Martin *et al. Genome Biology* (2022) 23:11
https://doi.org/10.1186/s13059-021-02582-x

Genome Biology

METHOD Open Access

Nanopore adaptive sampling: a tool for enrichment of low abundance species in metagenomic samples

Samuel Martin[1†], Darren Heavens[1†], Yuxuan Lan[1], Samuel Horsfield[1], Matthew D. Clark[2] and Richard M. Leggett[1*]

* Correspondence: richard.leggett@earlham.ac.uk
†Samuel Martin and Darren Heavens contributed equally to this work.
[1]Earlham Institute, Norwich Research Park, Norwich NR4 7UZ, UK
Full list of author information is available at the end of the article

Abstract

Adaptive sampling is a method of software-controlled enrichment unique to nanopore sequencing platforms. To test its potential for enrichment of rarer species within metagenomic samples, we create a synthetic mock community and construct sequencing libraries with a range of mean read lengths. Enrichment is up to 13.87-fold for the least abundant species in the longest read length library; factoring in reduced yields from rejecting molecules the calculated efficiency raises this to 4.93-fold. Finally, we introduce a mathematical model of enrichment based on molecule length and relative abundance, whose predictions correlate strongly with mock and complex real-world microbial communities.

Keywords: Nanopore, Adaptive sampling, ReadUntil, Metagenomics, Sequencing, Enrichment

Background

Whole genome shotgun sequencing of metagenomic samples has become a popular tool for understanding communities of mixed species. In particular, the ability to assemble individual species, or gene clusters such as antibiotic resistance genes, has the potential to shed new light on function, or to enable generation of reference sequences for unculturable organisms. With the increasing use of long read technologies, either on their own or combined in hybrid approaches with short-read technologies, metagenome assembled genome (MAG) contiguity and accuracy metrics have improved still further [1]. Such approaches have been applied widely including in the assembly of pathogen genomes from clinical samples [2], bacterial genomes and gene clusters from the human gut [3], the rumen microbiome of cattle [4], and a project which assembled tens of thousands of MAGs by re-analysing over 10,000 previously collected metagenomes [5]. Nevertheless, despite these successes, some doubts remain about the reliability of MAG approaches when faced with complex populations [6].

Metagenomic samples are composed of a range of different species at varying levels of abundance. In nature, abundance often follows a power law [7], and this can mean that sequencing of metagenomic samples produces data that results in deep coverage of some species with low or partial coverage of others. For rarer species, this is likely to result in much poorer assemblies and a reduction in the ability to distinguish between strains or related species. Effective enrichment strategies to maximise the sequence outputs of the rare species would address this weakness and biodiversity blindspot.

Reducing host DNA is an important consideration in diagnostic applications, especially in clinical settings. A number of approaches are available as commercial kits or detailed in published work including differential lysis, saponin-based depletion [8], osmotic lysis [9], or by enriching microbial DNA [10]. However, these approaches are not universally applicable and require sample-specific adaptation often with many additional steps.

Hybridisation has been used effectively to both deplete and enrich samples prior to sequencing with approaches to remove rRNA from total RNA or to enrich for molecules of specific sequence up to a few thousand base pairs long, such as RenSeq [11, 12] proving popular. A common feature of these methods is an extended and often complicated library construction protocol which involves multiple PCR amplification steps that limits the length of DNA that can be interrogated. This can result in amplification biases in the output data (including against longer molecules), and they require highly stringent hybridisation conditions coupled with accurate probe design to be effective.

More recent developments have come with Clone Adapted Template Capture Hybridisation (CATCH-Seq) which was developed to resolve target regions of interest and circumnavigate the need to design specific probes. Using a bacterial artificial chromosome (BAC) known to contain regions of interest, generic probes are generated from the BAC and then used to pull down fragments spanning 60 kbp (single BAC) up to several hundred kbp (multiple BACs) to target difficult to sequence regions and help identify structural variation. Later protocols such as HLS-CATCH [13] and nCATs [14] use Cas-9 nuclease with guide RNAs to target DNA molecules up to several million base pairs in length.

An alternative to lab-based depletion or enrichment approaches is promised by Oxford Nanopore Technologies' (ONT) adaptive sampling concept (sometimes called selective sequencing) which represents a form of software controlled enrichment. A programming interface known as "ReadUntil" enables control over individual pores and provides a mechanism for software to request ejection of the molecule currently being sequenced in a given pore. Thus, the first few hundred bases of a molecule can be examined and a decision made if the molecule is 'on target'. 'Off target' molecules are ejected by reversing the current across the pore, freeing the pore to capture a new molecule. In order for this to be effective, this must happen within a short time, so that the molecule can be ejected from the pore before most of it is sequenced. The longer the time taken for decision making and rejection, the lower the potential levels of enrichment possible.

Initially, ONT provided the ReadUntil programming interface and left it to third party developers to work out how to interrogate the raw pore signal to determine if a molecule was on-target. The first published implementation utilised a signal comparison algorithm known as dynamic time warping (DTW) to compare the signal from the pore with pre-computed signals for sequences of interest [15] (DTW is also used in

DySS [16]). However, this approach was computationally expensive, particularly for anything but relatively short reference sequences. Practical use was therefore limited due to the time required to make a decision when using larger reference databases. An alternative signal-based approach was provided by UNCALLED, which converted stretches of raw signal into k-mers and used higher probability k-mers as a query for a Ferragina-Manzini (FM) index search against a target database [17]. Whilst more efficient than the DTW approach, it still required significant computational resources. RUBRIC [18] abandoned signal-based comparison in favour of basecalling short (~150bp) portions of the start of reads and aligning to reference sequences using LAST [19]. However, this demonstrated limited enrichment and still required significant computational resources. More recently, ONT's provision of real-time GPU-based basecalling on GridION devices enabled the development of Readfish [20], which basecalls the first ~180 bases of sequence and aligns to references with minimap2 [21] in order to make a decision on accepting or rejecting a molecule. These solutions still required third party software in addition to ONT's own control software. From the November 2020 release of the GridION control software, adaptive sampling was built in as a user-selectable option, which has opened it up to much wider adoption. The software requires a user to upload a file of reference sequences and the system can be set to either deplete or enrich for these on a specified set of channels. In order to achieve this, the software basecalls the first few hundred bases of each read and compares it with the target reference sequences. Matching or unmatching sequences are rejected, depending on whether the software is set to enrich or deplete. A detailed algorithmic description of the GridION adaptive sampling implementation has not been released by ONT, but a Nanopore Community forum post by an ONT employee (https://community.nanoporetech.com/posts/beta-release-of-adaptive-s-7369) indicates minimap2 is used for read mapping. The implementation is achieved through ONT's ReadUntil interface, which is publicly available at https://github.com/nanoporetech/read_until_api, and is the same interface used by Readfish and other third party tools.

Adaptive sampling offers a potential solution to enrich for species of interest in metagenomic samples. It requires a simple library construction method and samples can be processed within an hour without the need for amplification. However, a challenge for microbiome research is the difficulty of extracting high molecular weight DNA from metagenomic samples. Their unknown nature and the likely presence of both Gram positive, Gram negative, and fungal species have led to the development of protocols such as the three peak extraction protocol where samples undergo three different methods involving either enzymatic, chemical, or physical disruption to try and preserve DNA molecule length and ensure that the DNA faithfully represents all the species present in the sample [22]. This has shown that DNA molecules > 20 kbp can be achieved, but for many scientists analysing microbiomes, bead beating is a necessity for DNA extraction due in part to the limitation of samples, the inability to effectively culture everything present, and, in some cases, the need for rapid diagnostic results. This approach can be completed inside 20 min but typically produces DNA molecules < 10 Kbp in length.

We wanted to investigate the effect of DNA molecule length on the efficiency and efficacy of adaptive sampling to determine its usefulness for both MAG and diagnostic applications. Here, we present a mathematical model which can predict the enrichment levels possible in a metagenomic community given a known relative abundance and read length distribution. Using a synthetic mock community, we demonstrate that the predictions of the model correlate well with observed behaviour and quantify the negative effect on flow cell yields caused by employing adaptive sampling.

Results

A mathematical model of enrichment potential for metagenomic samples

A number of factors will affect the potential enrichment achievable through adaptive sampling. Here, we derive a model to predict the theoretical enrichment of a set of taxa within a metagenomic community, based on the community composition and average read length. In the sections that follow, we show how this compares with real results achieved using the GridION. We begin with the assumptions (sequencing speed is 420 bases/s, the time taken to capture strands is 0.5 s and the response time once a strand is captured is 1.0 s) given in the worked example in the nanopore adaptive sampling information sheet [23].

We can consider two alternative measures of enrichment: enrichment by yield and enrichment by composition. We define enrichment by yield as the ratio of the yields (per unit time) of target sequence (species) with and without adaptive sampling. This measure is likely to be the main consideration for researchers wishing to target particular sequences—if the overall target yield is less during targeted sequencing, then a better strategy would be to perform deeper untargeted sequencing and bioinformatically filter for the sequences of interest. One key factor that affects a nanopore sequencing run's yield is the number of active pores. As the quality of pores before sequencing varies by flowcell, it is difficult to predict the yield of an experiment and compare adaptive sampling experiments between flowcells. Furthermore, the use of protein pores is known to degrade them over time, possibly from the electric potential [24], or from pore blockage [17]; thus, repeated potential flipping from adaptive sampling could further decrease active pores and yield.

Enrichment by composition is the ratio of the relative abundance of target sequence (species) with and without adaptive sampling. This shows us how much the abundance of a given species in a metagenomic sample can be changed simply by employing adaptive sampling. This measure is not affected by yield, so we are able to use it to compare different experiments using different flowcells. By estimating the composition of target sequences in the community, it is then possible to estimate the target yield for a particular experiment design (assuming all flow cells being equally productive). Below, we develop a model to predict enrichment by composition.

Let X be the set of taxa present in a sample, and for a taxon $x \in X$ define x_{ab} to be the abundance of x in terms of bases sequenced. This can be calculated by sequencing the sample without adaptive sampling and calculating the sequence length of all reads that belong to the taxon x. Then, the abundance of x can be given as this length divided by the total sequence length of all reads in the sample. For an experiment in which we enrich for x, let x_{ob} be the abundance of x observed in this experiment, calculated as before. Then, we say that the enrichment factor for x (or simply, the enrichment of x) denoted e_x, is given by

$$e_x = \frac{x_{\mathrm{ob}}}{x_{\mathrm{ab}}}$$

From this, it is clear that the enrichment must be less than x_{ab}^{-1}, since $x_{\mathrm{ob}} \leq 1$ (for example, a taxon at 50% abundance cannot have an enrichment factor greater than 2). However, this fails to take into account the fact that in order to determine whether a read should be accepted or rejected, it has to be sequenced to some extent, and so we propose that the maximum achievable enrichment is in fact lower than this.

For more generality, we will partition X into the taxa to enrich (also called the target taxa), denoted X_{e}, and those not to enrich (which by definition, will be depleted), denoted $X_d = X \backslash X_e$. Following the ONT info sheet [23], we estimate enrichment by the proportion of the total sequencing time that is spent sequencing the target taxa. We assume a constant sequencing rate throughout, denoted by S and in the units of number of bases sequenced per second. Let T be the proportion of sequencing time spent sequencing reads belonging to X_{e} without adaptive sampling and T_e the proportion of time spent sequencing reads belonging to X_{e} with adaptive sampling. Then, since the sequencing rate is assumed constant, we can estimate the enrichment of X_{e} as

$$e_{X_e} = \frac{T_e}{T}$$

To determine the values T_e and T, we will fix the following quantities. Let D be the time taken between a molecule entering a pore and a decision being made on whether it should be accepted or rejected. Let R be the average read size, and let C be the time taken for a pore to capture a new molecule after sequencing a molecule. First, we give an estimate for T. Denote by y the sum of abundances in X_{e}, that is

$$y = \sum_{x \in X_e} x_{ab}$$

Each molecule takes, on average, R/S seconds to pass through the pore, and then a further C seconds until a new molecule is captured. The proportion of molecules that we want to enrich (i.e. to not eject from the pore) is y, so we have

$$T = \frac{y(R/S)}{R/S + C} = \frac{yR}{R + CS} \tag{1}$$

For T_e, we make the following observation. For molecules belonging to X_{e} we still spend R/S seconds sequencing each molecule. For molecules belonging to X_{d}, however, we spend D seconds sequencing each molecule. Thus, the total sequencing time is given by $y(R/S) + (1-y)D + C$, and so

$$T_{\mathrm{e}} = \frac{y(R/S)}{y(R/S) + (1{-}y)D + C} = \frac{yR}{yR + (1{-}y)DS + CS} \tag{2}$$

Taking the quotient of (1) and (2) gives us the formula for enrichment

$$e_{X_e} = \frac{R/S + C}{y(R/S) + (1{-}y)D + C} = \frac{R + CS}{yR + (1{-}y)DS + CS} \tag{3}$$

This formula gives us the enrichment of the whole set X_{e}, but what if we want to determine the enrichment for a single taxon in X_{e}? It is an interesting feature of this model that it predicts enrichment to be the same for each taxon in X_{e}. To see this, note that if we were to determine the enrichment of a single taxon $x \in X_e$; then, in eq. (1) and (2), we would replace y in the numerator with x_{ab}, whilst the denominators remain the same. But then, in taking the quotient in eq. (3), these terms cancel.

If we wish to enrich for a taxon $x \in X$ (so that $X_e = \{x\}$), then we have that $y = x_{ab}$ and eq. (3) becomes

$$e_x = \frac{R + CS}{x_{ab}R + (1 - x_{ab})DS + CS}$$

We can rewrite the denominator as $x_{ab}(R + CS) + (1 - x_{ab})(DS + CS)$. Since $0 < x_{ab} < 1$ and $C, D, S > 0$ we have that $x_{ab}(R + CS) + (1 - x_{ab})(DS + CS) > x_{ab}(R + CS)$, and so

$$e_x = \frac{R + CS}{x_{ab}(R + CS) + (1 - x_{ab})(DS + CS)} < \frac{R + CS}{x_{ab}(R + CS)} = \frac{1}{x_{ab}}$$

Thus, our model predicts that enrichment of a single species will be lower than x_{ab}^{-1}, as discussed at the beginning of this section.

We created a Shiny web application which allows researchers to explore the potential enrichment that may be possible for their experiments. The app can be found at https://sr-martin.shinyapps.io/model_app/ and allows the user to adjust parameters such as the average read length to explore the effect on enrichment potential for species of varying abundance.

Starting relative abundance and molecule length determine the level of enrichment

We created a bacterial mock community consisting of seven species ranging in abundance from just over 1% to around 47% (Table 1) as determined during control sequencing runs. In order to observe the effect of molecule length on enrichment, we performed a series of experiments with different library preparations, each producing a different read length from the same input material (Table 2). For simplicity, we refer to the library by the mean read length generated during control runs; however, this value alone is insufficient to capture the sometimes complex read length distribution of the

Table 1 Relative abundance of the 7 bacteria used in the mock community, as determined from control runs. All were selected from the National Collection of Type Cultures and strain IDs are given. Percentages represent the percentage of sequenced bp aligned to reference genomes. Read counts give similar percentages and can be found in the spreadsheet Additional file 1

Species	NCTC ID	1.7kbp (%)	4.7kbp (%)	12.8kbp (%)	10.6kbp high to low (%)	10.6kbp low to high (%)
Achromobacter xylosoxidans	10807	36.98	36.72	42.67	30.13	28.99
Morganella morganii	235	37.29	34.68	32.05	38.30	37.68
Leminorella richardii	12151	12.07	11.89	11.25	7.50	7.19
Moellerella wisconsensis	12132	4.36	5.82	4.39	11.49	11.04
Pseudomonas aeruginosa	10332	5.23	5.32	5.91	4.39	4.79
Proteus vulgaris	13145	2.69	4.03	2.51	5.92	6.22
Streptococcus dysgalactiae	5370	1.34	1.55	1.19	2.20	2.58
Unmapped		0.03	0.01	0.04	0.07	1.50

Table 2 Read statistics for control runs for each library

Library	Control run reads	Median	Mean	N50
1.7kbp	213,035	1300	1696.5	2441
4.7kbp	47,772	4718	4686.7	6011
12.8kbp	17,086	6739	12,767.7	26,160
10.6kbp (High to Low)	32,345	2495	10,581.3	41,464
10.6kbp (Low to High)	41,864	2651	9845.5	35,428

library (Figs. 1a, b). For each library, we performed a control run in which we sequenced for approximately 1 h (enough for at least 17,086 reads, and averaging 70,420). We then enabled Adaptive Sampling and enriched for each bacterial genome, one by one, starting with the most abundant species and ending with the least abundant. For the library with a mean of 10.6kbp, we performed an additional "Low to High" run, in which the bacteria were enriched in reverse order, lowest abundance first. For both of these runs, we maintained half the pores as control pores throughout; for all other runs, we did not maintain control pores after the initial control run.

We calculated the enrichment by composition by dividing the relative abundance of a species with enrichment by the relative abundance without enrichment. As predicted by the model, the enrichment factor was higher for less abundant species, and for longer read lengths (Fig. 1c). Highest levels of enrichment were produced for *S. dysgalactiae*, with relative abundance changed from 1.19 to 16.52%. The effect on community composition can be seen in Fig. 1d.

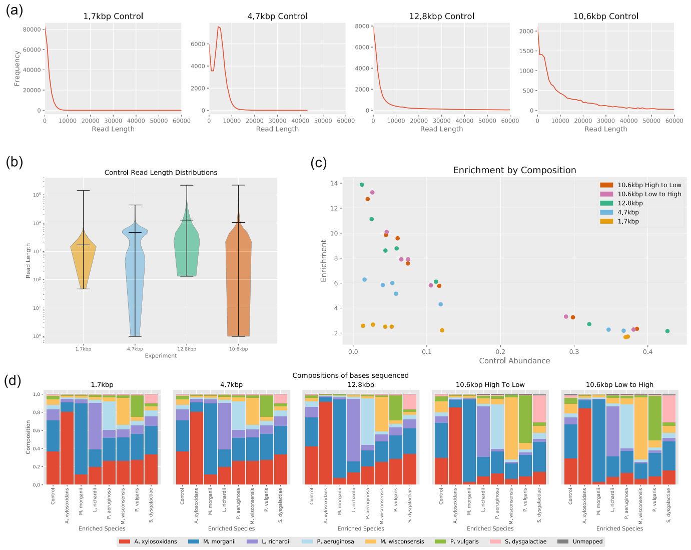

Fig. 1 **a** Read length distributions from control runs. Reads were binned by length into bins of size 1000 bp. Distribution for 10.6 kbp library taken from control run in high-to-low experiment. **b** Violin plots (log scale) of read length distributions from control runs. Distribution for 10.6 kbp library taken from control run in high-to-low experiment. Extrema and means shown in black. **c** Enrichment factor against relative abundance. Each point represents a species, with the position on the *x*-axis indicating the original relative abundance of the species and the position on the *y*-axis indicating the enrichment factor obtained. **d** Community composition for each enrichment target during the runs

Enrichment by composition approaches the maximum predicted by the model

We compared the model predictions with the results of the mock community runs. For the 1.7kbp, 4.7kbp, and 12.8kbp runs, we calculated enrichment by composition as the quotient of the abundance during enrichment and the abundance during the control run. For the 10.6kbp runs, enrichment by composition was calculated as the quotient of the abundance on enrichment channels (1–256) and abundance on control channels (257–512) for each species in the mock community. Following the ONT info sheet [23], we used the fixed values S = 400bps (bases per second), C = 0.5s, and D = 1.0s. Figure 2a overlays results from the experimental runs with predictions from the model.

We calculated the root-mean-square deviation of each data set from the values predicted by the corresponding model (Table 3). Our model predictions also correlated strongly with our observations (Pearson's r = 0.9825) as can be seen in Fig. 2b.

Enrichment by yield is significantly less than enrichment by composition

For each run, we also calculated enrichment by yield. For the 1.7kbp, 4.7kbp, and 12.8kbp runs, we calculated enrichment by yield as the quotient of the yield per hour per active channel during enrichment and the yield per hour per active channel during the control run. For the 10.6 kbp runs, enrichment by yield was calculated as the quotient of the yield per hour per active channel on enrichment channels (1–256) and yield per hour per active channel on control channels (257–512), for each species in the mock community. For the 1.7kbp run, yield of target sequences was slightly lower during adaptive sampling than during the control run (Fig. 3a). Normalising the yield by the number of active channels during the first 30 min of each experiment confirms this (Fig. 3c). For the 1.7kbp and 4.7kbp runs, we performed another control experiment after the adaptive sampling. Figure 3a–c indicate significantly reduced yield after adaptive sampling than the yield before adaptive sampling, particularly for the 1.7 kbp run.

Figure 3d summarises the levels of enrichment found for all bacteria in all runs. Highest enrichment of 4.93x was found for *P. aeruginosa* in the 10.6 kbp low to high run.

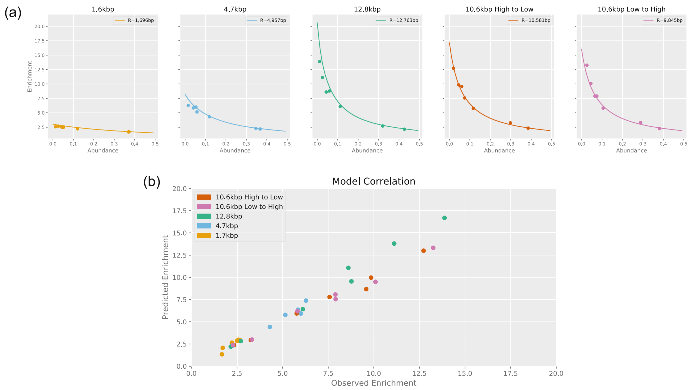

Fig. 2 a Scatterplots of enrichment vs abundance. Curves show enrichment values predicted by the model for average read lengths. **b** Correlation between observed enrichment values and predicted enrichment (Pearson's r of 0.9825)

Table 3 Root-mean-square deviation of experiments from model

Run name	Mean read length	Root-mean-square deviation
1.7kbp	1696	0.350
4.7kbp	4957	0.518
12.8kbp	12,763	1.821
10.6kbp high to low	10,581	0.392
10.6kbp low to high	9845	0.333

Reducing false negative identifications and associated pore ejections would significantly increase enrichment by yield

It is apparent from Fig. 2a that the observed enrichment for low abundance species during the 12.8kbp run was less than the model predicted. Figure 4a shows the distribution of read lengths for the control portion of this run.

During adaptive sampling, we expect to see distributions similar to these for species that are being enriched, and a sharp peak around 500 bp for all other species, which are depleted. However, we find that, when a species is being enriched, it also displays a peak around 500 bp, suggesting that target molecules are being rejected (Fig. 4a).

By parsing the logs provided by the GridION, we found that during adaptive sampling, approximately 36% (lowest 24.6%, highest 48.5%) of target molecules were being ejected from the pore. These are molecules that are misclassified as non-targets by the first fast mapping but subsequently classified as targets by post-enrichment alignment (Fig. 4b). We performed further analysis to determine why this was. We split the read sets first by their species classification, and then by the signal sent to the pore when they were sequenced. Thus, each species had a set of sequenced reads (true positives) and ejected reads (false negatives). First, we calculated the average quality score of each read (Fig. 4c). This shows that the average quality of TPs was higher than the average quality of FNs for each run. Next, we took the first 200 bp from each read and used BLAST [26] to map them against the reference genomes. By doing this, we are attempting to use just the sequence data that is available to MinKNOW when it makes a decision during sequencing on whether to sequence the molecule, or eject it from the pore. For each read, we took the mapping of its first 200 bp which had the highest identity and mapped to the correct genome and used these to calculate the average identity (Fig. 4d). We found that the TPs had a higher average identity than the FNs, although in this case the TPs for the 1.7kbp run had a lower identity than the FPs from the 12.8kbp run. To determine whether regions of low genome complexity can affect the FN rate, we mapped all FN reads to their true target genome. A heatmap showing the coverage of each genome by the FN reads is displayed in Fig. 4e and shows no obvious clustering.

Use of adaptive sampling has an effect on active pores but increases target yield and MAG assembly potential

Our experiments demonstrated continued enrichment over 8–9-h sequencing periods, but we wanted to see how the repeated rejection of molecules affected the lifespan of the pores and if enrichment was still worthwhile over longer periods of time. We ran two new flowcells in which we enriched for a low abundant species (*S. dysgalactiae*,

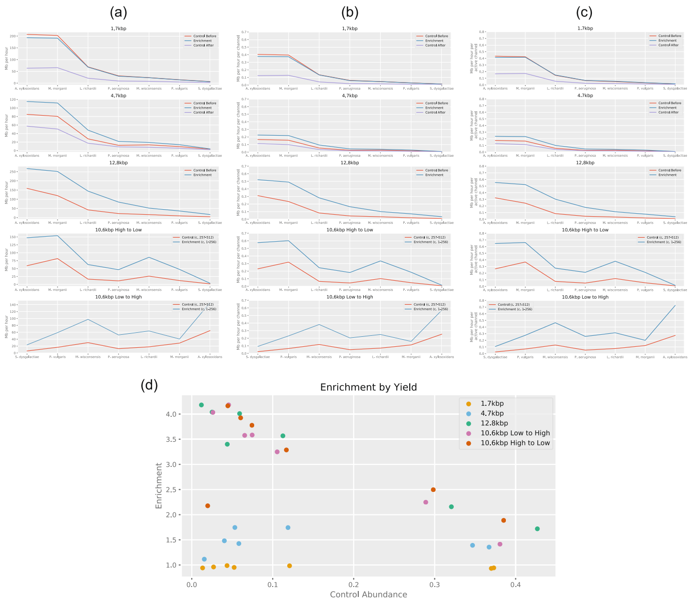

Fig. 3 a Yield of target sequences in Mb per hour during adaptive sampling (blue), control before/during (red), and control after (purple). **b** Yields per hour for all runs, normalised by channels used. **c** Yield of target sequences in Mb per hour per active channel during adaptive sampling. **d** Enrichment by yield values. Each experiment, except for the 1.7 kbp run, gave us increased yield when performing adaptive sampling

~2.6%) and a high abundant species (*M. morganii*, ~37.5%). We also ran two flowcells in which we depleted *M. morganii* (~ 37.5% of total) and depleted *M. morganii*, *A. xylosoxidans*, and *L. richardii* (together ~74.2% of total). With all four flowcells, half the channels were used as control channels in which no adaptive sampling took place. As previously, all four flowcells demonstrated increased yield of the target species but a decreased total yield (Fig. 5). The number of active channels was slightly higher for control channels than for enriched channels, but the difference was not large (Fig. 6b). Hourly yield for target species was consistently higher for the first 24 h with adaptive sampling (Fig. 6c). However, yield of target species declined at a greater proportionate rate on the adaptive sampling channels (down 36% from hour 1 to hour 6) than the control channels (down 25% between hour 1 and hour 6). By 50 h, hourly yield for adaptive sampling was similar to the control channels, but overall flow cell life was much declined by this point, in line with expectations for current nanopore flow cells. Time between target reads was reduced considerably in adaptive sampling channels over the control channels (Fig. 6d).

Reasoning that one mechanism of pore loss is clogging by DNA that cannot be ejected [17], we tested a nuclease flush during a sequencing run for a possible recoverative effect on pores used for adaptive sampling. We ran a new flowcell enriching for a single low abundant species (*S. dysgalactiae*, ~2.6%) for 6 h, carried out a nuclease flush, and then ran the flow cell for a further 6 hours. The flush appeared to result in an increased number of active channels for both the control and enriched portions of the flowcell (Fig. 7a). The effect on hourly yield was less clear (Fig. 7b).

In order to evaluate the effect of adaptive sampling on the potential for MAG assembly of a low abundance species, we took the reads available at 1-h intervals from the control channels and the enriched channels. Reads mapping to the *S. dysgalactiae*

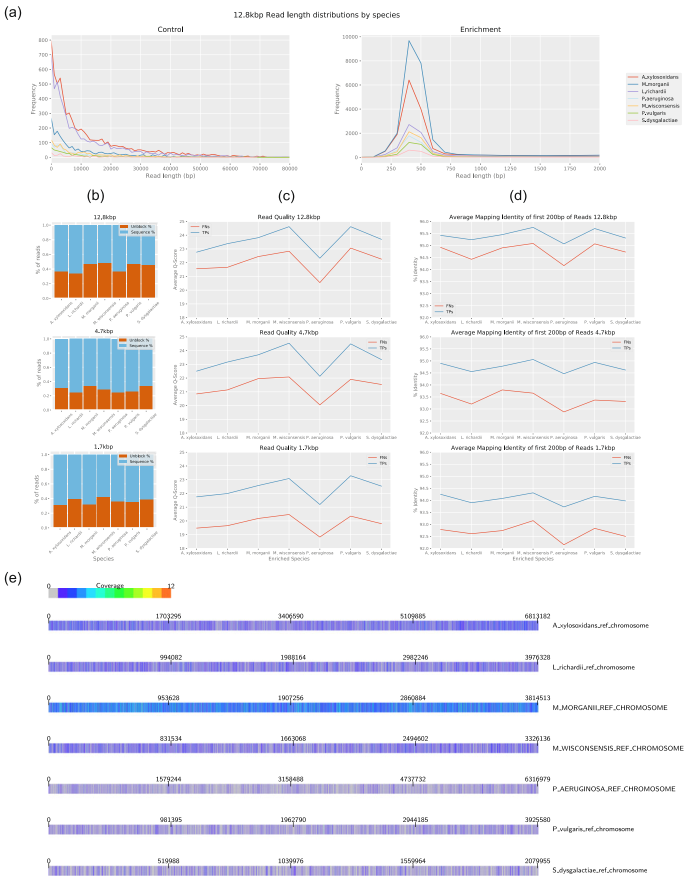

Fig. 4 a Distribution of read lengths during control portion and enrichment portions of 12.8 kbp run. Reads are split by species. **b** Proportion of target reads rejected during adaptive sampling. **c** Quality values of reads, split by species and TP/FN. **d** Average identity of mappings of first 200 bp of reads against reference genomes. The mapping to the correct genome with the highest identity was used to calculate the averages. **e** Coverage of target genomes by false negative reads (i.e. reads that were incorrectly ejected from the pore during adaptive sampling) during 12.8 kbp run. Image produced using the alignment visualisation software Alvis [25]

reference were used as the input to the Flye assembler [27]. For the enriched channels, a single contig, high accuracy assembly was produced with the data available at 2 h (Table 4, Fig. 6a). Subsequently, we also performed an assembly of the enriched channel with reads available at 1.5 h, and this also produced a single contig assembly. For the control channels, after 6 h, the *S. dysgalactiae* yield (32 Mbp) had not yet reached that produced by the enriched channels in 1.5 h (42 Mbp), which was also reflected in much lower contiguity (6 contigs vs 1 contig).

Enrichment using complex and real world communities

The experiments described have shown the effective enrichment of a single species in a mock microbial community. However, real metagenomic samples are often more complex, containing a larger number of species and possibly closely related strains of a single species. We hypothesised that using nanopore adaptive sequencing to enrich a single strain will also result in the enrichment of closely related strains. This could be

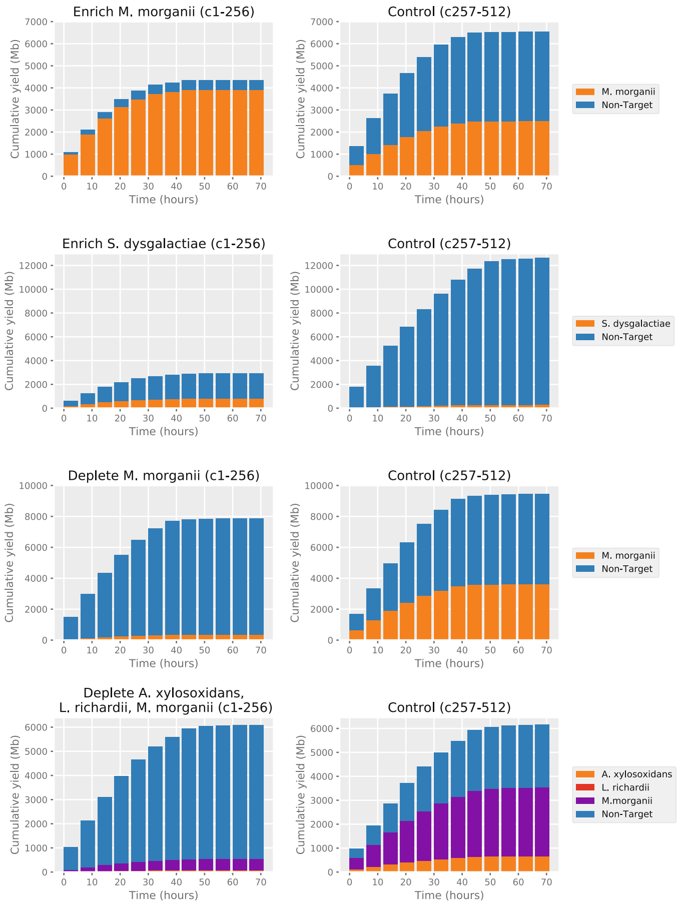

Fig. 5 Cumulative yields split by experiment channels and control channels

undesirable if the experimenter wishes to enrich only a single strain. However, it could be a useful attribute if there does not exist a good quality reference for the strain of interest or if multiple strains are of interest.

To determine the effectiveness of enriching a single strain, we sequenced the ZymoBIOMICS Gut Microbiome Standard, a more complex mock microbial community that includes 5 strains of *E. coli* (Table 5). We selected several species present in the mock at varying abundances, including all of the *E. coli* strains, and enriched each in turn as before. Each experiment lasted for approximately 1 h. Channels 1–256 were used for adaptive sequencing, whilst channels 257–512 were left without enrichment to provide control data. We also performed a 1-h control run at the beginning of the experiment.

Our control run gave us 106,955 reads with a total yield of 460.67 Mbp. The mean read length was 4.31 kbp, and the read N50 was 13.56 kbp. This mean read length is relatively low for adaptive sampling, so it was not expected to produce the highest levels of enrichment. However, individual taxa had varying mean read lengths (Table 5). The taxa chosen for enrichment were *A. muciniphila*, *P. corporis*, *S. enterica*, and each strain of *E. coli*. All were selected without knowledge of the taxa mean read lengths.

Figure 8 shows the observed enrichment for the selected taxa. Enrichment by composition was less than predicted for the five strains of *E. coli*; however, as Fig. 8c

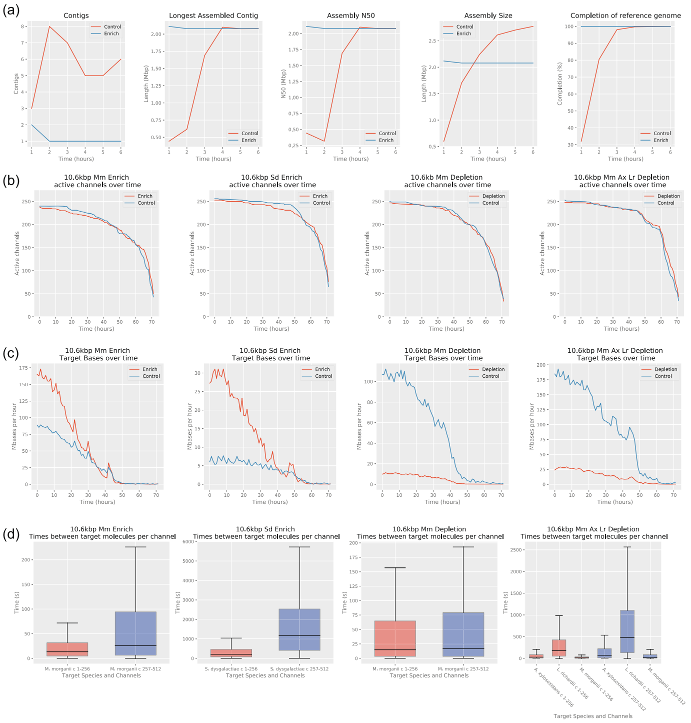

Fig. 6 a S. dysgalactiae assembly statistics for enriched and control channels. **b** Plots showing how the number of active channels varies with time. **c** Hourly yields from enriched/depleted channels vs control channels. **d** Times between consecutive target molecules on individual channels, split by enrich/deplete (channels 1–256, red) and control (channels 257–512, blue)

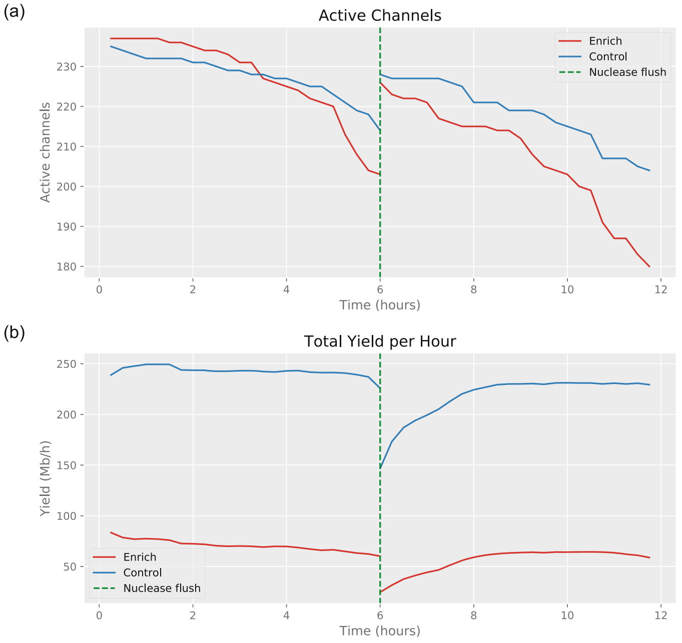

Fig. 7 Effect of nuclease flush on active channels and yield

Table 4 *S. dysgalactiae* assembly statistics for enriched and control channels

Enrich *S. dysgalactiae* (channels 1–256)							
Time (hours)	1	1.5	2	3	4	5	6
Reads	2066	3669	3810	5648	7385	8984	10,467
Total read length	22,620,532	41,656,686	41,901,525	60,708,472	78,890,074	94,518,641	108,992,425
Contigs	2	1	1	1	1	1	1
Total contig length	2,117,919	2,079,393	2,079,362	2,079,466	2,079,521	2,079,533	2,079,512
Contig N50	2,111,978	2,079,393	2,079,362	2,079,466	2,079,521	2,079,533	2,079,512
Longest contig	2,111,978	2,079,393	2,079,362	2,079,466	2,079,521	2,079,533	2,079,512
Aligned bases in ref	99.98%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%
Assembly time	00:05:31	00:09:00	00:08:47	00:11:39	00:14:43	00:17:19	00:19:20
Control (channels 257–512)							
Time (hours)	1	Not performed	2	3	4	5	6
Reads	466		902	1406	1797	2216	2609
Total read length	6,021,205		11,113,179	17,840,666	22,769,997	28,302,443	32,833,051
Contigs	3		8	7	5	5	6
Total contig length	600,309		1,705,413	2,242,365	2,610,844	2,703,245	2,772,594
Contig N50	442,714		315,659	1,691,837	2,100,413	2,077,669	2,079,314
Longest contig	442,714		616,463	1,691,837	2,100,413	2,077,669	2,079,314
Aligned bases in ref	32.05%		80.47%	98.06%	99.73%	99.88%	100.00%
Assembly time	00:02:13		00:03:23	00:04:34	00:05:47	00:06:32	00:07:07

indicates, each time we targeted a specific strain for enrichment, we observed enrichment for all five strains. In each case, except B766, the targeted strain had the highest enrichment. Enrichment for a single strain resulted in high % genome coverage (ranging between 97 and 99%+) for all strains, with the smaller genomes displaying higher coverage (Table 6).

We next sequenced a sample of garden compost. In order to determine a species to attempt enrichment on, the sample was first sequenced without enrichment. Classifying reads with MetaMaps [28] showed an extremely diverse community with 1502 defined genomes represented in 631 genera from the first 100,000 reads. The most abundant species (1.6%) was determined to be *Hydrogenophaga sp. PBC*, a Gram-negative bacterium previously isolated from wastewater [29]. We carried out a metagenomic assembly of all reads using metaFlye and classified these also. From the assembled contigs, 17 mapped to *Hydrogenophaga sp. PBC* and combined had a total length of 3.98 Mb, shorter than the reference sequence at 5.2 Mb (Table 7). We then carried out four runs of approximately 1 h—first a control run, then enrichment for the published reference genome, enrichment for the assembled contigs, and enrichment for both the reference and the assembled contigs. As before, in the enrichment runs, half the channels were used for control and half for enrichment. The control run produced 74,988 reads with a mean length of 6747 bp. Reads were classified with MetaMaps and enrichment calculated (Table 8). Using the published reference produced enrichment by yield of 2.09x. However, by using the assembled contigs, this was increased to 3.72x. This indicates the importance of using an enrichment reference that is extremely close to the genome of the organism being targeted, and an approach for doing so. Our model predicts enrichment by composition of 9.49, given the mean read length and control abundance,

Table 5 Observed composition of gut mock based on MinKNOW mappings. Species chosen for enrichment are highlighted in bold. Due to the large number of contigs (> 13,000) for *C. albicans* and *S. cerevisiae*, these references were not presented to the onboard alignment process, so data is not shown

Species	Manufacturer's theoretical abundance (%)	Observed composition by mapped bases (%)	Mean read length
Faecalibacterium prausnitzii	14	19.31	3428
Veillonella rogosae	14	15.04	14,353
Roseburia hominis	14	10.70	4412
Bacteroides fragilis	14	13.26	10,334
Prevotella corporis	6	6.50	9851
Bifidobacterium adolescentis	6	0.2	6699
Fusobacterium nucleatum	6	5.77	4928
Lactobacillus fermentum	6	7.38	866
Clostridioides difficile	1.5	1.72	8835
Akkermansia muciniphila	1.5	2.12	8765
Methanobrevibacter smithii	0.1	0.03	15,651
Salmonella enterica	0.01	0.03	8198
Enterococcus faecalis	0.0001	0.00	no reads
Clostridium perfringens	0.00001	0.00	no reads
Escherichia coli (JM109)	2.8	2.90	8782
Escherichia coli (B-3008)	2.8	3.44	7034
Escherichia coli (B-2207)	2.8	2.80	5895
Escherichia coli (B-766)	2.8	2.84	9007
Escherichia coli (B-1109)	2.8	3.10	8784
Candida albicans	1.5	n/a	n/a
Saccharomyces cerevisiae	1.4	n/a	n/a

which is slightly higher than the 8.58 obtained experimentally. This difference may be explained by the shorter than expected assembled reference (3.98 vs 5.2 Mb) which may mean target sequence is missing.

Discussion

Previous studies have demonstrated the use of bespoke third party adaptive sampling software to enrich for sequences within an organism, e.g. exons or loci of key variants. Here, we test ONT's own recent implementation of adaptive sampling in the GridION control software as a tool for enriching or depleting species in metagenomic samples. We describe a mathematical model that can predict enrichment potential for a species of given relative abundance and mean read length and show that enrichment by

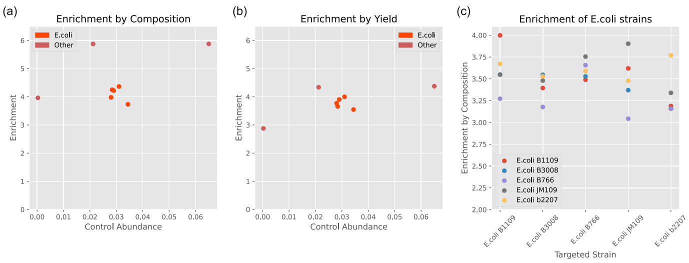

Fig. 8 Enrichment of selected taxa in complex microbial community. Strains of *E. coli* highlighted separately. **a** Enrichment by composition, includes enrichment curve as predicted by model. **b** Enrichment by yield. **c** Enrichment values for each strain of *E. coli* by targeted strain

composition in real experiments closely follows that predicted by the model (Pearson's *r* of 0.9825). Enrichment by yield, the value that is of most practical benefit to researchers, lags behind enrichment by composition, but we show that with longer read lengths, we were able to enrich relatively low abundance (~2%) organisms by almost 5x. High quality single contig MAG assemblies of the same species were possible within 1.5 h using adaptive sequencing and around 6 h without. Allowing for the fact that only half a flow cell was used for each assembly, we may reason these times could be cut in half. Adaptive sequencing could be fed with MAG sequences (from existing short read assemblies), verifying true assemblies, splitting chimeric MAGs [30] (using long nanopore reads as an orthogonal data type), or improving assemblies using long read scaffolding.

In a real, previously unsequenced, complex compost sample, we were able to demonstrate enrichment levels of almost 4x, despite a relatively suboptimal mean read length of 6747 bp. In terms of composition, we observed enrichment of 8.58, slightly lower than our model's prediction of 9.49. Our results demonstrated the importance of using a reference sequence that closely matches the target for enrichment, as better results were obtained by using our meta-assembled contigs instead of the closest published reference. The assembled contigs totalled 77% of the length of the published reference, so it is reasonable to assume that a more complete assembly would have produced greater enrichment still and could explain the difference between the observed enrichment by composition and the predicted enrichment by composition. The importance of the enrichment reference was also illustrated in the sequencing of the gut microbial standard containing 5 strains of *E. coli*. Targeting a specific strain also enriched the 4 other strains and, depending on the application, this could be a beneficial or detrimental

Table 6 Genome coverage (%) of each of the 5 *E. coli* strains. Each strain was enriched for in turn and the coverage of the enriched strain and the non-enriched strains was calculated

Strain enriched for	B1109	B2207	B3008	B766	JM109
Genome Length	4,765,434	5,111,512	4,739,263	5,062,632	4,497,410
B1109 genome coverage (%)	99.4735	98.5886	99.9077	97.6491	99.7461
B2207 genome coverage (%)	98.8860	99.5057	99.9448	98.5786	99.8991
B3008 genome coverage (%)	99.6358	99.1074	99.8742	98.5398	99.5258
B766 genome coverage (%)	99.2041	97.7140	99.8987	98.6976	99.4008
JM109 genome coverage (%)	99.7266	99.5035	99.9852	98.3076	99.8201

Table 7 The top 5 defined genomes determined by MetaMaps classification of the first 100,000 compost reads. Of the first 100,000 reads, 72,638 were long enough (> 1kb) for MetaMaps classification. Also shown are the number of contigs and total size of contigs mapping to the genome

Species	Reads per first 100k (72,638 > 1kb)	Abundance % (based on mapped reads > 1kb)	Number of contigs	Total length of contigs
Hydrogenophaga sp. PBC	1183	1.63	17	3,982,409
Pseudoxanthomonas suwonensis	600	0.83	34	3,552,404
Thauera humireducens	403	0.55	19	2,280,334
Sphingopyxis granuli	330	0.45	19	3,534,465
Devosia sp. A16	307	0.43	9	804,967

attribute. If sequence accuracy continues to increase, strain discrimination by adaptive sampling should too.

Two key factors affect the enrichment potential. Firstly, the initial abundance of the target species determines the theoretical maximum levels of enrichment. Secondly, the length of DNA molecules presented to the sequencing pores limits the efficiency of the process. For shorter molecules, the time taken to basecall, align, reject a molecule, and capture an alternative becomes significant compared to the time taken to sequence a typical read without adaptive sampling. It is for this reason that adaptive sampling of our 1.7 kbp mean library was not beneficial and possibly slightly detrimental to overall yield. Yet for longer reads, there are clear benefits to the adaptive sampling approach. Our data also indicate that there is still much potential to improve on current adaptive sampling implementations and to increase the enrichment by yield significantly. Our analysis showed that around 40% of on-target reads were falsely rejected. Where this value was highest, in the mean 10.6kbp run, we observed slightly lower enrichment by composition values than those predicted by our model and believe these incorrect rejections to be an explanation for the difference. Reducing incorrect ejections could increase target yield significantly and thus result in much higher enrichment factors. ONT provide two implementations of their GPU basecalling algorithm—a faster, less accurate one and a slower, more accurate approach. The adaptive sampling basecalling is performed with the higher speed, lower accuracy algorithm. When hardware progression or algorithmic improvements enable the use of the more accurate basecalling

Table 8 Yields of *Hydrogenophaga* from the enrichment runs. Normalised enrichment is obtained by dividing by the number of active channels during the 1 h run

Enriched for sequence	Control reads > 1kb	Enriched reads > 1kb	Control bp	Enriched bp	Enrichment by yield	Normalised enrichment by yield	Enrichment by composition
Hydrogenophaga reference	522	905	3,423,956	7,173,103	2.09	2.12	4.86
Hydrogenophaga contigs	499	1477	3,514,806	13,059,225	3.72	3.67	8.58
Hydrogenophaga both	537	1490	3,649,735	12,707,539	3.48	3.43	8.13

algorithm, this will likely bring a reduction in false negative ejections. Additionally, improvements to the alignment and decision making approaches employed, as well as to the underlying ReadUntil API, will also bring improvements.

In evaluating the use of adaptive sampling in a particular metagenomic application, a prime consideration will be the ability to prepare DNA that is long enough to derive meaningful enrichment. ONT and a number of users have demonstrated megabase read lengths from genomic samples [31, 32], but it is not possible to imagine such read lengths in complex metagenomic samples due to the need to lyse different cell types, including some that are particularly troublesome to break open. A move away from mechanical lysis approaches such as bead beating towards newer enzymatic techniques will be key. The desire to target particular species with known cell wall characteristics—e.g. for assembly—may mean that harsher lysis approaches are unnecessary for that species. Even with bead beating, we have previously demonstrated DNA extractions from faeces (not the simplest of samples) can produce nanopore data with mean read lengths as high as 8.1kbp [2], and in other experiments, we have generated mean read lengths of up to 15 kbp from soil metagenome samples (Heavens D, unpublished data). Our data indicates that significant enrichment is achievable at these lengths. If input material is not limited, then physical (bead or gel based) size selection can be used to increase the mean read length further. Even the 4.7 kbp run presented here produced enrichment of 2x, which could mean experiments cost half as much money or take half as long to complete. This reduction in time could be particularly important in clinical applications or for environmental pathogen detection.

Previous studies have shown a faster decline in active sequencing channels for flowcells undergoing adaptive sequencing, which may be due to the act of rejecting molecules or that the likelihood of clogging is statistically related to the number of molecules captured by a pore [17, 20]. Our own data shows a slight decline in active channels (Fig. 6b), but not as much as seen previously. Nevertheless, like others we find that overall yield including non-target yield is reduced compared to control channels (Fig. 5), particularly when enriching for lower abundance organisms. We find that a nuclease flush appears to have a restorative effect on active channel count, but the effect on yield was less clear. It is possible that 6 hours was too soon to derive much benefit and others have suggested flushing every 24 h [20].

Overall, our results show that adaptive sampling can increase target yield significantly in real terms, provided that molecules of a modest length are used. Given the strong effect of read (library molecule) length, it is likely that ONT ligation based libraries would outperform rapid based ones from the same material, as the transposase would decrease library molecule size. The use of adaptive sampling provides us with the benefits of library-based enrichment, without complex protocols or the bias that these may introduce. This is a significant advantage to researchers who may not have access to specialised laboratory equipment. Furthermore, it maintains the advantages of nanopore sequencing, e.g. speed, longer read lengths, detecting methylation and other epigenetic modifications using the raw nanopore signal, and the possibility of conducting experiments in-field.

We envisage several applications of adaptive sampling in the near future. One possibility is the targeting of molecules to close gaps in reference genomes. This could be achieved by enriching for molecules that align to sequences flanking gaps in the genome, and depleting everything else. Whilst we demonstrated over 4-fold enrichment in terms of yield, the potential read lengths for metagenomic applications are limited by the variety of DNA extraction methods required for the many cell types that may be present in the sample. Significantly higher average read lengths are possible for non-metagenomic samples, and so the potential for enrichment is greater.

Another possible application of adaptive sampling is the improvement of MAGs. In the “Use of adaptive sampling has an effect on active pores but increases target yield and MAG assembly potential” section, we demonstrated the improved time-to-assembly of a known bacteria using adaptive sampling, and in the “Enrichment using complex and real world communities” section, we demonstrated the enrichment of a species using contigs assembled from the same sample. In the future, we plan to develop a pipeline to assemble metagenomic reads de novo in real time during the experiment. Using adaptive sequencing, we could deplete molecules that cover regions that are already well assembled, or even enrich for reads with sequence at the ends of contigs and pointing into the unknown region, maximising the useful data to improve the assembly. Even rejected reads (~500bp) need not go to waste as these can be used for digital abundance measurements and for polishing assemblies. Existing software such as Readfish [20] already enables the updating of target regions during a run, thereby allowing continual adjustment of targets to refine the assembly as the sequencing progresses. We believe this would lead to improved MAG quality, particularly for low abundance species.

Conclusions

Through ONT’s adaptive sampling software, we demonstrated enrichment in terms of both yield and composition, in a synthetic mock metagenomic community and in a complex real sample. We found that enrichment was higher for lower abundant species, and for libraries with a higher average molecule length, showing that extraction methods that can preserve molecule length are key to obtaining the highest enrichment. We developed a mathematical model to estimate the enrichment by composition that can be expected based on experimental factors and showed that the model’s predictions correlated strongly with the observed data. We also observed that the occurrence of false negatives affects the achieved enrichment, but expect that improvements in hardware and software will minimise this in the future. By performing targeted enrichment on a low abundance species, we were able to significantly reduce the time taken to achieve a high-accuracy, single contig assembly, compared to non-targeted sequencing. Notably, we found that the repeated ejection of molecules from the pores had less effect on pore stability than has been previously reported. We conclude that adaptive sampling will prove to be a useful tool for many nanopore-based metagenomic studies.

Methods

Bacterial cell culture and DNA extraction

Seven bacterial strains were identified from the NCTC that had a fully assembled single chromosome genome, varying GC content and sizes with no plasmids. Full strain details and assemblies available at https://www.sanger.ac.uk/resources/downloads/

bacteria/nctc/. Bacteria were grown overnight in 3ml of 2xYT in a 5ml tube in an Eppendorf Thermomixer C at 37°C shaking at 400 rpm. Following the incubation, the tubes were spun at max speed in an Eppendorf 5427R centrifuge for 5 min to pellet the cells and the supernatant discarded.

For the Gram-positive bacteria, cell pellets were resuspended in 160 μl of Qiagen P1 buffer, transferred to a 1.5ml tube and then 20 μl of 100mg/ml lysozyme added, mixed and incubated for 30 min at 37°C shaking at 900 rpm in an Eppendorf Thermomixer C. To this, 20 μl of proteinase K was added and incubated for 30 min at 56°C shaking at 900 rpm. The tube was then cooled on ice, and 2 μl of RNase added and incubated at room temperature for 2 min.

To precipitate the DNA onto beads, 150 μl of ATL was added followed by 15 μl of MagAttract Suspension G and 280 μl of MB buffer. This was incubated for 3 min at room temperature shaking at 1400rpm. The beads were then pelleted on a magnetic particle concentrator (MPC), the supernatant discarded, and the beads washed twice with 700 μl MW1 buffer and twice with PE buffer resuspending the beads on each occasion.

Two 700 μl water washes were then performed whilst the beads remained on the MPC incubating for 1 min at a time. DNA was then eluted from the beads by mixing the beads for 3 min at room temperature shaking at 1400 rpm in 100 μl AE buffer.

For the Gram-negative bacteria, cell pellets were resuspended in 180 μl of ATL buffer, transferred to a 1.5ml tube, and 20 μl of proteinase K was added and incubated for 30 min at 56°C shaking at 900 rpm. The tube was then cooled on ice, and 2 μl of RNase added and incubated at room temperature for 2 min.

To precipitate the DNA onto beads, 150 μl of ATL was added followed by 15 μl of MagAttract Suspension G and 280 μl of MB buffer. This was incubated for 3 min at room temperature shaking at 1400rpm. The beads were then pelleted on a MPC, the supernatant discarded, and the beads washed twice with 700 μl MW1 buffer and twice with PE buffer resuspending the beads on each occasion.

Two 700 μl water washes were then performed whilst the beads remained on the MPC incubating for 1 min at a time. DNA was then eluted from the beads by mixing the beads for 3 min at room temperature shaking at 1400rpm in 100 μl AE buffer.

DNA extraction from Zymo Gut Mock cells and compost

DNA from the ZymoBIOMICS Gut Microbiome Standard cells (Zymo Research, Irvine, CA, USA) and the compost sample were extracted using the Quick-DNA HMW MagBead kit (Zymo Research). For the Zymo gut mock cells, a 200 μl aliquot was spun at 5000 rcf in an Eppendorf 5427 centrifuge for 1 min to pellet the cells. The supernatant was removed and retained for later. For the compost, 100 mg of material was resuspended in 200 μl DNA/RNA shield (Zymo Research) and then spun at 5000 rcf in an Eppendorf 5427 centrifuge for 1 min to pellet. The supernatant was removed and retained for later.

The cellular material was then resuspended in 200 μl PBS and 10 μl metapolyzyme (Sigma-Aldrich, St. Louis, MO, USA) (1mg/100 μl PBS) added and the cell mixture incubated for 2 h at 35°C. Post this initial incubation, the supernatant saved from earlier was added back along with, 20 μl 10% SDS and 20 μl Proteinase K (20mg/ ml), and this mixture then incubated for a further 30 min at 55°C. The lysed cells were then spun at 5000 rcf in an Eppendorf 5427 centrifuge for 10 min and 400 μl of the supernatant transferred to a new tube. To this, 800 μl of MagBead buffer and 50 μl of MagBeads added and the tube rotated for 10 min at room temperature on a lab rotator. The tube was then briefly spun and then placed on a magnetic particle concentrator to pellet the beads and the supernatant removed and discarded.

The beads were then resuspended in 100 μl elution buffer, 500 μl of MagBead buffer added and mixed gently, and the tube rotated for 10 min at room temperature on a lab rotator. The tube was then briefly spun and then placed on a magnetic particle concentrator to pellet the beads and the supernatant removed and discarded. The beads were then washed with 900 μl Prewash buffer and twice with wash buffer resuspending the beads on each occasion.

A 900 μl elution buffer wash was then performed whilst the beads remained on the MPC incubating for 1 min between adding and removing the buffer taking care not to disturb the bead pellets. DNA was then eluted from the beads by mixing the beads for 10 min at room temperature shaking at 350 rpm in 75 μl elution buffer.

DNA QC

DNA concentration was determined using the Life Technologies Qubit broad range and high sensitivity assay kits. A 1 μl aliquot of DNA was combined with 198 μl of the appropriate buffer and 1 μl of dye in a 0.5ml qubit tube, vortexed and left at room temperature for 2 min. DNA concentration was then measured on a Qubit 3 fluorometer. If DNA concentration between the high sense and broad range assays differed by more than 10%, then the extractions were repeated. DNA was then calculated by averaging the measurement from each assay.

To confirm molecule length extracted DNA was run on either the Agilent Tapestation or Agilent Femto Pulse. For the initial extractions, DNA was diluted, if required, to < 50ng/ μl and a 1 μl aliquot run on an Agilent Genomic Tape on a Tapestation instrument according to the manufacturer's instructions. For the second set of extractions, DNA was diluted to 0.25ng/ μl and a 1 μl aliquot run on an Agilent Femto Pulse instrument according to the manufacturer's instructions. Electropherograms for each bacterial species can be seen in Additional file 2: Figs. S1-S17.

Construction of the synthetic mocks

Two synthetic mocks consisting of all 7 species at 7 different proportions were constructed. For both mocks, we targeted 50% *A. xylosoxidans*, 25% *M. morganii*, 12% *L. richardii*, 6% *P. aeruginosa*, 4% *M. wisconsensis*, 2% *P. vulgaris*, and 1% *S. dysgalactiae* based on average Qubit measurements. The first was used for the 1.7kbp, 4.7kbp, and 12.8kbp runs and the second for the 10.6kbp runs.

To remove smaller molecules and improve average read lengths, a size exclusion step using the Sage Scientific BluePippin was performed. Four 5 μg aliquots of the unfragmented mock were run on a High Pass cassette on a BluePippin to remove molecules < 15 kbp according to the manufacturer's instructions, collecting the size selected material in 40 μl of running buffer.

To target read N50s around 6 kbp, a 5 μl aliquot of the unfragmented mock in 100 μl was placed in a G- tube and spun for 2x 1 min at 10000 rpm in an Eppendorf 5415 centrifuge. To confirm the size of the fragmented DNA, a 1 μl aliquot was run on a Agilent Tapestation genomic tape according to the manufacturer's instructions.

Library construction and sequencing

Libraries for the 4.5 kbp, 12.8kbp size exclusion, 10.6 kbp and 16.4 kbp runs were constructed using the Oxford Nanopore Technologies (ONT) SQK-LSK109 kit according to the manufacturer's instructions except that Kapa beads (Roche, UK) were used to perform the clean up steps rather than Ampure XP beads. To target average sequence reads of 1.7 kbp, 100ng of G-tube fragmented mock was used in a 10 μl reaction using the ONT RAD004 kit according to the manufacturer's instructions.

In all cases, final libraries were sequenced on individual R9.4.1 Rev D 106 flowcells on an ONT GridION.

When targeting successive enrichment of each individual species within the mock, runs were set up with no enrichment for the first hour to ascertain their baseline composition. At the end of the hour, the run was stopped and restarted enriching for the next target genome. This process was repeated and sequence data collected for 1 h until all seven targets had been selected. For the 1.7kbp, 4.7kbp, and 12.4kbp size exclusion runs, all 512 pores were chosen to enrich. For the 10.6kbp and 16.4kbp runs, pores 1 to 256 were chosen to enrich and pores 257 to 512 were chosen for controls.

Additional runs involved sequencing a 10.6 kbp library for 6 h enriching for *S. dysgalactiae* only followed by a nuclease flush and re loading the library and running for a further 6 h enriching for *S. dysgalactiae* only, running a 10.6 kbp library and enriching for *M. morganii* and collecting for 72 h, running a 10.6 kbp library and enriching for *S. dysgalactiae* and collecting for 72 h, running a 10.6kbp library and depleting for *M. morganii* and collecting for 72 h and running a 10.6 kbp library and depleting for *M. morganii*, *A xylosoxidans*, and *L. richardii* and collecting for 72 h. In each case, pores 1 to 256 were chosen to enrich and pores 257 to 512 were chosen for controls.

All sequencing data are available in the European Nucleotide Archive (http://www.ebi.ac.uk/ena) repository under accession number PRJEB44844. ONT run reports, along with a table providing direct links to the ENA runs can be found at https://github.com/richardmleggett/adaptive_sampling.

GridION adaptive sampling

For each adaptive sampling run, we supply MinKNOW with a reference file containing only the genome of the species we wish to target. This is the reference file that is used to perform the classification of the first ~450bps, upon which the molecule is either sequenced entirely or ejected from the pore. We also use MinKNOW's "align" function to align all reads to a reference file containing the genomes of all species in the sample. This mapping does not affect the decisions on sequencing or ejecting molecules and is the mapping we use for our classification. Because the initial classification used to inform the decision on whether to sequence or not must be done very quickly (before the molecule has passed through the pore), it does not necessarily coincide with the more thorough mapping done later. Misclassifications from the initial mapping have a moderate effect on the enrichment we observe.

Adaptive sampling model web app

A web application was created in R using the "Shiny" library, to allow researchers to see the effect experiment parameters will have on the predicted enrichment, as detailed in the "A mathematical model of enrichment potential for metagenomic samples" section. The app is available at https://sr-martin.shinyapps.io/model_app/, and the source code can be found in the github repository https://github.com/SR-Martin/Adaptive-Sequencing-Analysis-Scripts (GPL v3 license).

Bioinformatic analysis of mock communities

All bespoke scripts used in the analysis were written in Python and are freely available in the github repository https://github.com/SR-Martin/Adaptive-Sequencing-Analysis-Scripts (GPL v3 license).

Sequences were basecalled during the experiment on the GridION using the MinKNOW software. Mappings of each read to the reference genomes of the seven species in the mock community were also created by MinKNOW. The script analyse_RU.py was used to cross reference the mappings with the reads, and report read and bp statistics for each species, split by channels used for adaptive sampling and all others (when appropriate).

For the analysis of false negatives, the script RU_decision_stats.py was used to parse the adaptive sampling logs created by MinKNOW for each experiment. This script determines the signal sent to the pore for each read and uses these to split the read set into reads that have been ejected from the pore ("unblocked") and those that were sequenced. These read sets were then cross referenced with the file of mappings, and reads were manually binned by species and signal type. The script get_read_stats.py was used to obtain statistics for each read set.

The read length distributions in Fig. 1a and the control distributions in Fig. 4a were obtained by binning reads by length into bins of size 1000. For the enrichment distributions in Fig. 4a, reads were binned by length into bins of size 100.

In Fig. 3.c, yields were normalised by the number of active channels, where active channels were those that sequenced a molecule in the first 30 min of the experiment. For the plots of active channels over time (Fig. 6b and Fig. 7), a channel was defined as active from the beginning of the experiment, up until the time it sequenced its final molecule (as long as it sequenced at least one molecule). Active channels were counted using the script GetActiveChannels.py, with counts each hour for the 72-h experiments, and every 15 min for the 6-h nuclease flush experiment.

For Fig. 6d, the time between two successive target molecules was recorded for each channel using the script GetWaitingTimes.py. For Fig. 6c, the script GetTimeHist.py was used to get the target yield for channels 1–256 and 257–512 each hour. For the yield plot in Fig. 7, a different approach was taken to reduce the effect of the mux scans; the script GetTimeHistFlush.py was used to get the total yield for channels 1–256 and 257–512 in sliding windows every 15 min. For the first six 15-min intervals, the sliding window was the duration of the experiment up to that point. For the remaining intervals, the window was the 90 min before. The yield in each window was normalised by its duration. For Fig. 5, the script GetYieldByTarget.py was used to determine the yield each hour, split by channels 1–256 and 257–512, and split by reference.

All plots were created in Python using pandas and matplotlib in Jupyter Lab.

For Table 6, results were obtained using the Python script GetCoverageBySAM.py to parse the SAM file used for the analysis on each run. For each strain, an array represented the genome (per base), where the entries were either 0 (not covered) or 1 (covered). An alignment to a strain updated the corresponding array by adding 1s to the alignments reference position. Coverage was calculated by summing the array and dividing by its length.

MAG assembly

Reads mapping to *S. dysgalactiae* were binned by their start time, with bins containing reads that were sequenced in the first hour, the first two hours, etc. up to all 12 h, using the script GetReadsByTargetAndTime.py. After 6 h, a nuclease flush was performed. Each bin was assembled with Flye v2.8.1 using the command

```
flye --nano-raw <read bin> --genome-size 2.1m
```

Assembly statistics were collected with a custom script, and each assembly was compared to the reference genome using dnadiff (part of Mummer v3.23 [33]).

Bioinformatic analysis of compost sample

The first 100,000 reads from an initial control run were classified with MetaMaps v0.1 using the miniSeq+H database provided with the tool [28]. Default options were used, which means that MetaMaps only classifies reads > = 1 kbp in length. This gave 72,638 reads long enough, of which 57,825 were unmapped. To determine the number and abundance of species in the sample, the rows for the 'definedGenomes' AnalysisLevel were extracted from the .EM.WIMP file output by MetaMaps, then sorted in order of absolute read count. To determine the number of genera, the same approach was used, but instead the rows for the 'genus' AnalysisLevel were extracted.

All reads from the control run that passed QC were assembled with Flye using the options `--nano-raw` and `--meta`. The resulting contigs were classified with MetaMaps v0.1, again with default options.

The reference sequence for *Hydrogenophaga* sp. PBC was downloaded from RefSeq as accession NZ_CP017311.1, having been published in 2012 [29].

MetaMaps was again used to classify enriched reads. However, in order to determine the total size of sequence, it was necessary to write a program, MMParse, to parse MetaMaps .reads2Taxon file in order to identify if each read was classified as a descendent of *Hydrogenophaga* and to calculate the total sequence bp of such reads. The source code can be found at https://github.com/richardmleggett/MMParse (MIT license).

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s13059-021-02582-x.

Additional file 1. Alignment statistics. Excel spreadsheets showing detailed calculations of alignment statistics.

Additional file 2. DNA quantification. Figures S1-S17 showing TapeStation and Femto Pulse traces for bacterial DNA extractions.

Additional file 3. Review history

Acknowledgements

We thank Carl Mayers for early discussions about selective sequencing approaches. This research was supported in part by the NBI Computing Infrastructure for Science Group, which provides technical support and maintenance to Earlham Institute's high-performance computing cluster and storage systems.

We also thank E Yabuuchi, The Lister Institute, F W Hickamn-Brenner, O Lysenko, and F Griffith for depositing the bacterial strains used in this study with the NCTC and making them available to the scientific community.

Review history

The review history is available as Additional file 3.

Peer review information

Barbara Cheifet was the primary editor of this article and managed its editorial process and peer review in collaboration with the rest of the editorial team.

Authors' contributions

RML and MDC designed the study. DH and SH performed the experiments. SM developed the model. SM, YL, and RML performed the data analysis. RML, MDC, SM, DH, and YL wrote the manuscript. The authors read and approved the final manuscript.

Authors' information

Twitter handles: @richardmleggett (Richard M Leggett)

Funding

This work was supported by the Biotechnology and Biological Sciences Research Council (BBSRC), part of UK Research and Innovation, through Tools and Resources Development Fund award BB/N023196/1, Core Capability Grant BB/CCG1720/1, Core Strategic Programme Grant BB/CSP1720/1. The funders made no intellectual contribution to the design or execution of the experiments.

Availability of data and materials

Scripts and software used in the analysis are available in the github repositories https://github.com/SR-Martin/Adaptive-Sequencing-Analysis-Scripts [34] (GPL v3 license) and https://github.com/richardmleggett/MMParse [35] (MIT license), as described in the Methods. Additionally, the versions used in the manuscript have been deposited at Zenodo, doi:10.5281/zenodo.5557291 [36].

The sequence datasets generated and analysed during the current study are available in the European Nucleotide Archive (http://www.ebi.ac.uk/ena) repository under accession number PRJEB44844 [37].

Declarations

Ethics approval and consent to participate

Not applicable

Consent for publication

Not applicable

Competing interests

The authors have not received direct financial contributions from ONT; however, RML and MDC have received a small number of free flow cells as part of the MAP and MARC programmes. RML has also received travel and accommodation expenses to speak at ONT-organised conferences.

Author details

[1]Earlham Institute, Norwich Research Park, Norwich NR4 7UZ, UK. [2]Natural History Museum, London SW7 5BD, UK.

Received: 8 May 2021 Accepted: 20 December 2021
Published online: 24 January 2022

References

1. Overholt WA, Hölzer M, Geesink P, Diezel C, Marz M, Küsel K. Inclusion of Oxford Nanopore long reads improves all microbial and viral metagenome-assembled genomes from a complex aquifer system. Environ Microbiol. 2020;22(9):4000–13. https://doi.org/10.1111/1462-2920.15186.
2. Leggett RM, Alcon-Giner C, Heavens D, Caim S, Brook TC, Kujawska M, et al. Rapid profiling of the preterm infant gut microbiota using nanopore sequencing aids pathogen diagnostics. Nat Microbiol. 2020;5(3):430–42. https://doi.org/10.1038/s41564-019-0626-z.
3. Youngblut ND, de la Cuesta-Zuluaga J, Reischer GH, Dauser S, Schuster N, Walzer C, et al. Large-scale metagenome assembly reveals novel animal-associated microbial genomes, biosynthetic gene clusters, and other genetic diversity. mSystems. 2020;5(6):e01045–20.

4. Wilkinson T, Korir D, Ogugo M, Stewart RD, Watson M, Paxton E, et al. 1200 high-quality metagenome-assembled genomes from the rumen of African cattle and their relevance in the context of sub-optimal feeding. Genome Biol. 2020;21(1):229. https://doi.org/10.1186/s13059-020-02144-7.
5. Nayfach S, Roux S, Seshadri R, Udwary D, Varghese N, Schulz F, et al. A genomic catalog of Earth's microbiomes. Nat Biotechnol. 2021;39(4):499–509. https://doi.org/10.1038/s41587-020-0718-6.
6. Meziti A, Rodriguez-R LM, Hatt JK, Peña-Gonzalez A, Levy K, Konstantinidis KT. The reliability of metagenome-assembled genomes (MAGs) in representing natural populations: insights from comparing MAGs against isolate genomes derived from the same fecal sample. Appl Environ Microbiol. 2021;87(6):e02593–20. https://doi.org/10.1128/AEM.02593-20.
7. Matthews TJ, Whittaker RJ. On the species abundance distribution in applied ecology and biodiversity management. J Appl Ecol. 2014;52(2):443–54.
8. Charalampous T, Richardson H, Kay GL, Baldan R, Jeanes C, Rae D, et al. Rapid diagnosis of bacterial lower respiratory infection using nanopore-based clinical metagenomics. Nat Biotechnol. 2019;37(7):783–92. https://doi.org/10.1038/s41587-019-0156-5.
9. Marotz CA, Sanders JG, Zuniga C, Zaramela LS, Knight R, Zengler K, et al. Improving saliva shotgun metagenomics by chemical host DNA depletion. Microbiome. 2018;6(1):42. https://doi.org/10.1186/s40168-018-0426-3.
10. Feehery GR, Yigit E, Oyola SO, Langhorst BW, Schmidt VT, Stewart FJ, et al. A method for selectively enriching microbial DNA from contaminating vertebrate host DNA. PLoS One. 2013;8(10):e76096. https://doi.org/10.1371/journal.pone.0076096.
11. Jupe F, Witek K, Verwei W, Śliwka J, Pritchard L, Etherington GJ, et al. Resistance gene enrichment sequencing (RenSeq) enables re-annotation of the NB-LRR gene family from sequenced plant genomes and rapid mapping of resistance loci in segregating populations. Plant J. 2013;76(3):530–44. https://doi.org/10.1111/tpj.12307.
12. Witek K, Jupe F, Witek A, Baker D, Clark MD, JDG J. Accelerated cloning of a potato late blight–resistance gene using RenSeq and SMRT sequencing. Nat Biotechnol. 2016;34(6):656–60. https://doi.org/10.1038/nbt.3540.
13. Jiang W, Zhao X, Gabrieli T, Lou C, Ebenstein Y, Zhu TF. Cas9-Assisted Targeting of CHromosome segments CATCH enables one-step targeted cloning of large gene clusters. Nat Commun. 2015;6(1):8101. https://doi.org/10.1038/ncomms9101.
14. Gilpatrick T, Lee I, Graham JE, Raimondeau E, Bowen R, Heron A, et al. Targeted nanopore sequencing with Cas9-guided adapter ligation. Nat Biotechnol. 2020;38(4):433–8. https://doi.org/10.1038/s41587-020-0407-5.
15. Loose M, Malla S, Stout M. Real-time selective sequencing using nanopore technology. Nat Methods. 2016;13(9):751–4. https://doi.org/10.1038/nmeth.3930.
16. Masutani B, Morishita S. A framework and an algorithm to detect low-abundance DNA by a handy sequencer and a palm-sized computer. Bioinformatics. 2019;35(4):584–92. https://doi.org/10.1093/bioinformatics/bty663.
17. Kovaka S, Fan Y, Ni B, Timp W, Schatz MC. Targeted nanopore sequencing by real-time mapping of raw electrical signal with UNCALLED. Nat Biotechnol. 2021;39(4):431–41. https://doi.org/10.1038/s41587-020-0731-9.
18. Edwards HS, Krishnakumar R, Sinha A, Bird SW, Patel KD, Bartsch MS. Real-time selective sequencing with RUBRIC: read until with basecall and reference-informed criteria. Sci Rep. 2019;9(1):11475. https://doi.org/10.1038/s41598-019-47857-3.
19. Kielbasa SM, Wan R, Sato K, Horton P, Frith MC. Adaptive seeds tame genomic sequence comparison. Genome Res. 2011;21(3):487–93. https://doi.org/10.1101/gr.113985.110.
20. Payne A, Holmes N, Clarke T, Munro R, Debebe BJ, Loose M. Readfish enables targeted nanopore sequencing of gigabase-sized genomes. Nat Biotechnol. 2021;39(4):442–50. https://doi.org/10.1038/s41587-020-00746-x.
21. Li H. Minimap2: pairwise alignment for nucleotide sequences. Bioinformatics. 2018;34(18):3094–100. https://doi.org/10.1093/bioinformatics/bty191.
22. Quick J, Nicholls S, Loman N. The 'Three Peaks' faecal DNA extraction method for long-read sequencing. protocols.io; 2019. https://doi.org/10.17504/protocols.io.7rshm6e.
23. Nanopore Info Sheet: Adaptive Sampling. 2020. https://community.nanoporetech.com/info_sheets/adaptive-sampling/v/ads_s1016_v1_reva_12nov2020. Accessed 27 April 2021.
24. Laszlo AH, Derrington IM, Gundlach JH. MspA nanopore as a single-molecule tool: from sequencing to SPRNT. Methods. 2016;105:75–89. https://doi.org/10.1016/j.ymeth.2016.03.026.
25. Martins S, Leggett RM. Alvis: a tool for contig and read ALignment VISualisation and chimera detection. BMC Bioinformatics. 2021;22(1):124. https://doi.org/10.1186/s12859-021-04056-0.
26. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. J. Mol. Biol. 1990;215(3):403–10. https://doi.org/10.1016/S0022-2836(05)80360-2.
27. Kolmogorov M, Yuan J, Lin Y, Pevzner PA. Assembly of long, error-prone reads using repeat graphs. Nat Biotechnol. 2019;37(5):540–6. https://doi.org/10.1038/s41587-019-0072-8.
28. Dilthey AT, Jain C, Koren S, Phillippy AM. Strain-level metagenomic assignment and compositional estimation for long reads with MetaMaps. Nat Commun. 2019;10(1):3066. https://doi.org/10.1038/s41467-019-10934-2.
29. Gan HM, Chew TH, Tay YL, Lye SF, Yahya A. Genome sequence of Hydrogenophaga sp. strain PBC, a 4-aminobenzenesulfonate-degrading bacterium. J Bacteriol. 2012;194(17):4759–60. https://doi.org/10.1128/JB.00990-12.
30. Mineeva O, Rojas-Carulla M, Ley RE, Schölkopf B, Youngblut ND. DeepMAsED: evaluating the quality of metagenomic assemblies. Bioinformatics. 2020;36(10):3011–7. https://doi.org/10.1093/bioinformatics/btaa124.
31. Jain M, Koren S, Miga K, Quick J, Rand AC, Sasani TA, et al. Nanopore sequencing and assembly of a human genome with ultra-long reads. Nat Biotechnol. 2018;36(4):338–45. https://doi.org/10.1038/nbt.4060.
32. Payne A, Holmes N, Rakyan V, Loose M. BulkVis: a graphical viewer for Oxford nanopore bulk FAST5 files. Bioinformatics. 2018;35(13):2193–8.
33. Kurtz S, Phillippy A, Delcher AL, Smoot M, Shumway M, Antonescu C, et al. Versatile and open software for comparing large genomes. Genome Biol. 2004 Jan;5(2):R12. https://doi.org/10.1186/gb-2004-5-2-r12.
34. Martin S, Heavens D, Lan Y, Horsfield S, Clark MD, Leggett RM (2021). Adaptive Sequencing Analysis Scripts. GitHub. https://github.com/SR-Martin/Adaptive-Sequencing-Analysis-Scripts
35. Martin S, Heavens D, Lan Y, Horsfield S, Clark MD, Leggett RM (2021). MMParse. GitHub. https://github.com/richardmleggett/MMParse

36. Martin S, Heavens D, Lan Y, Horsfield S, Clark MD, Leggett RM (2021). Nanopore adaptive sampling: a tool for enrichment of low abundance species in metagenomic samples. Zenodo. https://zenodo.org/record/5557291
37. Martin S, Heavens D, Lan Y, Horsfield S, Clark MD, Leggett RM (2021). Oxford Nanopore sequence data from adaptive sampling experiments using a synthetic bacterial mock community. PRJEB44844. EMBL-EBI European Nucleotide Archive. https://www.ebi.ac.uk/ena/browser/view/PRJEB44844

Publisher's Note

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