bioRxiv preprint doi: https://doi.org/10.1101/021758; this version posted July 1, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Species level resolution of 16S rRNA gene amplicons sequenced through MinION™ portable nanopore sequencer

Alfonso Benítez-Páez*, Kevin Portune, and Yolanda Sanz

Affiliations:

Microbial Ecology, Nutrition & Health Research Unit. Institute of Agrochemistry and Food Technology Institute, National Research Council (IATA-CSIC). Valencia, Spain.

* Corresponding Author: C. Catedràtic Agustín Escardino, 7. 46980 Paterna-Valencia, Spain. Tel +34 963 900 022 ext 2129. E-mail abenitez@iata.csic.es.

Running Title: Species level resolution of 16S amplicon by using MinION sequencer

bioRxiv preprint doi: https://doi.org/10.1101/021758; this version posted July 1, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Abstract

The miniaturized and portable DNA sequencer MinION™ has been released to the scientific community in the framework of an early access programme to evaluate its scope in a wide variety of genetic approaches. Although this technology is under constant development to completely deliver error-free and high quality reads, it has demonstrated a great potential especially in wide-genome analyses. In this study, we tested the ability of the MinION™ to perform amplicon sequencing in order to design new approaches to study microbial diversity using nearly full-length 16S rDNA sequences.Using the R7.3 chemistry, we have generated more than 3.8 million events (nt) during a sequencing run. Such data was enough to reconstruct more than 90% of the 16S rRNA gene sequences for 20 different species present in the mock community used as a reference. After read mapping and 16S rRNA gene assembly, we could recover consensus sequences useful to make taxonomy assignments at the species level. Additionally, we could measure the relative abundance of all the species present in the mock community by detecting a homogeneous distribution for most of the species as expected.Despite the nanopore-based sequencing produces reads with lower per-base accuracy in comparison with platforms such as Illumina and 454, promising results were obtained from MinION™, indicating that this technology is helpful to perform microbial diversity analysis. With the imminent improvement of the nanopore chemistry, better results and global performance of the platform are expected to contribute to the specific detection of microbial species and strains in complex ecosystems.

Keywords: MinION, nanopore sequencer, 16S rDNAamplicon sequencing, microbial diversity, long-read sequencing

bioRxiv preprint doi: https://doi.org/10.1101/021758; this version posted July 1, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Introduction

The third generation of DNA sequencers is based on single-molecule analysis technology that constantly is under development to deliver error-free and high quality reads. Oxford Nanopore Technologies (ONT) released the first miniaturized and portable DNA sequencer to researchers in early 2014 in the framework of the MinION™ Access Programme. The MinION™ is a USB-stick size device operated from a computer via USB 3.0. Real-time data analysis can be visualized in terms of number of reads and length distribution. Nucleotide basecalling and quality assessment of reads require a further processing where data exchange of Hierarchical Data Format (HDF5) files, containing a large amount of numerical data, is indispensable. This data exchange is done via the Internet through the Metrichor platform, process that can optionally be launched after sequencing process itself. According to its theoretical capabilities, the MinION™ brings out new alternatives for genomic analyses of which the accomplishing of completely finished bacterial genomes is some of the most attractive, as it has been demonstrated recently by Quick and co-workers (Quick et al 2014). The moderate throughput that MinION™ exhibits in terms of number of reads when compared with other massive sequencing platforms is definitively compensated with its performance in terms of read length, thus making it possible to obtain reads of thousands nucleotides in length. Short-read length sequencing approaches have permitted delivering high quality but partial genome sequences with unsolved repetitive elements, thus making it impossible to study genetic variation or molecular evolution directly or indirectly associated to such elements. Therefore, long-read approaches offer new insights into genomic analyses facilitating consecution of finished genomes through hybrid assembly strategies (Utturkar et al 2014). Additionally to genome sequencing analysis, microbial diversity and taxonomy approaches are also deeply

bioRxiv preprint doi: https://doi.org/10.1101/021758; this version posted July 1, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

limited by short-read strategies. Early massive sequencing approaches producing effective 50nt (Genome Analyzer, SolexaIllumina) to 200nt (454 Roche) reads only permitted to accurately explore diversity at the phylum level. However, and thanks to the chemistry improvement of most common sequencing platforms, in recent years dozens of studies have presented a vast inventory of human- or environment-associated microbial communities reaching detail at the family or even genus level. To date, the paired-end short reads approaches for massive sequencing permits the analysis of sequence information of roughly 30% (~500nt) of the 16S rDNA, leaving taxonomic assignment of reads at the species level elusive. Therefore, implementation of long-read sequencing approaches to study the 16S rRNA genes should be determinant to design new studies conducted to evidence central role of precise bacterial species in a great variety of microbial consortia. As a consequence, we present a preliminary study of 16S rDNA amplicon sequencing from a mock microbial community composed of genomic DNA from 20 different bacterial species (BEI Resources) using MinION^{TM} to evaluate the scope of nanopore technology on bacterial diversity and taxonomic analysis performing sequence analysis of the nearly-full length bacterial 16S rRNAgenes.

Methods

Bacterial DNA and 16S rDNA Amplicons

Genomic DNA for the reference mock microbial community was kindly donated by BEI Resources (http://www.beiresources.org). This mock community (HM-782D) is composed of a genomic DNA mix from 20 bacterial strains containing equimolar ribosomal RNA operon counts (100,000 copies per organism per µL) as indicated by the manufacturer. According to BEI Resources instructions, 1 μL of mock community DNA was used for amplification of 16S rDNA genes. A 30 cycles PCR of 95º for 20 seg,

bioRxiv preprint doi: https://doi.org/10.1101/021758; this version posted July 1, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

47ºC for 30 seg, and 72ºC for 60 seg; was setup using with Phusion High-Fidelity Taq Polymerase (Thermo Scientific) and S-D-Bact-0008-c-S-20 and S-D-Bact-1391-a-A-17 primers targeting a high range of bacterial 16S rDNA (Klindworth et al 2012, Loy et al 2007). PCR reactions produced ~1.5Kbp blunt-end fragments, which were purified using Illustra GFX PCR DNA and Gel Band Purification Kit (GE Healthcare). Amplicon DNA was quantified using a Qubit 3.0 fluorometer (Life Technologies).

Amplicon DNA library preparation

The Genomic DNA Sequencing Kit SQK-MAP-005 was used to prepare the amplicon library to be loaded in MinION^{TM}. Approximately, 250 ng of amplicon DNA (0.25 pmol) was processed for end-repair using NEBNext End Repair Module (New England Biolabs) followed by purification using Agencourt AMPure XP beads (Beckman Coulter). Subsequently, we used 200 ng of the purified amplicon DNA (~0.2 pmol) to perform dA-tailing using the NEBNextdA-tailing module (New England Biolabs) in a total volume of 30 μl according to the manufacturer's instructions during 15 minutes at 37ºC. To the 30 μl dA-tailed ampliconDNA, 50 μl Blunt/TA ligase master mix (New England Biolabs), 10 μl of Adapter mix, and 2 μl HP adapter were added and the reaction was incubated at 16ºC for 15 minutes. The adaptor-ligated amplicon was recovered by using Dynabeads® His-Tag (Life Technologies) and washing buffer provided with the Genomic DNA Sequencing Kit SQK-MAP-005 (Oxford Nanopore Technologies). Finally, the sample was eluted from the Dynabeads® by adding 25 μl of elution buffer and incubating for 10 minutes before pelleting in a magnetic rack.

Flowcellset-up

bioRxiv preprint doi: https://doi.org/10.1101/021758; this version posted July 1, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

The brand sealed R7.3 flowcell was stored at 4°C until usage. The R7.3 flowcell was fitted to the MinION™ ensuring a good thermal contact with plastic screws. Priming of the R7.3 flow cell was done two times with premixed 71 μl nuclease-free water, 75 μL 2x Running Buffer, and 4 μL Fuel Mix. At least 10 minutes were needed to equilibrate the flowcell before every priming round and final DNA library loading.

Amplicon DNASequencing

The sequencing mix was prepared with 63 μl nuclease-free water, 75 μl 2x Running Buffer, , 8 μL DNA library, and 4μL Fuel Mix. A standard 48-hour sequencing protocol was initiated using the MinKNOW™ v0.50.1.15. Base-calling was performed through data transference using the Metrichor™ agent v2.29.1 and 2D basecalling workflow v1.16. During the sequencing run, one additional freshly diluted aliquot of DNA library was loaded after 12 hours of initial input.

Data analysis

Quality assessment of read data and conversion to fasta format was performed with *poretools*(Quick et al 2014) and *poRe*(Watson et al 2014) packages. Read mapping was performed against the reference 16S ribosomal RNA sequences (accessions NC_009085, NZ_ACYT00000000, NC_003909, NC_009614, NC_009617, NC_001263, NC_017316, NC_000913, NC_000915, NC_008530, NC_003210, NC_003112, NC_006085, NC_002516, NC_007493, NC_010079, NC_004461, NC_004116, NC_004350, and NC_003028) using the LAST aligner v.189 with parameters -q1 -b1 -Q0 -a1 -e45 which were configured to give the best balance between 16S rDNA assembly length and variants. LAST outputs were converted to sam files and processed with *samtools* (Li et al 2009) to build indexed bam files and obtain

bioRxiv preprint doi: https://doi.org/10.1101/021758; this version posted July 1, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

consensus sequences from alignments and variant calling. Read mapping stats from sam files were calculated with the *ea-utils* package and its *sam-stats* function (Aronesty 2011). Different comparisons, GC content correlations, and plots were performed and designed in R v3.2.0 (http://cran.r-project.org). The species coverage was calculated by obtaining fold-change (Log_2) of species-specific read counting against the expected average for the entire community. A coverage bias was assumed when coverage deviation was lower than -1 or higher than 1. The Simpson's reciprocal index was calculated with the general formula, $D = 1 / \Sigma p_i^2$, where p_i is the proportion of reads belonging to the *i*th species.

Results

The raw data collected in this experiment was obtained from MinKNOW software v0.50.1.15 (Oxford Nanopore Technologies) as fast5 files after conversion of electric signals into base calls via Metrichor Agent v2.29 and the 2D Basecalling workflow v1.16. Base called data passing quality control and filtering was downloaded and basic statistics of the experiment's data was assessed with *poretools* (Loman and Quinlan 2014) and *poRe* (Watson et al 2014) packages. Fasta sequences were filtered by size and then mapped against reference 16S rRNA gene sequences using common and publicly available sequence analysis tools (see methods). Mapping stats are depicted in Table 1 and fast5 raw data can be accessed at ENA (European Nucleotide Archive) under project PRJEB8730 (sample ERS760633). Only one data set was generated after a sequencing run of MinION™. After the sequencing process, we obtained 3,404 reads of which 58.5% were "template" reads (1,991), 23.8% were "complement" reads (812), and 17.7% were "2d" reads (601). Read lengths had a wide distribution ranging from reads with 12 nt to more than 50,000 nt in length with median of 1,100 nt. We

bioRxiv preprint doi: https://doi.org/10.1101/021758; this version posted July 1, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

hypothesize that extremely large reads could be products of multiple amplicon ligation. However, when we tried to perform alignment among reference sequences and large reads, we could not detect any matches (data not shown). Accordingly, we performed a filtering step only retaining 97% of the original dataset (3,297 reads) with a size range between 100 to 2,000nt in length for downstream analysis. When we performed initial analysis using separate sets of reads being "template", "complement", and "2d", we observed a detrimental effect of 2d reads in the quality of assembled sequences, thus obtaining a higher number of unnatural variants along several 16S sequences (~14%). This could be explained because more than 86% of 2d reads were considered as low quality, a fact critical for aims addressed in this study and regarding study of sequence variants after assembly to distinguish very close species as those included in the mock community and belonging to the *Streptococcus* (3 species) and *Staphylococcus* (2 species) genus. Given that this effect was even present for some species when the full set of "template", "complement", and "2d" reads were combined and used to accomplish the sequence analyses, , the dataset was finally reduced to contain information from "template" and "complement" reads (2,696 in total). Using reference sequence information for the mock microbial community analyzed, we reconstructed more than 90% of 16S rRNA gene sequences for all organisms included in the mock community (Table1). We observed that even at very low coverage as that retrieved for *Bacteroides vulgatus* 16S rRNA gene (Figure 1), it is possible to reconstruct almost 94% of the entire molecule; therefore, MinION^{TM} sequencing shows no size limitations additional to that associated to the PCR process itself. Indeed, the maximum size of amplicons sequenced in all cases was that expected according to PCR design (Table 1). In terms of coverage, we retrieved a notable lower number of 16S reads than expected from *B. vulgatus* species (Figure 1). We could not define if such lower coverage could be caused

bioRxiv preprint doi: https://doi.org/10.1101/021758; this version posted July 1, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

during PCR amplification even using high coverage primers (Klindworth et al 2012, Loy et al 2007), or if that bias coverage is the result of the sequencing process itself. Despite this, *B. vulgatus* 16S rDNA amplicon sequences were almost fully assembled with low amounts of variants after DNA read alignment and pileup (Table 1). Data produced by MinION™ was further assessed to theoretically calculate the level of diversity observed using the Simpson's reciprocal index. This diversity index was estimated to be 17.785, a value very close of the maximum expected, 20 for the mock community analyzed. Read mapping stats were analyzed in order to further measure the performance of MinION™ sequencing in microbial diversity analysis based on 16S rDNA sequences. GC content of reads produced by MinION™ showed an important and significant correlation (Pearson's $r = 0.47$, p-value ≤ 0.0376) against GC content of reference values (Figure 2A) which indicates that 16S rDNA GC content is fairly well replicated during sequencing. However, we found a 16S rDNA GC content bias to some extent in the reads obtained from MinION™ that in all cases exceeds the GC content of reference (Figure 2A). To test the probable influence of such bias and GC content itself in basecalling accuracy, we performed linear comparisons against mismatch rate, indels rate, and coverage deviation. We observed that coverage deviation (p-value ≤ 0.00003) and mismatch rate (p-value ≤ 0.00004) are influenced by read GC content (Figure 2B and 2C, respectively). In the first case, the influence of GC content on coverage deviation could have minimal effect because 95% of species analyzed show no more than 1-fold deviation. However, with GC bias detected in reads from the MinION™ sequencer, this effect could be magnified, especially in species where GC content is high. On the other hand, we found a strong correlation between GC content of reads and the mismatch rate retrieved from alignments which would insinuate again the GC content as a factor that influences 16S amplicon sequencing in the MinION™ platform.

bioRxiv preprint doi: https://doi.org/10.1101/021758; this version posted July 1, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Conversely, GC content did not appear to profoundly affect indel rate (Figure 2D). The complete assembly of the amplified 16S rRNA gene permitted the quantification of the level of sequence variants in the consensus sequence. These variants were recovered after a pileup of reads against reference sequences and they were variable in number with a median of 8 variants per 16S rRNA gene (Table 1). Such number of nucleotide substitutions means that approximately 0.5% of the 16S rDNA sequence assembled from MinION™ reads retained unnatural genetic variants directly generated from the sequencing process itself, theoretically leaving a bona fide identification and taxonomy assignment of 16S rDNA sequences at species level. In the worst cases where number of variants are meaningfully (~2.3% of the full assembly), like those observed for the *Acinetobacter baumannii* and *Bacillus cereus* (Table 1), direct BLAST comparisons of these assembled 16S rDNA sequences against the non-redundant reference database only matched those sequences with homologue sequences belonging to the same species, respectively. As expected, a homogenous distribution for strand mapping was observed roughly obtaining 50% of reads mapped against the forward strand and 50% of reads mapped against the complement strand, on average (Table 1).

Discussion

The microbial diversity analyses based on 16S rDNA sequencing are frequently used in biomedical research to determine dysbiosis associated in a great variety of gut-related human diseases. Identification of microbial species inhabiting different organs and cavities of human body relies on handling and processing of millions of DNA sequences obtained through the second generation of massive and parallel sequencing methods which still present limitations mainly in terms of DNA read length. Inability to fully determine 16S rDNA sequences during massive sequencing has led to the development

bioRxiv preprint doi: https://doi.org/10.1101/021758; this version posted July 1, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

of multiple algorithms dedicated to theoretically discern microbial species present in samples according to the sequence similarity degree, the Operational Taxonomy Units (OTUs). Despite high accuracy and a constant update of methods used in OTU-based approaches, available algorithms produce no consensus outputs leaving a high degree of uncertainty when the number of theoretical species and their abundance is being subject of study (He et al 2015, Koskinen et al 2014, Schmidt et al 2014a, Schmidt et al 2014b).

The third generation of sequencing methods based on single-molecule technology offers a new fashion to study the microbial diversity and taxonomic composition thanks to overcoming DNA read limitations at the expense of decreasing their throughput. MinION™ is one of these single-molecule methodologies which has demonstrated its capacity in genome sequencing (Ashton et al 2015, Quick et al 2014). Very recent reports have shown application of this technology in medical microbiology by using amplicon sequencing to potentially determine bacterial and viral infections (Kilianski et al 2015, Quick et al 2015). In this study, we have further explored the scope of MinION™ into microbial diversity studies by using amplicon sequencing of nearly full-size 16S rDNA from a mock bacterial community, obtaining a Simpson's reciprocal index diversity index close to expected. Our results indicate that MinION™ per-base accuracy (65-70%) is in concordance with previous results (Kilianski et al 2015, Mikheyev and Tin 2014, Quick et al 2014). We found that sequence coverage was close to expected in most of cases with only one exceptions, *B. vulgatus* (gene GC = 52%) which presented 1.84-fold less of the expected coverage. Although we could not demonstrate the definitive implication of the sequencing process itself in this coverage bias, this effect could be associated with the PCR process despite using "universal" primers with higher coverage among bacteria species during amplicon synthesis

bioRxiv preprint doi: https://doi.org/10.1101/021758; this version posted July 1, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

(Klindworth et al 2012). In any case, such coverage was enough to reconstruct 93% of the 16S rRNA gene with a low proportion of unnatural variants. We observed a general influence of GC content in the mismatch rate but not in the indel rate, suggesting that base miscalling could be associated with the amplicon GC content. Moreover, a slight correlation between the amplicon GC content observed and coverage bias was evidenced, indicating that GC content could be negatively affecting amplicon coverage to some extent. Although MinION™ replicated fairly well the GC content expected for every amplicon sequenced, we observed a slight overrepresentation of GC in all reads obtained. This over-calling of GC bases in 16S rDNA amplicons could additionally influence the issues stated above in a negative manner.

The R7.3 chemistry used in MinION™ allowed the acquisition of reads of moderate quality which were enough to reconstruct more than 90% of the 16S rDNA molecule in all 20 bacterial species analyzed. None of the 20 16S rDNA consensus sequences analyzed showed more than 3% of sequence variation, which can be considered as a threshold for canonical species identification. Therefore, the consensus sequence assembled were useful to get a reliable taxonomic identification at the species level. As expected, unnatural variants were associated with low coverage regions; therefore, increasing the sequencing coverage will reduce drastically the ambiguities at the assembled sequences. Notwithstanding, further analysis could help to understand if some coverage bias might be associated with certain taxonomic groups.

We have obtained promising results regarding the study of microbial communities by using 16S rDNA amplicon sequencing through MinION™ device. Despite the observed modest per-base accuracy of this sequencing platform, we were able to reconstruct

bioRxiv preprint doi: https://doi.org/10.1101/021758; this version posted July 1, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

nearly full-length16S rDNA sequences for 20 different species analyzed from a mock bacterial community while obtaining a modest coverage for some species. To date, MinION™ and nanopore technology have demonstrated a great potential in DNA sequencing allowing one to retrieve bacterial whole genome sequences with a minimum level of variation (Quick et al 2014). With the results presented here, we postulate that the MinION™ platform is a reliable methodology to study diversity of microbial communities permitting: i) a taxonomy identification at the species level through 16S rDNA sequence comparisons; and ii) a quantitative method to determine the relative species abundance. This type of analysis will likely become more accurate over time as the nanopore chemistry is improved in future releases together the implementation of the "What's In My Pot" (WIMP) Metrichor workflow aiming real-time taxonomic identification of sequences by comparison against different bacterial references databases (i.e. NCBI, SILVA, GreenGenes). Accordingly, sequence studies of the entire 16S rDNA molecule could permit a bypass of OTU-based analysis, thus making it feasible to obtain a direct inventory of bacterial species and relative abundance, as well as determine the key players at the species and/or strain level in different microbial communities of interest. Implications of the primary and secondary structure of 16S rDNA amplicons in the MinION™ sequencing performance must be further explored in order to evaluate, minimize, and correct for technical bias regarding quantitative approaches of microbial diversity studies.

Acknowledegements

Authors thank to the European 7th Framework Programme for funding to ABP and KP researchers who were supported by the EC Project no. 613979 (MyNewGut).

bioRxiv preprint doi: https://doi.org/10.1101/021758; this version posted July 1, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Conflict of Interest

ABP is part of the MinION™ Access Program supported by ONT. Sequencing kits used in this research were kindly donated by ONT.

bioRxiv preprint doi: https://doi.org/10.1101/021758; this version posted July 1, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

References

Aronesty E (2011). ea-utils: Command-line tools for processing biological sequencing data. http://code.google.com/p/ea-utils.

Ashton PM, Nair S, Dallman T, Rubino S, Rabsch W, Mwaigwisya S *et al* (2015). MinION nanopore sequencing identifies the position and structure of a bacterial antibiotic resistance island. *Nat Biotechnol* **33:** 296-300.

He Y, Caporaso JG, Jiang XT, Sheng HF, Huse SM, Rideout JR *et al* (2015). Stability of operational taxonomic units: an important but neglected property for analyzing microbial diversity. *Microbiome* **3:** 20.

Kilianski A, Haas JL, Corriveau EJ, Liem AT, Willis KL, Kadavy DR *et al* (2015). Bacterial and viral identification and differentiation by amplicon sequencing on the MinION nanopore sequencer. *Gigascience* **4:** 12.

Klindworth A, Pruesse E, Schweer T, Peplies J, Quast C, Horn M *et al* (2012). Evaluation of general 16S ribosomal RNA gene PCR primers for classical and next-generation sequencing-based diversity studies. *Nucleic Acids Res* **41:** e1.

Koskinen K, Auvinen P, Bjorkroth KJ, Hultman J (2014). Inconsistent Denoising and Clustering Algorithms for Amplicon Sequence Data. *J Comput Biol* **doi:10.1089/cmb.2014.0268**.

Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N *et al* (2009). The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25:** 2078-2079.

Loman NJ, Quinlan AR (2014). Poretools: a toolkit for analyzing nanopore sequence data. *Bioinformatics* **30:** 3399-3401.

Loy A, Maixner F, Wagner M, Horn M (2007). probeBase--an online resource for rRNA-targeted oligonucleotide probes: new features 2007. *Nucleic Acids Res* **35:** D800-804.

Mikheyev AS, Tin MM (2014). A first look at the Oxford Nanopore MinION sequencer. *Mol Ecol Resour* **14:** 1097-1102.

Quick J, Quinlan AR, Loman NJ (2014). A reference bacterial genome dataset generated on the MinION portable single-molecule nanopore sequencer. *Gigascience* **3:** 22.

Quick J, Ashton P, Calus S, Chatt C, Gossain S, Hawker J *et al* (2015). Rapid draft sequencing and real-time nanopore sequencing in a hospital outbreak of Salmonella. *Genome Biol* **16:** 114.

Schmidt TS, Matias Rodrigues JF, von Mering C (2014a). Ecological consistency of SSU rRNA-based operational taxonomic units at a global scale. *PLoS Comput Biol* **10:** e1003594.

bioRxiv preprint doi: https://doi.org/10.1101/021758; this version posted July 1, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Schmidt TS, Matias Rodrigues JF, von Mering C (2014b). Limits to robustness and reproducibility in the demarcation of operational taxonomic units. *Environ Microbiol* **doi:10.1111/1462-2920.12610**.

Utturkar SM, Klingeman DM, Land ML, Schadt CW, Doktycz MJ, Pelletier DA *et al* (2014). Evaluation and validation of de novo and hybrid assembly techniques to derive high-quality genome sequences. *Bioinformatics* **30:** 2709-2716.

Watson M, Thomson M, Risse J, Talbot R, Santoyo-Lopez J, Gharbi K *et al* (2014). poRe: an R package for the visualization and analysis of nanopore sequencing data. *Bioinformatics* **31:** 114-115.

bioRxiv preprint doi: https://doi.org/10.1101/021758; this version posted July 1, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Table 1. Statistics of the mapping process using 16S rDNA reads produced by MinION™.

Organism	Mapped Reads	Mapped Bases	Strand Mapping[a]	Max Length[b]	Mean Length	Variants	Consensus	rRNA gene[c]	Assembled 16S
A. baumannii	98	99,352	0.46:0.54	1,390	1,013	32	1,415	1,529	0.93
A. odontolyticus	79	73,480	0.49:0.51	1,377	930	8	1,407	1,528	0.92
B. cereus	144	151,668	0.46:0.54	1,419	1,053	31	1,415	1,508	0.94
B. vulgatus	33	29,499	0.58:0.42	1,346	893	25	1,403	1,510	0.93
C. beijerinckii	97	99,476	0.46:0.54	1,393	1,025	13	1,408	1,505	0.94
D. radiodurans	73	69,940	0.45:0.55	1,390	958	8	1,398	1,502	0.93
E. faecalis	149	153,581	0.50:0.50	1,398	1,030	8	1,444	1,549	0.93
E. coli	167	181,084	0.45:0.55	1,398	1,084	0	1,434	1,542	0.93
H. pylori	67	62,838	0.46:0.54	1,390	937	11	1,411	1,498	0.94
L. gasseri	123	128,120	0.51:0.49	1,407	1,041	0	1,467	1,579	0.93
L. monocytogenes	139	140,478	0.50:0.50	1,343	1,010	13	1,374	1,486	0.92
N. meningitidis	87	86,916	0.48:0.52	1,390	999	11	1,433	1,544	0.93
P. acnes	75	70,160	0.48:0.52	1,375	935	21	1,401	1,525	0.92
P. aeruginosa	113	120,520	0.55:0.45	1,398	1,066	14	1,425	1,536	0.93
R. sphaeroides	95	89,750	0.52:0.48	1,416	944	5	1,352	1,463	0.92
S. epidermidis	164	177,084	0.51:0.49	1,423	1,079	0	1,443	1,540	0.94
S. aureus	163	179,477	0.51:0.49	1,423	1,101	1	1,435	1,554	0.92
S. agalactiae	156	166,420	0.52:0.48	1,411	1,066	5	1,439	1,551	0.93
S. mutans	196	221,682	0.47:0.53	1,411	1,131	2	1,440	1,552	0.93
S. pneumoniae	154	168,657	0.52:0.48	1,411	1,095	2	1,442	1,560	0.92

a Proportion of reads mapped against the forward and complemented strand, respectively.

b Maximum length of reads mapped.

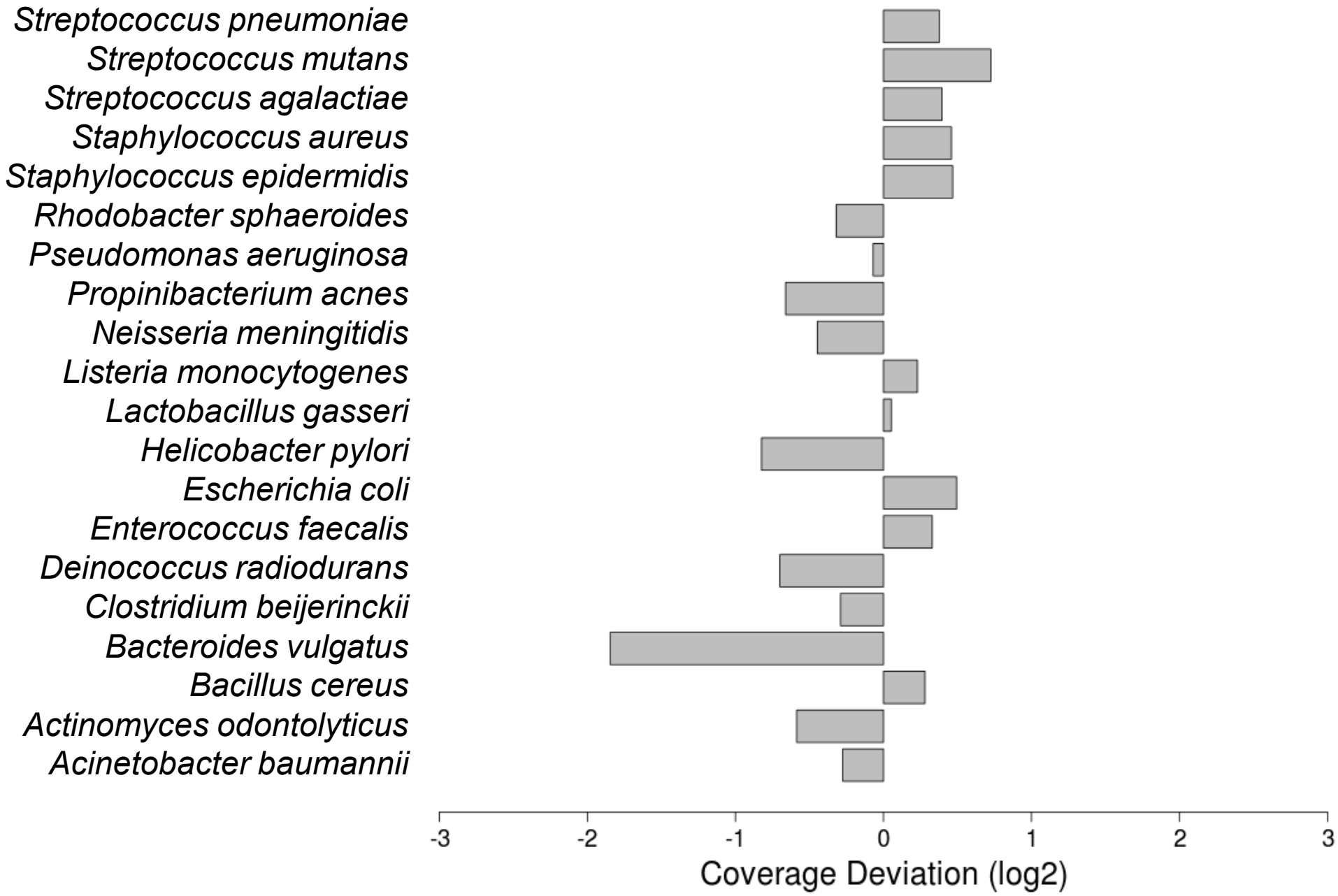
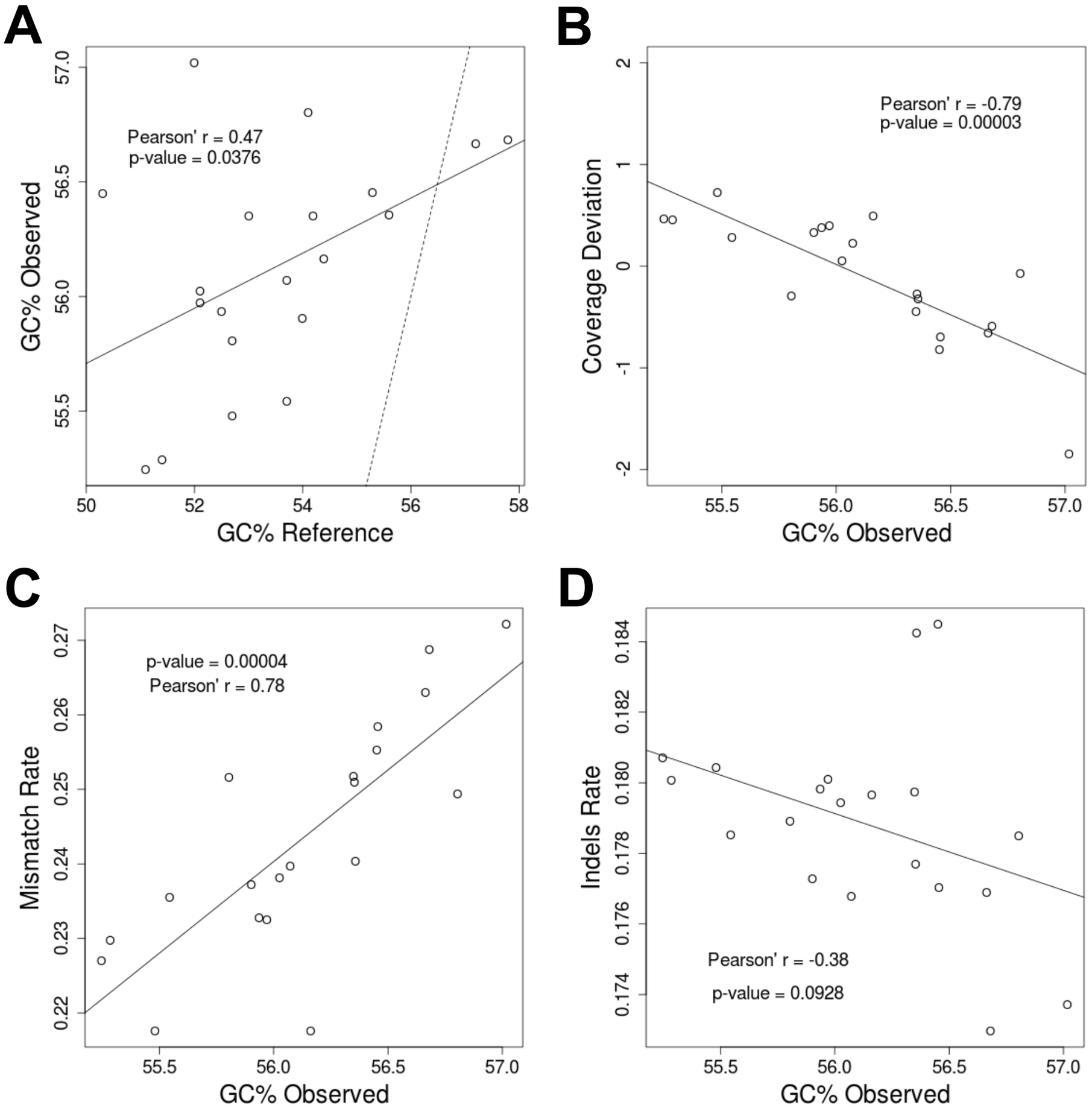
c Length of the 16S sequence used as reference.

bioRxiv preprint doi: https://doi.org/10.1101/021758; this version posted July 1, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Figure Legends

Figure 1. Species abundance in the mock community. Species coverage was calculated by obtaining fold-change (Log_2) of species-specific read counting against the expected average for the entire community. A coverage bias was assumed when coverage deviation was lower than -1 or higher than 1.

Figure 2. Per-base accuracy of the mapped reads. A - Scatter plot of the GC content observed in mapped reads against the GC obtained from the references sequence. The dashed line indicates correlation with a Pearson's r = 1. B - Correlation between GC content observed in mapped reads and coverage bias observed in Figure 1. C - Influence of the GC content observed in mapped reads on mismatch rates calculated after mapping. D - Scatter plot of the observed GC content of mapped reads and indels rates calculated after mapping. In all cases the Pearson's r coefficients and p-values supporting such correlations are presented inside the scatter plots and solid lines indicate the tendency of correlations.


Streptococcus pneumoniae
Streptococcus mutans
Streptococcus agalactiae
Staphylococcus aureus
Staphylococcus epidermidis
Rhodobacter sphaeroides
Pseudomonas aeruginosa
Propinibacterium acnes
Neisseria meningitidis
Listeria monocytogenes
Lactobacillus gasseri
Helicobacter pylori
Escherichia coli
Enterococcus faecalis
Deinococcus radiodurans
Clostridium beijerinckii
Bacteroides vulgatus
Bacillus cereus
Actinomyces odontolyticus
Acinetobacter baumannii
-3
-2
-1
0
1
2
3
Coverage Deviation (log2)

certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under
aCC-BY-NC-ND 4.0 International license.