SEAVENTzymes II: an integrated step-forward approach using whole-genome sequencing for the identification of industrial relevant enzymes from deep-sea vent prokaryotes

Ana Sofia Eria Oliveira, July 2017

Thesis summary to obtain the Master of Science Degree in Microbiology
Supervisors: Dr. Ricardo Pedro Moreira Dias and Prof. Isabel Maria de Sá Correia Leite de Almeida
Examination Committee: Chairperson Prof. Jorge Humberto Gomes Leitão, Supervisor Dr. Ricardo Pedro Moreira Dias and Member of the Committee Prof. Rodrigo da Silva Costa

Hydrothermal vents are underwater volcanic singularities that extrude superheated jets of enriched water from the ocean crust. They comprise some of the most extreme environments found on Earth. Prokaryotes that thrive in such environments are particularly interesting for bioprospecting since their enzymes should function under the similarly harsh conditions of industrial processes. Under the SEAHMA project (SEAfloor and subseafloor Hydrothermal Modeling in the Azores sea), 36 samples were taken from vents near Azores, from which 296 isolates were obtained and characterized. During the SEAVENTzymes project, aiming to identify industrial relevant biocatalysts, this collection of isolates was screened for the production of polysaccharide-degrading enzymes, lipases/esterases and peptidases. Phenotypic tests were useful to pinpoint promising aerobic mesophilic isolates. However, sequence-based screening, by degenerate-PCR, of the anaerobic thermophilic subset, fell short from expected, with virtually no genes identified. Here we performed whole-genome nanopore sequencing of a *Bacillus* sp. isolate to assess the potential of this methodology as an alternate approach for bioprospecting enzymes. From the sequencing data we were able to identify putative genes encoding peptidases, lipases, esterases and starch-, cellulose-, xylan-, mannan-, pectin- and chitin-degrading enzymes, in accordance with previous phenotypic assays. This was accomplished with low depth of sequencing - *ca.* 3.7-fold -, by annotating nanopore long reads (mean of 3.8 kilobases) directly, with no need for prior error correction or assembly. We propose that this approach can develop into a full pipeline for biotechnological potential assessment of isolates or samples, which could be implemented to revisit the SEAHMA collection.

INTRODUCTION

Deep-sea hydrothermal vents are underwater singularities driven by volcanic activity near the Earth's tectonic plate limits. As cold seawater percolates trough small crevices into the hot crust, where pressures can reach several hundred atmospheres, it heats up to 270-400ºC, and reacts with the surrounding rocks, losing oxygen, becoming strongly acidic (pH of 2-3) and getting enriched in reduced compounds, till it finally extrudes back from the crust as a superheated jet of water[1]. Here, we find the most extreme conditions on Earth, with sharp chemical and physical gradients that regardless, are able to support very rich communities of macro- and microorganisms, surpassing in biomass those of coastal or tropical systems[2]. Prokaryotes, in particular, are widely distributed and diverse in such environments. It is the interest in understanding these organisms and their diverse and extreme-resisting metabolic mechanisms that has been driving deep-sea exploration, in a major way due to the biotechnological potential that is anticipated[3].

Industrial processes entail extreme conditions that are somewhat similar to those found in deep-sea hydrothermal vents[4]. Prokaryotes thriving in vent environments are expected to hold naturally tailored enzymes with extreme-resisting characteristics, making them the ultimate frontier for industrial enzyme bioprospection[4]. Particularly, there is a large justified investment in the bioprospection of extreme-resisting versions of biomass-degrading enzymes, since these enzymes dominate the global enzyme market, having a central role in titan industries such as the food, feed, paper, textile chemical and pharmaceutical industries[4,5].

Portugal is a privileged place for industrial enzyme bioprospecting since it detains exclusive economic rights over a large fraction of the North Atlantic Ocean, enclosing multiple hydrothermal vent fields. Several projects have explored these vent fields, one such example is the SEAHMA project (SEAfloor and subseafloor Hydrothermal Modeling in the Azores sea). During this project, five hydrothermal fields near Azores, namely Lucky Strike, Menez Gwen, Menez Hom, Mount Saldanha and Rainbow, were visited by the research cruise SEAHMA-1. From a collection of 36 samples, 296 prokaryotes were obtained and further characterized by multiple fingerprinting approaches. The SEAVENTzymes project arose as the natural progression of the SEAHMA project. Its purpose was to search for

biotechnologically relevant enzymes in this privileged collection of hydrothermal vent prokaryotes. Specifically, it aimed for the bioprospection of novel hydrolytic enzymes with industrial applications (*e.g.* amylases, cellulases, xylanases, mannanases, pectinases, chitinases, proteases and lipases). Aerobic isolates were subjected to phenotypic assays to evaluate their potential. Conversely, thermophilic anaerobes, which require certain conditions that render the phenotypic screening unfeasible, were subjected to molecular-based screening by the use of degenerate PCR primers targeting the genes of interest. However, there were virtually no genes of interest amplified, even with multiple stages of PCR optimization. Thus, even though the phenotypic screening of the mesophilic aerobes yielded several positive results, overall, the project had limited success in the exploitation of the SEAVENTbugs collection.

After 13 years from the first instance of the SEAVENTzymes project, the interest in extreme-resisting enzymes still persists, but now, several technological advances have emerged. Fortunately, with the conservation and maintenance of the SEAVENTbugs collection, we are now able to revisit the project with a fresh approach.

Sequencing methods of bioprospecting offer great advantages over the screening approaches taken during the first SEAVENTzymes project. For instance, whole-genome sequencing acts as a window to the full genomic potential of an isolate or a sample, deeming the screening independent of multiple focused tests or enzyme expression conditions. Moreover, it can be applied independently of the growth requirements of the organism, which is a concern in the screening of vent microorganisms that require physico-chemical extremes incompatible with streamlined phenotypic assays.

Nanopore sequencing, in particular, brings an additional set of advantages to the sequencing field. In nanopore sequencing[6], biological engineered nanopores are embedded in an electrically resistant polymer membrane (Figure 1 A). When a voltage is applied across the membrane, ions in solution pass through the nanopores and create a current. Free-floating DNA molecules, driven by their charge, tend to cross the pores causing a disruption of this current. The changes in current are detected by electrodes and are recorded as squiggles (Figure 1 B), which in turn can be decoded into sequences. Being a third-generation technology, it brings two major improvements for whole-genome sequencing over the second-generation: the ability to sequence single molecules, avoiding the errors and biases introduced during PCR amplification, and, most importantly, the massive increase of read length.

Contrary to other technologies, the read lengths offered by nanopore sequencing have no theoretical instrument-imposed limitation. Nevertheless, the most distinctive characteristic of nanopore sequencing, and specifically the MinION device, is its portability and small footprint. The MinION is no larger than a smartphone (Figure 1 C) and runs off a personal computer. Furthermore, it enables real-time analysis, meaning that there is no need to wait till the end of an experiment to get access to sequence information.

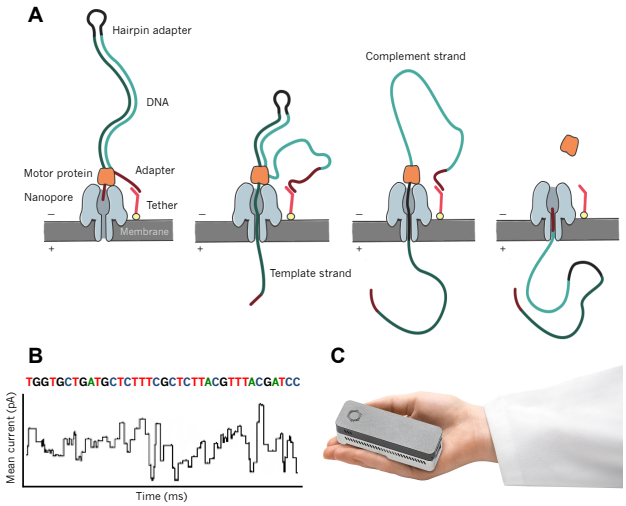

Figure 1 | **Schematic representation of nanopore-based sequencing of 2D reads (A) and squiggle lines resulting from a DNA molecule passing the nanopore (B); MinION sequencer picture – property of Oxford Nanopore Technologies (C).** For 2D nanopore sequencing a hairpin adapter is added during library preparation, which links both strands of a DNA molecule and allows for their contiguous translocation and sequencing, ultimately enabling the generation of a consensus sequence and increasing read quality. Besides the hairpin adapter, a leader adapter is also ligated to the DNA molecule with an attached motor protein, that controls the speed of translocation, and a tether that allows for the concentration of DNA in the membranes near the pores. As DNA molecules pass through the nanopores, disruptions of the baseline current are recorded by electrodes as squiggles, which can be decoded into sequences.

For the R7/R7.3 version of the nanopore technology, used in this work, there is a high error rate associated with the process of attributing bases to current squiggles, reaching up to 30%, and being mostly dominated by indel errors. To decrease the error rates, nanopore sequencing uses a hairpined sequence library that allows for the contiguous translocation of both forward and reverse strands of a single DNA molecule (Figure 1 A). Sequencing information from both strands can eventually be conjugated to generate a consensus sequence, which will have a higher quality score. The forward strand generates what is called a '1D Template' read, whilst the reverse strand generates a '1D Complement' read. The consensus sequence obtained from the joint analysis of paired template and complement reads is called a '2D' read.

Thus, nanopore sequencing produces single-molecule long reads, from real-time portable sequencing. From the nature of this technology, we could anticipate some competitive advantages for the bioprospecting of enzymes from hydrothermal vent microorganisms.

We propose that nanopore sequencing can be implemented as an alternate more advantageous

method for the bioprospection of industrial relevant enzymes from hydrothermal vent prokaryotes. Thus, this work aims to proof-of-concept the use of this methodology as a screening method, by first implementing it for the search of biomass-degrading enzymes on a single isolate of the collection, already characterized with phenotypic assays. For that purpose we will complete the following tasks:
(I) Reanalyze the results from the SEAVENTzymes project to choose a promising isolate.
(II) Use nanopore sequencing to perform whole-genome sequencing of the chosen isolate.
(III) Evaluate sequencing data quality and read processing needs.
(IV) Mine the sequencing data for biodegrading-enzymes with industrial potential and integrate the results with previous phenotypic results.

METHODS

Reanalysis of the screening results from the SEAVENTzymes project

All results from both growth and colorimetric phenotypic assays performed during the SEAVENTzymes project were integrated and subjected to a Principal Component Analysis (PCA) in NTSYSpc v2.21q (Exeter Software).

Isolate recovery, identification and selection

Isolates of the SEAVENTbugs collection were recovered by streaking 5 µl of the cryopreserved cultures onto plates of Marine Broth (Difco) Agar, incubating them at 22ºC for 3 to 5 days until growth was visible. Genomic DNA was extracted from the recovered isolates using a modified version of the Guanidium Thiocyanate Method[7]. The modifications concerned the initial stages of the protocol. Cells were resuspended in 250 µl of lysis buffer (50 mM Tris; 250 mM NaCl; 50 mM EDTA; 0.3% (w/v) SDS; pH 8.0) and 100 µl of microspheres. After a homogenization step in a vortex for 2 min, the cells were incubated at 65ºC for 30 min, followed by another 2 min of homogenization. 250 µl of GES (5 M Guanidium thiocyanate; 10 mM EDTA; 0.5% (w/v) Sarkosyl; pH 8.0) was added and at this stage the remaining steps of the original protocol were followed[7].
16S rRNA gene was partially amplified using the universal primers PA 5'AGAGTTTGATCCTGGCTCA G3' and 907r 5'CCGTCAATTCMTTTRAGTTT3'. Reactions were carried out in 50 µl, containing 1X PCR buffer, 2 mM of MgCl, 1 µM of each primer, 50 µM of each of the four dNTPs, 1 U of Taq polymerase (Invitrogen) and 1 µl of template DNA (50-100 ng). PCRs were run in a Biometra T Gradient thermal cycler, with the following PCR conditions: 3 min of initial denaturation at 94ºC, followed by 35 cycles of denaturation at 94ºC for 1 min, annealing at 55ºC for 1 min and extension at 72ºC for 1 min, with a final extension at 72ºC for 3 min. The amplification products were purified using the JetQuick PCR Product Purification Spin Kit (Genomed) and sequenced by Biopremier (Lisbon, Portugal). A phylogenetic reconstruction with both the isolates' partial 16S rRNA gene sequences, and their top BLAST hits was generated by MEGA software v7.0.16 using the neighbor-joining algorithm accompanied by a bootstrap analysis of 1000 fold.

Whole-genome nanopore sequencing

The MG SD 082 isolate was streaked onto a plate of Marine Broth (Difco) Agar and incubated for 72 hours at 22ºC. Cells were harvested and DNA extraction was performed with the Promega Wizard Genomic DNA Purification Kit. 2D-sequencing library preparation was performed with the Oxford Nanopore Technologies Genomic Sequencing Kit SQK-MAP-006, following the manufacturer's instructions, and employing the suggested fragmentation step with a Covaris g-tube. Two different DNA libraries were prepared from two independent cultures of the same isolate. Library 1 was performed exactly as the manufacturer's instructions. Library 2 was performed in a similar manner, with the exception that the cleaning step after the Covaris g-tube fragmentation was done with 0.6X by volume of magnetic Agencourt AMPure XP beads (Beckman Coulter) instead of 1X. For each of the two sequencing runs (Run 1 and Run 2) a new R7.3 flow cell was used and mounted into the MinION Mk I device, connected to a PC with an installment of the control software MinKNOW v0.51.2.40. The flow cell was primed as per the manufacturer's instructions. At this stage, the sequencing mix was immediately loaded into the flow cell and the '48 hours sequencing protocol' script was run on MinKNOW. The flow cell was topped-up with freshly prepared sequencing mix every 12 hours.

Sequencing data analysis

Basecalling of the sequencing data was performed in the Metrichor system EPI2ME v2.39.3.
The sequencing data of both runs was pooled together and repartitioned into three separate datasets: all 2D reads, 2D Pass reads and all 1D reads. Sequences were extracted in fasta format from basecalled fast5 files using Poretools v0.3.0.
RAST online server was used to determine the closest neighbor of the sequenced isolate by submitting only high quality reads - 2D Pass reads. The genome sequence of the closest neighbor, determined to be *Bacillus velezensis* strain FZB42 [NC_009725.1] (former *B. amyloliquefaciens* subsp. *plantarum* FZB42^T) was retrieved from the Genome database at NCBI as a fasta file and used as a reference for the purpose of comparing different subsets of the sequencing data.
Each of the three datasets, that is, 1D, 2D and 2D Pass reads, was subjected to independent correction, assembly and polish. Canu was used to correct each of the original datasets as suggested by

Canu developers. The value inputted for the expected size of the genome to be assembled was the one corresponding to the genome of the closest neighbor, as determined by RAST, *i.e.* 3.9 Mb. The threshold for minimum read length accepted was set to 100 bases, whilst the threshold for minimal overlap between reads was set to 50 bases. Following correction, canu –trim and canu –assemble commands were run to complete the assembly pipeline. The assembled datasets resulting from Canu were further polished using Nanopolish v0.2.0, as described by the developers.

At this stage, each original dataset (2D, 2D Pass and 1D), generated a set of three derived datasets, namely correction, assembly and polish. Each of the 12 datasets was subjected to a series of read and mapping quality assessments. Statistical analysis concerning the resulting metrics was performed in RStudio v1.0.143.

Read and contig metrics of each dataset were obtained by QUAST.

K(5)-mer composition of the chosen reference and each dataset was determined using the 'kmer' script from Poreminion v0.0.4. Based on the frequency tables of the *k*-mer counts, Kullback-Leibler divergence was calculated as a measure of entropy of one dataset with regard to the chosen reference, following the equation:

$$d^{KL}(S, R) = \sum_{i=1}^{1024} f_i^S \cdot \log_2\left(\frac{f_i^S}{f_i^R}\right)$$

where S represents the dataset in question, R represents the reference and f the relative frequency of the *k*-mer i in the total of 1024 possible *k*-mers of length 5.

Mapping potential of the different datasets was assessed based on the mapping of the reads against the chosen reference, using BLAST+. For the purpose of counting mapped reads and evaluate mapping statistics, only the highest scored mapping for each independent read was considered.

Finally, all datasets were subjected to RAST with standard parameters for determination of gene recall potential, by performing a sequence-based comparison with the reference in the SEED viewer.

Annotation and enzyme identification

All 2D reads were submitted to RAST online server for annotation, with standard parameters, taking advantage of the embedded RAST ORF finder. Pinpointing relevant industrial enzymes was done by manually curating the total set of annotations. The selected annotations were those concerning starch-, cellulose-, xylan-, mannan-, pectin-, chitin-degrading enzymes, lipases/esterases and proteases. The selected annotated sequences were further subjected to PSORTb v3.0.2.

For Blast2GO annotation, the 2D reads were first genecalled using Prodigal. The predicted protein sequences were fed to Blast2GO and a BLASTP was performed against the nr database with standard parameters. After BLAST was completed, and still in the Blast2GO interface, the results were mapped to GO terms and annotated. InterProScan was run and the annotations were recalculated in an integrated manner. Additionally, PSORTb was run inside the Blast2GO interface. To conclude the Enzyme codes were mapped to the previously determined GO terms. The results were manually curated to pinpoint the final set of annotations of interest, just as with RAST results. The coding sequences of interest of both annotation systems were further subjected to a BLASTP against both the MEROPS database, as well as the CAZy database, to confirm the annotation of peptidases and carbohydrate-active enzymes, respectively. Annotation dereplication was performed by manual curation of repetitive annotations. BLAST was used to assist this process, by evaluating if the original reads yielding equally annotated ORFs were indeed mapping to the same coordinates and genes of the reference genome.

RESULTS AND DISCUSSION

Integrating all phenotypic screening results from the SEAVENTzymes project by PCA allows to pinpoint promising isolates

To choose a single isolate to work towards our aim of testing whole-genome nanopore sequencing screening capabilities, we reviewed the screening results from the SEAVENTzymes project. During the SEAVENTzymes project, mesophilic aerobic isolates were subjected to phenotypic screening for the detection of biomass-degrading enzymes with industrial potential. Two different methods were applied: (I) growth assays with the target-enzyme substrate as the sole source of a nutrient, and (II) colorimetric assays based on commercial chromogenic substrates for the target-enzymes.

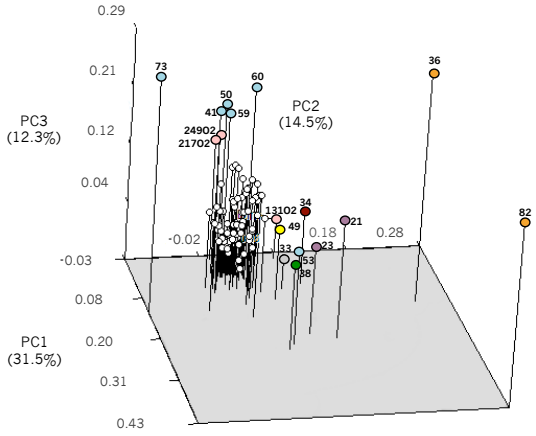

Figure 2 | **Projection of the isolates on the principal component space constructed from the integrated analysis of all results from the phenotypic screening performed during the SEANVENTzymes project.** Percent variance of the system associated with each principal component is shown. Only the set of isolates that are distinguishable from the main cluster have their code names indicated, which were shorten for clarity. Isolates with the same color circles were found to be clustered together by fingerprinting analysis (data not shown).

Even though there was a lack of association between the two screening methods (data not shown), both of the methods had high reproducibility in discerning positive from negative isolates (over 94%). Thus, they are most likely portraying different aspects of the enzyme production capability of each isolate. For the purpose of selecting a promising isolate with overall biomass-degrading capabilities, we integrated all results by PCA.

In Figure 2 we see that most isolates are grouped in a focalized cluster. Only a smaller set of isolates was distinguishable from the main group (shown in colored circles), which indicates that these isolates have unique responses to the phenotypic assays. When analyzing the phenotypic results for each of these isolates evidenced by PCA we found that they indeed represent some of the overall best producers of enzymes in the collection, in terms of total enzyme number and level of production/activity (data not shown). Since the purpose was to select a single isolate, for a more informed decision, these pre-selected isolates based on PCA were further analyzed by partial 16S rRNA gene sequencing, in an attempt to taxonomically position them. We restricted our analysis to three groups of isolates that had shown to be coherently clustered together by PCA and by fingerprinting analysis (data not shown).

Promising isolates belong to the *Bacillus*, *Rheinheimera* and *Vibrio* genera

The 16S rRNA gene of isolates pre-selected based on PCA was partially sequenced and a phylogenetic reconstruction is shown in Figure 3.

The isolates MG SD 082 and MG SD 036 belong to the *Bacillus* genus and seem to be indistinguishable by comparison of partial 16S rRNA gene sequence from the *B. amyloliquefaciens subsp. plantarum* FZB42^T and two other *Bacillus* strains, one belonging to *B. subtilis* and the other to *B. methylotrophicus*. Discrimination of species within the *Bacillus* genus has been proven difficult by 16S rRNA gene sequence. Here, we were also unable to identify the isolates at the level of species. The close clustering of *B. amyloliquefaciens* subsp. *plantarum* strain FZB42^T with *B. methylotrophicus* stains can be explained in the light of a recent publication in the International Journal of Systematic and Evolutionary Microbiology by Dunlap *et al*. (2016)[10]. Dunlap *et al*. revealed that the type strains of *B. methylotrophicus* KACC 13015^T, *B. velezensis* NRRL B-41580^T and *B. amyloliquefaciens* subsp. *plantarum* FZB42^T, are likely later heterotypic synonyms of *B. velezensis*, and should be reclassified as such. The fact that MG SD 082 and MG SD 036 16S rRNA gene sequences are closely clustered with those of both *B. amyloliquefaciens* subsp. *plantarum* FZB42^T and two strains of *B. methylotrophicus*, may indicate that the isolates are closely related to the restructured *B. velezensis* specie (*post hoc* confirmed using nanopore-sequencing whole-genome data).

MG CR 021 and MG CR 23 fit into the *Vibrio* genus and belong to at least two different *Vibrio* species, since they were separated into two different clusters.

Finally, operational cluster 3 (in blue) belongs to the *Rheinheimera* genus and all isolates seem to be very closely related between them and with the *Rheinheimera aquimaris* type strain.

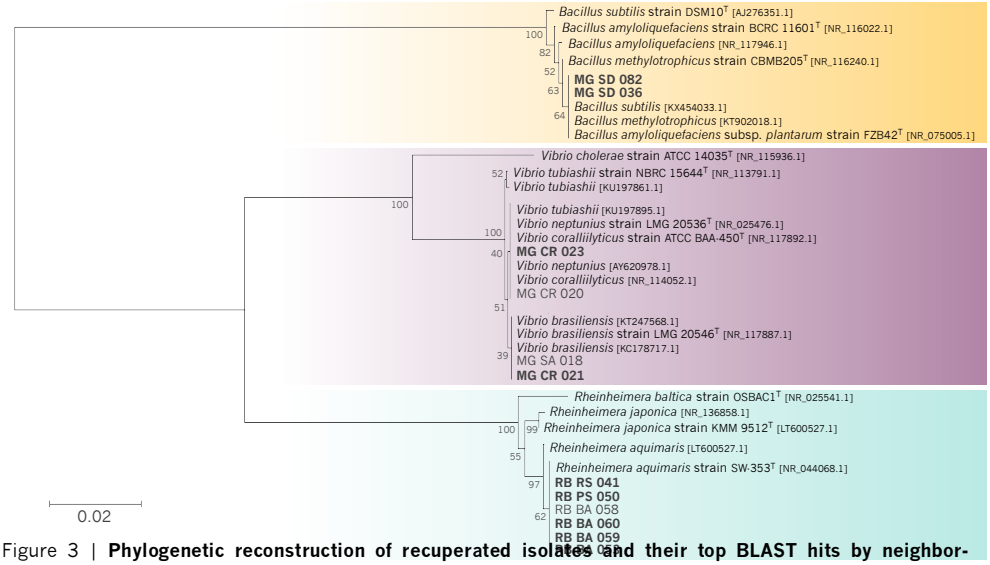

Figure 3 | **Phylogenetic reconstruction of recuperated isolates and their top BLAST hits by neighbor-joining clustering of their 16S rRNA partial gene sequences.** Percent bootstrap values, derived from 1000-fold sampling, are indicated near the respective nodes. Isolates selected based on PCA analysis are indicated in bold. The type strain of the type species of each represented genus was also included, namely *Bacillus subtilis* strain DSM10^T, *Vibrio cholerae* strain ATCC 14035^T and *Rheinheimera baltica* strain OSBAC1^T.

Bacillus sp. MG SD 082 demonstrated its ability to produce polysaccharide-, lipid- and peptide-degrading enzymes by phenotypic assays

Considering both the phenotypic results and the 16S rRNA gene based identification, the chosen isolate to be subject to nanopore sequencing was MG SD 082, a *Bacillus* sp. recovered from the seafloor sediments of the Menez Gwen hydrothermal vent field. As seen in Figure 4, the MG SD 082 isolate seems to produce endo-hydrolytic enzymes acting on starch, cellulose, xylan, mannan and casein, as evidenced by the colorimetric assays. The production of starch-, xylan- and mannan-degrading enzymes was further confirmed by growth assays. The NAUCr(ES)/NAUCr(BM), where NAUCr stands for relative Net Area Under Curve, calculated for the growth in media with cellulose and casein was 5 and 2, respectively. Although these values indicate that the growth in media with the enzyme substrate (ES) was

higher that the growth in base media (BM) alone, they still fall under the defined threshold for positive results based on replicate analysis. Growth assays further evidenced the production of chitin-degrading enzymes and lipases, which were however, not observable by colorimetric assays.

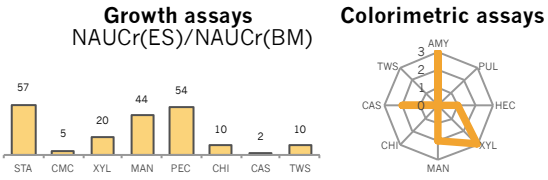

Figure 4 | **Growth and colorimetric screening results obtained during the SEAVENTzymes project for the selected isolate *Bacillus* sp. MG SD 082.** Bar graph represents the results for growth assays in media with starch (STA), carboxymethylcellulose (CMC), xylan (XYL), mannan (MAN), pectin (PEC), chitin (CHI), casein (CAS) and a mixture of 'tween' 20 and 'tween' 80 (TWS). Radial graph represents the results from the colorimetric screening where 0 represents a negative result, 1 represents a weak positive result, 2 an evident positive result and 3 a strong positive result. Colorimetric assays were performed with AZCL-amylose (AMY), AZCL-pullulan (PUL), AZCL-hydroxyethylcellulose (HEC), AZCL-xylan (XYL), AZCL-glucomannan (MAN), chitin-azure (CHI), AZCL-casein (CAS) and a mixture of 'tween' 20 and 'tween' 80 plus calcium chloride (TWS).

Thus, the MG SD 082 isolate was selected not only because it presented consistent promising results in both growth and colorimetric assays, but also because it belongs to a genus that is recurrent in seafloor sediments and well known for its biotechnological utility and production of industrial relevant enzymes - the object of study of this work.

Independent sequencing runs differ in yields and read length distributions

Two 2D-nanopore-sequencing runs were performed with the *Bacillus* sp. MG SD 082 DNA.

Run 1 sequenced a total of 44.47 Mb of 1D data, whilst Run 2 sequenced only 31.01 Mb, equating to 12.25 Mb and 9.34 Mb of 2D consensus data, respectively. The differences in total data yield between the two runs are likely related with the amount of available pores of the flow cells used, which is known to affect throughput. The number of working pores of the flow cells varies greatly as a result of the manufacture of the flow cell itself and the storage conditions to which it was subjected[6]. It is expected that a flow cell with higher number of working nanopores would produce more data. Run 1 was performed with 27% functional nanopores (559), whilst Run 2 used as little as 16% (334), thus explaining the lower throughput of Run 2. But despite the differences between the two runs, the maximum data yield obtained was still bellow some of the yields reported using the same R7.3 chemistry[8].

Run 1 also demonstrated a 2D read length distribution much more skewed towards smaller read lengths that Run 2. Indeed, the mean 2D read length for Run 1 was 2.97 Kb, statistically different from the mean of Run 2, which was 6.60 Kb (Mann-Whitney test, p=2.91x10-81, α=0.05). This implies a high concentration of low molecular weight 2D molecules in the Run 1 sequencing library, which might have been a result of unwanted fragmentation of the DNA prior to library preparation. Run 2 however, had the expected average 2D read length of 6.60 Kb. This was accomplished with a minor tweak during library preparation, where larger fragments were size selected by using a limiting proportion of DNA sequestering beads before the adapter linkage.

Nevertheless, both runs still sequenced 1D reads that reached more than 100 Kb in length, revealing the long-read capability of this technology.

In terms of 2D mean quality scores, the two runs generated distributions with similar means (8.8 for Run 1 and 8.7 for Run 2). It seems quality depends more on the chemistry and basecalling of the technology and less on library or flow cell variability.

Thus, overviewing the sequencing metrics revealed that the two runs generated considerably different data yields and read length distributions.

2D reads represent a smaller but higher-quality fraction of the nanopore-sequencing data

Since the overall yield of either run was lower than what expected, the data from both experiments was pooled together to create a richer dataset. Figure 5 presents summary metrics for the pooled data.

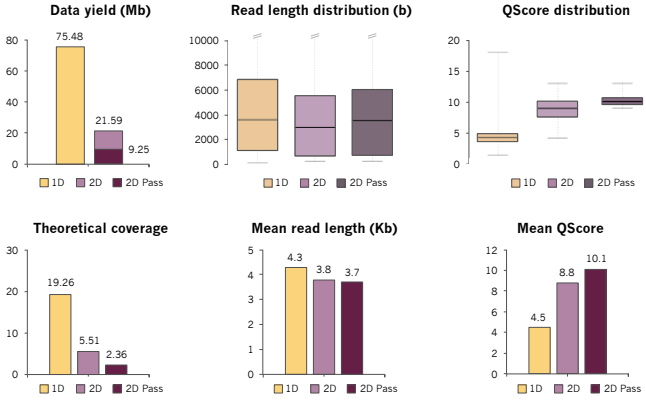

Figure 5 | **Yield, read and quality metrics of the repartitioned datasets 1D, 2D and 2D Pass.** Theoretical coverage was calculated as the ratio of data yield by the size of the genome of *B. velezensis* strain FZB42. Boxplots are represented with Spear whiskers that extend to minimum and maximum values. For 2D read length distributions the maximum values are omitted for clarity purposes. Qscore of a read refers to the per base quality score mean.

2D nanopore sequencing generates three sets of data, one of which, 2D Pass, (2D reads with QScores ≥9) is automatically filtered with the intention of constituting the usable higher-quality dataset. Yet, we have found in preliminary tests that using all 2D data, rather than just 2D Pass, can increase several fold the coverage of the dataset. Additionally, we have seen that a large portion of 1D reads does not get transformed into 2D consensus. That means that a great fraction of the information portrayed in 1D data gets lost when selecting to used only the consensus data. It would be of interest to take advantage of this untapped potential of 1D data, since it represents the largest share of the actual

generated data by nanopore sequencing.

We characterized each dataset to understand their usefulness for our ultimate goal of mining industrial enzymes. We anticipated that there were three main characteristics of the data that should impact their suitability for our intended purpose, namely coverage of the genome, read length and read quality.

We have found that either dataset has mean read lengths (between 3.7-4.3 Kb) sufficient to span entire bacterial genes - assuming an average gene size of 1 000-1 200 bases -, the highest theoretical coverage is achieved with 1D data (19.26-fold), and that 2D Pass data offers the highest quality data (10.1 on average). These datasets are most likely going to lead to very dissimilar responses to downstream enzyme mining systems.

Here, we also evaluated the need for data processing for the purpose of mining enzymes, by comparing the original datasets with their corrected, assembled and polished versions.

For the purpose of comparing datasets based on read/contig lengths, we used NG50 and LG50 assembly quality metrics. The best datasets are those that have the highest NG50 lengths and the lowest LG50, having a high contiguity.

After processing the data, 1D non-processed reads still comprise the highest amount of sequencing data (Table 1), representing the highest theoretical coverage we could achieve with the data generated. Furthermore, it offers 12 408 reads with more than 1 Kb in length, providing a large amount of sequences, which could, in theory, span entire bacterial genes.

1D reads correction decreased greatly the amount of 1D data to 14%. The effect of correction in the 2D and 2D Pass datasets also had the same general steep, with 2D data being reduced to 74% and 2D Pass data to 92%. This is a predictable response since 2D Pass reads have higher quality scores and, as such, should require a less aggressive correction.

Table 1 | **Read/contig metrics of the 1D, 2D and 2D Pass datasets and their corrected, assembled and polished versions.**

Dataset	Total (b)	Reads/ contigs	> 1 Kb	NG50	LG50
1D	75.48 M	18 604	12 408	85 087	8
1D corrected	10.22 M	2 890	1 663	74 291	10
1D assembled	618.65 K	260	154	na	na
1D polished	618.65 K	260	154	na	na
2D	21.59 M	5 557	3 539	14 929	79
2D corrected	15.87 M	3 487	2 315	13 655	99
2D assembled	6.64 M	1 915	1 128	9 329	118
2D polished	6.64 M	1 915	1 128	9 329	118
2D Pass	9.25 M	2 157	1 350	11 541	125
2D Pass corrected	8.50 M	1 771	1 184	11 236	140
2D Pass assembled	4.49 M	1 321	817	7 175	185
2D Pass polished	4.49 M	1 321	817	7 175	185

na – not applicable.

Further assembling the 1D corrected reads led to a decrease of data to levels that were not sufficient for 1-fold coverage of the genome. From the 2D data and 2D Pass data, we would expect an improvement in NG50 and LG50, by boosting the contiguity of the data with the assembly of reads, but this did not happen. Further polishing the assembled data did not seem to have any impact on the datasets either.

Both the correction and assembly algorithms are lossy processes that tend to eliminate low quality regions of the data, depending greatly on data coverage to surpass that limitation. Here, it seems that the aggressive trimming, combined with the generally low coverage of the data, trumped the benefits of assembling the reads, and did not improve contiguity of the datasets.

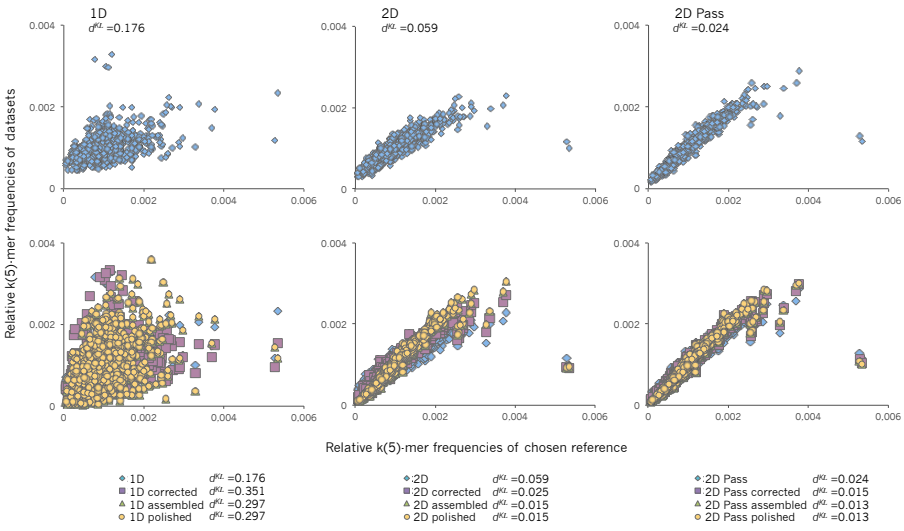

Figure 6 | **K(5)-mer relative frequencies comparison between each dataset and the chosen reference B. velezensis FZB42.** Each point in the graphs represents one of the 1024 possible *k*(5)-mers traced as its relative frequency in the specific dataset (y-axis) versus its relative frequency in the reference genome (x-axis). Kullback-Leibler divergence (d^{KL}) was used as a numeric measure of entropy of the dataset when compared to the reference. The two points that are consistently furthest away from the dispersions represent the *k*-mers 'AAAAA' and 'TTTTT', which are homopolymers known to be underrepresented in the nanopore-sequencing data.

Overall, we found that the non-processed 1D dataset offers the best theoretical coverage and read-length distribution metrics. However, read length and data yield are not sufficient indicators of the value of a dataset. Aiming to evaluate sequence accuracy, we compared the *k*(5)-mer composition of each dataset with the chosen reference (Figure 6). *K*-mer frequency comparison allows for an appreciation of the differences between the sequences of a dataset and the reference, without the need for alignment. The entropy of the comparison can be calculated as a Kullback-Leibler divergence (d^{KL}).

1D data is highly divergent from the reference (d^{KL}=0.176). Moreover, 1D data processing worsens the overall quality of the data (d^{KL}>>0.176). As expected by the quality score i

mpact. Yet, the decrease in data amount was less

distribution, 2D Pass data shows the lowest divergence in relation to the reference (d^{KL}=0.024), particularly when processed till the assembled stage (d^{KL}=0.013). This decrease in entropy may be due to the elimination of lower quality data from the dataset and/or by improving accuracy from consensus calling aligned reads.

To conclude, although 1D reads constitute larger amounts of data, equating to a higher theoretical coverage and high number of gene-size sequences, this data is very dissimilar from the expected true sequence information. Contrariwise, 2D reads, either Pass or not, are a much smaller fraction of the sequencing data with less gene-size sequences, but seem to be highly similar with the expected original sequence. The accuracy of the data can eventually be a better fit for the purpose of mining genes. What remains to be answered is if, for the intended purpose, the increase in accuracy obtained by processing 2D and 2D Pass data compensates the loss of information caused by the aggressive trimming algorithms.

Low-coverage non-processed 2D nanopore-sequencing data offers high gene recall

A straightforward way to evaluate usefulness of the datasets for enzyme mining is to compare their mapping and gene-recalling statistics (Figure 7).

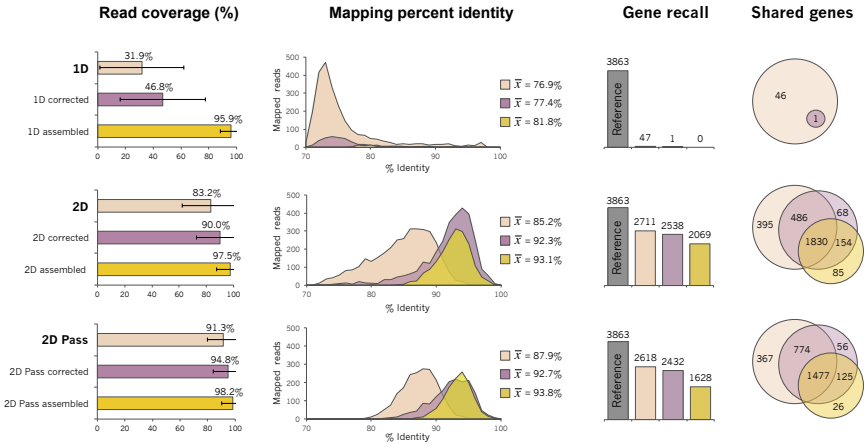

Figure 7 | **Read mapping coverage, distribution of mapping percent identity and gene recall of each dataset.** These metrics were calculated using as reference the genome of *B. velezensis* strain FZB42. Results from the polished assembly were omitted since values were equal to the assembled datasets. 'Error-bars' in read coverage graphs represent standard deviation. Read coverage refers to the extension of the read that mapped to the reference. Gene recall corresponds to the number of genes of the reference that were found in the particular dataset.

On average, only 31.9% of the extension of a particular 1D read is mapped, showing mean percent identities of 76.9%. It seems like 1D reads are mosaic in nature, harboring hotspots of higher fidelity that are able to map to the reference. This data profile eventually led to the low gene recall of 1D data to only 47 genes, *i.e.* only 47 genes of the reference were found in the dataset. Low quality reads with high error rates, particularly with the indel rich profile reported for nanopore sequencing[6] can create frameshifts that hinder the genecalling of the data. Again, we can see how processing 1D reads was detrimental for their usefulness, reducing the number of genes of the reference that were recalled from 47 to only 1 after correction and 0 after assembly. 2D data offers the highest number of mapped reads of all tested datasets (3 543), with mappings spanning almost the entirety of the read (83.2%). Average identity was found to be 85.2%, but values go as low as 70%. Regardless, 2D reads had the highest gene recall, with 2 711 genes found from the total 3 863 of the reference genome. Note that, although we had estimated a theoretical coverage of 5.51-fold for 2D data, when true depth of sequencing was examined in SAMtools, it only reached a value of 3.7-fold per base on average (data not shown). On further inspection, we found that there were a total of 208 Kb - 5% of the genome - that were not covered in any instance by this dataset. This alone does not explain why it failed to recall 1 152 genes - 30% of the total genes. Undercalling of genes might be a consequence of the error rate of the data, that, based on mapping identity assessment, is approximately 15%, which is in accordance with what has been reported[6].

Correction of 2D reads shifted the mapping identity and read coverage up, reaching a mean of 92.3% and 90.0%, respectively. Yet, the loss of data in the correction process, discussed before, led to a decrease in the number of genes recalled (2 538). Although the processed dataset recalled fewer genes, the overall higher accuracy and higher mapping percent identity led to the identification of 222 genes that were not disclosed in the original 2D dataset. The same applies for 2D assembled data. Assembling the data further enabled the calling of 85 new genes, which may be a result of the higher accuracy of the dataset by consensus calling of aligned reads and/or an eventual assembly of reads disclosing previously interrupted genes.

Note that the difference in terms of amount of data between the 2D dataset and the 2D Pass dataset is of 12.34 Mb. Yet, they differ in gene recall by only 93 genes. That means that the majority of genes called in 2D data were actually coming from reads with quality scores above 9. Having said that, in certain applications, one must weight the benefits of using 2D data versus 2D Pass data. 2D data offers an ever so slight increase in gene recall associated with a major increase in the amount of data to be processed.

Since the increase in computational effort was not limiting in the specific context of this work, the 2D dataset was chosen to be subjected to mining for industrial relevant enzymes.

***Bacillus velezensis* MG SD 082 2D nanopore-sequencing data allows direct annotation of polysaccharide-, lipid- and peptide-degrading enzymes in accordance with phenotypic assays**

As seen in Figure 8, the ORFs called by Prodigal, and submitted to Blast2GO reached a total of 21 348. RAST however, identified in the same data 51 481 ORFs, more than twice as much as Prodigal. By the difference in amount of called ORFs, it can already be foreseen that the genefinders employed for the Blast2GO and RAST annotation are very different in their predictions and are most likely going to lead to very different annotation results. At the end of the annotation process, Blast2GO had attributed functional annotations to 2 493 ORFs and RAST to 10 860. In a first glance both systems were able to assign putative biological functions to an extensive set of ORFs, including some non-ribosomal peptide synthetases (data not shown). In the set of RAST annotations we found 381 entries, which may represent biomass-degrading enzymes of industrial interest. Blast2GO only generated 126.

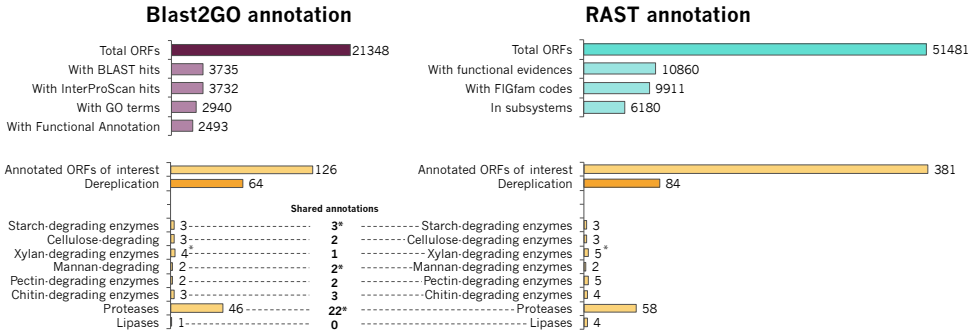

Figure 8 | **Annotation metrics and overview from the analysis of the Bacillus velezensis MG SD 082 2D-nanopore-sequencing data using Blast2GO and RAST.** The original dataset consisted of 5 557 non-processed 2D reads, amounting to 21.59 Mb. Blast2GO was used in combination with Prodigal genecaller'*' indicates the presence of extracellular endo-hydrolytic enzymes.

Since 2D data has the redundant nature of a non-assembled dataset, at this stage we dereplicated the selected ORFs by eliminating repetitive annotations in different reads. Furthermore, we found that several identical annotations were emerging in the same reads. When investigated further we understood that in a particular read, a gene was being annotated in fractions, even though the reads were spanning the entirety of the gene. This is the unwanted result of using error-prone reads, and specifically indel-prone reads. Since the annotation depends on the alignment of protein sequences rather than DNA sequences, it is more sensitive to frameshift-like errors, which can drastically change the resulting predicted protein sequence.

For the purpose of counting ORFs of interest, same-read replicated annotations were subtracted. At this point, we constructed a set of relevant annotations for each annotation system. Blast2GO in combination with Prodigal generated a total of 64 annotations which fit into the industrial enzymes category, whereas RAST revealed 84. Both annotation systems shared a total of 37 annotations. Thus, by applying both annotation systems, we were able to identify, in the *Bacillus velezensis* MG SD 082 whole-genome nanopore-sequencing data, evidences for the production of a total of 111 putative industrial relevant enzymes capable of acting on the degradation of starch, cellulose, xylan, mannan, pectin, chitin, proteinaceous compounds and lipids, which seems to be in accordance with the phenotypic assays performed during the SEAVENTzymes project in Figure 4. As an example, mining the whole-genome nanopore-sequencing data unveiled a putative extracellular α-amylase. The production of this endo-acting extracellular enzyme would generate the positive result observed in the colorimetric assays with AZCL-amylose, since the enzyme can act on the internal linkages of the cross-linked substrate. Furthermore, two other cytosolic enzymes with the capability to act on starch utilization were identified, namely an α-glucosidase (EC 3.2.1.20) and an oligo-1,6-glucosidase (EC 3.2.1.10), which release monomers of glucose from their action on starch-derived oligosaccharides. The combination of these enzymes reflects the ability of the isolate to degrade starch into glucose, and explains the results obtained in the growth assays performed with starch as the sole source of carbon.

There was a specific result in the colorimetric assays for which we could not detect the responsible enzyme – an endo-acting cellulase. Rather, we identified some enzymes with potential to act on the external ends of cellulose. In an attempt to find this enzyme we further submitted all ORFs called by RAST and Prodigal from the 2D data, 2D corrected and 2D assembled, to a dbCAN BLASP against the CAZy database of carbohydrate-active enzymes. There was still no evidence of such enzyme. It could have easily been a miscalled gene that was obscured by erroneous data. However, further investigation revealed that neither of the two genes coding for putative cellulases of the reference genome were covered by the 2D reads. Thus, the absence of the enzyme is most likely a result of low coverage sequencing, that is, assuming that the genome of the chosen reference is any indication of the genome of the MG SD 082 isolate.

FURTHER CONSIDERATIONS

In this work, we indeed proof-of-concept the use of whole-genome nanopore sequencing to evaluate the biotechnological potential of a *Bacillus velezensis* isolate from hydrothermal vent sediments, with regard to industrial relevant enzymes.

We assayed the potential of different possible

datasets of the nanopore-sequencing technology, either processed or non-processed, for the purpose of mining enzymes, in terms of overall genome coverage, read/contig length, general quality/accuracy, and gene recall amenability. In the end, we found that, from low-coverage sequencing, non-processed long 2D reads enabled direct annotation with the highest gene recall. In this dataset we were able to find evidences for several enzymes of interest in accordance with previous phenotypic results, despite the lower-throughput and less-than-optimum error rates of the used R7.3 version of the technology. Although not explored in depth in this work, ultimately, the same whole-genome nanopore-sequencing data also enabled the identification of the isolate at the species level – *Bacillus velezensis* - and unveiled several other ORFs of biotechnological interest that transcended our initial set of industrial relevant enzymes (*e.g.* biosynthetic clusters of secondary metabolites). The fact that we stumbled upon such genes, reflects one of the major advantages of sequencing-based strategies over the phenotypic assays. We can easily unveil a large and diverse set of determinants of interest by mining the same sequencing data, with no need for a specific assay for each group. Furthermore, we were able to identify a much larger collection of relevant enzymes than both phenotypic screening approaches together. Albeit, the products of the predicted genes still have to be heterologously expressed for confirmation.

Just as other sequence-based technologies, the ability to mine for enzymes in nanopore-sequencing data depends on our current understanding of sequencing information and knowledge of enzymes and their function**.** Fortunately, even though some genes may escape us under our current knowledge base, the sequence data obtained has permanent character and it can be revisited again, as new methods of studying and understanding these sequences develop and disclose new opportunities and potential in "old" data. Overall, even with the current limitations of sequence-based methods, the MinION revealed itself a useful and accessible sequencing platform. Its portability and real-time potential was not explored directly in this work, nor its implementation with metagenomic samples but, it is these aspects, accompanied by the generation of very long reads, that deem this technology so interesting for bioprospecting deep-sea vent microorganisms. The study of microorganisms from deep-sea environments, or other remote locations, typically entails the collection, preservation and transport of environmental samples to laboratories. However, this paradigm has several disadvantages, being the most relevant the potential loss or corruption of unique samples. This may represent an irreparable damage to a project since the deployment of sampling procedures in remote locations is many times limited to brief opportunity windows or even singular visits. Additionally, since the sampling is so divorced from the analysis step, the exploration of these locations becomes a reactive practice. 'In-field' sequencing, enabled by the real-time portable character of nanopore sequencing, would be useful to, for instance, reiterate sampling in response to opportunities unveiled by sequencing whilst still in the field, supporting a more proactive approach. Thus, this technology has the ability to change the paradigm of deep-sea exploration and as it evolves it promises to expedite screening methods to quasi real-time.

Indeed, in this work we only proof-of-concept that long 2D reads generated by the nanopore sequencer can be annotated directly for bioprospecting purposes with no need for data processing. But it is this independency of data processing that would eventually allow the implementation of real-time annotation, by enabling mining of 2D reads as soon as they are sequenced by the device.

From our analysis, we propose that whole-genome nanopore sequencing has the capability to become a relevant system for the biotechnological potential assessment of prokaryotic isolates or samples from deep-sea hydrothermal vents or other remote environments.

As for the SEAVENTbugs collection, we have still not grasped all its potential. We are now in a position where we can implement this technology to screen all isolates of interest, or even the metagenome of the preserved SEAHMA samples. Furthermore, the sequencing data generated can be useful to assist in the following stages of the bioprospection project, by enabling the well-informed design of cloning experiments.

Future projects should implement on this system, and evaluate metagenome sequencing, develop real-time annotation pipelines and finally deploy such methodologies to actual remote locations.

REFERENCES

1. Allsopp, M., *et al.,* 2009. State of the World's Oceans, *Amsterdam, The Netherlands: Springer,* pp.1-10.
2. Jørgensen, B.B. & Boetius, A., 2007. Feast and famine - microbial life in the deep-sea bed. *Nature Reviews Microbiology*, 5(10), pp.770-781.
3. Podar, M. & Reysenbach, A.-L., 2006. New opportunities revealed by biotechnological explorations of extremophiles. *Current Opinion in Biotechnology*, 17(3), pp.250-255.
4. Elleuche, S. *et al*., 2014. Extremozymes-biocatalysts with unique properties from extremophilic microorganisms. *Current Opinion in Biotechnology*, 29, pp.116-123.
5. Dalmaso, G.Z.L., Ferreira, D. & Vermelho, A.B., 2015. Marine extremophiles a source of hydrolases for biotechnological applications. *Marine Drugs*, 13(4), pp.1925-1965.
6. Brown, C., 2016. Inside the SkunkWorx. Plenary lecture in *London Calling 2016* by Oxford Nanopore Technologies, London, United Kingdom, May 26th-27th, 2016. Available at https://vimeo.com/168687629. Accessed September 3rd, 2016.
7. Pitcher, D.G., Saunders, N.A. & Owen, R.J., 1989. Rapid extraction of bacterial genomic DNA with guanidium thiocyanate. *Letters in Applied Microbiology*, 8(4), pp.151-156.
8. Dunlap, C.A. *et al*., 2016. *Bacillus velezensis* is not a later heterotypic synonym of *Bacillus amyloliquefaciens*; *Bacillus methylotrophicus*, *Bacillus amyloliquefaciens* subsp. *plantarum* and '*Bacillus oryzicola*' are later heterotypic synonyms of *Bacillus velezensis* based on phylogenomics. *International Journal of Systematic and Evolutionary Microbiology*, 66, pp.1212-1217.