



HAL
open science

A Sample-to-Report Solution for Taxonomic Identification of Cultured Bacteria in the Clinical Setting Based on Nanopore Sequencing

Stefan Moritz Neuenschwander, Miguel Angel Terrazos Miani, Heiko Amlang, Carmen Perroulaz, Pascal Bittel, Carlo Casanova, Sara Droz, Jean-Pierre Flandrois, Stephen Leib, Franziska Suter-Riniker, et al.

► To cite this version:

Stefan Moritz Neuenschwander, Miguel Angel Terrazos Miani, Heiko Amlang, Carmen Perroulaz, Pascal Bittel, et al.. A Sample-to-Report Solution for Taxonomic Identification of Cultured Bacteria in the Clinical Setting Based on Nanopore Sequencing. *Journal of Clinical Microbiology*, 2020, 58 (6), pp.1128. 10.1128/JCM.00060-20 . hal-03105483

HAL Id: hal-03105483

<https://hal.science/hal-03105483v1>

Submitted on 23 Jul 2024

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.



Distributed under a Creative Commons Attribution 4.0 International License

1 A SAMPLE-TO-REPORT SOLUTION FOR TAXONOMIC IDENTIFICATION OF CULTURED
2 BACTERIA IN THE CLINICAL SETTING BASED ON NANOPORE SEQUENCING

3

4 Stefan Moritz Neuenschwander¹, Miguel Angel Terrazos Miani¹, Heiko Amlang¹, Carmen Perroulaz¹,
5 Pascal Bittel¹, Carlo Casanova¹, Sara Droz¹, Jean-Pierre Flandrois², Stephen L. Leib¹, Franziska Suter-
6 Riniker¹, Alban Ramette^{1,*}

7

8 ¹University of Bern, Institute for Infectious Diseases, Bern, Switzerland

9 ² University of Lyon, CNRS, UMR 5558, Laboratoire de Biométrie et Biologie Evolutive, Villeurbanne,
10 France

11 * Corresponding author: Alban Ramette, University of Bern, Institute for Infectious Diseases,

12 Friedbühlstrasse 51, CH-3001 Bern, Switzerland. alban.ramette@ifik.unibe.ch

13

14

15

16 **Abstract**

17 Amplicon sequencing of 16S rRNA gene is commonly used for the identification of bacterial isolates in
18 diagnostic laboratories, and mostly relies on the Sanger sequencing method. The latter, however, suffers
19 from a number of limitations with the most significant being the inability to resolve mixed amplicons
20 when closely related species are co-amplified from a mixed culture. This often leads to either increased
21 turnaround time or absence of usable sequence data. Short-read NGS technologies could solve the mixed
22 amplicon issue, but would lack both cost efficiency at low throughput and fast turnaround times.
23 Nanopore sequencing developed by Oxford Nanopore Technologies (ONT) could solve those issues by
24 enabling flexible number of samples per run and adjustable sequencing time. Here we report on the
25 development of a standardized laboratory workflow combined with a fully automated analysis pipeline
26 *LORCAN* (Long Read Consensus ANalysis), which together provide a sample-to-report solution for
27 amplicon sequencing and taxonomic identification of the resulting consensus sequences. Validation of
28 the approach was conducted on a panel of reference strains and on clinical samples consisting of single
29 or mixed rRNA amplicons associated with various bacterial genera by direct comparison to the
30 corresponding Sanger sequences. Additionally, simulated read and amplicon mixtures were used to
31 assess *LORCAN*'s behaviour when dealing with samples with known cross-contamination level. We
32 demonstrate that by combining ONT amplicon sequencing results with *LORCAN*, the accuracy of Sanger
33 sequencing can be closely matched (>99.6% sequence identity) and that mixed samples can be resolved
34 at the single base resolution level. The presented approach has the potential to significantly improve the
35 flexibility, reliability and availability of amplicon sequencing in diagnostic settings.

36

37 Introduction

38 The sequencing of the 16S rRNA gene is essential to describe the diversity of the human
39 microbiome (1, 2). Yet, the frequency of the use of 16S sequencing for species identification from
40 cultured isolates in clinical laboratories is decreasing (3), despite the usefulness of 16S rRNA gene
41 sequencing to provide taxonomic classification for isolates that do not match recognized biochemical
42 profiles, that only produce low identification score according to commercial systems, or that are not
43 typically associated with human pathogens (3, 4). In the clinical microbiology laboratory, amplicon
44 sequencing of 16S rRNA gene mostly relies on the Sanger sequencing method, which is based on chain
45 termination via fluorescently labelled deoxyribonucleotides (dNTPs), capillary electrophoresis and
46 fluorescence measurement (5). Although the Sanger method is still the gold-standard for validating the
47 accuracy of sequences from specific genes, when compared to more recent technologies, the method has
48 a number of significant shortcomings: During a sequencing run, each capillary is limited to the
49 production of one single sequence with a maximal length of about 1000 bp (6), resulting in low
50 throughput, and high sequencing costs. Furthermore, the sequencing machines are comparably large
51 and require maintenance, limiting their suitability for all types of laboratory settings. The most
52 important limitation of the Sanger method is, however, its limited ability to produce complete sequence
53 information when diverse amplicons are present (7). Under routine diagnostic conditions, this
54 frequently leads to either increased turnaround time or lack of results (8), leading to potential delays or
55 inaccuracies in patient treatment and management.

56 Next generation sequencing technologies (i.e. second-generation sequencing technologies, such as
57 provided by Illumina) might overcome most of these limitations, but are not designed for the analysis of
58 small numbers of pure amplicons. Even the smallest and fastest available 500 and 600 cycles Illumina
59 kits show runtimes of >24 hours, with associated running costs of several hundred euro regardless of the

60 numbers of samples processed, limiting their usefulness for the fast and flexible identification of small
61 batches of samples (company information). The third-generation single-molecule sequencing technology
62 provided by Oxford Nanopore Technologies (ONT) might offer the necessary flexibility in throughput
63 and is capable of producing reads with lengths of several hundred to several hundred-thousand bases at
64 competitive costs (9). Furthermore, ONT sequencers are small devices, virtually maintenance free and
65 affordable for small laboratories. Despite the constant improvement over the last years in read accuracy
66 (with read accuracy of about 96% currently), the remaining sequencing errors in single nanopore reads
67 do not yet allow for an analysis at the read level. *De novo* assembly or consensus generation from
68 individual ONT reads are therefore commonly used to generate sequences that are virtually free from
69 substitution errors (10). Additionally, "polishing" tools can be applied to remove remaining non-random
70 errors such as indels in homopolymer regions from the generated consensus sequences (10-13).
71 Resulting sequences can then be directly substituted to Sanger sequences in existing classification
72 pipelines or, due to the added flexibility in read length, may provide far higher resolution if the analyses
73 are based on full-length marker genes or entire operons (14). One obstacle for a broad adoption of
74 nanopore sequencing in routine diagnostic laboratories is the added bioinformatic complexity as
75 compared to established Sanger sequencing workflows. Furthermore, available workflows are often
76 limited to the analysis of pure amplicons (10-13), include complex modifications of the ONT laboratory
77 workflows (15, 16), or lack published validation by using samples other than mock communities (17, 18).

78 Here, we developed a complete workflow based on standard ONT protocols and a fully automated
79 analysis pipeline *LORCAN* capable of producing high-quality consensus sequences and thorough
80 taxonomic analysis from pure and low-complexity cultures. The foreseen end-users of the workflow are
81 clinical bacteriology laboratories. As such, tuneable workflow parameters were evaluated with
82 amplicons generated from reference strains of pathogenic genera (*Bacteroides*, *Eggerthella*, *Enterococcus*,
83 *Klebsiella*, *Mycobacterium*, *Campylobacter*, *Pseudomonas*) and validated on bacterial cultures obtained from

84 patient material over several months. Furthermore, we explored the robustness of *LORCAN*'s consensus
85 generation and species identification by analysing artificial mixtures of reads at different levels of
86 genetic distances.

87

88 **Methods**

89 **Samples, DNA extraction, PCR amplification**

90 Bacterial isolates all originated from the Institute for Infectious Diseases (IFIK, Bern) Biobank. The IFIK
91 provides the entire spectrum of medical microbiological diagnostic services to the largest Swiss hospital
92 group (Inselgruppe) and other regional hospitals. The diagnostic division of IFIK (clinical microbiology)
93 is ISO/IEC 17025 accredited to perform routine bacterial diagnostics from clinical samples. ATCC strains
94 were obtained from LGC Standards (Wesel, Germany) and were grown on solid media as recommended
95 by the manufacturer.

96 Overnight-grown bacterial cultures were harvested from agar plates and dissolved in 300 µl of Tris-
97 EDTA (pH 8.0). DNA was extracted with a NucliSense Easymag (bioMérieux, Switzerland) robot
98 according to the manufacturer's protocol. 16S rRNA gene PCR was performed with the primer sets
99 16S_f: 5'-AGAGTTTGATCMTGGCTCAG-3' and 16S_r: 5'-TACCGCGGCWGCTGGCACRDA-3' (general
100 bacteria) and mbak_f: 5'-GAGTTTGATCCTGGCTCAGGA-3' and mbak_r: 5'-
101 TGCACACAGGCCACAAGGGA-3' (*Mycobacteria*) supplemented with the universal tails 5'-
102 TTTCTGTTGGTGCTGATATTGC-3' (ONT forward primer), 5'-ACTTGCCTGTCGCTCTATCTTC-3'
103 (ONT reverse primer), 5'-TGTAACGACGGCC AG-3' (M13f, Sanger forward primer) or 5'-
104 CAGGAAACAGCTATGAC-3' (M13r, Sanger reverse primer). PCR reactions (25 µl) for general bacteria
105 and *Mycobacteria* were assembled, respectively, with 1 and 2.5 ng DNA template, 10 µl of a 1.25 and 2.5

106 μ M primer working solution, both with 12.5 μ l Q5 Master-Mix. Amplification was performed in a
107 GeneAmp 9700 Thermocycler (Thermo Fisher Scientific Inc., MA, USA) with the following program:
108 98°C for 1 min; 30 cycles of: 98°C for 10 s, 63°C for 15 s, 72°C for 30 sec; 72°C for 2 min. PCR products
109 were purified with CleanNGS beads (CleanNA, Waddinxveen, NL) according to the manufacturer's
110 instructions with the following modifications: After the washing step an additional 3 sec centrifugation
111 step was introduced and the purified DNA was eluted in 80 μ l of Tris-HCl (0.01M, pH 8.0). Fragment
112 size of the amplicons was analysed using the TapeStation D1000 assay (Agilent, Santa Clara, CA USA),
113 concentrations were measured with the Qubit dsDNA BR assay (Thermo Fisher Scientific), and the
114 purity of the DNA was analysed with a Nanodrop spectrophotometer (Thermo Fisher Scientific).
115 Samples with DNA concentrations <1.05 nM were excluded from the analysis.

116 **Library preparation**

117 A typical library consisted of the pooling of PCR amplicons from 2 to 15 clinical samples and 1
118 positive control (*Mycobacteria intracellulare*, amplified with general bacterial primers). Library
119 preparation was performed with the kits EXP-PBC096, SQL-LSK109 (Oxford Nanopore Technologies,
120 OX, UK), using the supplementary reagents NEBNext End repair/ dA-tailing Module (E7546, New
121 England Biolabs, ON, CA), NEB Blunt/TA Ligase Master Mix (M0367, New England Biolabs), Taq 2X
122 Master Mix (NEB M0270, New England Biolabs), CleanNGS beads (CleanNA). All modifications made
123 to the manufacturer's protocol (PCR barcoding (96) genomic DNA,
124 PBAC96_9069_v109_revK_14Aug2019) are described in the following section (see also **Figure 1A**), for
125 a detailed protocol see **Supplementary Text S1**): AMPure beads were substituted with CleanNGS beads
126 and the Hula-Mixer (Thermo Fisher Scientific) parameters "Orbital: 40 rpm, 07 s; Reciprocal: 89 deg, 2 s;
127 Vibro: 5 deg, 2 s; Vertical position" were used. Barcoding-PCR reactions (12 cycles) were set up with 25.2
128 nmol of template per reaction. Raw barcoded PCR products were quantified with the Qubit dsDNA BR

129 assay and pooled at equal molar proportions. Products containing less than 0.57 pmol DNA were
130 excluded from the analysis, If the total amount of DNA in a pooled library was below 9.23 pmol, "place-
131 holder" (filling) barcoded samples were added to the pooled library to avoid flow cell underloading (see
132 example of calculations and adjustments in **Supplementary Text S1**). Place-holder barcoded samples
133 were produced in advance from the same template as the positive controls, with 15 instead of 12
134 barcoding PCR cycles. Resulting PCR products were quantified with Qubit and stored at -20°C. The
135 pooled library was purified (CleanNGS beads, 50 µl elution volume), and quantified with the Qubit
136 dsDNA BR assay. The purified library pools were diluted to 140 nM before proceeding to the "End
137 Preparation" step of the protocol.

138 **Sequencing**

139 ONT-sequencing was performed on a GridION X5 instrument (Oxford Nanopore Technologies) with
140 real-time base calling enabled (*ont-guppy-for-gridion* v.1.4.3-1 and v.3.0.3-1, fast base calling mode).
141 Sequencing runs were terminated after production of 1 million reads or when sequencing rates dropped
142 below 20 reads per second. Purified PCR products were submitted to Sanger sequencing at Microsynth
143 (Balgach, Switzerland).

144 **Bioinformatic analyses**

145 *LORCAN pipeline description.* *LORCAN* was developed to facilitate reproducible ONT sequencing
146 based marker gene analysis in diagnostics facilities. The pipeline written in *Perl* 5, *R* and *BASH*, runs on
147 Linux servers or workstations. The code is publicly available (19) and is based on publicly available,
148 third-party dependencies (**Table S1**). Major steps of the workflow are described in the following section
149 (numbers correspond to the steps in **Figure 1B**): Step 1) Basecalled reads are demultiplexed and adapters
150 trimmed (*Porechop* (20), parameters: --format fasta, --discard_unassigned, --require_two_barcodes). Step

151 2) Reads are filtered by length, keeping only those with lengths of -20 to +100 bases (lower boundary
152 adjustable) around the modal sequence length (custom *Perl* and *R* scripts; **Figure 1B**). Step 3) Reads are
153 mapped to a non-redundant reference database (*minimap2* (21); see database preparation below). Step 4)
154 Reads are extracted, binned by taxonomic level (here species) and remapped to the reference sequence
155 that obtained the highest number of mapped reads among all sequences of the corresponding species
156 (*minimap2*, *SAMtools* (22), *SeqKit* (23)). Step 5) Consensus sequences are derived using a 50% majority
157 rule consensus. Step 6) The 10 closest reference sequences are selected by sequence similarity to the
158 consensus sequence (*BLASTN*, *BLAST+*, (24)). Step 7) Phylogenetic trees for each consensus sequence
159 with its 10 closest references are created (*MAFFT* (25) with parameters *-maxiterate 1000 -localpair; Gblocks*
160 (26) with parameters *-t=d, IQ-TREE* (27) with parameters *-m GTR+I+G -bb 1000 -czb*). Parameters of all
161 software are also provided in the *LORCAN* GitHub repository.

162 **Database preparation.** Reference databases used by *LORCAN* are non-redundant and assay specific.
163 Detailed instructions for database creation are provided online at:
164 <https://github.com/aramette/LORCAN/>. In short, the reference database (in this study: *leBIBI* SSU-rDNA-
165 *mk37_stringent*, <https://umr5558-bibiserv.univ-lyon1.fr/BIBIDOCNEW/db-BIBI.html>; (28)) was trimmed
166 to the region of interest (amplified region minus primers) and de-replicated (*Mothur* (29)), and sequence
167 names were simplified (custom *Perl* scripts). The names of identical sequences are saved to a file during
168 the dereplication step. The resulting non-redundant database is then used to generate a custom *BLAST*
169 database which is used in *LORCAN* pipeline.

170 **Sanger sequence analyses.** Forward and reverse sequences were assembled into consensus sequences
171 using *SeqMan Pro* (DNASar, Madison, WI, USA), primers were trimmed manually, and ambiguous
172 bases were resolved based on visual inspection of the chromatograms. Consensus sequences were
173 taxonomically classified using the online tool *leBIBI QBPP* (28, 30).

174 *SNV discrimination and performance with mixed samples.* Amplicons produced from pure samples
175 were quantified (Qubit dsDNA BR assay). Mixtures of pure amplicons were produced at defined ratios
176 before library preparation to produce libraries of heterogeneous ("mixed") samples. Artificial read
177 mixtures were also produced *in silico* by mixing reads originating from pure amplicon samples. Those
178 reads were obtained from the *LORCAN* output directories (output file 1_fasta/BC*.mode_closest.fasta,
179 produced by step 2; **Figure 1B**) and sampled using *Seqtk subseq* (v.1.3-r106, <https://github.com/lh3/seqtk>)
180 to produce different proportions of original, pure amplicons. Reads from mixed amplicon samples were
181 fed back into *LORCAN* and detected species compositions were extracted from the resulting *LORCAN*
182 reports. Sequence identities between the paired *Mycobacterium* species were determined based on
183 pairwise alignment of the amplified region using *Multalin* (version 5.4.1,
184 <http://multalin.toulouse.inra.fr/multalin/>; (31)).

185 *Influence of database completeness on consensus accuracy.* Amplicons from a set of seven ATCC
186 reference strains were ONT sequenced and analysed with *LORCAN* using the full non-redundant leBIBI
187 16S rRNA database, restricted to the region amplified by the general bacterial primer set. The resulting
188 top consensus sequences were extracted, combined with the above-mentioned database. The resulting
189 sequence dataset was aligned (*MAFFT* v7.313, FFT-NS-1, progressive method) and pairwise distances
190 were calculated (*Mothur* v. 1.40.5, *dist.seqs*, calc=eachgap, countends=F, cutoff=0.20). For each consensus
191 sequence, 10 subsets of sequences with minimal distances below thresholds ranging from 0 to 0.1 were
192 extracted (*Seqtk subseq*), and minimal distances between each dataset and the corresponding consensus
193 sequence were analysed. The seven read sets (ATCC strains) were re-analysed with *LORCAN* and the
194 corresponding subsetted databases to produce consensus sequences. Top consensus sequences from
195 each sample-database combination were extracted, combined with the consensus sequences generated
196 with the full database, and aligned (*MAFFT* v7.313, L-INS-I, iterative refinement method (<16) with local

197 pairwise alignment information). Pairwise distances were analysed as described above and distances
198 between the consensus sequences generated from the full and the subsetted databases were extracted .

199 **Data availability**

200 All reads and consensus sequences corresponding to the data presented in Table 1 and the *LORCAN*-
201 derived consensus sequences used as references in Figure 3 were deposited to the European Nucleotide
202 Archive, under the project reference PRJEB34167, or made available as supplementary multi-FASTA
203 files.

204

205 **Results**

206 We present a standardized laboratory workflow, accompanied by a fully automated analysis pipeline,
207 which together provide a sample-to-report solution for taxonomic identification of bacterial cultures
208 based on amplicon sequencing of their 16S rRNA genes (Figure 1). The laboratory workflow, which was
209 tested and adjusted for parallel processing of up to 16 samples done manually by a single person
210 (theoretically scalable up to 96 samples using automation), includes stringent quality control steps to
211 guarantee consistent results. The whole procedure has been running under ISO/IEC 17025 accreditation
212 standards since January 2019 in our microbial diagnostic department. The analysis pipeline is based on
213 publicly available software components and runs on Linux servers or workstations. It automates quality
214 control, demultiplexing, consensus sequence generation, taxonomic analysis based on the highly curated
215 leBIBI 16S database, as well as report generation (text, PDF; see example report as Supplementary
216 Information). Turnaround time from raw amplicons to PDF reporting is about 8 hours (consisting of 6
217 hours wet lab, 1 hour sequencing, and 1 hour bioinformatic analysis). Validation of the sequencing
218 results was conducted by direct comparison to Sanger sequencing with real clinical samples consisting
219 of pure or mixed rRNA amplicons belonging to several bacterial genera (*Bacteroides*, *Eggerthella*,

10/25

220 *Enterococcus*, *Klebsiella*, *Mycobacterium*, *Campylobacter*, *Pseudomonas*) of expected amplicon sizes of 500 bp
221 (longer amplicons of ca. 900 bp were also successfully analysed with the proposed pipeline; data not
222 shown). Additionally, we created artificial read mixtures from closely related bacterial species to assess
223 the workflow's performance and robustness when confronted with contaminated samples. We
224 demonstrated that by combining ONT sequencing and *LORCAN*, the accuracy of Sanger sequencing can
225 be closely matched (>99.6% sequence identity on average) and that mixed samples can be resolved at the
226 single base resolution level.

227 **Validation of SNV discrimination and analysis of mixed samples.** To test the ability of *LORCAN* to
228 resolve mixed samples, artificial mixtures were created by mixing either amplicons (**Figure 2A**), or reads
229 produced from pure samples (**Figures 2B, 2C, S1 and S2**). The taxonomic identity of all involved strains
230 was successfully recovered by *LORCAN*. The slightly lower amplicon length of *Pseudomonas aeruginosa*
231 compared to *Staphylococcus aureus* and *Enterococcus faecalis* resulted in a slight underrepresentation of the
232 latter in the mixtures (**Figures 2B**) due to the narrow size window chosen for read size selection (the
233 lower boundary of the size window around the modal read length is adjustable in the *LORCAN*
234 command line). The mixture of two *Mycobacterium* species (97.6 % sequence identity in the amplified
235 region; **Figure 2C**) were accurately reproduced.

236 **Influence of database completeness on consensus accuracy and taxonomic classification.** We analysed
237 the influence of reference database completeness on the resulting consensus quality and accuracy by
238 creating incomplete reference databases, from which we excluded reference sequences if they were too
239 close to the ideal reference sequence, and then performed *LORCAN* analysis with each of these truncated
240 databases in turn. The genetic distances of the closest reference sequences in the reference database
241 strongly influenced the accuracy of the resulting consensus sequences. For instance, *Enterococcus faecalis*
242 showed the lowest consensus accuracy at 95% database identity (**Figure 3**). This was caused by gaps in
243 the closest reference sequence available. For databases with closest identities $\leq 94\%$, the reference

244 sequence with the identified gaps was absent and consensus quality increased again (**Figures S3 and**
245 **S4**). Classification at the species level was, however, virtually unaffected in pure amplicons. The
246 *Eggerthella lenta* dataset contained a contamination of *Pseudomonas stutzeri* reads (0.8% of all reads),
247 which did not influence classification when reference sequences that enabled a mapping of *Eggerthella*
248 *lenta* reads were available. In the absence of sufficiently close reference sequences, the sample was
249 misidentified (**Figure 3A**). Information provided in the *LORCAN* report did, however, reveal that the
250 *Pseudomonas stutzeri* consensus sequence was only based on 20 out of 850 reads, which therefore
251 indicated a likely case of sub-optimal taxonomic classification.

252 *Validation of sequence consensus generated by the combination of nanopore sequencing and*
253 **LORCAN**. The comparison of 78 *LORCAN* generated consensus sequences from 14 sequencing runs
254 (including 49 clinical samples and 15 ATCC reference strains) to their corresponding Sanger sequences
255 revealed an average sequence identity of $99.6\% \pm 0.6$ (standard deviation). The positive control
256 (originating from the same pool of amplicons) that was systematically sequenced in these 14 runs
257 showed an average identity of $99.8\% \pm 0.2$ to its corresponding Sanger sequence. All reference strains
258 were correctly identified at the species level by *LORCAN*. Identification by LeBIBI QBPP resulted in
259 assignment of the expected species (lowest patristic distance) or the placement of the expected species in
260 the proximal cluster of the query sequence (in the phylogenetic tree) in all but two cases. In these cases,
261 the analysed strains were placed in close neighbourhood of the expected species in the phylogenetic tree
262 produced by LeBIBI QBPP (**Table 1, Figure S5**).

263 *Comparison of sequencing costs* . Costs per sample, of the Sanger method were the lowest across
264 different sequencing technologies (**Figure 4**), provided the analysed amplicons are pure and short
265 enough to be covered by a single sequence at sufficient quality. Among the analysed NGS methods,
266 nanopore sequencing was by far the most cost-effective option particularly at throughputs of 24 to 48
267 samples. The high costs per sample of Illumina are mainly caused by the non-reusable sequencing

268 cartridges (the full costs apply, regardless of the number of processed samples) and the comparably high
269 prices of the library preparation kits.

270 *Effects of parameter modifications on LORCAN results.* We studied the influence of the read size
271 fraction (relative to the modal read length) and the number of input-reads on LORCAN consensus
272 quality. In short, optimal results were obtained when reads shorter than 20 bases below the modal read
273 length were excluded from the analysis (Figure S6). Further, we found 100 reads to be sufficient for the
274 generation of high quality of consensus sequences (Figure S7, S8, S9). The required number of input
275 reads may vary with the taxonomic complexity of the analysed samples and the resolution required by
276 the operator. From a theoretical viewpoint (Figure 1B; step 2), a total of 3,000 size-selected reads may
277 allow for the creation of high-quality consensus sequences and reliable species identification for species
278 contributing $\geq 3.3\%$ of those 3,000 selected reads (i.e. when setting a minimum reference mapping depth
279 of 100 reads in LORCAN, which corresponds to the minimum number recommended of reads for
280 reliable consensus creation; Figure S7). In most cases, however, even when a sample may consist of
281 amplicons derived from a unique species, not all reads are assigned to the target species (e.g. due to read
282 errors and/or the presence of highly similar sequences associated with other species). Furthermore,
283 demultiplexing and size selection could result in significant reduction of available reads. For illustrative
284 purposes, during our last 11 sequencing runs consisting of 89 samples (including place-holder samples;
285 see paragraph "Library preparation" in the Methods section), an average of $639,944 \pm 267,704$ basecalled
286 reads were produced, while multiplexing on average 8 ± 3 barcoded samples per sequencing run. Read
287 demultiplexing produced thereafter an average of $46,571 \pm 22,129$ reads per library (i.e. in 58% of all
288 reads both index sequences have been identified and assigned to the same barcode). This comparably
289 high read loss resulted from the stringent demultiplexing parameters used (detection of both 5' and 3'
290 barcodes required, exclusion of reads with internal barcodes), which may effectively prevent crosstalk
291 between libraries (32). Subsequent size selection (read length -20 to +100 bp around the modal sequence

292 length) resulted in an average of $43,265 \pm 21,305$ reads per barcode that were available for further
293 processing. Samples producing more than 3,000 reads of the expected amplicon size were further down-
294 sampled at a threshold of 3000 reads (adjustable *LORCAN* parameter), resulting in an average number of
295 used reads of $3,008 \pm 6$ reads per sample. All samples, controls and place-holders processed in these 11
296 sequencing runs were successfully taxonomically identified. Although species identification could have
297 been achieved with a lower number of reads per sample, sequence production was fast (i.e.
298 approximately 1-2 hours for 1 million reads), and even if flow cells may have been reused up to four
299 times, the maximal sequencing capacity of the flow cells was never utilized (**Table S2**).

300

301 Discussion

302 We present here the first sample-to-report solution for marker-gene based taxonomic identification of
303 bacterial cultures specifically designed for clinical applications. We extensively tested the influences of
304 various analysis parameters and therefore provide a basis for optimal tuning of the *LORCAN* pipeline to
305 specific requirements. We demonstrated that reads significantly shorter than the modal read length
306 showed reduced mappability to reference sequences and that resulting consensus sequences were of
307 reduced quality. No such observations were made when using reads from longer length fractions
308 (**Figure S6**). Therefore, we excluded reads that were significantly shorter than the mode of the read
309 length distribution (by 20 bases) from the analysis with the corresponding command line parameter in
310 *LORCAN*. With these parameters being set this way, accurate consensus sequences ($\geq 99\%$ identity to
311 Sanger sequences produced from the same DNA) were reliably produced with as few as 100 size-filtered
312 reads per sample (**Figure S7**), confirming previous findings (33).

313 Applicability to samples consisting of mixed amplicons was a key requirement during development
314 of *LORCAN* as contaminations are not rare in bacterial cultures derived from clinical samples. To
315 exclude sources of variation due to fluctuations in wet laboratory processes, we analysed artificially

316 mixed amplicons based on pure reads generated from pure amplicons. *LORCAN* showed high
317 robustness against such mixture events and was capable of quantitatively representing read
318 compositions in mixed samples, as long as the analysed gene region and the used database provide the
319 required taxonomic resolution. Nevertheless, we consider our presented approach as semi-quantitative
320 as biases inherent to DNA extraction and amplicon generation might occur. In addition, the presence of
321 near-identical reference sequences belonging to different species can result in elevated levels of
322 background due to miss-assignment of a fraction of the reads. Although we could observe a likely bias
323 due to this phenomenon (**Figure S1**), the bias did not prevent the correct taxonomic identification of the
324 most abundant species in any of our experiments. Furthermore, this bias can be mitigated by choosing
325 longer amplicons, and the planned improvement in read quality by ONT will likely improve
326 discrimination under such conditions.

327 A number of studies on ONT-based marker gene analysis have been published over the past years,
328 covering a range of different laboratory and computational approaches aiming to obtain high quality
329 sequences from ONT reads. Most computational workflows either include reference-based consensus
330 generation or *de novo* assembly, in combination with additional error correction steps. They were
331 reported to perform similarly in terms of the accuracy of the produced sequences (12, 13, 15, 17, 33). *De*
332 *nov*o approaches are preferable when reference sequences are missing, however, so far the only studies
333 demonstrating "reference-free" consensus generation from complex samples (e.g. mock communities)
334 relied on rather laborious wet-lab procedures such as rolling cycle amplification or unique tagging of the
335 individual amplicons before sequencing (15, 16). Unlike previous studies we specifically designed our
336 workflow for clinical routine applications. Compatibility with mixed samples and time/cost efficacy
337 were therefore key requirements and comprehensive reference databases were readily available. We
338 therefore chose a reference-based approach allowing us to separate reads originating from mixed
339 cultures while using standard ONT protocols. Furthermore, and in contrast to most previous studies, we

340 omitted consensus error correction which is commonly applied to remove homopolymer errors from
341 consensus sequences and assemblies produced from nanopore reads (12, 13), because we did not detect a
342 negative influence of the latter errors in our taxonomic classification approach.

343 The strengths of our proposed approach is that overall the procedure is faster, more flexible, and
344 more cost effective than Sanger or Illumina-based approaches, as it relies on both straightforward ONT
345 protocols and automated sample analysis up to result reporting. In addition, nanopore sequencing is
346 compatible with any amplicon size, which is a clear advantage over other existing sequencing
347 technologies, and also allows the processing and resolution of mixed amplicon samples as demonstrated
348 here. Finally, even when the reference sequence database is incomplete or lacks closely related reference
349 sequences, we showed that the approach is robust and provides correct taxonomic identification of the
350 bacterial species.

351 Our approach has several limitations. i) The taxonomic resolution is inherently limited by the
352 choice of a single-gene based approach. Commonly used 16S rRNA gene regions, for example, have been
353 reported to allow for genus identification in >90% of cases, for species identification in 65 to 83% of cases
354 and to result in unsuccessful identification in 1 to 14% of all analysed isolates (8, 34, 35). Other
355 approaches, such as MALDI-TOF (matrix-assisted laser desorption/ionization time-of-flight) mass
356 spectrometry may complementarily provide fast and reliable identification of clinically-relevant
357 microorganisms (36). Yet, MALDI-TOF may also suffer from sub-optimal identification due to
358 limitations, including insufficient representation of reference species profiles in available commercial
359 databases, absence of newly discovered species, and the existence of several commercial systems (37-39).
360 ii) The dependency on database quality and completeness in the *LORCAN* reference-based approach for
361 consensus building was explored extensively by using modified databases which lacked reference
362 sequences closely related to the analysed strains: Not surprisingly, consensus accuracy was strongly
363 affected, and *LORCAN* required reference databases of high quality and completeness to reliably reach

364 sequence qualities on par with the quality obtained by the Sanger method. Even if databases contained
365 sequences with up to 99% identity to the analysed species, further improvements could often be made
366 by adding closer reference sequences (**Figure 3**). Importantly though for clinical diagnostics, taxonomic
367 identification based on the produced consensus sequence was far less affected by database
368 completeness: Even consensus sequences produced with distant reference sequences ($\leq 90\%$ identity to
369 the query sequence, using an incomplete database) allowed for reliable bacterial species identification,
370 when the generated consensus was compared to a complete database. This finding indicates a high
371 reliability of the taxonomic identification despite the database dependency of the approach. This was
372 confirmed by extensive validation in our diagnostics department, which was based on the parallel
373 sequencing and analysis of clinical samples using both Sanger and nanopore sequencing over several
374 months, which overall showed average sequence identities of 99.6% (and 99.8% for positive controls
375 sequenced conjointly with the clinical samples). iii) Finally, the wet laboratory procedures still take
376 several hours, and would need to be optimized to allow fast and efficient processing of several samples
377 via automation or via simplified steps.

378 In conclusion, we demonstrate that the combination of nanopore sequencing and *LORCAN* pipeline
379 offers a significant improvement over the well-established Sanger or short-read sequencing approaches,
380 in terms of reliability (robustness against contaminated samples) and flexibility (read length limited by
381 PCR only), while offering comparable turnaround time, cost and reproducibility of the results. The
382 described workflow has great potential to be successfully introduced in the routine of diagnostics
383 department and may thus facilitate custom amplicon sequencing and further taxonomic identification of
384 bacterial pathogens.

385

386

387 Acknowledgements

388 We thank Christian Baumann for his excellent technical assistance, and John W Looney for his help in
389 the preparation of technical documents.

390

391 Funding

392 The project was financed by the Institute for Infectious Diseases, University of Bern, Switzerland.

393

394 Conflict of interest

395 AR received travel grants from Oxford Nanopore Technologies to attend scientific conferences. The
396 sponsor had no role in the design, execution, interpretation, or writing of the study.

397

398 References

- 399 1. Maruvada P, Leone V, Kaplan LM, Chang EB. 2017. The Human Microbiome and Obesity:
400 Moving beyond Associations. *Cell Host Microbe* 22:589-599.
- 401 2. Durban A, Abellan JJ, Jimenez-Hernandez N, Ponce M, Ponce J, Sala T, D'Auria G, Latorre
402 A, Moya A. 2011. Assessing gut microbial diversity from feces and rectal mucosa. *Microb*
403 *Ecol* 61:123-33.
- 404 3. Janda JM, Abbott SL. 2007. 16S rRNA gene sequencing for bacterial identification in the
405 diagnostic laboratory: pluses, perils, and pitfalls. *J Clin Microbiol* 45:2761-4.
- 406 4. Srinivasan R, Karaoz U, Volegova M, MacKichan J, Kato-Maeda M, Miller S, Nadarajan R,
407 Brodie EL, Lynch SV. 2015. Use of 16S rRNA gene for identification of a broad range of
408 clinically relevant bacterial pathogens. *PLoS One* 10:e0117617.
- 409 5. Zhang J, Fang Y, Hou JY, Ren HJ, Jiang R, Roos P, Dovichi NJ. 1995. Use of non-cross-
410 linked polyacrylamide for four-color DNA sequencing by capillary electrophoresis
411 separation of fragments up to 640 bases in length in two hours. *Anal Chem* 67:4589-93.
- 412 6. Heather JM, Chain B. 2016. The sequence of sequencers: The history of sequencing DNA.
413 *Genomics* 107:1-8.
- 414 7. Tenney AE, Wu JQ, Langton L, Klueh P, Quatrano R, Brent MR. 2007. A tale of two
415 templates: automatically resolving double traces has many applications, including
416 efficient PCR-based elucidation of alternative splices. *Genome Res* 17:212-8.

- 417 8. Mignard S, Flandrois JP. 2006. 16S rRNA sequencing in routine bacterial identification: a
418 30-month experiment. *J Microbiol Methods* 67:574-81.
- 419 9. Nicholls SM, Quick JC, Tang S, Loman NJ. 2019. Ultra-deep, long-read nanopore
420 sequencing of mock microbial community standards. *Gigascience* 8.
- 421 10. Srivathsan A, Baloglu B, Wang W, Tan WX, Bertrand D, Ng AHQ, Boey EJH, Koh JY,
422 Nagarajan N, Meier R. 2018. A MinION-based pipeline for fast and cost-effective DNA
423 barcoding. *Mol Ecol Resour* doi:10.1111/1755-0998.12890.
- 424 11. Vaser R, Sovic I, Nagarajan N, Sikic M. 2017. Fast and accurate de novo genome assembly
425 from long uncorrected reads. *Genome Res* 27:737-746.
- 426 12. Menegon M, Cantaloni C, Rodriguez-Prieto A, Centomo C, Abdelfattah A, Rossato M,
427 Bernardi M, Xumerle L, Loader S, Delledonne M. 2017. On site DNA barcoding by
428 nanopore sequencing. *PLoS One* 12:e0184741.
- 429 13. Maestri S, Cosentino E, Paterno M, Freitag H, Garces JM, Marcolungo L, Alfano M, Njunjic
430 I, Schilthuisen M, Slik F, Menegon M, Rossato M, Delledonne M. 2019. A Rapid and
431 Accurate MinION-Based Workflow for Tracking Species Biodiversity in the Field. *Genes*
432 (Basel) 10.
- 433 14. Somerville V, Lutz S, Schmid M, Frei D, Moser A, Irmeler S, Frey JE, Ahrens CH. 2019.
434 Long-read based de novo assembly of low-complexity metagenome samples results in
435 finished genomes and reveals insights into strain diversity and an active phage system.
436 *BMC Microbiol* 19:143.
- 437 15. Calus ST, Ijaz UZ, Pinto AJ. 2018. NanoAmpli-Seq: a workflow for amplicon sequencing for
438 mixed microbial communities on the nanopore sequencing platform. *Gigascience* 7.
- 439 16. Karst SM, Ziels RM, Kirkegaard RH, Albertsen M. 2019. Enabling high-accuracy long-read
440 amplicon sequences using unique molecular identifiers and Nanopore sequencing.
441 *bioRxiv* doi:10.1101/645903:645903.
- 442 17. Benitez-Paez A, Portune KJ, Sanz Y. 2016. Species-level resolution of 16S rRNA gene
443 amplicons sequenced through the MinION portable nanopore sequencer. *Gigascience*
444 5:4.
- 445 18. Kai S, Matsuo Y, Nakagawa S, Kryukov K, Matsukawa S, Tanaka H, Iwai T, Imanishi T,
446 Hirota K. 2019. Rapid bacterial identification by direct PCR amplification of 16S rRNA
447 genes using the MinION nanopore sequencer. *FEBS Open Bio* 9:548-557.
- 448 19. Ramette A. 2019. GitHub Repository for LORCAN Pipeline. Available online:
449 <https://github.com/aramette/LORCAN/> Accessed 27/08/2019.
- 450 20. Wick RR. Available online: <https://github.com/rrwick/Porechop>.
- 451 21. Li H. 2018. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics*
452 34:3094-3100.
- 453 22. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin
454 R, Genome Project Data Processing S. 2009. The Sequence Alignment/Map format and
455 SAMtools. *Bioinformatics* 25:2078-9.
- 456 23. Shen W, Le S, Li Y, Hu F. 2016. SeqKit: A Cross-Platform and Ultrafast Toolkit for
457 FASTA/Q File Manipulation. *PLoS One* 11:e0163962.
- 458 24. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. 1990. Basic local alignment search
459 tool. *J Mol Biol* 215:403-10.

- 460 25. Katoh K, Standley DM. 2013. MAFFT multiple sequence alignment software version 7:
461 improvements in performance and usability. *Mol Biol Evol* 30:772-80.
- 462 26. Castresana J. 2000. Selection of conserved blocks from multiple alignments for their use in
463 phylogenetic analysis. *Mol Biol Evol* 17:540-52.
- 464 27. Nguyen LT, Schmidt HA, von Haeseler A, Minh BQ. 2015. IQ-TREE: a fast and effective
465 stochastic algorithm for estimating maximum-likelihood phylogenies. *Mol Biol Evol*
466 32:268-74.
- 467 28. Flandrois JP, Perriere G, Gouy M. 2015. leBIBIQBPP: a set of databases and a webtool for
468 automatic phylogenetic analysis of prokaryotic sequences. *BMC Bioinformatics* 16:251.
- 469 29. Schloss PD, Westcott SL, Ryabin T, Hall JR, Hartmann M, Hollister EB, Lesniewski RA,
470 Oakley BB, Parks DH, Robinson CJ, Sahl JW, Stres B, Thallinger GG, Van Horn DJ, Weber
471 CF. 2009. Introducing mothur: open-source, platform-independent, community-
472 supported software for describing and comparing microbial communities. *Appl Environ*
473 *Microbiol* 75:7537-41.
- 474 30. leBIBI-QBPP. <https://umr5558-bibiserv.univ-lyon1.fr/lebibi/lebibi.cgi>, database
475 procaryota_SSU-rDNA-16S_TS-stringent, version 2019/Feb/07 14:40. Accessed
- 476 31. Corpet F. 1988. Multiple sequence alignment with hierarchical clustering. *Nucleic Acids*
477 *Res* 16:10881-90.
- 478 32. Xu Y, Lewandowski K, Lumley S, Pullan S, Vipond R, Carroll M, Foster D, Matthews PC,
479 Peto T, Crook D. 2018. Detection of Viral Pathogens With Multiplex Nanopore MinION
480 Sequencing: Be Careful With Cross-Talk. *Frontiers in microbiology* 9:2225-2225.
- 481 33. Pomerantz A, Penafiel N, Arteaga A, Bustamante L, Pichardo F, Coloma LA, Barrio-
482 Amoros CL, Salazar-Valenzuela D, Prost S. 2018. Real-time DNA barcoding in a
483 rainforest using nanopore sequencing: opportunities for rapid biodiversity assessments
484 and local capacity building. *Gigascience* 7.
- 485 34. Woo PC, Ng KH, Lau SK, Yip KT, Fung AM, Leung KW, Tam DM, Que TL, Yuen KY. 2003.
486 Usefulness of the MicroSeq 500 16S ribosomal DNA-based bacterial identification system
487 for identification of clinically significant bacterial isolates with ambiguous biochemical
488 profiles. *J Clin Microbiol* 41:1996-2001.
- 489 35. Drancourt M, Bollet C, Carliz A, Martelin R, Gayral JP, Raoult D. 2000. 16S ribosomal
490 DNA sequence analysis of a large collection of environmental and clinical unidentifiable
491 bacterial isolates. *J Clin Microbiol* 38:3623-30.
- 492 36. Keys CJ, Dare DJ, Sutton H, Wells G, Lunt M, McKenna T, McDowall M, Shah HN. 2004.
493 Compilation of a MALDI-TOF mass spectral database for the rapid screening and
494 characterisation of bacteria implicated in human infectious diseases. *Infect Genet Evol*
495 4:221-42.
- 496 37. Sandalakis V, Goniotakis I, Vranakis I, Chochlakis D, Psaroulaki A. 2017. Use of MALDI-
497 TOF mass spectrometry in the battle against bacterial infectious diseases: recent
498 achievements and future perspectives. *Expert Rev Proteomics* 14:253-267.
- 499 38. Psaroulaki A, Chochlakis D. 2018. Use of MALDI-TOF mass spectrometry in the battle
500 against bacterial infectious diseases: recent achievements and future perspectives. *Expert*
501 *Rev Proteomics* 15:537-539.

- 502 39. Tsuchida S. 2018. Application of MALDI-TOF for Bacterial Identification. The Use of Mass
503 Spectrometry Technology (MALDI-TOF) in Clinical Microbiology,. doi:10.1016/b978-0-
504 12-814451-0.00007-1:101–112.
505

506

507 **Figure legends**

508

509 **Figure 1.** **A)** Overview of the wet laboratory workflow. **B)** Steps of the *LORCAN* analysis and **C)**
510 corresponding sections of the generated report. Step 1: Demultiplexing and adapter trimming. Step 2:
511 Read filtering by size. Step 3: Mapping to a reference database. Step 4: Read extraction, binning by
512 species and re-mapping. Step 5: Consensus calling. Step 6: Selection of the closest references by BLAST.
513 Step 7: Taxonomic tree building.

514 **Figure 2.** Taxonomic analysis of amplicon mixtures by *LORCAN*. **A)** Amplicons from *Staphylococcus*
515 *aureus*, *Enterococcus faecalis* and *Pseudomonas aeruginosa* mixed after PCR amplification, and **B)** mixed *in*
516 *silico* from reads obtained from pure amplicons. Standard deviations indicate the variability across three
517 independent replicate samples. None of the observed ratios was significantly different from the expected
518 ratios (Chi-square test for expected probabilities; $P>0.99$). **C)** *in silico* mixtures of *Mycobacterium gordonae*
519 and *Mycobacterium avium*.

520 **Figure 3.** Influence of reference database completeness on consensus sequence accuracy. Each consensus
521 sequence was compared to a consensus sequence produced with a perfectly matching reference
522 sequence. Additionally, each consensus sequence was identified by *BLAST* similarity search against the
523 full reference database. The uneven spacing of the data points reflects the database composition after
524 subsetting. Missing values are a result of insufficient numbers of reads mapping to the reference
525 database. **A)** Filled circles indicate correct taxonomic identification of the ATCC strains. The low
526 identities and unsuccessful identification of *Eggerthella lenta* are a result of a low-level contamination in
527 combination with unsuccessful mapping of the *Eggerthella* reads. **B)** The diameter of the circles is
528 proportional to the number of reads mapped and further used in the consensus generation step
529 (obtained from the *LORCAN* output). Additional detail is provided in **Table S3** and **Figure S10**.

530 **Figure 4.** Cost estimate based on current list prices in Switzerland (currency CHF, December 2019):
531 Prices for Illumina and Nanopore sequencing include reagents and consumables; prices for Sanger
532 sequencing correspond to the rates at a large local service provider. The lines of MiniSeq and MISEQ v3
533 are confounded in the Figure. Detail is provided in **Table S4.**

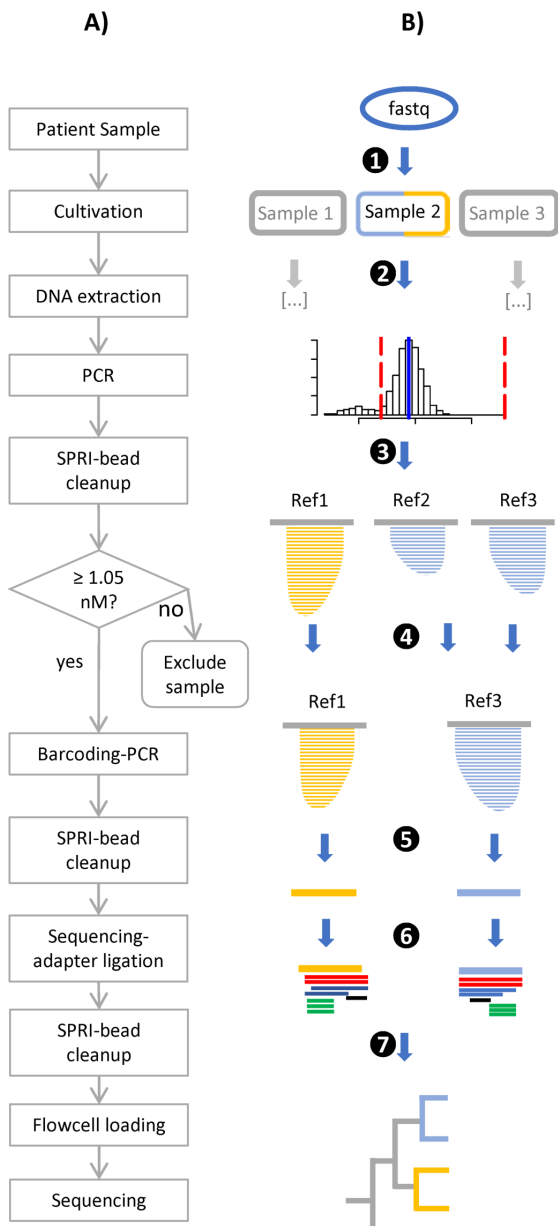
534 **Table 1.** Validation of taxonomic classification of ATCC reference strains. Samples were analysed in
 535 parallel by Sanger sequencing and with the LORCAN approach. The resulting consensus sequences were
 536 submitted to the online taxonomic identification platform leBIBI QBPP.

ATCC strain		LORCAN top consensus sequence		SANGER consensus sequence	LORCAN vs. Sanger consensus sequences
Reference number	Taxonomy	LORCAN taxonomy	leBIBI QBPP Taxonomy ¹⁾	leBIBI QBPP Taxonomy ¹⁾	Identity [%]
33560	<i>C. jejuni</i> subsp. <i>jejuni</i>	<i>Campylobacter jejuni</i>	[<i>Campylobacter lari</i> subsp. <i>concheus</i> , <i>Campylobacter jejuni</i> subsp. <i>jejuni</i> *, <i>Campylobacter jejuni</i> subsp. <i>doylei</i>] (and 2 others)	[<i>Campylobacter lari</i> subsp. <i>concheus</i> , <i>Campylobacter jejuni</i> subsp. <i>jejuni</i> *, <i>Campylobacter jejuni</i> subsp. <i>doylei</i>] (and 2 others)	99.77
43504	<i>Helicobacter pylori</i>	<i>Helicobacter pylori</i>	[<i>Helicobacter pylori</i> *]	[<i>Helicobacter pylori</i> *]	99.54
29212	<i>Enterococcus faecalis</i>	<i>Enterococcus faecalis</i>	[<i>Enterococcus faecalis</i> *]	[<i>Enterococcus faecalis</i> *]	100.00
25922	<i>Escherichia coli</i>	<i>Escherichia coli</i>	[<i>Escherichia marmotae</i> , <i>Escherichia fergusonii</i>] <i>Shigella flexneri</i> *	[<i>Shigella flexneri</i>]	99.57
49247	<i>Haemophilus influenzae</i>	<i>Haemophilus influenzae</i>	[<i>Haemophilus influenzae</i> *]	[<i>Haemophilus influenzae</i> *]	98.94
49226	<i>Neisseria gonorrhoeae</i>	<i>Neisseria gonorrhoeae</i>	[<i>Neisseria gonorrhoeae</i> *]	[<i>Neisseria gonorrhoeae</i> *]	100.00
27853	<i>Pseudomonas aeruginosa</i>	<i>Pseudomonas aeruginosa</i>	[<i>Pseudomonas tropicalis</i> *, <i>Pseudomonas aeruginosa</i> , <i>Pseudomonas hussainii</i>]	[<i>Pseudomonas tropicalis</i> *, <i>Pseudomonas indica</i> , <i>Pseudomonas aeruginosa</i>]	99.78
25923	<i>Staphylococcus aureus</i>	<i>Staphylococcus aureus</i>	[<i>Staphylococcus aureus</i> subsp. <i>anaerobius</i> *]	[<i>Staphylococcus argenteus</i> , <i>Staphylococcus aureus</i> subsp. <i>aureus</i> , <i>Staphylococcus schweitzeri</i>] (and 2 others)	99.79
49619	<i>Streptococcus pneumoniae</i>	<i>Streptococcus pneumoniae</i>	[<i>Streptococcus pneumoniae</i> *, <i>Streptococcus pseudopneumoniae</i>]	[<i>Streptococcus mitis</i> , <i>Streptococcus pneumoniae</i> *]	99.79
29741	<i>Bacteroides thetaiotaomicron</i>	<i>Bacteroides thetaiotaomicron</i>	[<i>Bacteroides thetaiotaomicron</i> *]	[<i>Bacteroides thetaiotaomicron</i> *]	99.78
43055	<i>Eggerthella lenta</i>	<i>Eggerthella lenta</i>	[<i>Eggerthella lenta</i> *]	[<i>Eggerthella lenta</i> *, <i>Eggerthella timonensis</i>]	99.32
51299	<i>Enterococcus faecalis</i>	<i>Enterococcus faecalis</i>	[<i>Enterococcus faecalis</i> *]	[<i>Enterococcus faecalis</i> *]	100.00
8176	<i>Moraxella catarrhalis</i>	<i>Moraxella catarrhalis</i>	[<i>Moraxella canis</i> , <i>Moraxella catarrhalis</i> *, <i>Moraxella nonliquefaciens</i>]	[<i>Moraxella canis</i> , <i>Moraxella catarrhalis</i> *]	100.00
BAA-1705	<i>Klebsiella pneumoniae</i>	<i>Klebsiella pneumoniae</i>	[<i>Klebsiella variicola</i> , <i>Klebsiella quasivariicola</i> *]	[<i>Klebsiella pneumoniae</i> subsp. <i>rhinoscleromatis</i> *, <i>Klebsiella quasipneumoniae</i> subsp.	98.93

13637	<i>Stenotrophomonas maltophilia</i>	<i>Stenotrophomonas maltophilia</i>	[<i>Stenotrophomonas maltophilia</i> *	<i>quasipneumoniae</i> [<i>Stenotrophomonas maltophilia</i>]	100.00
-------	-------------------------------------	-------------------------------------	---	---	--------

¹⁾ Square brackets indicate proximal clusters. Asterisks indicate closest sequences based on patristic distances.

537



```

=====
***      REPORT OF SEQUENCED PCR AMPLICONS (NANOPORE)      ***
=====
Analysis of barcoded sample (BC38): Mix1:1:1:1:1:16s
Date of report: Thu Oct 24 14:18:19 2019
=====
Total number of reads in the input fasta file: *3000*
The reference database *BiB116SLong* contains *69566* sequences
=====
    
```

```

== A) Read counts
=====
Total: 3013 reads aligned to references,
100.0% (3013 reads) were kept after applying
100.0% (3013 reads) were kept after applying
    
```

```

== B) Selection of taxonomic groups based on mapped reads
=====
- (1143,37.9%) Staphylococcus aureus
- (1120,37.2%) Enterococcus faecalis
- (725,24.1%) Pseudomonas aeruginosa
    
```

```

== C) STATISTICS ABOUT CONSENSUS SEQUENCES =====
Len  N  - Taxonomic levels
-----
505  0  0 Staphylococcus_aureus
517  0  0 Enterococcus_faecalis
493  0  1 Pseudomonas_aeruginosa
=====
    
```

```

== D) CONSENSUS SEQUENCES
=====
== Reference group: Staphylococcus aureus
GCTGGCGGCGTGCCTAATACATGCAAGTCGAGCGACGACGAGAAGCTTGCTTCTCTGA
TGTTAGCGCGGACGGGTGAGTAAACAGTGGATAACCTACCTATAAGACTGGGATAACTT...
== Reference group: Enterococcus faecalis
GCTGGCGGCGTGCCTAATACATGCAAGTCGAGCGCTTCTTCTCCGAGTGCCTGCACT
CAATTGGAAGAGGAGTGGCGGACGGGTGAGTAAACAGTGGGTACCTACCCATCAGAGG...
    
```

```

== E) Checking the taxonomic classification of the obtained consensus sequences
...
100.000/505/505/0/0 Staphylococcus_aureus-v-N-URS000078968B=Bacteria-Firmicut...
99.802/505/506/0/1 Staphylococcus_aureus-v-N-URS0000745EFS=Bacteria-Firmicut...
99.802/505/506/0/1 Staphylococcus_aureus_subsp._aureus-v-N-URS00005E8D17=Bac...
...
    
```

```

== F) PHYLOGENETIC TREE =====
=====
**Staphylococcus_aureus-v-N-URS0000699A=Bacteria-Firmicutes-Bacilli-Bacillales-Staphylococaceae...
|
|-----**Enterococcus_faecalis-v-N-URS00005D399=Bacteria-Firmicutes-Bacilli-Lactobacill...
|
|----- (53)
|-----**BC38_using_Enterococcus_faecalis_consensus_517bp_30_Oct_24_2019
|
|-----**BC38_using_Staphylococcus_aureus_consensus_505bp_30_Oct_24_2019
    
```

