Benchmarking the MinION: Evaluating long reads for microbial profiling

Robert Maximilian Leidenfrost1*, Dierk-Christoph Pöther2, Udo Jäckel2 & Röbbe Wünschiers1

Nanopore based DNA-sequencing delivers long reads, thereby simplifying the decipherment of bacterial communities. Since its commercial appearance, this technology has been assigned several attributes, such as its error proneness, comparatively low cost, ease-of-use, and, most notably, aforementioned long reads. The technology as a whole is under continued development. As such, benchmarks are required to conceive, test and improve analysis protocols, including those related to the understanding of the composition of microbial communities. Here we present a dataset composed of twelve different prokaryotic species split into four samples differing by nucleic acid quantification technique to assess the specificity and sensitivity of the MinION nanopore sequencer in a blind study design. Taxonomic classification was performed by standard taxonomic sequence classification tools, namely Kraken, Kraken2 and Centrifuge directly on reads. This allowed taxonomic assignments of up to 99.27% on genus level and 92.78% on species level, enabling true-positive classification of strains down to 25,000 genomes per sample. Full genomic coverage is achieved for strains abundant as low as 250,000 genomes per sample under our experimental settings. In summary, we present an evaluation of nanopore sequence processing analysis with respect to microbial community composition. It provides an open protocol and the data may serve as basis for the development and benchmarking of future data processing pipelines.
sequencers using commercially available standards\cite{21}. We investigated four mixed microbial DNA samples differing by the employed DNA-quantitation technique and their composition using the MinION sequencer. The samples were composed of DNA covering up to five orders of magnitude in genome amounts from twelve bacterial species. The aim of the study includes an establishment of a suitable classification pipeline and an assessment of the accuracy of the MinION in samples with unknown microbial composition.

Results and discussion

Raw dataset description. Using the MinION DNA-sequencing platform we generated app. 809k reads in Fast5 file format, equal to an estimate of 8.15 Gbp in a single run within 36 hours (see Supplementary Fig. S1,S2). We could observe increased yield for each pore group switch (a.k.a remux) and output of constant quality on a uniform read length distribution for our sequencing run (see Supplementary Figs. S3–S5). Approximately 807k reads equal to 7.06 Gbp were successfully basecalled and demultiplexed generating an overall yield of 662k reads equivalent to 6.17 Gbp for downstream analysis. Samples one to four, corresponding to the four barcodes used, are composed of app. 142k (#1, heterogeneous sample quantified by ddPCR), 262k (#2, heterogeneous sample quantified by Qubit), 110k (#3, equimolar sample quantified by ddPCR) and 148k (#4, equimolar sample quantified by Qubit) reads, respectively. A total of only four reads were not properly demultiplexed by Porechop, i.e. assigned to a barcode not present in the library. A total of app. 140k reads were demultiplexed as “unclassified” by Porechop, i.e. not assigned to any barcode. All reads not assigned to barcodes #1, #2, #3 or #4 (corresponding to the four samples) were discarded and thus excluded from downstream analysis (Table 1).

Data classification and validation. The use of Centrifuge with nanopore read datasets has been demonstrated before\cite{22,23}. The application of Kraken and Kraken2 on nanopore data has also been described, albeit within different experimental settings, such as the taxonomic classification of reads of well characterized isolates\cite{24} or the taxonomic classification of complete assemblies\cite{25}. Taxonomic classification performed by either, Centrifuge, Kraken or Kraken2 allowed for the heterogeneously concentrated samples (samples #1 and #2, adjusted by ddPCR and Qubit, respectively) an initial choice of five out of twelve strains based on the available Krona plots (see Supplementary data S1). For the samples with equimolar genomic concentration (samples #3 and #4), a selection of twelve strains was immediately possible (Fig. 1). Generally, despite the differences in the underlying software and databases/indices, we could observe substantial agreement\cite{26} between the results obtained from Kraken, Kraken2 and Centrifuge with their respective databases as tested by Fleiss Kappa (lowest 0.778, highest 0.931).

Quantitation by ddPCR delivers slightly different results than quantitation by fluorometry such as Qubit\cite{26,27}. This is due to e.g. different basemap compositions, staining efficiencies or denaturation of DNA prior to droplet generation. Thus, we investigated, if the slight difference between these two quantitation approaches (Qubit vs. ddPCR) were also determinable by nanopore-based DNA-sequencing. Indeed, differences in quantitation, which resulted in different volumes necessary for sample preparations, corresponded to different amount of reads for that specific organism to the same extent (see Supplementary Fig. S6).

Unblinding the ground truth to the sequencing laboratory revealed a correct, that is true positive, selection of all twelve strains in samples of equimolar genomic concentration, as well as a correct selection of five out of twelve strains in the two samples with different genomic concentration. The five strains selected from the heterogeneously concentrated samples made up 99.38% of the genomes calculated to be available in the actual samples of different genomic concentration. This corresponds to a concentration of 2.5 million to 50 million genomes per species and sample. Notably, read classification matching the ground truth on genus level was possible for up to
99.27% (Centrifuge) between all samples, whereas read classification matching the ground truth on species level was up to 92.78% (Centrifuge) across all samples (Table 2). Generally, accuracy and deviation metrics (root mean squared deviation (RMSD) and mean absolute error (MAE)) on genus level were better than on species level. Comparing Centrifuge, Kraken and Kraken2 running their precompiled databases/indices, Centrifuge was able to assign the highest fraction of reads to the theoretically expected genera and species across all samples. Also, Centrifuge performed best with respect to both measures of deviation (RMSD, MAE), whereas Kraken 2 was superior over Kraken. However, beyond the accuracy of each classifier, computational aspects need to be considered. Especially, when limited computational resources are available, such as in field applications, Kraken 2 offers superior processing speed and lower memory consumption compared to Centrifuge and Kraken28.

Precision and recall per species and genus reached generally high values on read level (see Supplementary Table S3, S4). For genera with very low abundance, drops in precision could be observed (see Supplementary Table S3). Reads wrongly classified on species level were, e.g., attributable to close relatives, such as Bacillus species to Bacillus licheniformis, Enterobacter cloacae to Enterobacter hormaechei, et cetera, or exhibited differences in read abundance as compared to true positive hits, which is similar to findings reported by Deshpande et al.19 despite a different sequencing and analysis approach. This is also reflected by the lower values of recall for these species on read level (see Supplementary Table S4). The necessity for accurate databases and unified nomenclature is discussed elsewhere29–32 and has been shown to affect classification of nanopore data18. These results indicate that classification is, as of yet, more reliable on genus level than on species level.

Figure 1. Centrifuge, Kraken and Kraken 2 classification results on genus and species level for equimolar (sample/barcode 3, adjusted by ddPCR and sample/barcode 4, adjusted by Qubit) and heterogeneously concentrated samples (sample/barcode 1, adjusted by ddPCR and sample/barcode 2, adjusted by Qubit) of 12 target strains. Theoretical values and validation by NanoOK (alignments with minimap2) are given for comparison.
Table 2. Taxonomic assignment accuracy and corresponding deviation metrics (RMSD and MAE) for Centrifuge, Kraken and Kraken 2 across all four samples, on genus and species level, respectively. Centrifuge has highest accuracy for all samples, genus level classification metrics are superior compared to corresponding species level classification.

| Sample | Software     | Genus Accuracy (%) | RMSD | MAE | Species Accuracy (%) | RMSD | MAE |
|--------|--------------|--------------------|------|-----|----------------------|------|-----|
| 1 (heterogenous, adjusted by ddPCR) | Centrifuge | 99.27              | 0.0585 | 0.0286 | 84.10              | 0.0718 | 0.0374 |
|        | Kraken       | 97.60              | 0.0589 | 0.0293 | 77.98              | 0.0847 | 0.0426 |
|        | Kraken 2     | 98.57              | 0.0587 | 0.0290 | 81.18              | 0.0775 | 0.0398 |
| 2 (heterogenous, adjusted by Qubit) | Centrifuge | 98.96              | 0.0221 | 0.0123 | 85.15              | 0.0494 | 0.0256 |
|        | Kraken       | 97.06              | 0.0238 | 0.0140 | 79.54              | 0.0669 | 0.0341 |
|        | Kraken 2     | 98.06              | 0.0228 | 0.0132 | 82.47              | 0.0576 | 0.0296 |
| 3 (equimolar, adjusted by ddPCR) | Centrifuge | 99.26              | 0.0469 | 0.0322 | 92.78              | 0.0530 | 0.0417 |
|        | Kraken       | 97.10              | 0.0459 | 0.0332 | 86.89              | 0.0600 | 0.0469 |
|        | Kraken 2     | 98.37              | 0.0464 | 0.0326 | 89.15              | 0.0568 | 0.0450 |
| 4 (equimolar, adjusted by Qubit) | Centrifuge | 99.08              | 0.0287 | 0.0224 | 91.49              | 0.0396 | 0.0332 |
|        | Kraken       | 96.14              | 0.0290 | 0.0234 | 85.75              | 0.0518 | 0.0390 |
|        | Kraken 2     | 98.14              | 0.0287 | 0.0228 | 88.03              | 0.0466 | 0.0368 |

25,000 to 500,000 genomes per sample, using Krona plots. The remaining three strains adjusted to the range of 500 to 5,000 genomes per sample could not be reliably retrieved from the two samples with heterogeneous genomic concentrations (Fig. 2). Their presence was obfuscated by the filter process, i.e. they were as abundant as falsely classified reads and, subsequently, a clear discrimination allowing selection and classification was impossible. With the experimental settings and proceeding as described here, this suggests a dynamic range of detection falsly classified reads and, subsequently, a clear discrimination allowing selection and classification was impossible.

These results showed good consistency with a) the output from the NanoOK analysis by direct comparison (Table 3, see Supplementary Table S5), where at least 99.21% of all available reads could be aligned to selected references and b) the theoretical expectation. Moreover, mean coverages reported by NanoOK indicate potential for de novo genome assemblies (Fig. 3). Full genomic coverage realistically permitting de novo assembly was achieved for strains down to a concentration of 250,000 genomes per sample (see Supplementary Table S5). At comparable sequencing times, we anticipate the concentration level required to achieve full genomic coverage to be even lower for libraries that are not multiplexed.

Despite the error rates currently accompanying MinION sequencing, these results clearly illustrate the viability and possibilities of long reads for direct taxonomic classification and abundance estimation with currently available bioinformatics pipelines.

Conclusion

We present a MinION DNA sequence read dataset to facilitate the Nanopore community to improve and develop new bioinformatics pipelines aimed at the understanding of microbial diversity. Continual benchmarking using updated sequencing methods and chemistries in metagenome analyses is required. With the presented detailed methodology, as a whole, this study follows the FAIR Guiding Principles for scientific data management and stewardship by contributing (F)indable and (A)ccessible data under bioproject accession PRJNA545964 and corresponding signal level data that is (R)eusable for the fast-paced development of third generation sequencing and downstream bioinformatics in a metagenomics context.

Based on the dataset, we present a simple and straightforward analysis pipeline to investigate the composition of microbial communities. Given our experimental approach we were able to achieve highly accurate taxonomic classification of low abundant (25,000 genomes/sample) organisms to at least genus level. Full genomic coverage was achieved for species with an abundancy of 250,000 genomes per sample and sufficient coverage for de novo assembly could be obtained.

While there is no standardized approach for the characterization of bacterial communities, molecular tools are considered powerful to gain knowledge and insight into these, and nanopore sequencing is no exception to this point. In summary, the presented benchmark provides insight into nanopore data and data processing for the taxonomic classification of microbial communities. Hence, this study contributes to the toolsets and development of processing pipelines available to elucidate microbial diversity.

Material and methods

The overall experimental design is setup as follows: Bacteria cultivation, DNA extraction, quantification and creation of mock samples were performed by the Unit for Biological Agents, Federal Institute for Occupational Safety and Health (BAuA). Samples were shipped to the sequencing team (Mittweida UAS). The sequencing team performed library preparation, sequencing and downstream processing unaware of the samples’ actual respective compositions (Fig. 4).
**Sample cultivation and preparation.** DNA from twelve bacterial strains was extracted to form a mock community sample (Table 4) for benchmarking the MinION sequencing platform using the following criteria: (A) Each strain is the type strain of the bacterial species and is available from the Leibniz Institute DSMZ - German Collection of Microorganisms and Cell Cultures GmbH (DSMZ), the National Collection of Type Cultures (NCTC) or the American Type Culture Collection (ATCC). (B) Each strain has a reference sequence deposited at the National Center for Biological Information (NCBI). (C) Each strain has several assemblies of the same species available at the NCBI. (D) The sequencing laboratory is blind to both, the selection itself and the actual composition of the strains selected.
Bacteria were grown overnight as follows: *Dickeya solani*\(^7\) (Todd-Hewitt + 0.5% yeast extract (THY), 28 °C), *Serratia fonticola*\(^7\) (DSMZ-Medium 1, 28 °C), *Bacillus licheniformis*\(^7\) (DSMZ-Medium 1, 37 °C), *Corynebacterium glutamicum*\(^7\) (THY, 28 °C), *Micrococcus luteus*\(^7\) (THY, 28 °C), *Cronobacter sakazakii*\(^7\) (DSMZ-Medium 1, 28 °C), *Achromobacter xylosoxidans* subsp. *xylosoxidans*\(^7\) (DSMZ-Medium 1, 28 °C), *Chromobacterium violaceum*\(^7\) (DSMZ-Medium 1, 28 °C), *Enterobacter hormaechei* subsp. *steigerwaltii*\(^2\) (CASO, 37 °C), *Staphylococcus saprophyticus* subsp. *saprophyticus*\(^3\) (DSMZ-Medium 92, 37 °C) and *Xanthomonas campestris*\(^7\) (DSMZ-Medium 1, 28 °C). DNA of 1 ml of the cell suspension derived from liquid culture or resuspended colonies in PBS was extracted using a modified protocol of the GenElute Plant Genomic DNA Miniprep Kit (Sigma Aldrich,\(^2\)). DNA concentrations were quantified using the Qubit BR assay in a Qubit 1.0 fluorometer according to the manufacturer's protocol. Subsequently, ddPCR targeting the 16S rRNA-gene
was conducted with app. less than 40,000 target genes according to the manufacturer’s instructions (Bio-Rad) using the ddPCR Supermix for Probes (no dUTP). Final concentrations of oligonucleotides were 0.4 pmol/µL 1055Falt (ATGGRTGTCGTCAGCT), 0.2 pmol/µL 1392 R (ACGGGCGGTGTGTAC) and 0.1 pmol/µL 1115IB (FAM-CAACGAGCG-ZEN-CAACCC-3IABkFQ) adopted from Rothrock et al.38. Droplet generation was conducted according to manufacturer’s instructions in a QX200 Droplet Generator and amplified in a T100 Thermal Cycler. PCR conditions were initial denaturation at 95 °C for 10 min, and 30 cycles of denaturation at 95 °C for 30 s, annealing at 57 °C for 45 s, extension at 72 °C for 45 s with a ramp rate of 1 °C/s, followed by a final extension at 98 °C for 10 min and cooling to 12 °C. Droplet evaluation was performed in a QX200 Droplet Reader with QuantaSoft-Software.

Based on Qubit and ddPCR quantitation, the nucleic acids were adjusted to different genomic concentrations ranging from 5 to 5*10^5 genomes/µl (samples #1 and #2, corresponding to sequencing library barcodes #1 and #2), or to equimolar genomic concentration of 5*10^4 genomes/µl (samples #3 and #4 corresponding to sequencing library barcodes #3 and #4).

Samples were shipped on ice by public postal services.

Library preparation and sequencing. A sequencing library was prepared according to manufacturer’s instructions. The Ligation Sequencing Kit (SQK-LSK108, Oxford Nanopore Technologies (ONT)) and the Native Barcoding Expansion 1–12 kit (EXP-NBD103, ONT), barcoding each of the samples (barcodes #1, #2, #3, #4), were used with the following exceptions: Shearing times were prolonged and an optional FFPE DNA repair step (M6630, New England Biolabs (NEB)) was included. The incubation times during the end-repair/dA-tailing (E7546, NEB) were extended from five to 20 minutes for both, the 20 °C and 65 °C incubation steps. Qubit checkpoint measurements were performed according to the library preparation protocol (see Supplementary Table S1). Pooling of the barcoded samples was performed 'as is' instead of protocol-given 'equimolar' . Sequencing was then performed on a R9.4 flowcell (FLO-MIN106, ONT, >1200 pores, see Supplementary Table S2) with MinKNOW (version 2.1.12, ONT) at room temperature.

Base calling and demultiplexing. Upon conclusion of sequencing, raw data in Fast5 file format were transferred to our server (4.17.2-1-ARCH, 20 cores with 2 threads each, 256 GB RAM) and basecalled using the Albacore software (version 2.0.2, ONT) with barcoding option. Subsequently, barcodes were removed from basecalled output and subsequently sorted utilizing Porechop (version 0.2.3, standard settings, https://github.com/rrwick/Porechop). Basecalled and demultiplexed sequencing data quality was assessed with NanoPack (version 1.13.0, https://github.com/wdecoster/NanoPlot)39.

Data classification and validation. Taxonomic classification was performed with standard parameters (Centrifuge “-k 1”) on native reads using Centrifuge (precompiled index: "Bacteria, Archaea (compressed), 2018-4-15")42, as well as Kraken (precompiled database: "DustMasked MiniKraken DB 8GB")40 and Kraken2 (precompiled database: MiniKraken2_v1_8GB)43 and the results were visualized with Krona41 and R42–45. The interactive and intuitive Krona visualization was used to manually select up to twelve bacterial strains. The corresponding genome reference sequences were obtained from NCBI Reference Sequence Database46 (accessed on 2018-07-31).

NanoOK (version 1.34)47 was utilized for an assessment of the read dataset against the selection of NCBI genome reference sequences, using minimap2 aligner (version 2.11)48. To create the minimap2 index, the reference sequences obtained from NCBI Reference Sequence Database were concatenated into a single FastA file. Statistics and additional visualizations were computed with R42–45,49,50. We calculated the accuracy of the classification performed by Centrifuge, Kraken and Kraken 2 on each sample the proportion of reads assigned to the
Table 4. Strain community overview: Overview of the strains selected to compose the microbial community with accessions and genomic specifications shown. 

known input organism at the genus and species level out of the total number reads given any assignment at that rank\(^{18}\). To calculate a corresponding estimate of the accompanying error, the mean absolute error, as well as root mean squared deviation of classified to theoretically present fractions on genus and species level were computed. On read level, precision and recall for genus and species identification were computed\(^{32}\) for Centrifuge, Kraken 2\(^{18}\) vs Kraken 2. To calculate a corresponding estimate of the accompanying error, the mean absolute error, as well as root mean squared deviation of classified to theoretically present fractions on genus and species level were computed\(^{32}\) for Centrifuge, Kraken 2 vs Kraken 2.

Data availability

The data sets supporting the results of this article are available in the under bioproject accession PRJNA545964 (https://www.ncbi.nlm.nih.gov/sra/PRJNA545964) and as Zenodo deposit 3600229 (10.5281/zenodo.3600229).

Received: 13 November 2019; Accepted: 4 March 2020; Published online: 20 March 2020

References

1. Handelsman, J. Metagenomics: application of genomics to uncultured microorganisms. Microbiology and molecular biology reviews: MMRR 68, 669–685, https://doi.org/10.1128/MMRR.68.4.669-685.2004 (2004).
2. Schloss, P. D. & Handelsman, J. Metagenomics for studying unculturable microorganisms: cutting the Gordian knot. Genome biology 6, 229, https://doi.org/10.1186/gb-2005-6-8-229 (2005).
3. Tringe, S. G. & Rubin, E. M. Metagenomics: DNA sequencing of environmental samples. Nature reviews. Genetics 6, 805–814, https://doi.org/10.1038/nrg1709 (2005).
4. Shendure, J. & Ji, H. Next-generation DNA sequencing. Nature biotechnology 26, 1135–1145, https://doi.org/10.1038/nbt1486 (2008).
5. Shendure, J. & Lieberman Aiden, E. The expanding scope of DNA sequencing. Nature biotechnology 30, 1084–1094, https://doi.org/10.1038/nbt.1421 (2012).
6. Quick, J. et al. Real-time, portable genome sequencing for Ebola surveillance. Nature 530, 228–232, https://doi.org/10.1038/53016996 (2016).
7. Jain, M. et al. Linear assembly of a human centromere on the Y chromosome. Nature biotechnology 36, 321–323, https://doi.org/10.1038/nbt.4109 (2018).
8. Bronzato Badial, A. et al. Nanopore Sequencing as a Surveillance Tool for Plant Pathogens in Plant and Insect Tissues. Plant disease 102, 1648–1652, https://doi.org/10.1094/PDIS-04-17-0488-RE (2018).
9. Laver, T. et al. Assessing the performance of the Oxford Nanopore Technologies MinION. Biomolecular detection and quantification 3, 1–8, https://doi.org/10.1002/bdq.2015.02.001 (2015).
10. Wick, R. R., Judd, L. M. & Holt, K. E. Performance of neural network basecalling tools on Oxford Nanopore sequencing. Genome biology 20, 129, https://doi.org/10.1186/s13059-019-1727-y (2019).
11. Bleidorn, C. Third generation sequencing: technology and its potential impact on evolutionary biodiversity research. Systematics and Biodiversity 14, 1–8, https://doi.org/10.1002/1477-2888.2015.1099575 (2016).
12. Walter, M. C. et al. MinION as part of a biomedical rapidly deployable laboratory. Journal of biotechnology 250, 16–22, https://doi.org/10.1016/j.jbiotec.2016.12.006 (2017).

| Species | Strain ID | Accession | DSMZ | ATCC | Refseq ID | Size Mbp | GC (%) | 16S rRNA-genes | genomes in ranged sample |
|---------|-----------|-----------|------|------|-----------|----------|--------|---------------|------------------------|
| Dickeya solani | Gibson 46 | 13 | 14580 | 10341 | NC_006270.3 | 4.22 | 46.19 | 7 | Pos | 2.50E + 07 | 18.7500 |
| Xanthomonas campestris | P25 | 3586 | 33913 | — | NC_003902.1 | 5.08 | 65.07 | 2 | Neg | 2.50E + 05 | 0.1875 |
| Staphylococcus saprophyticus subsp. saprophyticus | S-41 | 20229 | 15305 | 13634 | NC_007350.1 | 2.52 | 33.24 | 6 | Pos | 2.50E + 03 | 0.0018 |
| Corynebacterium glutamicum | 534 | 20300 | 13032 | — | NC_003450.3 | 3.31 | 53.81 | 6 | Neg | 5.00E + 07 | 37.5000 |
| Micrococcus luteus | 2665 | 20030 | 4698 | 2665 | NC_012803.1 | 2.5 | 73.00 | 2 | Pos | 5.00E + 05 | 0.3750 |
| Enterobacter hormaechei subsp. Steigerwaltii | EN-562 | 16691 | — | — | NZ_CP017179.1 | 4.78 | 55.55 | 8 | Neg | 2.50E + 06 | 1.8750 |
| Cronobacter sakazakii | CDC 4562–70 | 4485 | 29544 | 11467 | NC_007351.1 | 0.038 | 30.75 | 6 | Pos | 5.00E + 07 | 37.5000 |
| Achromobacter xylosoxidans subsp. xylosoxidans | KM543 | 2402 | 27061 | 10807 | NC_LN381029.1 | 6.81 | 67.38 | 3 | Neg | 5.00E + 04 | 0.0375 |
| Paeucibacillus odorifer | TOD45 | 15391 | BAA-93 | — | NC_CP009428.1 | 6.81 | 44.21 | 10 | Pos | 5.00E + 03 | 0.0037 |
| Chromobacterium violaceum | MK | 30191 | 12472 | 9757 | NC_005085.1 | 4.75 | 64.83 | 8 | Pos | 2.50E + 04 | 0.0188 |
| Dickeya solani | IPO2222 | 28711 | — | — | NZ_CP015137.1 | 4.92 | 56.21 | 7 | Neg | 5.00E + 02 | 0.0004 |
| Serratia fonticola | CUETM 77.165 | 4576 | 29844 | 12965 | NC_003902.1 | 5.08 | 65.07 | 2 | Neg | 2.50E + 02 | 0.0004 |

**Table 4.** Strain community overview: Overview of the strains selected to compose the microbial community with accessions and genomic specifications shown.
31. Lydon, K. A. & Lipp, E. K. Taxonomic annotation errors incorrectly assign the family Pseudoalteromonadaceae to the order Vibrionales.

28. Wood, D. E., Lu, J. & Langmead, B. Improved metagenomic analysis with Kraken 2.

29. Yarza, P.

26. Bhat, S.

25. Landis, J. R. & Koch, G. G. The measurement of observer agreement for categorical data.

23. Sanderson, N. D.

24. Tyler, A. D.

22. Sedlacek, F. J., Lee, H., Darby, C. J. & Schatz, M. C. Piercing the dark matter: bioinformatics of long-range sequencing and mapping. Nature reviews. Genetics 19, 329–346, https://doi.org/10.1038/s41576-018-0034-3 (2018).

21. Nichols, S. M., Quick, J. C., Tang, S. & Loman, N. J. Ultra-deep, long-read nanopore sequencing of mock microbial community standards. GigaScience 8, 3, https://doi.org/10.1093/gigascience/giy034 (2019).

20. Brown, B. L., Watson, M., Minot, S. S., Rivera, M. C. & Franklin, R. B. MinION™ nanopore sequencing of environmental metagenomes: a synthetic approach. GigaScience 6, 1–10, https://doi.org/10.1093/gigascience/gix07 (2017).

19. Deshpande, S. V. et al. Offline Next Generation Metagenomics Sequence Analysis Using MinION Detection Software (MINDS). Genes 10, https://doi.org/10.3390/genes10080578 (2019).

18. Brown, B. L., Watson, M., Minot, S. S., Rivera, M. C. & Franklin, R. B. MinION™ nanopore sequencing of environmental metagenomes: a synthetic approach. GigaScience 6, 1–10, https://doi.org/10.1093/gigascience/gix07 (2017).

17. Benítez-Paz, A., Portune, K. J. & Sanz, Y. Multi-locus and long amplicon sequencing approach to study microbial diversity at species level using the MinION™ portable nanopore sequence. GigaScience 6, 1–12, https://doi.org/10.1093/gigascience/giz043 (2017).

16. Benítez-Paz, A. & Sanz, Y. Multi-locus and long amplicon sequencing approach to study microbial diversity at species level using the MinION™ portable nanopore sequence. GigaScience 6, 1–12, https://doi.org/10.1093/gigascience/giz043 (2017).

15. Schäfer, J., Weiß, S. & Jäckel, U. Preliminary Validation of a Method Combining Cultivation and Cloning-Based Approaches to Quantify and Identify Culturable Airborne Bacteria. Genome biology 26, 1721–1729, https://doi.org/10.1186/s13059-019-1891-9 (2019).

14. Varza, P. et al. Uniting the classification of cultured and uncultured bacteria and archaea using 16S rRNA gene sequences. Nature reviews. Microbiology 12, 635–645, https://doi.org/10.1038/nrnmicro3330 (2014).

13. Walper, S. A. & Lipp, E. K. Taxonomic annotation errors incorrectly assign the family Pseudoalteromonadaceae to the order Vibrionales.

12. Coster, W., de, D’Hert, S., Schultz, D. T., Cruts, M. & van Broeckhoven, C. NanoPack: visualizing and processing long-read sequencing data for microbial profiling. Available at https://zenodo.org/record/3600229 (2020).

11. Wood, D. E., Lu, J. & Langmead, B. Improved metagenomic analysis with Kraken 2. Genome biology 20, 257, https://doi.org/10.1186/s13059-019-1891-9 (2019).

10. Varza, P. et al. Uniting the classification of cultured and uncultured bacteria and archaea using 16S rRNA gene sequences. Nature reviews. Microbiology 12, 635–645, https://doi.org/10.1038/nrnmicro3330 (2014).

9. Edgar, R. C. Accuracy of taxonomy prediction for 16S rRNA and fungal ITS sequences. PeerJ 6, e4652, https://doi.org/10.7717/peerj.4652 (2018).

8. Lydon, K. A. & Lipp, E. K. Taxonomic annotation errors incorrectly assign the family Pseudoalteromonadaceae to the order Vibrionales in Greengenes: implications for microbial community assessments. PeerJ 6, e5248, https://doi.org/10.7717/peerj.5248 (2018).

7. McDannell, A. R. et al. Comprehensive benchmarking and ensemble approaches for metagenomic classifiers. Genome biology 18, 182, https://doi.org/10.1186/s13059-017-1299-7 (2017).

6. Wilkinson, M. D. et al. The FAIR Guiding Principles for scientific data management and stewardship. Scientific data 3, 160018, https://doi.org/10.1038/sdata.2016.18 (2016).

5. Leidenfrost, R., Pother, D.-C., Jäckel, U. & Wünschiers, R. Nanopore raw signal data to Benchmarking the MinION: Evaluating long-read sequencing for microbial profiling. Available at https://zenodo.org/record/3600229 (2020).

4. Wood, D. E. & Salzberg, S. L. Centrifuge: rapid and sensitive classification of metagenomic sequences. Genome research 26, 1721–1729, https://doi.org/10.1101/gr.120641.116 (2016).

3. Sanderson, N. D. et al. Real-time analysis of nanopore-based metagenomic sequencing from infected orthopaedic devices. BMC genomics 19, 714, https://doi.org/10.1186/s12864-018-5094-y (2018).

2. Tyler, A. D. et al. Evaluation of Oxford Nanopore’s MinION Sequencing Device for Microbial Whole Genome Sequencing Applications. Scientific reports 8, 10931, https://doi.org/10.1038/s41598-018-29334-5 (2018).

1. Landis, J. R. & Koch, G. G. The measurement of observer agreement for categorical data. Biometrics 33, 159–174 (1977).
Author contributions
The conception and design of the study were done by all authors. D.P. and U.J. performed bacteria growth, DNA extraction, ddPCR based DNA quantification and sample preparation. R.L. received the samples, prepared barcoded sequencing libraries, performed sequencing and downstream sequence and statistical analyses. The latter were supervised by R.W. Result interpretation and manuscript preparation were done by all authors. All authors have edited and approved the final manuscript.

Competing interests
The authors declare no competing interests.

Additional information
Supplementary information is available for this paper at https://doi.org/10.1038/s41598-020-61989-x.

Correspondence and requests for materials should be addressed to R.M.L.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher’s note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article’s Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article’s Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit http://creativecommons.org/licenses/by/4.0/.

© The Author(s) 2020