniques to
427 derive high-quality genome sequences. *Bioinformatics* **30**, 2709-2716,
428 doi:10.1093/bioinformatics/btu391 (2014).
- 429 15 Koren, S. & Phillippy, A. M. One chromosome, one contig: complete microbial genomes from
430 long-read sequencing and assembly. *Curr Opin Microbiol* **23**, 110-120,
431 doi:10.1016/j.mib.2014.11.014 (2015).
- 432 16 Brown, S. D. *et al.* Comparison of single-molecule sequencing and hybrid approaches for
433 finishing the genome of *Clostridium autoethanogenum* and analysis of CRISPR systems in
434 industrial relevant Clostridia. *Biotechnol Biofuels* **7**, 40, doi:10.1186/1754-6834-7-40 (2014).

- 435 17 Chin, C. S. *et al.* Nonhybrid, finished microbial genome assemblies from long-read SMRT
436 sequencing data. *Nat Methods* **10**, 563-569, doi:10.1038/nmeth.2474 (2013).
- 437 18 Feng, Y., Zhang, Y., Ying, C., Wang, D. & Du, C. Nanopore-based fourth-generation DNA
438 sequencing technology. *Genomics Proteomics Bioinformatics* **13**, 4-16,
439 doi:10.1016/j.gpb.2015.01.009 (2015).
- 440 19 Loman, N. J., Quick, J. & Simpson, J. T. A complete bacterial genome assembled de novo using
441 only nanopore sequencing data. *Nat Methods* **12**, 733-735, doi:10.1038/nmeth.3444 (2015).
- 442 20 Butt, S. L. *et al.* Rapid virulence prediction and identification of Newcastle disease virus
443 genotypes using third-generation sequencing. *Virology journal*, doi:10.1101/349159 (2018).
- 444 21 Phan, H. T. T. *et al.* Illumina short-read and MinION long-read WGS to characterize the molecular
445 epidemiology of an NDM-1 *Serratia marcescens* outbreak in Romania. *J Antimicrob Chemother*,
446 doi:10.1093/jac/dkx456 (2017).
- 447 22 Kislyuk, A. O. *et al.* A computational genomics pipeline for prokaryotic sequencing projects.
448 *Bioinformatics* **26**, 1819-1826, doi:10.1093/bioinformatics/btq284 (2010).
- 449 23 Bankevich, A. *et al.* SPAdes: a new genome assembly algorithm and its applications to single-cell
450 sequencing. *J Comput Biol* **19**, 455-477, doi:10.1089/cmb.2012.0021 (2012).
- 451 24 Wick, R. *Fitlong: quality filtering tool for long reads* <<https://github.com/rrwick/Fitlong>> (2007).
- 452 25 Wick, R. R., Judd, L. M., Gorrie, C. L. & Holt, K. E. Unicycler: Resolving bacterial genome
453 assemblies from short and long sequencing reads. *PLoS Comput Biol* **13**, e1005595,
454 doi:10.1371/journal.pcbi.1005595 (2017).
- 455 26 Sommer, D. D., Delcher, A. L., Salzberg, S. L. & Pop, M. Minimus: a fast, lightweight genome
456 assembler. *BMC Bioinformatics* **8**, 64, doi:10.1186/1471-2105-8-64 (2007).
- 457 27 Simpson, J. *Nanopolish: Signal-level algorithms for MinION data*
458 <<https://github.com/jts/nanopolish>> (2018).
- 459 28 Simao, F. A., Waterhouse, R. M., Ioannidis, P., Kriventseva, E. V. & Zdobnov, E. M. BUSCO:
460 assessing genome assembly and annotation completeness with single-copy orthologs.
461 *Bioinformatics* **31**, 3210-3212, doi:10.1093/bioinformatics/btv351 (2015).
- 462 29 Hoffmann, M. *et al.* Tracing Origins of the Salmonella Bareilly Strain Causing a Food-borne
463 Outbreak in the United States. *J Infect Dis* **213**, 502-508, doi:10.1093/infdis/jiv297 (2016).
- 464 30 Kurtz, S. *et al.* Versatile and open software for comparing large genomes. *Genome biology* **5**,
465 R12, doi:10.1186/gb-2004-5-2-r12 (2004).
- 466 31 Wilm, A. *et al.* LoFreq: a sequence-quality aware, ultra-sensitive variant caller for uncovering
467 cell-population heterogeneity from high-throughput sequencing datasets. *Nucleic acids research*
468 **40**, 11189-11201, doi:10.1093/nar/gks918 (2012).
- 469 32 Walker, B. J. *et al.* Pilon: an integrated tool for comprehensive microbial variant detection and
470 genome assembly improvement. *PLoS one* **9**, e112963, doi:10.1371/journal.pone.0112963
471 (2014).
- 472 33 Zankari, E. *et al.* Identification of acquired antimicrobial resistance genes. *J Antimicrob*
473 *Chemother* **67**, 2640-2644, doi:10.1093/jac/dks261 (2012).
- 474 34 Timme, R. E. *et al.* Benchmark datasets for phylogenomic pipeline validation, applications for
475 foodborne pathogen surveillance. *PeerJ* **5**, e3893, doi:10.7717/peerj.3893 (2017).
- 476 35 Huang, W., Li, L., Myers, J. R. & Marth, G. T. ART: a next-generation sequencing read simulator.
477 *Bioinformatics* **28**, 593-594, doi:10.1093/bioinformatics/btr708 (2012).
- 478 36 Li, H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. *arXiv*
479 *preprint arXiv:13033997* (2013).
- 480 37 Guindon, S. *et al.* New algorithms and methods to estimate maximum-likelihood phylogenies:
481 assessing the performance of PhyML 3.0. *Syst Biol* **59**, 307-321, doi:10.1093/sysbio/syq010
482 (2010).

- 483 38 Di Tommaso, P. *et al.* Nextflow enables reproducible computational workflows. *Nat Biotechnol*
484 **35**, 316-319, doi:10.1038/nbt.3820 (2017).
- 485 39 Lim, J. Y., Yoon, J. & Hovde, C. J. A brief overview of Escherichia coli O157:H7 and its plasmid
486 O157. *J Microbiol Biotechnol* **20**, 5-14 (2010).
- 487 40 Franzin, F. M. & Sircili, M. P. Locus of enterocyte effacement: a pathogenicity island involved in
488 the virulence of enteropathogenic and enterohemorrhagic Escherichia coli subjected to a
489 complex network of gene regulation. *BioMed research international* **2015**, 534738,
490 doi:10.1155/2015/534738 (2015).
- 491 41 Baranzoni, G. M. *et al.* Characterization of Shiga Toxin Subtypes and Virulence Genes in Porcine
492 Shiga Toxin-Producing Escherichia coli. *Front Microbiol* **7**, 574, doi:10.3389/fmicb.2016.00574
493 (2016).
- 494 42 Edgar, R. & Bibi, E. MdfA, an Escherichia coli multidrug resistance protein with an extraordinarily
495 broad spectrum of drug recognition. *J Bacteriol* **179**, 2274-2280 (1997).
- 496 43 Ring, N. *et al.* Resolving the complex Bordetella pertussis genome using barcoded nanopore
497 sequencing. *Microb Genom* **4**, doi:10.1099/mgen.0.000234 (2018).
- 498 44 Tyson, J. R. *et al.* MinION-based long-read sequencing and assembly extends the Caenorhabditis
499 elegans reference genome. *Genome Res* **28**, 266-274, doi:10.1101/gr.221184.117 (2018).
- 500 45 Ondov, B. D. *et al.* Mash: fast genome and metagenome distance estimation using MinHash.
501 *Genome biology* **17**, 132, doi:10.1186/s13059-016-0997-x (2016).
- 502 46 Quick, J. *et al.* Rapid draft sequencing and real-time nanopore sequencing in a hospital outbreak
503 of Salmonella. *Genome biology* **16**, 114, doi:10.1186/s13059-015-0677-2 (2015).
- 504 47 Lemons, B., Khaing, H., Ward, A. & Thakur, P. A rapid method for the sequential separation of
505 polonium, plutonium, americium and uranium in drinking water. *Appl Radiat Isot* **136**, 10-17,
506 doi:10.1016/j.apradiso.2018.02.008 (2018).

507

508 **Figure Legends**

509 **Figure 1: Polishing Results of the MinION-only Assemblies Using Multiple Rounds of Nanopolish.**

510 Due to the errors remaining in the MinION-only assemblies, a signal-level consensus software,
511 Nanopolish, was used to increase the assembly accuracy. The overall accuracy, the Benchmarking
512 Universal Single-Copy Orthologs (BUSCO) completeness, BUSCO Fragmented, BUSCO Missing, number of
513 indels per kb, and number of SNPs per kb are shown after 0, 1, 2, 3 and 4 rounds of Nanopolish. After
514 two rounds of polishing, the overall accuracy and the number of Indels and SNPs per kb did not
515 considerably change.

516

517 **Figure 2: MinION Assembly of the *E. coli* chromosome.**

518 The *E. coli* O157:H7 chromosome was sequenced and assembled into a final consensus of 5,482,542
519 nucleotides. The annotation of the genome provided the location of 5,748 coding sequences (CDS), 106
520 tRNAs, 29 rRNAs, 6 regulatory regions, and 1 repeat regions. For imaging purposes, only the 6 regulatory
521 regions (green), the one repeat region (brown) and the CDS of two virulence factors (yellow) are shown
522 magnified. The LEE (locus of enterocyte effacement) is highlighted at position 4,603,699 to 4,636,299,
523 and the Shiga Toxin subunits are shown at position 3,181,004 to 3,180,992.

524

525 **Figure 3: MinION Assembly of the *E. coli* pO157.**

526 The *E. coli* pO157 plasmid was sequenced and assembled into a final consensus of 94,503 nucleotides.
527 The annotation shows all 124 coding sequences (CDS) in yellow. The CDS of three well-known virulence
528 factors are highlighted: hemolysin (*ehx*) at position 16,584 to 19,578, catalase-peroxidase (*katP*) at
529 position 76,704 to 78,356, and the type II secretion system (T2SS) at position 64,056 to 85,694.

530

531 **Figure 4: Maximum likelihood phylogenetic SNPs trees of Salmonella reference datasets.**

532 A) Maximum likelihood SNPs tree provided with the reference datasets from Timme et al., 2017 that
533 was used for phylogenetic pipeline validation for foodborne pathogen surveillance³⁴. B) PhyML v3.0 was
534 used to generate a maximum likelihood SNP tree for comparison purposes with twenty-two *Salmonella*
535 reference datasets³⁴ and by replacing the CFSAN000189 data with SNPs from the 30 mins, 240 mins,
536 and 1500 mins un-polished, MinION-only assemblies obtained in this study, which are enclosed in a
537 rectangle and are shown in red. The HYK85 model was used with 500 bootstraps, and branches with less
538 than 60% support were collapsed. There were a total of 64 positions in the final dataset. The tree is
539 drawn to scale, with branch lengths measured in the number of substitutions per site.

540 **Tables:**

541 **Table 1: Comparison of the Final Raw Data from MinION and Illumina**

Sequence Method	Average Read Length	Total Bases	Min Read Length	Max Read Length	Average Read Quality	Read Number	Mean Depth
MiSeq (<i>Salmonella</i>)	149.51	288,633,579	35	151	36.66	1,930,511	57.72
MinION (<i>Salmonella</i>)	8638.36	2,879,148,408	113	120,119	19.36	333,298	599.06
MiSeq (<i>E.coli</i>)	242.61	556,035,081	35	251	34.96	2,291,825	111.2
MinION (<i>E.coli</i>)	8979.55	3,860,389,678	85	112,643	19.38	429,909	692.19

542 a. MiSeq Quality Standards = $Q \geq 30$

543 b. MinION Quality Standards = $Q \geq 10$

544

545 **Table 2: Assembly Data for MinION sequencing of *Salmonella***

546

Duration (min)	Reads	Subsampled Reads	Assembly Size	Circular Contigs ^a	Linear Contigs	Longest Contig	Longest Circular Contig	NG50 ^b	Average identity in %	Reference Coverage
15	7229	7229	1135723	0	19	163786	0	0	99.13	24.36
30	14888	14888	4577215	0	18	841969	0	471499	99.55	95.38
60	29132	29132	4722179	1	0	4722179	4722179	4722179	99.79	98.4
120	51226	51226	4805334	2	0	4723663	4723663	4723663	99.84	100
240	84156	28492	4806150	2	0	4724389	4724389	4724389	99.87	100
480	132137	20193	4806518	2	0	4724724	4724724	4724724	99.87	100
960	248910	16221	4806892	2	0	4725103	4725103	4725103	99.89	100
1500	333298	15249	4806995	2	0	4725191	4725191	4725191	99.89	100

547 a. Two circular contigs indicates both the chromosome and the plasmid

548 b. NG50 - 50% of the entire assembly is contained in contigs or scaffolds equal to or larger than this value

549

550 **Table 3: Assembly Data for MinION sequencing of *E. coli***

551

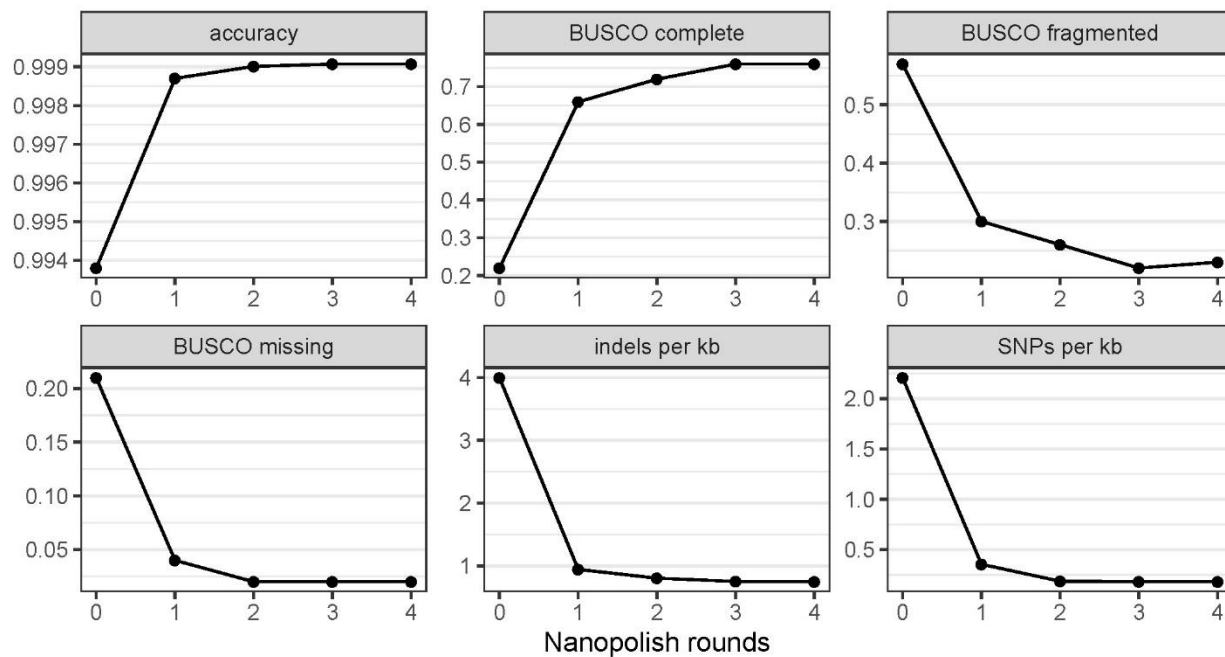
Duration (min)	Reads	Subsampled Reads	Assembly Size	Circular Contigs ^a	Linear Contigs	Longest Contig	Longest Circular Contig	NG50 ^b	Average identity in %
15	8731	8731	1352560	0	19	154626	0	0	99.18
30	18053	18053	5141583	0	14	1565772	0	518218	99.63
60	35335	35335	5481126	1	0	5481126	5481126	5481126	99.82
120	62415	60362	5570410	1	1	5481662	5481662	5481662	99.87
240	103681	19589	5577045	2	0	5482542	5482542	5482542	99.89
480	164641	15265	5577346	2	0	5482831	5482831	5482831	99.90
960	317698	12941	5577818	2	0	5483284	5483284	5483284	99.91
1500	429909	12403	5577934	2	0	5483397	5483397	5483397	99.91

552 a. Two circular contigs indicates both the chromosome and the plasmid

553 b. NG50 - 50% of the entire assembly is contained in contigs or scaffolds equal to or larger than this value

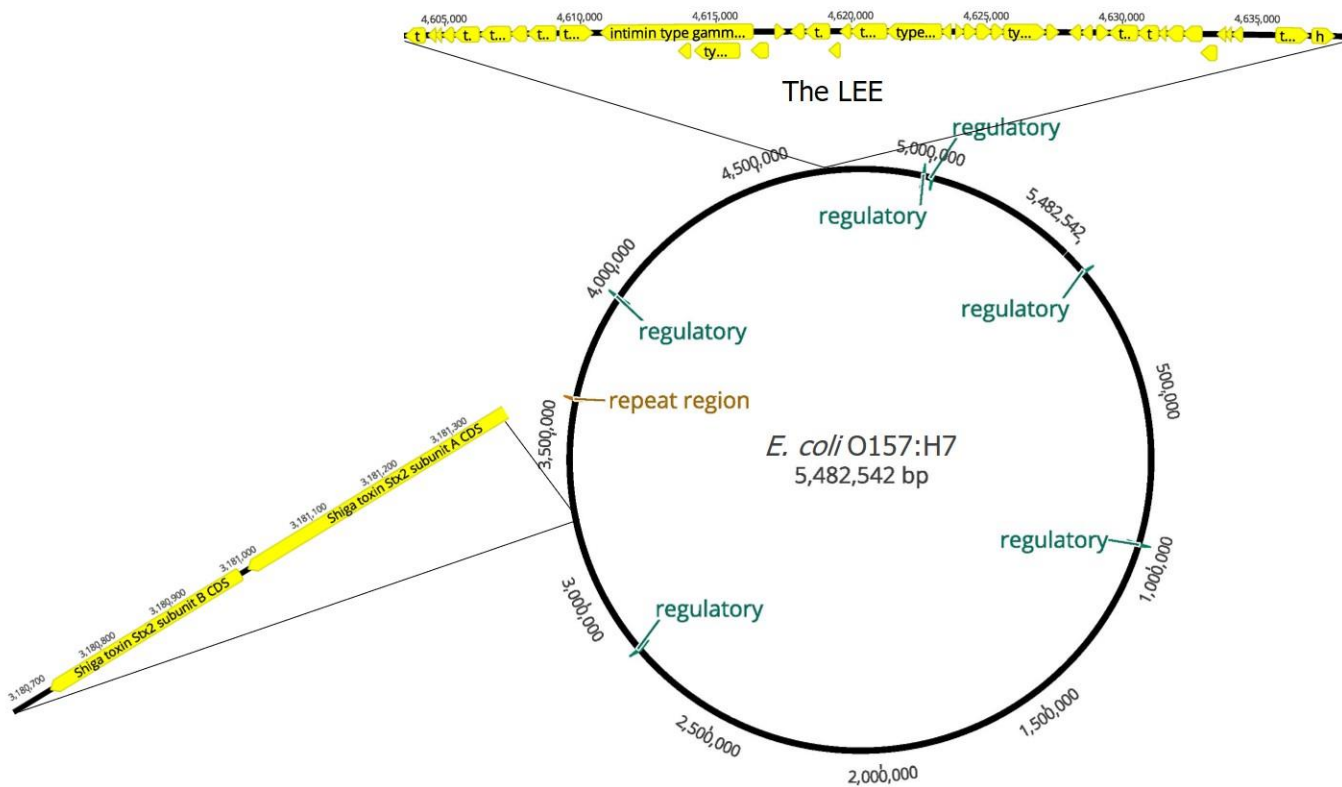
554 **Figures:**

555 **Figure 1: Polishing Results of the MinION-only Assemblies Using Multiple Rounds of Nanopolish**



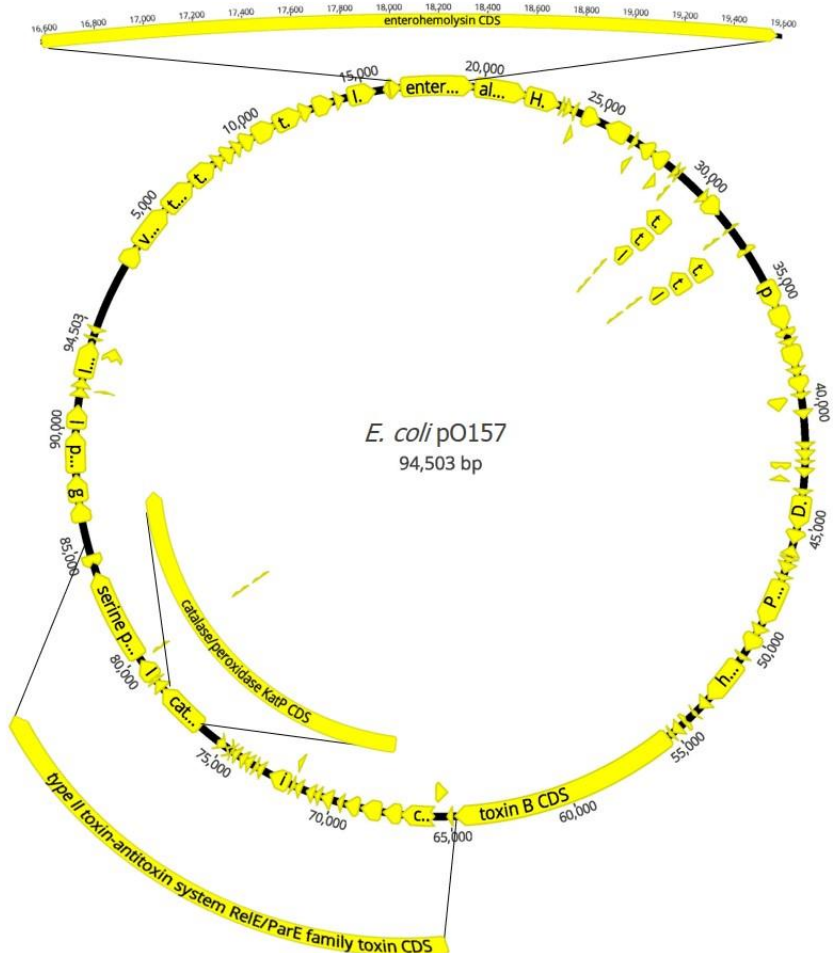
556

557 **Figure 2: MinION Assembly of the *E. coli* Chromosome**



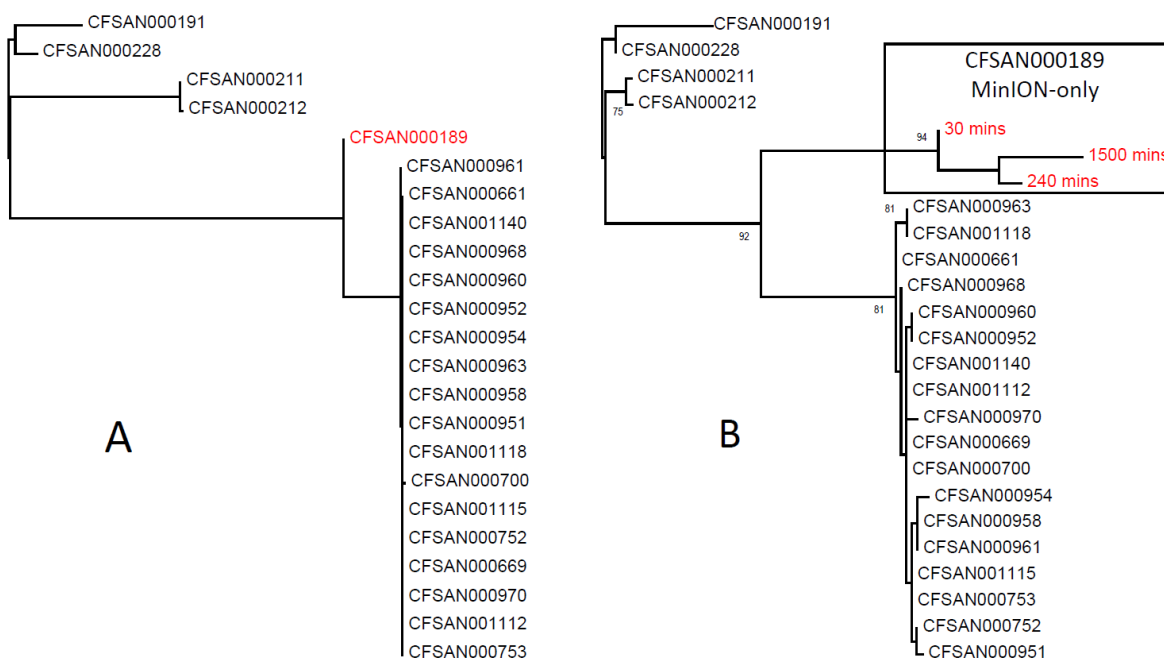
558

559 **Figure 3: MinION Assembly of the *E. coli* pO157**



560

561 **Figure 4: Maximum likelihood phylogenetic SNPs trees of Salmonella reference datasets.**



562

0.1

0.1