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.

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

Stefan Moritz Neuenschwander[1], Miguel Angel Terrazos Miani[1], Heiko Amlang[1], Carmen Perroulaz[1], Pascal Bittel[1], Carlo Casanova[1], Sara Droz[1], Jean-Pierre Flandrois[2], Stephen L. Leib[1], Franziska Suter-Riniker[1], Alban Ramette[1,*]

[1] University of Bern, Institute for Infectious Diseases, Bern, Switzerland

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

[*] Corresponding author: Alban Ramette, University of Bern, Institute for Infectious Diseases, Friedbühlstrasse 51, CH-3001 Bern, Switzerland. alban.ramette@ifik.unibe.ch

Abstract

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

Introduction

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

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

Journal of Clinical Microbiology

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

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

Journal of Clinical Microbiology

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

Methods

Samples, DNA extraction, PCR amplification

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

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

106 µM primer working solution, both with 12.5 µl 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 µl of Tris-HCI (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

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

Sequencing

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

Bioinformatic analyses

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

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

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

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

Journal of Clinical Microbiology

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

Journal of Clinical Microbiology

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*,

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

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

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

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

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

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

Journal of Clinical Microbiology

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

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

Journal of Clinical Microbiology

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

Discussion

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

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

Journal of Clinical Microbiology

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 *novo* 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

Journal of Clinical Microbiology

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

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

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

Journal of Clinical Microbiology

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

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

Journal of Clinical Microbiology

Acknowledgements

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

Funding

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

Conflict of interest

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

References

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

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

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

Figure legends

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

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

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

Journal of Clinical Microbiology

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	*C. jejuni subsp. jejuni*	*Campylobacter jejuni*	*[Campylobacter lari subsp. concheus, Campylobacter jejuni subsp. jejuni*, Campylobacter jejuni subsp. doylei] (and 2 others)*	*[Campylobacter lari subsp. concheus, Campylobacter jejuni subsp. jejuni*, Campylobacter jejuni subsp. doylei] (and 2 others)*	*99.77*
43504	*Helicobacter pylori*	*Helicobacter pylori*	*[Helicobacter pylori*]*	*[Helicobacter pylori*]*	*99.54*
29212	*Enterococcus faecalis*	*Enterococcus faecalis*	*[Enterococcus faecalis*]*	*[Enterococcus faecalis*]*	*100.00*
25922	*Escherichia coli*	*Escherichia coli*	*[Escherichia marmotae, Escherichia fergusonii] Shigella flexneri**	*[Shigella flexneri]*	*99.57*
49247	*Haemophilus influenzae*	*Haemophilus influenzae*	*[Haemophilus influenzae*]*	*[Haemophilus influenzae*]*	*98.94*
49226	*Neisseria gonorrhoeae*	*Neisseria gonorrhoeae*	*[Neisseria gonorrhoeae*]*	*[Neisseria gonorrhoeae*]*	*100.00*
27853	*Pseudomonas aeruginosa*	*Pseudomonas aeruginosa*	*[Pseudomonas tropicalis*, Pseudomonas aeruginosa, Pseudomonas hussainii]*	*[Pseudomonas tropicalis*, Pseudomonas indica, Pseudomonas aeruginosa]*	*99.78*
25923	*Staphylococcus aureus*	*Staphylococcus aureus*	*[Staphylococcus aureus subsp. anaerobius*]*	*[Staphylococcus argenteus, Staphylococcus aureus subsp. aureus, Staphylococcus schweitzeri*] (and 2 others)*	*99.79*
49619	*Streptococcus pneumoniae*	*Streptococcus pneumoniae*	*[Streptococcus pneumoniae*, Streptococcus pseudopneumoniae]*	*[Streptococcus mitis, Streptococcus pneumoniae*]*	*99.79*
29741	*Bacteroides thetaiotaomicron*	*Bacteroides thetaiotaomicron*	*[Bacteroides thetaiotaomicron*]*	*[Bacteroides thetaiotaomicron*]*	*99.78*
43055	*Eggerthella lenta lenta*	*Eggerthella lenta*	*[Eggerthella lenta*]*	*[Eggerthella lenta*, Eggerthella timonensis]*	*99.32*
51299	*Enterococcus faecalis*	*Enterococcus faecalis*	*[Enterococcus faecalis*]*	*[Enterococcus faecalis*]*	*100.00*
8176	*Moraxella catarrhalis*	*Moraxella catarrhalis*	*[Moraxella canis, Moraxella catarrhalis*, Moraxella nonliquefaciens]*	*[Moraxella canis,Moraxella catarrhalis*]*	*100.00*
BAA-1705	*Klebsiella pneumoniae*	*Klebsiella pneumoniae*	*[Klebsiella variicola, Klebsiella quasivariicola*]*	*[Klebsiella pneumoniae subsp. rhinoscleromatis*, Klebsiella quasipneumoniae subsp.*	*98.93*

				quasipneumoniae]	
13637	*Stenotrophomonas maltophilia*	*Stenotrophomonas maltophilia*	*[Stenotrophomonas maltophilia*]*	*[Stenotrophomonas maltophilia]*	*100.00*

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

537