MinION™ nanopore sequencing of environmental metagenomes

Citation for published version:

Digital Object Identifier (DOI):
10.1093/gigascience/gix007

Link:
Link to publication record in Edinburgh Research Explorer

Document Version:
Publisher's PDF, also known as Version of record

Published In:
GigaScience

Publisher Rights Statement:
© The Author 2017. Published by Oxford University Press. This is an Open Access article distributed under the terms of the Creative Commons Attribution License (http://creativecommons.org/licenses/by/4.0/), which permits unrestricted reuse, distribution, and reproduction in any medium, provided the original work is properly cited.

General rights
Copyright for the publications made accessible via the Edinburgh Research Explorer is retained by the author(s) and / or other copyright owners and it is a condition of accessing these publications that users recognise and abide by the legal requirements associated with these rights.

Take down policy
The University of Edinburgh has made every reasonable effort to ensure that Edinburgh Research Explorer content complies with UK legislation. If you believe that the public display of this file breaches copyright please contact openaccess@ed.ac.uk providing details, and we will remove access to the work immediately and investigate your claim.
MinION™ nanopore sequencing of environmental metagenomes: a synthetic approach

Bonnie L. Brown1,*, Mick Watson2, Samuel S. Minot3, Maria C. Rivera1 and Rima B. Franklin1

1Virginia Commonwealth University, Department of Biology, 1000 W Cary Street, Richmond, VA 23284, USA,
2The Roslin Institute, University of Edinburgh, Division of Genetics and Genomics, Easter Bush, Midlothian,
EH25 9RG, UK and 3One Codex, 165 11th St, San Francisco, CA 94103, USA

*Correspondence address: Bonnie L. Brown, Virginia Commonwealth University, Department of Biology, 1000 W Cary Street, Richmond, VA 23284, USA.
Tel: 804-828-1562; Fax: 804-828-0506; E-mail: blbrown@vcu.edu

Abstract

Background: Environmental metagenomic analysis is typically accomplished by assigning taxonomy and/or function from whole genome sequencing or 16S amplicon sequences. Both of these approaches are limited, however, by read length, among other technical and biological factors. A nanopore-based sequencing platform, MinION™, produces reads that are ≥1 × 10^4 bp in length, potentially providing for more precise assignment, thereby alleviating some of the limitations inherent in determining metagenome composition from short reads. We tested the ability of sequence data produced by MinION (R7.3 flow cells) to correctly assign taxonomy in single bacterial species runs and in three types of low-complexity synthetic communities: a mixture of DNA using equal mass from four species, a community with one relatively rare (1%) and three abundant (33% each) components, and a mixture of genomic DNA from 20 bacterial strains of staggered representation. Taxonomic composition of the low-complexity communities was assessed by analyzing the MinION sequence data with three different bioinformatic approaches: Kraken, MG-RAST, and One Codex. Results: Long read sequences generated from libraries prepared from single strains using the version 5 kit and chemistry, run on the original MinION device, yielded as few as 224 to as many as 3497 bidirectional high-quality (2D) reads with an average overall study length of 6000 bp. For the single-strain analyses, assignment of reads to the correct genus by different methods ranged from 53.1% to 99.5%, assignment to the correct species ranged from 23.9% to 99.5%, and the majority of misassigned reads were to closely related organisms. A synthetic metagenome sequenced with the same setup yielded 714 high quality 2D reads of approximately 5500 bp that were up to 98% correctly assigned to the species level. Synthetic metagenome MinION libraries generated using version 6 kit and chemistry yielded from 899 to 3497 2D reads with lengths averaging 5700 bp with up to 98% assignment accuracy at the species level. The observed community proportions for “equal” and “rare” synthetic libraries were close to the known proportions, deviating from 0.1% to 10% across all tests. For a 20-species mock community with staggered contributions, a sequencing run detected all but 3 species (each included at <0.05% of DNA in the total mixture), 91% of reads were assigned to the correct species, 93% of reads were assigned to the correct genus, and >99% of reads were assigned to the correct family. Conclusions: At the current level of output and sequence quality (just under 4 × 10^3 2D reads for a synthetic metagenome), MinION sequencing followed by Kraken or One Codex analysis has the potential
to provide rapid and accurate metagenomic analysis where the consortium is comprised of a limited number of taxa. Important considerations noted in this study included: high sensitivity of the MinION platform to the quality of input DNA, high variability of sequencing results across libraries and flow cells, and relatively small numbers of 2D reads per analysis limit. Together, these limited detection of very rare components of the microbial consortia, and would likely limit the utility of MinION for the sequencing of high-complexity metagenomic communities where thousands of taxa are expected. Furthermore, the limitations of the currently available data analysis tools suggest there is considerable room for improvement in the analytical approaches for the characterization of microbial communities using long reads. Nevertheless, the fact that the accurate taxonomic assignment of high-quality reads generated by MinION is approaching 99.5% and, in most cases, the inferred community structure mirrors the known proportions of a synthetic mixture warrants further exploration of practical application to environmental metagenomics as the platform continues to develop and improve. With further improvement in sequence throughput and error rate reduction, this platform shows great promise for precise real-time analysis of the composition and structure of more complex microbial communities.

Keywords: MinION™, Oxford Nanopore Technologies; Metagenome; Whole-genome sequencing; Long-read sequencing

Introduction

Environmental metagenomics, employing whole genome sequence analysis to identify ecologically and epidemiologically important components of sediments, soils, waters, and surfaces, is rapidly evolving through advances in both hardware and software [1]. Knowledge of the consortia that inhabit these ecosystems allows for better understanding of the organisms and their ecological roles, provides for the development of effective strategies to mitigate ecosystem damage, and facilitates evaluation of the responses of species to environmental change. One common approach in environmental metagenomics involves sequencing and subsequent annotation of whole genome nucleic acid fragments (whole genome sequencing [WGS]) extracted directly from environmental samples to discover major microbial members of the ecosystem; if sequenced deeply enough, rare species can be detected [2]. For well-studied members of the microbial community, such metagenomic data also can be used to characterize the functional potential of complex communities.

One technique for characterizing environmental metagenomes is to use short-read high-throughput sequencing followed by mapping the reads to reference genomes. Profiling the taxonomic composition of the community also can be accomplished by the analysis of the distribution of k-mers (e.g., using Kraken or One Codex). Although these methodologies are very powerful due to the depth of sequencing, the capacity to resolve the taxonomy of the community to the species level is limited by read length. One approach to overcome this limitation is to assemble short reads into contigs prior to analysis and annotation. If assembled correctly, the longer sequence lengths of the contigs have a greater chance of accurately identifying the members of the community; however, due to the mixed nature of the samples, such assembly approaches are challenged by many artifacts including chimeric contigs that inappropriately combine sequence reads from multiple species. The high information content of very long reads such as those provided by MinION™ (Oxford Nanopore Technologies, Inc., Oxford, UK) has the potential to overcome some of the limitations of short reads by allowing for longer alignments that potentially can contribute to higher taxonomic specificity, functional characterization, and resolution. Although conceived almost two decades ago [3], nanopore-based whole-molecule sequencing has only recently become available to MinION™ Access Programme (MAP) participants for exploration and practical application [4]. Data generated by early access MinION™ flow cells have been assessed for WGS [5–9], gene expression and transcriptome studies [10–12], clinical applications such as inferring antibiotic resistance of bacterial strains and the detection of influenza and Ebola virus [13–15], bacterial and viral serotyping [16], and clinical metagenomes of viral pathogens [17]. Efforts to use this technology to study diverse environmental communities have been limited [18] and there has not been, to our knowledge, any cross-validation of the results or any systematic assessment to determine the best data analysis strategies for nanopore-based environmental metagenomics. To investigate the potential of this platform for broader applications, we performed a set of experiments to quantify the ability of MinION™ long-read sequence data to accurately characterize the taxonomic composition and structure of metagenomes by assessing its performance in the characterization of low complexity synthetic metagenomes.

Data description

The raw MinION data [19] collected during sequencing by MinKNOW software (versions 0.49.2.9 through 0.51.3.40 b201605171140) were immediately uploaded as FAST5 packets to Metrichor Agent (r7.3 2D basecalling, ver rx-2.22-44717-dg-1.6.1-ch-1.6.3; Mk1 2D base-calling, ver WIMP Bacteria k24 for SQK-MAP006), after which base-called data [19] were returned to the host computer, also in the form of FAST5 files. The programs poRe [20], Poretools [21], and NanoOK [22] were used to extract and characterize the numbers of reads and channels, after which only the 2D reads were stored in FASTQ and FASTA files for downstream analyses. The base-called data sets were scrutinized by methods commonly employed in metagenome analysis of short reads including MG-RAST [23], which assigns taxonomy based on predicted proteins and rRNA genes. The data sets also were analyzed by tools that have been shown to work for long-read data including: (1) WIMP [24], which assigns taxonomy by comparing read sequences against a database of bacteria; (2) Kraken [25], which uses exact alignments of k-mers and indexes more than 5000 genomes and plasmids; (3) One Codex [26], which uses exact k-mer alignment to classify sequences against a reference database of ~40,000 complete microbial genomes (including bacteria, viruses, fungi, protists, and archaea); and (4) by principal components analysis (PCA) based on the frequency of 5-mers in each read followed by annotation of reads with the top BlastN [27] hit (carried out in R [28]). Specific parameters are described in Methods.
Table 1: Identity of single-species used in this study as determined by Sanger sequencing of 16S rDNA amplicons from different DNA preparations of each species.

<table>
<thead>
<tr>
<th>Culture *</th>
<th>Final sequence length (bp)</th>
<th>%</th>
<th>Sequence matches in BlastN organism</th>
</tr>
</thead>
<tbody>
<tr>
<td>Escherichia coli</td>
<td>1440–1696 *</td>
<td>98</td>
<td>E. coli numerous strains</td>
</tr>
<tr>
<td>Microcystis aeruginosa</td>
<td>1418</td>
<td>90</td>
<td>M. aeruginosa NIES-843 and NIHEIS-2549, and M. paniformis FACHB-1757</td>
</tr>
<tr>
<td>Pseudomonas fluorescens</td>
<td>1478–1570 *</td>
<td>96</td>
<td>P. fluorescens AS06 and LBUM221</td>
</tr>
<tr>
<td>Synechococcus elongatus</td>
<td>1431–1719 *</td>
<td>99</td>
<td>S. elongatus PCC 7942, PCC 6301, UTEX 2973</td>
</tr>
</tbody>
</table>

*Multiple DNA preparations from bacterial cultures were used during the progress of the study, and each was tested, yielding for each strain slightly different final 16S sequence lengths, but the same BLAST matches.

Results

MinION™ WGS libraries were generated from 1 μg of fresh DNA isolates (see Methods) of separate cultures of two Proteobacteria, Escherichia coli and Pseudomonas fluorescens; and two Cyanobacteria, Microcystis aeruginosa, and Synechococcus elongatus; and from two different DNA mixtures of these four species. One mixture combined an equal mass of genomic DNA (gDNA) from each of the four species. The other mixture was created by combining 33% mass of gDNA from each of three species and only 1% of gDNA mass from the other species. The preparation of these libraries yielded sufficient pre-sequencing Mix for multiple loads of each flow cell. An additional library was derived from a commercially prepared 20-species mock community. Because only 100 ng of material was provided by the supplier, genome preamplification using phi29 polymerase was required to generate sufficient mass of DNA to create the sequencing library (see Methods).

To assess the purity of the cultures used in this study, we used the Sanger method to sequence full-length (~1500 bp) 16S amplicons from each (Table 1). Inspection of those data revealed varying degrees of genomic uniqueness at the species level. For the strain of M. aeruginosa used in this study, the top 16S hit had a low sequence identity to any reference sequence in the database (90%). In contrast, the input strain of S. elongatus was 99% identical to two different species of Synechococcus (S. elongatus and S. UTEx 2973). In addition, whole-genome alignment indicated that the input strain of P. fluorescens was highly similar to multiple species of Pseudomonas. However, all of the input organisms were distinct at the genus level; thus, that taxonomic level was used for downstream analysis of the single-species and ‘equal’ and ‘rare’ synthetic samples.

MinION sequencing of the single-species libraries generated up to 31 × 10^3 reads (0.2–1.1 × 10^3 2D reads that passed the quality filter) ranging from as short as 5 bp to as long as 267 × 10^3 bp (data include both 2D pass and fail reads), and the resulting average length of single-species read subjected to downstream analysis was 6 × 10^3 bp. Using MG-RAST, Kraken, and One Codex, up to 99.5% of the high-quality 2D reads obtained from the sequencing of the single-species libraries of E. coli, P. fluorescens, S. elongatus, and M. aeruginosa were taxonomically assigned to the corresponding input taxa (Table 3). The least accurate assignments were for M. aeruginosa, where at best 58% of 2D reads were correctly assigned to the level of species, although more than one-half of the misassigned reads were to closely related cyanobacteria genera and other prokaryotes known to break down microcystin [29] (data not shown). All three methods of analysis assigned sequence reads of the P. fluorescens single-species library to Stenotrophomonas. Over all of these analyses, MG-RAST generally showed the lowest rate of correct taxonomic assignment and, although One Codex and Kraken provided similar results, Kraken showed a lower rate of correct assignment for M. aeruginosa (85%) compared to One Codex (95%).

In the second round of validation, using three synthetic communities containing mixtures of the previously described species, 6–12 × 10^3 reads (0.7–1.3 × 10^3 2D reads) were generated per run, ranging in length from 0.6 to 56.8 × 10^3 bp (Table 2). For the two communities comprised of equal DNA contribution from four bacteria (25% each species), WGS proportions accurately aligned with the known proportions 87% to 99% of the time when analyzed using Kraken or One Codex and 65% to 85% using MG-RAST (Table 3). Specifically, taxonomic assignment of reads obtained from the sequencing of the equal mixture of four species (25% each) using version 5 chemistry and run on an

---

Table 2: Details of MinION™ WGS output for single-species and synthetic mixtures. Sequencing experiments used the MinION device and new R7.3 flow cells. Libraries were prepared with kit SQK-MAP005 as indicated by (5) and SQK-MAP006 chemistry, indicated by (6). Columns relating to 2D indicate bi-directional reads with quality above Q9.

<table>
<thead>
<tr>
<th>Experiment (chemistry)</th>
<th>Pores with reads</th>
<th>Run time (h) *</th>
<th>Total bp (Mbp)</th>
<th>Total reads</th>
<th>Number of 2D pass reads</th>
<th>Mean 2D read length (bp)</th>
<th>MG-RAST accession</th>
<th>ENA accession</th>
</tr>
</thead>
<tbody>
<tr>
<td>Single species</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>E. coli (5)</td>
<td>430</td>
<td>42</td>
<td>83.6</td>
<td>26 590</td>
<td>1112</td>
<td>5274</td>
<td>4629367.3</td>
<td>ERR1713483</td>
</tr>
<tr>
<td>P. fluorescens (5)</td>
<td>453</td>
<td>48</td>
<td>119.4</td>
<td>25 228</td>
<td>777</td>
<td>7784</td>
<td>4629445.3</td>
<td>ERR1713487</td>
</tr>
<tr>
<td>M. aeruginosa (5)</td>
<td>377</td>
<td>18</td>
<td>40.8</td>
<td>22 760</td>
<td>568</td>
<td>5676</td>
<td>4629369.3</td>
<td>ERR1713486</td>
</tr>
<tr>
<td>S. elongatus (5)</td>
<td>367</td>
<td>23</td>
<td>18.3</td>
<td>6163</td>
<td>224</td>
<td>5101</td>
<td>4629381.3</td>
<td>ERR1713489</td>
</tr>
<tr>
<td>Mixtures</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Equal (6)</td>
<td>129</td>
<td>24</td>
<td>26.5</td>
<td>10 592</td>
<td>714</td>
<td>5527</td>
<td>4614572.3</td>
<td>ERR1713484</td>
</tr>
<tr>
<td>Equal (6)</td>
<td>437</td>
<td>44</td>
<td>77.1</td>
<td>12 174</td>
<td>1348</td>
<td>5202</td>
<td>4685746.3</td>
<td>ERR1713485</td>
</tr>
<tr>
<td>Rare (6)</td>
<td>449</td>
<td>18</td>
<td>39.0</td>
<td>6728</td>
<td>899</td>
<td>6194</td>
<td>4685745.3</td>
<td>ERR1713488</td>
</tr>
<tr>
<td>Staggered (6)</td>
<td>300</td>
<td>33</td>
<td>39.0</td>
<td>14 711</td>
<td>3497</td>
<td>2612</td>
<td>4705090.3</td>
<td>ERR1713490</td>
</tr>
</tbody>
</table>

*Runs were set to either 24 or 48 h and were allowed to continue until either sufficient sequence data were collected or until the 2D pass rate was greatly reduced.
Table 3: Taxonomic assignment accuracy of metagenomic reads across three analysis methods.

<table>
<thead>
<tr>
<th>Experiment</th>
<th>Accuracy of assignment to known genus (%)</th>
<th>MG-RAST</th>
<th>Kraken</th>
<th>One Codex</th>
</tr>
</thead>
<tbody>
<tr>
<td>Single species</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>E. coli (5)</td>
<td>74.4 b</td>
<td>99.5</td>
<td>98.7</td>
<td></td>
</tr>
<tr>
<td>P. fluorescens (5)</td>
<td>84.9 b</td>
<td>84.6 b</td>
<td>84.2 b</td>
<td></td>
</tr>
<tr>
<td>M. aeruginosa (5)</td>
<td>53.1</td>
<td>85.8</td>
<td>95.1</td>
<td></td>
</tr>
<tr>
<td>S. elongatus (5)</td>
<td>87.9</td>
<td>98.1</td>
<td>97.6</td>
<td></td>
</tr>
<tr>
<td>Mixtures</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Equal (5)</td>
<td>65.0 b</td>
<td>97.6</td>
<td>87.4 c</td>
<td></td>
</tr>
<tr>
<td>Equal (6)</td>
<td>85.9</td>
<td>98.0</td>
<td>98.7</td>
<td></td>
</tr>
<tr>
<td>Rare (9)</td>
<td>92.9</td>
<td>99.1</td>
<td>98.7</td>
<td></td>
</tr>
</tbody>
</table>

*15% of reads assigned to Shigella.  
97-15% of reads assigned to Stenotrophomonas.  
8% of reads assigned to Stenotrophomonas.  

Accuracy was calculated as the proportion of reads assigned to the known input organism at the genus level out of the total number reads given any assignment at that rank.

Figure 1: Result of “What’s in my pot” analysis of a mixture with equal DNA mass from four bacterial strains. Rendering of real-time analysis using WIMP to compare the input reads against their own more concise reference databases, providing an assignment for the most likely origin of each individual sequence.

Discussion

Sequencing of whole genome libraries can enhance environmental metagenomic analysis by providing more precise identification of the composition and structure of the community than is possible by amplicon sequencing of marker genes (e.g., 16S) [2, 30]. Typical environmental samples contain tens of thousands to millions of organisms, yet the resulting metagenomes almost certainly underrepresent this diversity and, often due to short-read strategy, the resulting data sets can be confidently assigned only to higher taxonomic levels [31, 32]. One strategy to improve the accuracy of taxonomic assignment is to carefully assemble metagenomic data, which despite the potential for chimeric contig formation has been shown to greatly enhance species call correctness [33]. However, even with enhanced sequencing and bioinformatic strategies, many public database accessions contain sequences that are not innate to the species that was analyzed; these include symbionts, parasites, pathogens, and sequencing linkers/primers/adapters (unbeknownst to those who have accessed the data) that can lead to false discovery rates [34]. Contaminated and misannotated reference sequences can affect environmental metagenome analyses that are derived from short reads to a greater extent than would be expected from analyses based on long reads. Long reads can circumvent these issues [31, 35, 36], so long as much of the genome for each component organism is represented in the sequencing library and there are few errors in the sequences and the reference database. The results reported here allow us to consider the potential utility of MinION long read sequencing and subsequent bioinformatic analysis for shotgun environmental metagenomics.

The primary challenge of microbial metagenomic sequence analysis using long reads is the comparison of input sequences against a large reference database of whole genomes from bacteria, viruses, fungi, etc. Although a number of algorithms have been developed for alignment of long, error-prone reads [37, 38], those sensitive algorithms are not optimized for the challenge of comparison against the large and ever-expanding universe of microbial genomes. The bioinformatic methods used in this analysis, MG-RAST, Kraken, One Codex, and WIMP, each compare the input reads against their own more concise reference databases, providing an assignment for the most likely origin of each individual sequence.

We found that for low complexity synthetic communities, long reads generated by MinION provided sufficiently precise sequence data to assign organisms represented at or above 1%.
Figure 2: PCA of normalized 5-mer frequency (i.e., percentage) within each MinION™ read for a mixture with equal DNA mass from four bacterial strains and a mixture with one rare component. (A) Sequencing run with equal DNA mass from four species. (B) sequencing run with three equally represented (33% DNA mass each) and one rare (1% DNA mass) species included in the DNA pool. None: read had no BlastN hits. Other: read had BlastN hits but not one of the four species included in the mix.

Table 4: Known composition of 20-species mock staggered community compared with analysis results for WIMP and One Codex. “nd”: not detected; “–” indicates that these species are included in the genus sum shown directly above.

<table>
<thead>
<tr>
<th>Organism</th>
<th>Operon count/mL</th>
<th>Quantity pg/mL</th>
<th>% DNA in template</th>
<th>WIMP % species</th>
<th>WIMP % genus</th>
<th>One Codex % species</th>
<th>One Codex % genus</th>
</tr>
</thead>
<tbody>
<tr>
<td>Acinetobacter baumannii</td>
<td>10 000</td>
<td>8.2</td>
<td>0.24</td>
<td>0.14</td>
<td>0.14</td>
<td>0.29</td>
<td>0.29</td>
</tr>
<tr>
<td>Actinomyces odontolyticus</td>
<td>100000</td>
<td>1</td>
<td>0.03</td>
<td>nd</td>
<td>nd</td>
<td>nd</td>
<td>nd</td>
</tr>
<tr>
<td>Bacillus cereus</td>
<td>100 000</td>
<td>45</td>
<td>1.33</td>
<td>0.53</td>
<td>0.53</td>
<td>0.75</td>
<td>0.75</td>
</tr>
<tr>
<td>Bacteroides vulgatus</td>
<td>100000</td>
<td>0.8</td>
<td>0.02</td>
<td>0.1</td>
<td>0.1</td>
<td>0.07</td>
<td>0.12</td>
</tr>
<tr>
<td>Clostridium beijerinckii</td>
<td>100 000</td>
<td>44</td>
<td>1.30</td>
<td>0.19</td>
<td>0.19</td>
<td>0.29</td>
<td>0.29</td>
</tr>
<tr>
<td>Deinococcus radiodurans</td>
<td>100000</td>
<td>1</td>
<td>0.03</td>
<td>0.05</td>
<td>0.05</td>
<td>0.07</td>
<td>0.06</td>
</tr>
<tr>
<td>Enterococcus faecalis</td>
<td>100000</td>
<td>0.7</td>
<td>0.02</td>
<td>nd</td>
<td>nd</td>
<td>nd</td>
<td>nd</td>
</tr>
<tr>
<td>Escherichia coli</td>
<td>1 000 000</td>
<td>680</td>
<td>20.04</td>
<td>45.61</td>
<td>45.66</td>
<td>52.15</td>
<td>52.52</td>
</tr>
<tr>
<td>Helicobacter pylori</td>
<td>10 000</td>
<td>8.6</td>
<td>0.25</td>
<td>1.68</td>
<td>1.68</td>
<td>3.43</td>
<td>2.72</td>
</tr>
<tr>
<td>Lactobacillus gasseri</td>
<td>100000</td>
<td>3.2</td>
<td>0.09</td>
<td>0.14</td>
<td>0.14</td>
<td>0.22</td>
<td>0.23</td>
</tr>
<tr>
<td>Listeria monocytogenes</td>
<td>100000</td>
<td>5</td>
<td>0.15</td>
<td>0.38</td>
<td>0.38</td>
<td>0.58</td>
<td>0.52</td>
</tr>
<tr>
<td>Neisseria meningitidis</td>
<td>100000</td>
<td>5.8</td>
<td>0.17</td>
<td>0.24</td>
<td>0.24</td>
<td>0.44</td>
<td>0.41</td>
</tr>
<tr>
<td>Propionibacterium acnes</td>
<td>10 000</td>
<td>8.8</td>
<td>0.26</td>
<td>0.48</td>
<td>0.48</td>
<td>0.07</td>
<td>0.64</td>
</tr>
<tr>
<td>Pseudomonas aeruginosa</td>
<td>100 000</td>
<td>160</td>
<td>4.71</td>
<td>1.25</td>
<td>1.25</td>
<td>3.07</td>
<td>3.18</td>
</tr>
<tr>
<td>Rhodobacter sphaeroides</td>
<td>1 000 000</td>
<td>1,400</td>
<td>41.25</td>
<td>1.01</td>
<td>1.01</td>
<td>1.46</td>
<td>1.27</td>
</tr>
<tr>
<td>Staphylococcus aureus</td>
<td>100000</td>
<td>59</td>
<td>1.74</td>
<td>0.38</td>
<td>0.38</td>
<td>3.88</td>
<td>1.31</td>
</tr>
<tr>
<td>Staphylococcus epidermidis</td>
<td>1 000 000</td>
<td>510</td>
<td>15.03</td>
<td>7.67</td>
<td>7.72</td>
<td>6.65</td>
<td>–</td>
</tr>
<tr>
<td>Streptococcus agalactiae</td>
<td>100000</td>
<td>32</td>
<td>0.94</td>
<td>0.96</td>
<td>0.96</td>
<td>0.95</td>
<td>16.97</td>
</tr>
<tr>
<td>Streptococcus mutans</td>
<td>1 000 000</td>
<td>420</td>
<td>12.38</td>
<td>10.17</td>
<td>10.17</td>
<td>19.50</td>
<td>–</td>
</tr>
<tr>
<td>Streptococcus pneumoniae</td>
<td>100000</td>
<td>0.6</td>
<td>0.02</td>
<td>nd</td>
<td>nd</td>
<td>nd</td>
<td>–</td>
</tr>
<tr>
<td>Other</td>
<td>0</td>
<td>0</td>
<td>29.02</td>
<td>70.98</td>
<td>74.63</td>
<td>91.23</td>
<td>92.76</td>
</tr>
</tbody>
</table>

*Theoretical copy number provided by BEI Resources certificate of analysis.

Proportion of individual species within the mock community.

Of these, 12.7% were correctly assigned to genus, 86.4% were Enterobacteriaceae, and only 0.7% were misclassifications. Of these, 86.4% were Enterobacteriaceae and only 0.7% were misclassified.

Overall, MG-RAST provided the lowest level of accuracy and detected multiple organisms that were not a part of the known input set. This is not surprising given that MG-RAST is optimized for analyzing short-read, low-error data. Kraken and One Codex performed similarly for the single-species samples except in the case of M. aeruginosa, in which case One Codex correctly identified this taxon at a higher rate than Kraken (95% vs 85%). For the equal mixture with the version 5 chemistry, Kraken showed a higher rate of correct assignment than One Codex (97.6% vs 87.4%), although the two methods were generally comparable (actually One Codex was slightly more accurate) for the equal mixture when using version 6 of the
MinION chemistry. An unexpected finding of this study was the detection by all three methods of *Stenotrophomonas* in the *P. fluorescens* single-species sample. Interestingly, *Stenotrophomonas* was classified as *Pseudomonas* when it was first discovered, based on similar metabolic capabilities, and was later moved to its own genus based on molecular data [39]. Our 16S sequences derived from laboratory cultures used in this study did not identify *Stenotrophomonas*, suggesting that its identification in the mixed metagenomes is not a result due to a contaminant but rather, an artifact caused by assigning taxonomy to reads with multiple sequencing errors. Also contributing to its identification is the fact that both *Pseudomonas* and *Stenotrophomonas* share functional phenotypic characteristics, indicating they may share homologous genes coding for those characteristics. The sharing of homologous genes, similar GC contents (both species genomes have 66% GC), and the higher error rate are the most likely factors responsible for the assignment of *Pseudomonas* sequence reads to *Stenotrophomonas*.

The fact that the estimated proportions of community members in synthetic mixtures were not precise despite careful DNA quantitation could indicate differences across library preparation (all libraries were prepared by BLB), reagent kits, flow cells, MinKNOW control scripts, the quality of DNAs used to create the synthetic metagenomes, and the methods used for quantification (Qubit for the home-grown mixtures and UV spectrophotometry for the 20-species mixture). Because DNA quality is of paramount importance for MinION sequencing, PreCR (used in the version 5 protocol) or FFPE Repair Mix (used in the version 6 protocol) was included in the preparation of all libraries. The potential for profound effects related to library preparation recently was examined by Jones and collaborators [30], leading to the recommendation that studies of complex metagenomes should be based on PCR-free approaches. The current data indicate that the MinION lends itself well to a PCR-free approach, but its utility for the analysis of complex metagenomes is presently limited by the small number of reads that pass the quality filtering process. The current study also provides data for considering alternatives to PCR for amplification, in this case GenomiPhi™, which was used to generate sufficient DNA for one library in the current study (“Staggered”). This method is optimized for linear DNA and was intended to generate unbiased copies of the 20-species genomes. Nevertheless, the φ29 preamplification step is one possible reason for the overrepresentation of *E. coli* and underrepresentation of *R. sphaeroides* in the sequencing of the 20-species mock community. Also, a consequence of φ29 preamplification combined with putative differences in DNA quality, chimeric amplicons (known to occur with φ29 amplification of microbial communities [40]) could have been formed predominantly from higher quality *E. coli* DNA repriming itself [41], leading to overrepresentation of the *E. coli* component. Notably, a novel low input DNA approach recently reported [42] could enhance MinION analyses of samples with low DNA yields. Although the preamplification step is the most likely culprit, an additional effect that could contribute to incongruence of known and estimated proportions in the 20-species mock community is that organisms for which there are many accessions in the public databases provide for more precise classification (e.g., NCBI has more than $6 \times 10^5$ *E. coli* complete genome accessions) and that vice versa, organisms with relatively few accessions (e.g., NCBI has only 116 *R. sphaeroides* complete genome accessions) result in less precise classification.

Despite the rather small number of 2D reads that were observed to pass the quality filter across all MinION runs, there was a strong biological signal in the data (Fig. 2). Thus, as investigators have found MinION useful for single genome introspection [6, 9, 15], 16S and other amplicon resolution [16, 43], cDNA
sequencing [11], and assembly [5, 44, 45], our findings imply that this platform has immediate utility for analysis of very simple mixtures (e.g., serum testing for pathogens). Over the 18-month period of MinION use for this set of experiments, 2D pass rates increased from 2% to 24%. Because the rate of improvement is concurrent with Moore’s Law [46], we speculate that future improvements will make the MinION platform very useful in the analysis of complex metagenomic samples in the near future. The cloud-based WIMP base-calling and taxon prediction program associated with the device provides a method of real-time analysis of metagenomic data. However, because we had no control over the comparative database, the cloud implementation of WIMP was less flexible for environmental metagenomic analysis than Kraken or One Codex, and we note that use of an incomplete database can lead to false positives and negatives. By the time of submission of this study, the R7.3 flow cells and sequencing chemistry were no longer available. Subsequent versions of the platform have shown dramatically lower error and higher throughput. This study nevertheless provides a baseline for considering nanopore metagenomes and provides an impetus for further development of MinION output and data analysis, specifically with regard to evaluation of the informative value of 1D reads, scrutiny of reference data, alternative alignment algorithms, and more sophisticated k-mer analyses. As the quality rate for this platform improves, the potential will increase for MinION to accurately resolve the diversity and composition of many of the taxa in an environmental metagenome.

Methods

To set a baseline of expectations for MinION metagenomic analysis, we performed single-species sequencing runs with four organisms. Cell cultures at log phase were harvested by spinning 15-mL culture tubes at 3000 × g for 30 min, and DNA was isolated using the PowerSoil DNA kit (MoBio, Carlsbad, CA, USA) according to the manufacturer’s instructions. Nucleic acid quality and quantity were checked via Nanodrop 2000 and Qubit, whereafter 1 μg of DNA was used to prepare sequencing libraries. For the first two mixtures, equal portions of DNAs from all four organisms (250 ng each) were used (‘equal’) and, for the third mixture (‘rare’), equivalent amounts of three of the species were used (330 ng each) and M. aeruginosa was included as only 1% of the mixture (10 ng). An additional preparation of a mock community containing DNA of 20 bacterial species in staggered amounts was obtained from a commercial source (Catalog # HM-783D, BEI Resources, ATCC, Manassas, VA, USA). This mock community preparation was chosen because it previously has been used to test the ability of the R7.3 version MinION to study microbial diversity via 16S amplicon approach [43]. However, because sequencing libraries for this study required 1 μg of DNA to generate sufficient starting material, 1 μL of the mock community sample (5.5 ng of template, the amount recommended by the supplier for a typical reaction) was preamplified using Φ29 enzyme from the GenomiPhi V3 kit (25-6601-24, GE Healthcare Bio-Sciences, Pittsburgh, PA, USA) according to the manufacturer’s recommendations. This version of Φ29 enzyme was chosen for isothermal preamplification due to the high-fidelity proof-reading aspects of its replication process [47].

The composition of each microbial mixture was calculated on the basis of the relative DNA mass contributed from each organism. Due to the random nature of shotgun sequencing, this library construction strategy is expected to result in a relative proportion of reads sequenced from each organism that corresponds to the relative input mass. In other words, the relative genome size of each organism should not have impacted the relative proportion of reads recovered from each organism.

Sequencing libraries were prepared for R7.3 flow cells run on an original MinION device using the Genomic DNA Sequencing Kit SQK–MAP005 (version 5 chemistry) according to the base protocol from Oxford Nanopore with slight modifications [48] and for flow cells run using the Nanopore Sequencing Kit SQK–MAP006 (version 6 chemistry) according to the manufacturer’s recommendations. The steps for library SQK–MAP005 preparation included in this order: shearing 1 μg in a Covaris g-TUBE (Covaris, Inc., Woburn, MA, USA) at 2000 × g for 2 min, treatment with PreCR (New England Biolabs, Beverly, MA, USA), cleanup with 1 × AMPure beads (Agencourt, Beckman Coulter, Brea CA, USA), end-repair with NEBNext End Repair Module (New England Biolabs), cleanup with 0.5 × AMPure beads, dA-tailing with NEBNext dA-Tailing Module (New England Biolabs), ligation to a cocktail of both the leader and hairpin sequencing adapters (Oxford Nanopore Technologies) using Blunt TA Ligase (New England Biolabs), cleanup using his-tag Dynabeads (Life Technologies, Carlsbad, CA, USA), and recovery of the presequencing mix in 25 μL of Elution Buffer (Oxford Nanopore Technologies). After priming the flow cell with EP solution according to the manufacturer’s recommendations, an initial 6-μL aliquot of the presequencing mix (at 10–20 ng/μL) was combined with 141 μL EP Solution and 3 μL Fuel Mix and applied to the flow cell. Thereafter, at 6- to 8-h intervals, additional presequencing mix aliquots (held on ice) combined with EP Solution and Fuel Mix were added to the flow cell at times roughly coinciding with reprogrammed pore “remux,” which is a process that adjusts the bias voltage and mux channels to maximize yield performance. Modified scripts (J. Tyson, personal communication) caused the MinION device to perform four remux steps at 8-h intervals to maintain regular increases in data (Fig. 4).

Steps for library SQK–MAP006 preparation included in this order: shearing in a Covaris g-TUBE (Covaris, Inc.) at 2000 × g for 2 min, treatment with PreCR (New England Biolabs), cleanup with 1 × AMPure beads (Agencourt, Beckman Coulter), combined end-repair and dA-tailing with NEBNext Ultrall End Repair/dA-Tailing Module (New England Biolabs), cleanup with 1 × AMPure beads, ligation to a cocktail of both the leader and hairpin sequencing adapters (Oxford Nanopore Technologies) using Blunt TA Ligase (New England Biolabs), addition of a tether to the hairpin segment, cleanup using MyOne Streptavidin C1 Beads (Life
Technologies), and recovery of the presequencing mix in 25 μL of Elution Buffer (Oxford Nanopore Technologies). After priming the flow cell with running buffer and fuel according to the manufacturer’s recommendations, an initial 6-μL aliquot of the presequencing mix (at 10–20 ng/μL) was combined with 75 μL Running Buffer, 65 μL water, and 4 μL Fuel Mix and applied to the flow cell. Thereafter, at 8-h intervals, additional presequencing mix aliquots (held on ice) were combined with Running Buffer and Fuel Mix and added to the flow cell at times roughly coinciding with reprogrammed pore remux (modified scripts from J. Tyson, personal communication). Modified remux scripts were not used for the final MinION run (staggered community analysis), because that run was controlled by a new version of MinION.

WGS data (2D FASTQ) from the MinION R7.3 flow cells were accessed on the MG-RAST server [23] and annotated based on their predicted proteins and rRNA genes using the BLAT annotation algorithm [49] against the MSNR protein Db, screened to remove any sequences matching H. sapiens (none found) and without dereplication or dynamic trimming. Although optimized for short read data, the MG-RAST tools were implemented, because they allow query of a suite of comprehensive nonredundant genetic databases and because this server provides a means to share both raw data and computational results. Raw read counts were later accessed from MG-RAST using the API endpoint for organism summaries. The recommended parameters “hit_type = single”, “source = RefSeq”, and “evalue = 15” were used to generate the appropriate read-level abundance information. The same read sets (2D FASTA) also were analyzed by Kraken [25] using the default k-mer size, minimizers, and other parameters, and accessing a local database created from archaea, bacteria, fungi, virus, protozoa, human, and invertebrate genomes. The Kraken tool was implemented, because it is much faster than MG-RAST and allowed use of a smaller, more targeted reference database. The results were translated (kraken-translate) and summarized (kraken-report) to provide full taxonomic names for each classified sequence. Metagenomic analysis using One Codex was performed by uploading the 2D FASTQ data to the One Codex platform at https://app.onecodex.com. This cloud-based k-mer method was selected, because it is reportedly more accurate than either the MG-RAST or the Kraken tools and because like MG-RAST, it provides for community access to the data and analytical results. Because of the high error rate of the R7.3 version MinION nucleotide data, the unfiltered One Codex results were used for this analysis, which do not include an automated error-filtering step. The One Codex read-level classification results were accessed by selecting the “unfiltered” option in the web-based results display and downloading a data table for each sample to generate appropriate read-level abundance information for tabulation.

Comparative data sets were generated for each of the four single species templates using full-length ~1500-bp Sanger sequencing of a 16S amplicon [50]. Reads from the 16S analysis were subjected to BlastN for taxonomic assignment.

Availability of supporting data

The datasets supporting the results of this article are available in the GigaDB repository [19], on the MG-RAST server 4629367.3, 4629445.3, 4629369.3, 4629381.3, 4614572.3, 4685746.3, 4685745.3, 4705090.3, and at the European Nucleotide Archive as primary accessions PRJEB8672 and PRJEB8716. One Codex results are available at https://app.onecodex.com/projects/bb_minion_env.

Abbreviations

2D: refers to sequences where both the template and the complement were completed (bidirectional) and passed the Metrichor quality threshold (Q9); gDNA: genomic DNA isolates from putatively pure cultures of bacterial strains; MAP: MinION™ Access Programme; PCA: principal component analysis; WGS: whole genome sequencing

Availability and requirements

- Project name: Experimental Metagenome on MinION
- Project home page: https://github.com/mw55309/MinION_SynthMetagenome
- Operating system: Unix
- Programming language: Bash and R
- Other requirements: Unix
- License: N/A

Competing interests

BLB, MW, MCR, and RBF are enrolled in the Oxford Nanopore MAP and received free materials for this research. SSM is an employee of One Codex.

Author contributions

BLB conceived of the study, performed the DNA extraction and sequencing, directed the data analysis, and drafted the manuscript. MW provided bioinformatic analyses and statistical analyses. MCR participated in study design, sequence alignment, and bioinformatic analysis. RBF participated in study design, sequencing, data analysis, and manuscript preparation. SSM performed some of the bioinformatic analyses and data interpretation. All authors read and approved the final manuscript.

Acknowledgements

This work was supported by the Virginia Commonwealth University Department of Biology (BLB, MCR, RBF), and by GenEco, LLC, Richmond, Virginia to BLB. Funding for MW was from the Biotechnology and Biological Sciences Research Council including Institute Strategic Programme and National Capability grants (BBSRC; BBS/E/D/20310000, BB/J004243/1, BB/M020037/1). The authors acknowledge M. Kensey Barker (VCU) for assistance with Kraken. Arwyn Edwards and Kevin Keegan provided thorough review and suggestions for improvement of the manuscript. The following reagent was obtained through BEI Resources, NIAID, NIH as part of the Human Microbiome Project: Genomic DNA from Microbial Mock Community B (Staggered, Low Concentration), v5.2L, for 16S rRNA Gene Sequencing, HM-783D.

References

1. Mendoza MLZ, Sicheritz-Ponten T, Gilbert MTP. Environmental genes and genomes: understanding the differences and challenges in the approaches and software for


39. Hugh R. Note: Pseudomonas maltophilia sp. nov., nom. rev. Int