TÉCNICO LISBOA

# 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**

Thesis to obtain the Master of Science Degree in

**Microbiology**

Supervisors: Doctor Ricardo Pedro Moreira Dias

Professor Isabel Maria de Sá Correia Leite de Almeida

## Examination Committee

Chairperson: Professor Jorge Humberto Gomes Leitão

Supervisor: Doctor Ricardo Pedro Moreira Dias

Member of the Committee: Professor Rodrigo da Silva Costa

**July 2017**

# Acknowledgements

This dissertation is the end product of the second year of the Microbiology Master of Science Degree of the University of Lisbon. This joint degree encompasses the Instituto Superior Técnico, the Faculty of Sciences, the Faculty of Medicine and the Faculty of Veterinary Medicine. The work pertained in this document however, was achieved with my integration in the Bugworkers group, at the Tec Labs, a center associated with the Faculty of Sciences, where I was supervised by both Professor Rogério Tenreiro and Doctor Ricardo Dias.

I know Professor Rogério Tenreiro since my bachelor degree and I chose this dissertation hoping that some of his knowledge and analytical skills would rub off on me. Nevertheless, during this year, what I found most impressive was how he manages to harness the most of every person that works with him. He has a seamless way of minimizing one's flaws and maximizing their qualities. Most importantly, he is able to align his priorities as a group leader with the personal priorities of each individual; that is what I imagine to be a good recipe for a highly motivated team. For some reason that really resonated with me and I hope to keep it in mind latter on. I am grateful that he trusted me with this project, but mostly, I appreciate all of his patience.

Next, but not least, I want to express my sincere gratitude to Doctor Ricardo Dias. There was no single interaction with Ricardo where I did not learn something completely new. He is an incredibly bright person that influenced my opinion on many things, academic and otherwise, and it is with no doubt that I say that our conversations ultimately shaped the tone of this dissertation. I greatly appreciate all the hours and resources that he placed at my disposal, which were many. I value the leap of faith he took by leaving the MinION in my hands and I thank him for his forbearance and most of all his ever-enduring support.

I would also like to thank Professor Isabel Sá Correia, my internal supervisor, for reviewing this dissertation and ensuring a smooth transition between the writing process and the defense and submission stage.

I am also thankful to Professor Ana Tenreiro and Professor Lélia Chambel. I was never shy to ask them for their help and I thank them for always being so accessible. I have great consideration for their opinion and support.

I could not forget the keepers of the lab. I thank CLO Filipa Antunes for her everyday effort to maintain the lab organized so that we can work in the best conditions possible. To our former technician, Cláudia Ramalho, I would like to say that I appreciate her drive and eagerness to help and I am grateful for all her aid during this year.

To Pedro Teixeira, the PhD student who was always available to indulge in my esoteric questions, I give him my utmost respect. I am also grateful to Ana Filipa Soares for the great company and support, particularly on the weekends. It is always good to have her around.

By a stroke of luck my friends ended up also being my partners in crime. It was great to share the same lab, learn and grow next to them. In more than one way they all contributed to the completion of this dissertation and I am truly grateful. Thank you. To Tatiana Cordeiro for always being so genuine, to Catarina Rocha for always being so truthful and to João Melo for always letting me release my weekly frustrations by bickering with him on a regular basis. I hope you all know that I shared the happiness of your accomplishments as if they were my own.

I am also grateful to my friend Rita Grenho for her enduring support and for allowing me some comic relief.

Finally, I am grateful to my parents, Jorge and Toya, for their unlimited love and care. They trusted me and supported me all throughout my education. I will never be able to repay everything they did for me, but I will try.

# Abstract

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.

Key-words: hydrothermal vents, prokaryotes, bioprospecting, industrial relevant enzymes, whole-genome sequencing, nanopore sequencing

# Resumo

As fontes hidrotermais são singularidades vulcânicas submarinas que expelem jactos de água superaquecida e enriquecida através da crosta oceânica. Aqui reúnem-se as condições mais extremas da Terra. Os procariotas que persistem nestes sistemas são particularmente interessantes para a bioprospecção de enzimas, já que estas deverão ser funcionais sob as condições igualmente extremas de certos processos industriais.

Durante o projeto SEAHMA (SEAfloor and subseafloor Hydrothermal Modeling in the Azores sea), 36 amostras foram recolhidas de fontes hidrotermais nos Açores, de onde 296 isolados foram obtidos e caracterizados. No projeto SEAVENTzymes, o potencial industrial destes isolados foi avaliado pela pesquisa de enzimas degradativas de polissacáridos, lipases/esterases e peptidases. Os testes fenotípicos permitiram identificar isolados aeróbios mesofílicos promissores. Porém, o *screening* molecular, por PCR com *primers* degenerados, feito aos isolados anaeróbios termofílicos, ficou aquém do expectável, não identificando genes de interesse.

Neste trabalho sequenciámos DNA genómico por *nanopore sequencing* de um isolado do género *Bacillus,* para avaliar o potencial desta metodologia como alternativa para a bioprospecção de enzimas. A partir dos dados de sequenciação identificámos genes putativos codificantes para peptidases, lipases, esterases e enzimas degradativas de amido, celulose, xilano, manano, pectina e quitina, concordantes com os testes fenotípicos. Isto foi possível com uma profundidade de sequenciação baixa de 3,7 vezes, pela anotação direta de *reads* longas (média de 3,8 kilobases), sem necessidade de correção ou *assembly*. Propomos que esta abordagem poderá transformar-se num sistema integrado para a avaliação do potencial biotecnológico de isolados ou amostras, podendo ser implementado para revisitar a coleção SEAHMA.

Palavras-chave: fontes hidrotermais, procariotas, bioprospecção, enzimas industriais, sequenciação de DNA genómico, *nanopore sequencing*

# Contents

# List of Tables

# List of Figures

# List of Abbreviations

| | |
|---|---|
| AZCL | AZurin-dyed Cross-Linked |
| BLAST | Basic Local Alignment Search Tool |
| EC | Enzyme Commission number |
| EEZ | Exclusive Economic Zone |
| EMEPC | Estrutura de Missão para a Extensão da Plataforma Continental |
| GO | Gene Ontology |
| MAP | MinION Access Program |
| NAUC | Net Area Under Curve |
| NCBI | National Center for Biotechnology Information |
| NGS | Next-Generation Sequencing |
| ORF | Open Reading Frame |
| PCA | Principal Component Analysis |
| PCR | Polymerase Chain Reaction |
| RAPD | Random Amplified Polymorphic DNA |
| RAST | Rapid Annotation using Subsystem Technology |
| ROV | Remotely Operated Vehicle |
| SDS-PAGE | Sodium Dodecyl Sulfate - PolyAcrylamide Gel Electrophoresis |
| SEAHMA | SEAfloor and subseafloor Hydrothermal Modeling in the Azores sea |
| TGGE | Temperature Gradient Gel Electrophoresis |
| UPGMA | Unweighted Pair Group Method with Arithmetic mean |

# Chapter 1. Introduction

## 1.1 An overview on Marine Biodiversity and Biotechnology

Covering over 70% of Earth's surface, the marine environment has had a major role in the history of life. Life is thought to have emerged in the early oceans, 3.8 *Ga* ago. In their depths a series of evolutionary events followed, from the development of a nucleus, to the development of multicellularity, the capture of organelles and the emergence of sexual reproduction (Boeuf 2011). Therefore, it is in the study of the oceans we can unveil many secrets that will help us better understand the history of the organisms. Although this is the case, marine organisms have still not been as extensively studied as their terrestrial counterparts (Webb 2009). A good reflection of this disproportion is the fact that, from all known and described species until 2006, only around 15% were marine – *ca.* 275000 species in a total of 1.8 million (Bouchet 2006).

Much has been done to estimate the number of species that indeed exist in the sea. Nevertheless, with this underlying lack of understanding on marine life, it is still common to find statements that portray the marine ecosystem as having somewhat lower biodiversity than the land (Appeltans *et al.* 2012). To clarify, biodiversity can be defined as all variation having an hereditary basis, at any level of organization - from the genes of a species, to the species of a community or even the communities composing the ecosystems (Wilson 1997). This idea that the marine ecosystem has lower diversity is perpetuated by the depiction of the ocean as a homogenous and continuous mass of water, corroborated by its seemingly stable salt concentration[1]. In such case, the continuity of the oceans could account for the lower diversity found, justified by the lack of boundaries which are known to favor isolation and speciation (Boeuf 2011). However, with advances in oceanographic mapping and sampling, it becomes increasingly apparent that this is not the case. Rather than a continuum, the oceans may instead have different contiguous but contrasting habitats that add up to a very rich marine species pool (Karl 2007).

Furthermore, our lesser extent of knowledge on ocean life is also incredibly biased, which can additionally contribute to an underestimation of the overall biodiversity of the sea. For instance, the best-known taxa are those of commercial importance (Fautin *et al.* 2010), followed by marine model

[1] The oceans have an overall salinity fluctuating between 3.2% and 3.8% (w/v) (Boeuf 2011).

organisms such as sponges, corals and sea urchins, that researchers study mainly within the context of morphological development (Thakur *et al.* 2008). Fautin *et al*. (2010) go a step further and claim that as a generalization, knowledge of a species is positively correlated with its size whilst negatively correlated with its distance from shore and depth. Simply put, much is known about bigger organisms living in the shore or surface of the sea but there is a major gap of information with regard to microscopic organisms in the oceans' depths. This gap of information appears even more concerning knowing that these inconspicuous microeukaryotes and prokaryotes actually represent the majority of existing organisms in the sea and together constitute more than half of Earth's total biomass (Thakur *et al.* 2008; Boeuf 2011).

There are two main reasons that justify this lack of knowledge on deep-sea microorganisms. Firstly, deeper regions of the ocean are hardly accessible and thus their exploration becomes expensive and unfeasible on a systematic basis (National Research Council US 2002). Lastly, several studies of microbial diversity depend on the culture of the organisms. These studies most likely misevaluate the existing biodiversity since the culturable fraction of marine microorganisms is estimated to be less than 1% (Staley & Konopka 1985; Connon & Giovannoni 2002). Although the methodologies used to culture these organisms can be improved, it is unlikely that the full set of marine species can be brought to pure culture. It is at this point that culture-independent molecular methods can enlighten us.

DNA fingerprinting is an example of a culture-independent approach that can be used to evaluate marine microbial communities (Ferrera, Banta & Reysenbach 2014), exploring either specific enzyme-restricted DNA fragments (Moeseneder *et al.* 1999) or specific Polymerase Chain Reaction (PCR) amplification products. These techniques allow the indirect appreciation of DNA sequence diversity in the sample. That is, each variant sequence is electrophoretically separated as a unique band, generating a profile of bands that reflects the sequence diversity present in the sample. By using DNA sequences known to vary in a species-specific manner, we can attempt to translate sequence diversity into species diversity.

The most widely used fingerprinting method for the assessment of sample biodiversity is PCR-based Denaturing Gradient Gel Electrophoresis (DGGE) (Fischer & Lerman 1979; Muyzer, Waal & Uitterlinden 1993). Specific amplicons, typically taxonomic/phylogenetic marker sequences[2], are separated by gel electrophoresis through an increasing denaturing gradient. Alternatively, a temperature gradient can be used at which point it is called Temperature Gradient Gel Electrophoresis (TGGE) (Rosenbaum & Riesner 1987). Each band of the generated profile should represent a different variant of the marker sequence, separated based on their intrinsic denaturing response conferred by their specific nucleotide composition. Eventually, DNA can be purified from the excised gel bands and sequenced for taxonomic identification and phylogenetic studies (Ferrera, Banta & Reysenbach 2014).

[2] 'Taxonomic/phylogenetic marker sequences' is a concept referring to DNA sequences which vary in a way that has been shown to allow either taxonomic discrimination or evolutionary history retracing, respectively. Some examples might be 16S rRNA or 18S rRNA genes, D1/D2 domain of the 26S rRNA gene or Internal Transcribe Spacers (ITS) between rRNA genes (Muyzer, Waal & Uitterlinden 1993; Ferreira *et al*. 2015; Liu J. *et al*. 2015).

In the last fifteen years there has been a gradual shift in paradigm and a different culture-independent approach has been taking the lead. Several research groups started to take advantage of Next-Generation Sequencing (NGS)[3] to evaluate community structure and diversity in a so-called Metagenomic approach (Oulas *et al.* 2015). Marine metagenomics in particular, has gained major momentum with initiatives such as the Global Ocean Sampling Project (Venter *et al.* 2004; Rusch *et al.* 2007), the Tara Ocean Project (Sunagawa *et al.* 2015) and the Ocean Sampling Day (Kopf *et al.* 2015), which have generated enormous quantities of data that still feed and propel the fields of marine microbial diversity, ecology and biodiscovery (Dupont *et al.* 2015; Farrant *et al.* 2016).

The metagenomic approach admits either the sequencing of specific marker amplicons or, by contrast, the whole-genome sequencing of total environmental DNA, *i.e.* the random/non-directed sequencing of sample DNA. Contrary to fingerprinting, sequencing-based methods directly compare sample sequences with known sequences gathered in established databases. In such an approach, a microbial community can be described in terms of richness and relative abundances of different taxa, but also in terms of overall metabolic capabilities. Thus, it allows for major insights not only into the community's diversity, but also on its ecology and biotechnological potential (National Research Council US 2002; Edwards R. *et al.* 2006).

Numerous studies have been using variations of these culture-independent methods and several of them estimated that a major part of the detected microbial species in the samples (up to 80%) had not yet been cultured (National Research Council US 2002; Zinger, Gobet & Pommier 2012). Nevertheless, it is important to note that even in molecular-based estimates, there are several biases that can occur. Sample handling (Tzeneva *et al.* 2009), DNA extraction methods (Ínceošlu *et al.* 2010) and PCR-primer selection (Fredriksson, Hermansson & Wilén 2013) can distort diversity assessments; there might even exist differential PCR amplification efficiencies for different variant molecules (Arezi *et al.* 2003; Gonzalez *et al.* 2012). Moreover, the formation of chimeric[4] molecules seems to have a worrisome impact on the apparent biodiversity of a sample. By recognizing PCR-chimers as novel organisms there is a possibility of inflating the number of taxa of a sample by more than twice its actual value (Haas *et al.* 2011; Boers, Hays & Jansen 2015). Being aware of these limitations however, allows the conscientious analysis of their effects and provides motivation for the development of alternative methods.

Overall, there are many aspects of marine research that imply a misevaluation of this ecosystem's biodiversity. Consequently, we still fail to grasp how much of the oceans' potential we have seized and how much we might yet access. In a strictly utilitarian sense, more biodiversity equates to more exploitable chemical diversity (Webb 2009). From an environmental perspective, having a clearer understanding of what exactly is out there will lead to a more mindful appreciation of the importance of the marine ecosystems, with consequences on governmental policies for marine habitat preservation and exploration (Beaumont *et al.* 2008). Webb (2009) expresses how difficult it is

[3] NGS refers to the group of modern sequencing technologies that are quicker, cheaper and more high-throughput than standard Sanger sequencing.
[4] Chimers, in this context, are PCR artifacts where hybrid products result from the amplification of multiple parent sequences.

to conserve habitats that are largely unknown but how attaching specific values and services to ecosystems tends to foster awareness for the significance of their biodiversity. Hence, understanding marine biodiversity will not only augment the universe for biotechnological exploration but also raise alertness for ecosystem preservation.

For instance, a relatively small number of marine organisms has already originated more than 12 000 novel chemicals (National Research Council US 2002). Marine species sense the world and communicate via chemical cues and these constitute what chemical ecologist Mark E. Hay (2009) calls “the language of life in the sea”. They mediate many interactions, determining mating, feeding and habitat choices, symbiotic and commensal associations and competitive exchanges, not only in animals, but also in plants and microorganisms (Hay 2009). This alone gives us a glimpse on how the enormous wealth of biological diversity in the sea may represent a massive treasure of valuable chemicals waiting for discovery.

Nevertheless, the marine ecosystem is considered the most underutilized reservoir of biological active compounds (National Research Council US 2002), even though it is the largest ecosystem on Earth and it has exceptional conditions to allow the presence of unique compounds. These compounds may have many biological activities with potential applications on human healthcare as novel drugs (Piel *et al.* 2004; Fiedler *et al.* 2008) or even in the food and feed industries as ingredients for food processing and storage (Ayadi *et al.* 2009; Fung, Hamid & Lu 2013). Complex biomolecules, such as marine biopolymers, can be used for the development of implantable prosthetics (Waite & Tanzer 1981; Kim & Venkatesa 2015) and biocatalysts like enzymes are already used in many processing steps of diverse industrial activities (Lundberg *et al.* 1991; Kim & Venkatesa 2015).

However, to harness this biological potential from the sea we depend on Marine Biotechnology. Biotechnology in itself is described as “any technological application that uses biological systems, living organisms, or derivatives thereof, to make or modify products or processes for a specific use” (Kim & Venkatesa 2015). It is normally expected that the products that advent from biotechnology have a genuine contribution on human life or are cost efficient/sustainable versions of a product or process, having a positive impact on the economy or on the environment. Thus, when we speak of Marine Biotechnology we are referring to the “biotechnology that is carried out using biological resources which have come from the marine environment” (Burgess 2012) with the purpose of creating new value.

Marine Biotechnology has already demonstrated its capability to create new value across a whole spectrum of applications, from biomedicine to industrial processes or environmental conservation. Yet, it is still not a mature economically significant field (Kim & Venkatesa 2015). There are, however, several lines of work suggested to have a substantial impact on the revitalization of marine biotechnology. These are (I) the exploration of unexamined habitats, (II) the focus on microorganisms and finally (III) the improvement of paradigms for screening useful marine products (National Research Council US 2002), all of which are integrated into this dissertation.

## 1.2 Oasis of the Marine Ecosystem: Deep sea and Hydrothermal Vents

The deep sea was, for many years, considered a biological inert territory, composed by extensive flat plains, occasionally inhabited by a sea monster of some sort. For that reason, it was generally understudied. It was only during the 19$^{th}$ century that the deep sea started to be explored, correlating with the advent of transoceanic communication technologies and the need to assemble and maintain submerse telegraphic cables. At this point, seabeds with more than 5000 m depth were discovered and life at higher depths started to be evidenced. The deep sea shifted from being depicted as an inert plain to being known as a complex assembly of different formations, with several geological and biological idiosyncrasies (Mills 1983; Etter & Hess 2015).

But discovering life in such depths brought more questions than answers, particularly with regard to the functioning of these deep ecosystems. There was no light penetrating deeper than 200 m and thus there was no known primary production to efficiently feed the trophic necessities of deep-sea communities. Researchers wondered about the mechanisms that would allow the proliferation of such communities and for quite a while deep-sea ecosystems were intuitively portrayed as having very poor biomasses (Danovaro, Snelgrove & Tyler 2014).

This notion changed with the geophysical expedition of 1977 to the Galapagos Ridge. In this expedition, the study of the seabed was enabled by the use of Alvin, a Remotely Operated Vehicle (ROV), very efficient in sampling and image detection. This submersible dived to the deep seabed and, unexpectedly, revealed an oasis of luxurious and exotic communities of giant clams, tubeworms and microbial mats surrounding venting fluids of 370ºC (Corliss *et al.* 1979). This expedition represented not only the discovery of hydrothermal vents but also the shattering of the notion that life was scarce in the deep sea (Mills 1983).

At this time the Plate Tectonics Theory had already been conceptualized (1960-1970). This theory enabled the reasoning of several geological phenomena that justified the heterogeneity of the deep sea and predicted the occurrence of these hot vent systems (Briggs 1987).

According to this theory, Earth is covered by an arrangement of continental and ocean crust plates that move on its surface with speeds of 1-10 cm per year. The forces generated when these plates diverge or collide lead to the formation of different seascapes. The mid-oceanic ridges are the result of the divergence of the plates, where lava wells up from the Earth's hot mantle, forming fresh basaltic ocean crust. These zones are normally associated with the formation of volcanic-type singularities. The new oceanic crust pushes the older regions slowly away from the ridges to the subduction zones, where plates converge and collide. In these regions the oldest portion of the plates, packed with sediments of sinking debris, end up being pushed back into the hot mantle. The compression over the sediments and their geothermal modification lead to the extrusion of gases as seep systems or the formation of mud volcanoes and chimneys. Hence, these tectonic movements are the great propellers of the general seabed and deep-sea heterogeneous configuration (Briggs 1987).

Hydrothermal vents are an example of a geological formation associated with these tectonic movements. They are driven by subsurface volcanic activity; they can occur in subduction zones or fracture zones of tectonic plates, where there are geothermal anomalies, but typically they appear where the Earth plates are actively spreading from one another. That is, hydrothermal vents are particularly concentrated along spreading ridges (Figure 1.2.1), specifically the Earth's mid-oceanic ridges, that form what looks like a 60 000 km seem of geological activity that runs through the planet (Allsop *et al.* 2009). However, the manifestation of these vents depends on a very specific combination of tectonic forces and volcanic activities, and for that reason they only appear at irregular intervals along the ridges. There are some shallow vents at depths of 100-500 m or even shallower, such as the D. João de Castro Bank in the Azores, that is only 20 m deep (Cardigos *et al.* 2005). However, they are much more common in water depths of 850-4 000 m, and are hence called deep-sea hydrothermal vents (OSPAR Commission 2010).

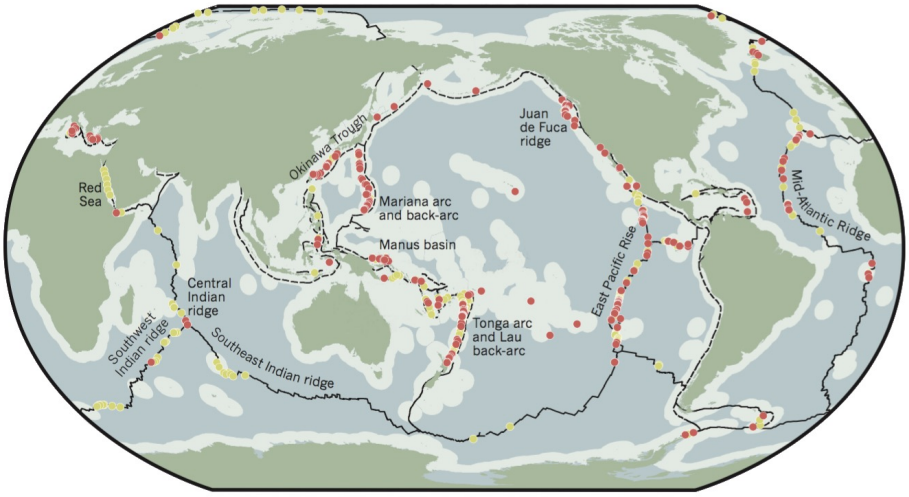


Figure 1.2.1 | **Distribution of hydrothermal vent fields along the Earth's tectonic plates limits.** Black lines represent the Earth's tectonic plates limits whilst red dots represent active vents and yellow dots represent unconfirmed vents. Adapted from Van Dover 2011.

Hydrothermal vents result from the percolation of seawater into the Earth's crust (Figure 1.2.2) through small pores or crevices formed during the cooling process of lava flows. The water chemically reacts with the hot crust, losing oxygen, becoming strongly acidic (pH of 2-3) and getting enriched in compounds such as metals and metallic sulfides, sulfates and gases such as carbon dioxide, hydrogen sulfide, methane and molecular hydrogen. As it heats, this enriched water starts to rise back to the seabed. The water can reach 270-400ºC and extrude from the crust as a superheated jet of water. As the jet of superhot water reaches the surrounding cold sea water, it cools, and the minerals and salts that were once dissolved start to precipitate as black clouds. This phenomenon is what gives the name of 'black smokers' to some vents (Ramirez-Llodra, Shank & German 2007). Some of the minerals that precipitate, end up forming tall chimneys, with heights of tens of meters, around the vents and surrounding sediments. Other minerals and components can form cloud-like plumes that can disperse through water creating a likewise very rich microenvironment. Contrary to a black smoker, in a white smoker, metallic sulfides, particularly iron sulfide, are precipitated while they are still within the rocks owing to somewhat lower temperatures. The extruded fluids in these hydrothermal

vents are therefore cloudier and whiter, hence the name of 'white smokers' (OSPAR Commission 2010).

Typically, each vent has a lifespan of a few decades to a full century and they undergo temporal and spatial evolution, which is reflected as changes in the physical and chemical properties of the vent discharges (Van Dover & Lutz 2004).

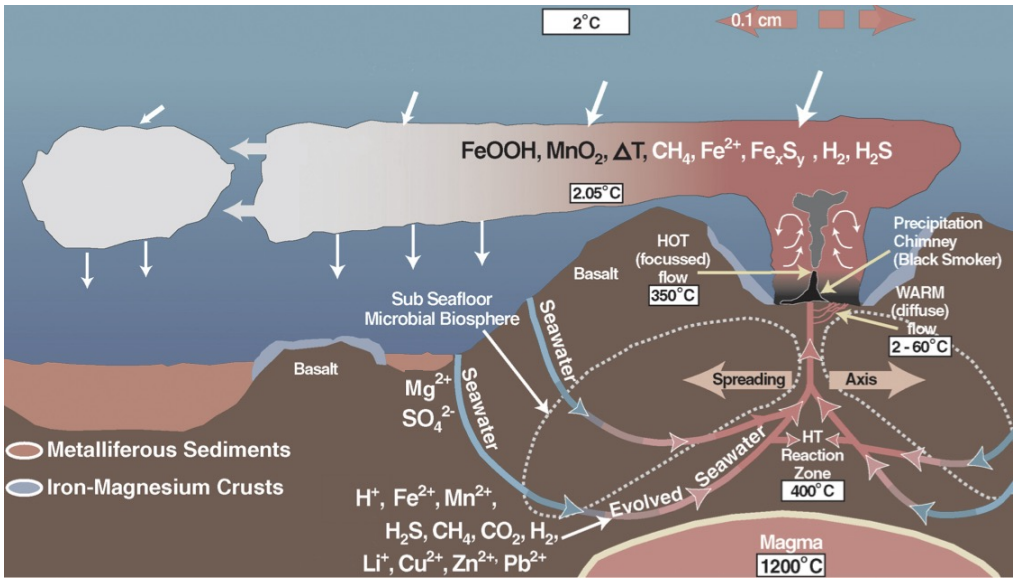


Figure 1.2.2 | **Schematic representation of a hydrothermal vent formation near a spreading ridge.** Seawater percolates through crevices near an active ridge and its progressively enriched and heated until it rises back and extrudes from the basaltic crust. When the enriched heated water enters in contact with the cold seawater, mineral compounds precipitate forming tall chimneys around the vent exit whilst reduced compounds and gases disperse through the column of water as plumes. Adapted from original diagram by Gary Massoth/PMEL (Pacific Marine Environmental Laboratory).

All differences between vents, in terms of temperature, rock composition, vent fluid chemistry, life stage, depth and proximity to the euphotic/aphotic regions[5], add to the particularities of the ecosystems that are formed around each specific vent. However, there are some general patterns that are identifiable in most vent systems. For instance, in the surroundings of vents there are very rich and observable communities of macroorganisms that are often endemic, at taxonomical levels higher than species, normally at the genus or family level (Glowka 2003). These macrofaunal communities are typically constituted by dense beds of *Bathymodiolus* mussels, small shrimps of the *Chorocaris, Microcaris* and *Rimicaris* genera, crabs of the *Segonzacia* genus and large colonies of giant tubeworms from genera like *Riftia*.

Hydrothermal vents have high biomasses but do not have the highest macroorganism biodiversity (Glowka 2003; Ramirez-Llodra, Shank & German 2007), although certainly it is much higher than originally expected, and thus the luxurious quality that is attributed to these communities. In contrast, hydrothermal vents host one of the highest levels of microorganism diversity on the planet (Synnes 2007).

[5] The ocean can be subdivided vertically into the euphotic region, that extends from the surface of the sea to the point where the sunlight available is less that 1% of maximum sunlight measured on the surface, at which point it starts the aphotic region (approximately 200 m) (Lee *et al.* 2007).

A vent is a system with very sharp chemical and physical gradients and for that reason it harbors a diverse range of habitats for microorganisms to flourish. For instance, there are several ecological niches that are formed in the superhot reduced metal-enriched fluids, or the cool oxidized water, or even at the mixing points of both. Furthermore, microorganisms can grow in mats on the surfaces of vent chimneys, within vent plumes, in the subsurface of deep-sea sediments surrounding vents or in conjunction with vent macroorganisms, for example, in symbiotic relationships (Jeanthon 2000).

As many diverse niches as there might be, in most ecosystems, there are a minor number of microbial lineages that dominate the overall community. Conversely, there are also low abundance lineages, the so-called 'rare biosphere', and these lineages are the ones that account for most of the diversity of the system (Anderson & Sogin & Baross 2015). Overall, the most widespread clades surrounding vents belong to the *Proteobacteria*, particularly the *Epsilon-*, *Alpha-* and *Gamma-* classes (Nakagawa & Takai 2008; Dick & Tebo 2010; Cerqueira *et al*. 2015; Jebbar *et al.* 2015; Zhang J. *et al.* 2015). Yet, the highest temperature centers of the vent structures are mostly archaea-dominated, mainly thermophilic *Crenarchaeota* and *Euryarchaeota* (Flores *et al.* 2012; Anderson, Sogin & Baross 2015; Jebbar *et al.* 2015). Contrariwise, in the subsurface and in the sediments, these sulfate-reducing, methanogenic or methane oxidizing archaea do not constitute the highest fraction of the microbial population. In this specific niche there is a higher representation of *Firmicutes* and *Alphaproteobacteria*, closely related to *Bacillus firmus* and *Rhizobium radiobacter*, respectively (Jørgensen & Boetius 2007; Edwards K.J., Wheat & Sylvan 2011; Jebbar *et al*. 2015).

Indeed, the seabed subsurface is a very particular niche. With an average overlying water column of 3 800 m, hydrostatic pressure reaches 380 atm – up to 1 000 atm at maximum depths (Bell & Heuer 2012). Combine that with the lithostatic pressure of the covering sediments, the lack of light to sustain primary activity and an average subsurface temperature of 2ºC and you have convincing arguments to believe that there is no life in the subseafloor. Curiously, subsurface sediments are the largest compartments of the global biosphere, representing 1/10 to 1/3 of the total living biomass (Teske & Sørensen 2008). Parkes, Cragg and Wellsbury (2000) estimated these values by doing systematic cell counts in deep-sea cores using fluorescence microscopy. Eventually microorganisms were found almost 1.9 km under the subsurface, which were at least to some extent metabolically active, as supported by highly sensitive fluorescence techniques (Parkes *et al.* 2014).

The question that persists is how are these communities of macro- and microorganisms able to flourish in what was thought to be a very carbon- and energy-poor location. Until the discovery of hydrothermal vents, photosynthesis was the major metabolic process known for primary production, ensuring the sustainability of life on Earth (Corliss *et al.* 1979). It uses light as the source of energy, and carbon dioxide as the inorganic source of carbon, to create primary biomass. The deep sea belongs to what is known as the aphotic region of the sea, where light does not reach and photosynthesis is out of the picture.

It has been suggested that hydrothermal vent communities depend on chemical energy, rather than light, to sustain their biomasses (Corliss *et al.* 1979). Near the vents, when the reduced

superheated discharges mix with the surrounding oxidized water, an interface of energetically favorable redox couples is formed. This imminent chemical energy becomes accessible to chemoautotrophic microorganisms for the fixation of inorganic carbon (Martin *et al.* 2008). In this case, they would act as the basis of the complex trophic networks that are found in vent ecosystems, creating a bridge between the physical-chemical environment and the biological entities. Indeed, it is very common to find chemoautotrophic prokaryotes in symbiosis with macroorganisms such as crustaceans, tubeworms and mussels, predominantly closer to vent emissions, enabling both parties to reliably establish near the available energy sources (Nakagawa & Takai 2008).

The export of organic particles from the euphotic zone also constitutes an additional contribution of carbon and energy to the heterotrophic deep-sea life, although its magnitude is still not quite understood. It alone could not explain how the deep sea could support biomasses that exceed coastal or even tropical systems (Jørgensen & Boetius 2007). Only 0.5-2% of the carbon formed in the euphotic zone reaches the sediments, since it is mostly consumed in the higher layers of the ocean (Brandt 2008). Studies that follow the deposition of phytodetritus by imagery methods have shown that it happens rather quickly and in massive manners in hadal[6] regions (Rice *et al.* 1986). But doubts remain regarding the frequency of this flow, and there are still no evidences that suggest that it has a consistent nature. This rather small carbon flux can be augmented by the downslope lateral transportation of organic matter from higher continental margins (Canals *et al.* 2006).

Eventually, the sinking of large animal carcasses or even wood also delivers high quality and fresh organic material that seems to be rather relevant, with significant consequences on the distribution and biomass of deep-sea organisms. For instance, certain symbiotic prokaryotes can digest precipitated or sunken material, such as cellulolytic compounds, allowing their hosts to exploit sources that could not be used otherwise (Danovaro, Snelgrove & Tyler 2014).

Furthermore, marine viruses may also play a relevant role in the sustainability of deep-sea communities. They are very common in the water column, and the deep sea is not exempt. It is estimated that 80% of prokaryotic organisms perish as a result of phage infection. They then are made available to the general food web as labile organic detritus that can be consumed by macrofaunal organisms (Brandt 2008).

To conclude, the deep sea, and hydrothermal vents in particular, comprise the most extreme environments found on Earth, with temperatures shifting from 2ºC to 400ºC, pressures of several hundred atmospheres and wide ranges of pH and salinity, together with a complex assortment of energy sources (Martin *et al.* 2008). Thus the organisms near hydrothermal vents and the metabolic strategies that they employ are widely diverse. It is the interest in understanding these organisms and their physiological mechanisms that has been driving deep-sea exploration, in a major way due to the biotechnological potential that is anticipated (Podar & Reysenbach 2006).

[6] The hadal zone refers to the column of water that has more than 6 000 m depth (Danovaro, Snelgrove & Tyler 2014).

## 1.3 Deep-sea vent prokaryotes as a major source of industrial relevant enzymes

Awareness towards environmental protection has grown in recent years, and with it comes the need for the development of sustainable industrial processes. This need has been the primary driver for the gradual replacement of chemical routes with enzyme catalysis in industry (Elleuche *et al.* 2014). Enzyme catalysis offers several advantages over chemical processes. It is clean and ecologically friendly, and it takes place in milder reaction conditions, thus reducing energy requirements (Dionisi, Lozada & Olivera 2012). Additionally, it is highly specific and can overcome the low selectivity and undesired formation of byproducts of several chemical processes (Demirjian, Móris-Varas & Cassidy 2001). Currently, there are already more than 500 industrial products manufactured using enzymes (Dalmaso, Ferreira & Vermelho 2015) and the enzyme global market is expected to reach $7.1x10^9$ US dollars by 2018 (BCC Research 2014).

While enzymes are able to function in much milder settings than some chemical reactions, several industrial practices still entail harsh conditions, such as extremely high pressures, acidic or alkaline pH, temperatures up to 140ºC, or near the freezing point of water (Elleuche *et al.* 2014). Mesophilic enzymes are often not well suited for these applications due to lack of stability (Demirjian, Móris-Varas & Cassidy 2001). In this case, enzymes stable at extreme conditions, *i.e.* extremozymes, would offer superior results over their mesophilic counterparts (Elleuche *et al.* 2014).

For instance, enzymes that have high catalytic efficiency at low temperatures can shorten the processing times of practices under cold settings, while maintaining quality of heat-sensible products (Vester, Glaring & Stougaard 2014). Accordingly, enzymes functional at higher temperatures bring advantages to industrial processes, since maintaining high temperatures can increase solubility of many polymeric substrates, decrease viscosity, increase bioavailability, and decrease the risk of contamination (Urbieta *et al.* 2015). Furthermore, enzymes that are stable at high salt concentrations can perform biocatalysis in low-water industrial scenarios (Demirjian, Móris-Varas & Cassidy 2001). Thus, extremozymes occupy an important place in the multimillion-dollar enzyme market with applications spanning numerous industrial sectors (Podar & Reysenbach 2006).

It generally holds true that the enzymes of an organism are adapted to function optimally at or near its normal growth conditions. Thus, the range of conditions at which enzyme activity might be detected should only be limited by the range of extremes at which life can be found. Note that some extreme-surviving organisms show other physiological adaptations to extreme conditions, yet it is still common for their enzymes to show a certain intrinsic stability (Hough & Danson 1999).

Stability to extreme conditions seems to be encoded in the gene sequence, as evidenced by studies focusing on thermostable enzymes (Adams, Perler & Kelly 1995; Hough & Danson 1999). There is high similarity between the sequence of mesophilic and extremophilic variant proteins, with multiple but subtle differences that lead to a generalized change in structure. Several structural features have been shown to be implicated in the stability of extremozymes, such as changes in specific amino-acid residues, changes in the size of loops, extent of secondary structure formation,

changes in the hydrophobic interactions at enzyme subunits interfaces, number of ion pairs, ratio of surface area to volume, and size of amino or carboxyl termini (Hough & Danson 1999). However, it seems there are no universal rules that justify the stability of these enzymes, with significant differences between the conclusions reached in different studies (Adams, Perler & Kelly 1995; Hough & Danson 1999). Therefore, direct engineering approaches aiming to optimize available mesophilic enzymes, by changing just a few specific amino acids, are unlikely to lead to a dramatic increase in the stability of an enzyme. A more attractive and valuable alternative might be the bioprospection[7] for novel enzyme variants from organisms living in extreme environments (Hough & Danson 1999).

Deep-sea hydrothermal vents comprise extremes of temperature, pressure, pH, and salinity, somewhat similar to those found in industrial processes. Additionally, they comprise a wide variety of energy sources, enabling diverse life forms to persist by exploiting different metabolic strategies and adaptations. Sea vents thus, are a rich source of naturally tailored enzymes for extreme environments, making these systems the ultimate frontier for industrial enzyme bioprospection.

Table 1.3.1 | **Examples of classes of extremophiles, their environments and applications.** This table was constructed based on Hough & Danson 1999 and Van den Burg 2003.

| Extremophile | Favorable environment for growth | Applications |
|---|---|---|
| Hyperthermophile | Temperatures above 80°C | ◆Proteases for hydrolysis in the food, feed, brewing and baking industries and detergent formulation.<br>◆Glycosyl hydrolases for processing starch, cellulose, chitin, pectin and textiles. |
| Thermophile | Temperatures between 55°C and 80°C | ◆Xylanases for paper bleaching.<br>◆Lipases and esterases for detergent formulations and stereo-specific reactions.<br>◆DNA polymerases for molecular biology; dehydrogenases for oxidation reactions in the fine chemical industry. |
| Psychrophile | Temperatures between -2°C and 20°C | ◆Hydrolases for detergent formulation.<br>◆Proteases for dairy processing.<br>◆Amylases for the baking industry.<br>◆Celullases for the feed industry.<br>◆Lipases for the food and cosmetic industries.<br>◆Dehydrogenases for the engeneiring of different biosensors. |
| Piezophile | Pressures above 400 atm | ◆Food processing and antibiotic production. |
| Halophile | 2 to 5M of NaCl | ◆Proteases for peptide synthesis.<br>◆Dehydrogenases for biocatalysis in organic media. |
| Acidophile | pH below 4 | ◆Amylases for starch processing.<br>◆Cellulases and proteases as feed additives.<br>◆Oxidases for desulfurization of coal acting as sulfur dioxide emission control. |
| Alkaliphile | pH above 9 | ◆Proteases and cellulases for detergent formulations and the feed and food industries. |

Although mesophilic organisms exist in such environments, extremophilic organisms are the ones that should have actual physiological adaptations to such harsh conditions (Adams, Perler & Kelly 1995). To clarify, an extremophile is here regarded as an organism that thrives in, or may require, physical or geochemical extreme conditions (Zhang C. & Kim 2010), and it may fall into several different classes (Table 1.3.1). While these extreme conditions were first defined as those that were detrimental to the majority of life on Earth, evidences show that a great portion of biomass

[7] Bioprospection is the systematic search for valuable products from biological resources.

actually exists in such conditions. The point to be made is that the term 'extremophile' is used in this dissertation just as an operational concept, as suggested by Costa (2015), since it is based on an anthropomorphic point of view of which conditions are extreme, rather than taking into consideration an organism's usual habitat.

When extremophiles started to be unveiled, they were viewed as exotic organisms, studied by only a few research groups. Now, although they still hold their eccentric status, they are more often than not used as sources of novel enzymes (Demirjian, Móris-Varas & Cassidy 2001). Enzymes have been isolated and purified from several groups of organisms, including animals, plants and microorganisms. However, extremozymes are found generally in prokaryotic lineages (Zhang C. & Kim 2010).

Although, for instance, thermal tolerance has been suggested for hydrothermal vent macrofauna, in reality the majority lives below 20ºC. *In vitro* collagen denaturation experiments do not support the idea of cellular adaptation of macroorganisms to life at extreme temperatures (Van Dover & Lutz 2004). *Rimicaris exoculate*, a rather frequent shrimp species around hot vents, does not even tolerate temperatures above 33ºC (Ravaux *et al.* 2003). This means that the main adaptive response of macroorganisms to extreme temperatures may actually be behavioral rather that biochemical (Van Dover & Lutz 2004).

By contrast, prokaryotes tolerate much broader ranges of temperature. The record is held by an archaeon isolated from a vent in Juan de Fuca Ridge, Strain 121, with a doubling time of 24 hours at 121ºC (Kashefi & Lovley 2003). Through the course of evolution, prokaryotes developed a plethora of biochemical characteristics that make them especially able to thrive in a variety of habitats, including the extremes found in deep-sea hydrothermal vents (Figure 1.3.1). Therefore, prokaryotes, and particularly those flourishing on hydrothermal systems, have been the focus of biodiscovery programs aiming to isolate biomolecules with extreme-resisting features. The diversity of prokaryotes in these systems and their catalytic capabilities has not been completely explored and expands continuously as new studies are completed, making it still the most desirable source of biotechnological-level enzymes.

Only a few microbial extremozymes entered the enzyme market till now. Probably the most relatable example for biological sciences researchers is the thermostable DNA polymerase from the bacterium *Thermus aquaticus* used in PCR. However, this organism was firstly found in a hot spring in Yellowstone National Park rather than in a marine environment. The most famous example of a marine derived microbial enzyme is the DNA polymerase from the hyperthermophilic archaeon *Pyrococcus furiosus*. Thermostable DNA polymerases (EC 2.7.7.7)[8] play a major role in widespread molecular biology applications such as DNA amplification and sequencing, but there are other enzymes that can be bioprospected from marine environments that have equally useful applications in molecular biology (Egorova & Antranikian 2005). These might be for example thermostable DNA

[8] EC number, or Enzyme Commission Number, is a classification scheme for enzymes, based on the chemical reactions they catalyze, regulated by the Nomenclature Committee of the International Union of Biochemistry and Molecular Biology (NC-IUBMB), which maintains an updated list of all the enzymes described in the ExplorEnz database (http://www.enzyme-database.org) (McDonald & Tipton 2013).

ligases (EC 6.5.1.1/ EC 6.5.1.2) or even type II restriction endonucleases (EC 3.1.21.4) (Morgan, Xiao & Xu 1998; Egorova & Antranikian 2005).

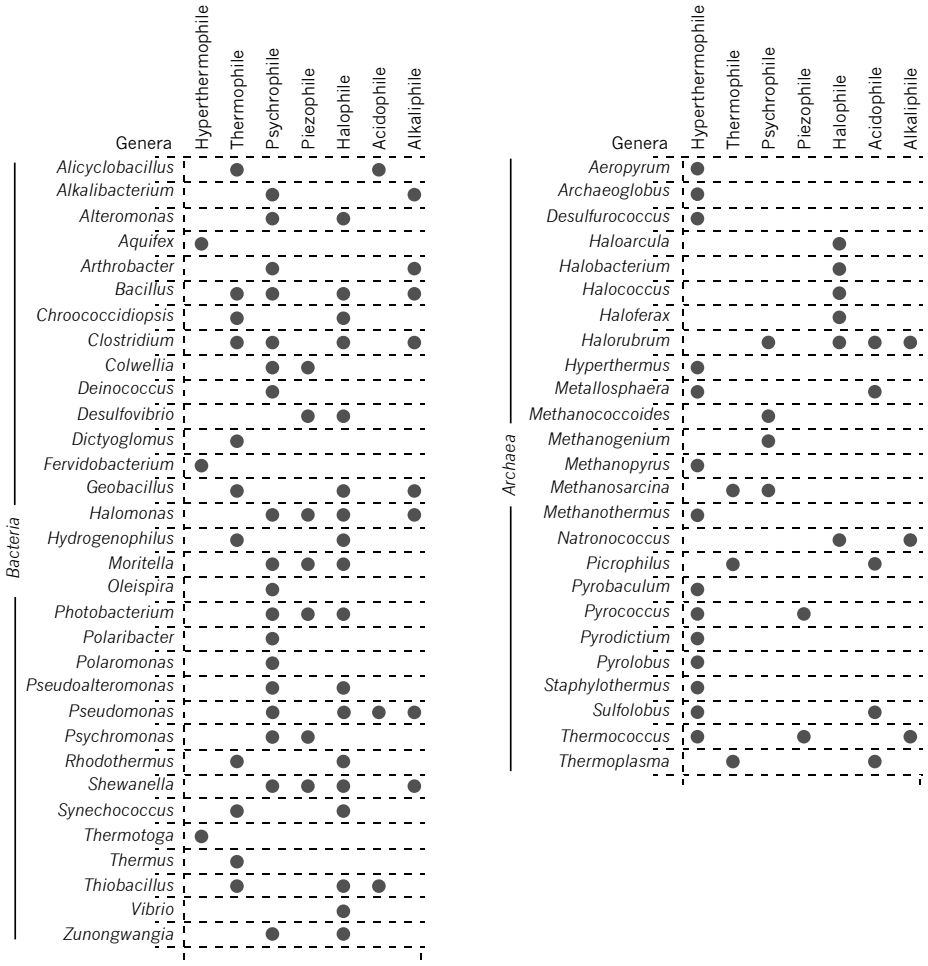


Figure 1.3.1 | **Distribution of extremophilic characteristics in different prokaryotic genera.** The extremophilic characteristics indicated appear in at least one species of each genus. This figure was constructed based on information retrieved from Johnson 1998; Barbieri *et al.* 1999; Cui *et al.* 2006; Romano *et al.* 2006; DasSarma, Coker & DasSarma 2009; Rani, Souche & Goel 2009; Sucharita *et al.* 2009; Poli *et al.* 2012; Yamauchi *et al.* 2013; Dalmaso, Ferreira & Vermelho 2015; Preiss *et al.* 2015.

However interesting these molecular biology enzymes might be, the global enzyme market is rather dominated by biomass-degrading enzymes (Appendix A). Specifically, from more than 3 000 enzymes known and used, approximately 65% are polysaccharide-degrading hydrolases[9] (EC 3.2.1.-, *e.g*. amylases, cellulases, xylanases, mannanases, pectinases and chitinases), peptide hydrolases (EC 3.4.-.-) and lipolytic enzymes (EC 3.1.1.-) (Dalmaso, Ferreira & Vermelho 2015).

Extremophilic versions of biomass-degrading enzymes have attracted much interest in several economically relevant industries such as the food, feed, paper, textile, chemical and pharmaceutical industries (Elleuche *et al.* 2014). They offer flexibility with their potential application in extreme settings. It is this imminent potential and expected economical value that justifies the investment in the bioprospection for these enzymes in extreme-resisting and metabolically versatile deep-sea vent prokaryotes.

[9] Hydrolases are enzymes that catalyze the hydrolysis of chemical bonds of a substrate and are classified as EC 3 (Dalmaso, Ferreira & Vermelho 2015).

## 1.4 Bioprospecting for new enzymes from deep-sea vent prokaryotes

The classical route for microbial enzyme bioprospection involves the isolation of a microorganism from a sample, its growth as a pure culture, the screening for the desired activity, followed by the purification of the enzyme and its eventual characterization (Figure 1.4.1). Indeed, isolation and identification of strains for long constituted the obligate primary step for the development of new products from microbial origin. The ability to culture strains and to characterize their physiology and biochemistry is still regarded as an important advantage for bioprospection, particularly in the development of whole-cell applications (Joint, Mühling & Querellou 2010). However, the extent of biodiversity that is accessible in culture form is known to be limited (Connon & Giovannoni 2002). Rich media traditionally used to isolate microorganisms often selects for fast-growing lineages, meaning that several organisms with potential interest can remain undisclosed due to lack of suitable culturing methods. Our inability to culture reflects our lack of understanding regarding the ecological and nutritional requirements of the target organisms (Dionisi, Lozada & Olivera 2012).

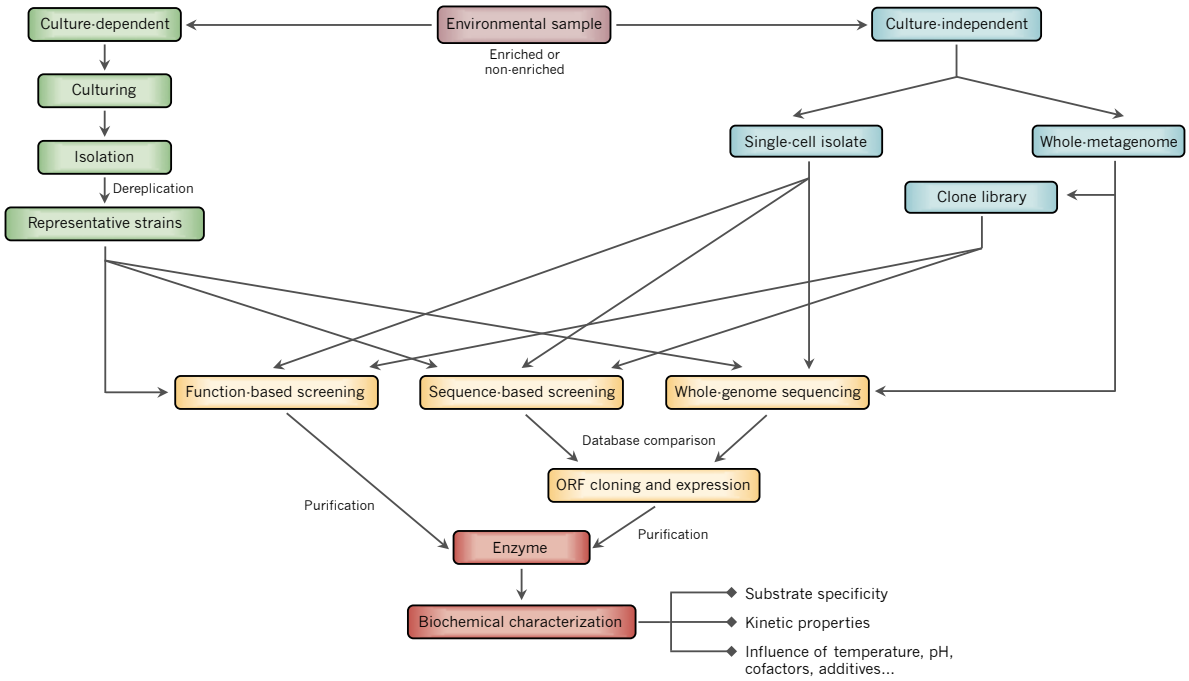


Figure 1.4.1 | **Schematic representation of possible pathways for bioprospecting new enzymes.**

Sea vent microorganisms are particularly challenging to culture since there is an immense complexity involved in their ecosystems. Although many of the organisms in hydrothermal sea vents are mesophilic, several others require extreme conditions to grow, that ultimately are difficult to reproduce in a lab, hindering the chances for successful isolation (Dionisi, Lozada & Olivera 2012). A common approach to increase the number of different isolates obtained from a sample is to slightly tweak or fine-tune growth conditions, for instance by modifying media composition, nutrient sources, pH, incubation temperatures, or the levels of certain gases. Yet, some fastidious groups will still likely be missed (Urbieta *et al.* 2015).

High-throughput cultivation techniques have emerged as a more sophisticated alternative, overcoming some of the limitations associated with traditional culturing methods. Such techniques are, for instance, dilution to extinction and microdroplet encapsulation. Dilution to extinction entails low-density partitioning of cells into, for example, microwell plates. It takes advantage of the observation that the number of culturable strains tends to increase as inoculum density decreases. In theory, it works by reducing interspecific competitive interactions and consequently increasing the probability of isolating weak slow growing competitors (Connon & Giovannoni 2002). Alternatively, microdroplet encapsulation works by enclosing single-cells in a droplet of agarose. The agarose acts as a porous matrix that traps the cell, isolating it, while still enabling the diffusion of nutrients and waste compounds in and out of the droplet, supporting the development of single colonies (Dionisi, Lozada & Olivera 2012). Currently, a second generation of high-throughput methods has already been established for *in situ* isolation and cultivation. The isolation chip (iChip), for instance, is a device composed of several hundred miniature diffusion chambers, which trap in a gel matrix individual microorganisms, allowing them to grown clonally while the chip is exposed to their natural and possibly complex environment (Nichols *et al.* 2010).

When bioprospecting for specific enzymes, pre-enrichment of the sample prior to isolation can provide an attractive means of enhancing the proportion of screening hits. That is, applying an enrichment condition can purposely create a bias on the isolates obtained, selecting for organisms with the desired characteristics in detriment of other lineages. This leads to an inevitable loss of sample diversity but generally improves the discovery efficiency of specific target enzymes. The enrichment can be based on diverse criteria, from nutritional to physical or chemical conditions (Srivastava, Ghosh & Pal 2013). For the discovery of biomass-degrading enzymes, just as in Klippel *et al.* 2014, adding the enzyme substrate in the media is a common enrichment practice.

At the end of the isolation process there should be a rather large collection of isolates that not always represent the community in a non-redundant manner. Dereplication is the term used to describe the process of differentiating strains in order to select only a representative set. This allows to expedite the screening process, thereby minimizing costs and time in sorting large collections of isolates (Goodfellow & Fiedler 2010). Fingerprinting methods are useful for dereplication in discovery programs. For example, RAPD (Random Amplified Polymorphic DNA) fingerprinting uses random primers and low annealing temperatures (*ca*. 37ºC) to randomly amplify several regions of the genome that when electrophoretically separated create a profile that should be identical for clonal isolates (Olive & Bean 1999). Another alternative is to use fingerprinting methods where primers are directed to specific repetitive sequences dispersed in the genome, such as csM13-PCR. In this case, primers are directed to the core sequence of the bacteriophage M13 that is known to appear in multiple copies in the genomes of eukaryotes and prokaryotes (Meyer *et al.* 1993).

Following the construction of a non-redundant collection of isolates, the next step in industrial enzyme bioprospection is the screening for the activity of interest. Traditionally, phenotypic screening, or function-based screening, has been the method of choice and can be done either with whole-organisms or their extracts. The screening methodology depends greatly on the target activity and it is generally a specific and focused test. For instance, when searching for biomass-degrading enzymes,

protein extracts can be screened using spectrophotometric assays, where enzyme activity can be followed by the rate of substrate consumption or product formation, measured as changes in how much light the assay solution absorbs, as in Gobalakrishnan & Sivakumar 2016. Alternatively, in colorimetric assays, enzyme activity is detected by a change of color in the solution, normally by using modified substrates (Nyyssönen *et al.* 2013). When using whole-organisms, the assays can eventually search for growth (Carrasco *et al*. 2016) or halo formation (Ashwini *et al.* 2013) in media supplemented with the target-enzyme's substrate.

A rather novel approach is taking advantage of droplet microfluidics in a similar way as microdroplet encapsulation for isolation purposes. Single-cells are incorporated into droplets with fluorogenic substrates, acting as miniature test tubes where each individual cell can be assayed. The encapsulation in the droplet confines the cell and the fluorescent product to the droplet, and detection and sorting can be done using *e.g.* flow cytometry (Sjostrom *et al.* 2014).

Function-based screening does not need prior knowledge of gene sequence and has the advantage that positive isolates are indeed functional and bear the desired potential (Dionisi, Lozada & Olivera 2012). Moreover, the assays can be adapted to directly pinpoint enzymes with the desired physico-chemical optima. However, these phenotypic approaches are limited to enzymatic activities for which dedicated screening systems can be developed (Podar & Reysenbach 2006).

In the case of sea vent microorganisms, particularly extremophiles, the conditions required for their growth can entail temperatures above 80°C, often anaerobic environments, and media with extreme pH or up to 5 M of sodium chloride. These conditions are incompatible with streamlined phenotypic screenings and with the standard large-scale procedures for production of enzymes. In these cases, sequence-based screening is an alternate and valuable option.

In contrast with function-based screening, the application of sequence-driven approaches involves the use of specific primers or probes, designed based on conserved regions of the genes of interest. Gene-directed PCR has been extensively used to probe for specific biodegradative capabilities in microorganisms (Cottrell *et al.* 2000). However, as a tool for enzyme discovery, it has some major drawbacks. For instance, the design of primers depends on existing gene sequence information and skews the screening in favor of similar sequence types. If the goal is to specifically search for variants of known genes, this might not be a problem. Furthermore, most of the time only a segment of the target gene will be amplified, and additional steps are required to access the full-length gene. Recovery of the flanking regions can also take advantage of PCR-based strategies, such as inverse PCR[10] (Cowan *et al*. 2005).

In sequence-based screening, although the putative function of a gene product can be deduced by sequence comparisons, ultimately the analysis of the expressed product is required. Due to the mentioned problems associated with large-scale culture of extremophiles and production of

[10] Inverse PCR is used when an internal section of a target region is known but the flanking regions are not. Genomic DNA is digested into fragments of a few kilobases by a moderate frequency (6-8 bases) restriction enzyme that does not cut in the known region. Under low DNA concentrations, self-ligation of fragments is induced to give a circular DNA product. PCR is carried out with primers complementary to end-sections of the known sequence so that the full circular fragment is amplified, containing the up- and down-flanking regions (Ochman, Gerber & Hartl 1988).

extremozymes, most applications have to rely on heterologous expression of target genes in more manageable hosts (Cowan *et al*. 2005). It is now well established that expression of recombinant extremozymes in mesophilic hosts, such as *Escherichia coli*, can be done successfully, maintaining their unique properties. However, they are often expressed at very low levels (Vieille & Zeikus 2001). The development of alternative host expression systems is still an area of much needed input (Srivastava, Ghosh & Pal 2013).

The identification of an enzyme is by no means the end of the bioprospecting process. The properties of the enzyme, such as substrate specificity, dependency of cofactors, optimum temperature and pH, activity kinetic parameters, stability to temperature or pH gradients and susceptibility to biosurfactants, must be determined to appraise its potential for the desired application (Soni, Soni & Goyal 2007). This stands true for all bioprospecting approaches.

As our access to genetic information becomes easier, faster and cheaper, new routes to screen for target enzymes emerge (Egan, Thomas & Kjelleberg 2008). Recently, there has been an immense effort to sequence several microbial genomes changing the paradigm of how these organisms are evaluated as sources of promising enzymes. Genomes can be directly interrogated for specific activities based on comparisons to known orthologous gene sequences (Hernández-González & Olmedo-Álvarez 2016). Most importantly, data acquired from whole-genome sequencing enables simultaneous analysis of the potential of an organism for a large set of activities of interest, without the need for a screening method directed for each specific one.

Screening genomic data can depend on primary sequence or motif comparison, or on the evaluation of predicted protein structures and putative catalytic sites matching known enzymes (Cowan *et al*. 2005). The main limitation of this approach is, again, the bias in the detection, favoring variants of known proteins, which represent only a small portion of the total genes within a genome. There are attempts to assign functions to genes with no database homologues, however these systems are in a state of infancy, having only limited success till now (Ijaq *et al.* 2015).

For the purpose of enzyme mining, there should be a careful cost and value analysis of finished genome versus draft sequencing. For many purposes, finishing a genome might be an unnecessary extravagance. It is a costly and time-consuming process that needs a recommended coverage of at least 30-fold, and consists of assembly and gap closure, rigorous quality control steps and error resolution (MacLean, Jones & Studholme 2009). Conversely, draft sequencing may result in missing genes and misassembled regions but can still give us a significant amount of information regarding protein-encoding potential, in a much quicker and less expensive way. As sequencing technologies advance however, complete genome sequencing may soon become common practice.

Note that, genome sequencing is not always a hypothesis driven approach, but rather an exploratory activity that should nevertheless be considered an asset, due to its potential to increase our knowledge base (National Research Council US 2002). Sequencing enables the generation of data that feeds hypothesis driven and biotechnological research.

Bioinformatics is a necessary means to understand and analyze genomic data. It is needed in several steps of the analysis pipeline and overall acts as the interface between the data obtained from

sequencing technologies and the actual knowledge that can be retrieved from it.

Assembly of reads obtained from sequencing technologies is typically one of the first steps in genome analysis and represents a major computational hurdle, particularly for short-read sequencing. This step often limits the biotechnological value of the obtained information. However, if assembly is done with at least some success, enzymes of interest might be identified (Dionisi, Lozada & Olivera 2012). For that purpose, the complex collection of sequences obtained needs to be converted into information that better describes the metabolic capabilities of the organisms, which is accomplished with sequence annotation pipelines (Sakharkar & Chow 2008). These pipelines comprise several software modules, and eventually input from experts, that are able to extract useful information from what seems like simple strings of letters (Stothard & Wishart 2006).

Usually prior to annotation, gene prediction algorithms are applied. Software such as Glimmer (Delcher *et al.* 1999), GeneMark (Besemer & Borodovsky 2005) or Prodigal (Hyatt *et al*. 2010) scan sequences for regions that are likely to encode proteins or functional RNA products - the so-called Open Reading Frames (ORFs) -, based on the current underlying knowledge of gene structure (*e.g.* start and stop codons, regulatory motifs, length, sequence periodicities and sequence entropy).

The identified ORFs are then compared against databases of DNA or proteins in an attempt to identify related entries. If similarity is recognized, based on primary sequence, predicted structure, or gene context, depending on the software, information about the function is transferred to the new sequence, annotating it. Additionally, several other types of information can be appended, such as protein chemical or structural properties and metabolic pathways.

Some of the existing annotation packages work as web-based services and others can be downloaded and run locally on a computer (Stothard & Wishart 2006). The degree to which the annotation procedure is automated also varies. Anyone looking for more flexibility or control over the annotation process can build their own pipeline by merging freely available analysis modules and databases. With the imminent increase of genomic information, the tendency is to rely on completely automated systems (Stothard & Wishart 2006; Médigue & Moszer 2007). Blast2GO and RAST (Rapid Annotation using Subsystem Technology) are two examples of annotation systems that are useful for microbial genomics, the first being a more flexible software and the latter a complete automated server-based system.

Functional annotation in Blast2GO (Conesa *et al.* 2005) is based on homology transfer. The sequences are passed to BLAST – Basic Local Alignment Search Tool (Altschul *et al.* 1990), for comparison against available sequence databases, such as the National Center for Biotechnology Information (NCBI) nr[11], RefSeq or UniProt. From there, Blast2GO extracts information from the top similar hits. This information is transferred to the new sequences and includes Gene Ontology[12] (GO) terms, common enzyme name, Enzyme Commission (EC) numbers and KEGG [13] pathways.

[11] 'nr' stands for 'non-redundant'.
[12] Gene ontology is an extensive and structured vocabulary scheme used for the description of gene product functions (Hill *et al*. 2008).
[13] KEGG, standing for 'Kyoto Encyclopedia of Genes and Genomes' is a database resource integrating several classification tables of genes and enzymes with metabolic pathways (Kanehisa & Goto 2000).

Additionally, Blast2GO enables InterPro searches (Hunter *et al.* 2009) directly from its interface. InterPro conjugates several different databases and provides functional analysis of proteins by classifying them into families and identifying known signatures using predictive models. InterPro IDs can be mapped to GO terms, which can further be merged with BLAST-derived GO terms to provide integrated annotation results. Blast2GO also offers other tools such as PSORTb (Yu *et al.* 2010), which predicts the cellular localization of bacterial proteins. Overall, Blast2GO allows for multiple annotation strategies, taking advantage of several software packages to piece together information concerning the submitted data, which can be either protein, gene or raw genomic sequences.

RAST (Overbeek *et al.* 2014) is a very quick and completely automated online annotation service for bacterial and archaeal genomes. It starts by taking DNA sequences, which are scanned for ORFs by an internal algorithm. Afterwards, rather than BLASTing all entries against a database, it does homology search in a strategic manner, using subsets of protein sequences at each step. For that, RAST takes advantage of a library of protein families, referred to as FIGfams, integrated into a collection of manually curated subsystems. Each subsystem is a set of related functional roles and each FIGfam is a collection of globally similar protein sequences sharing the same function within a specific subsystem. RAST starts by estimating the closest phylogenetic neighbors of the query genome. Once the neighboring genomes have been determined, it collects the set of FIGfams that are present in these genomes. This constitutes the set of FIGfams that are likely to be found in the new genome and the set to which predicted ORFs are going to be primarily compared with. The putative ORFs that remain to be annotated after this stage are compared to the entire collection of FIGfams. Finally, and only then, if still not annotated, the ORFs are BLASTed against a large non-redundant protein database. The service identifies protein-encoding, rRNA and tRNA genes, assigns functions to the genes, predicts which subsystems are represented in the genome and uses this information to reconstruct the metabolic network.

In the initial stages of the field of Genomics, the study of microbial genomes was, as many other methods, dependent on the culture of the microorganism. Currently, we have reached a stage where it is no longer necessary to cultivate a microorganism to be able to access its genomic information and recognize its potential. There are essentially two very different approaches that can be taken: Single-cell NGS and Metagenomics.

Single-cell NGS allows the study of genomes at the level of individual cells. It is useful for the study of some extremophiles, which cannot be cultured despite efforts with other methodologies. For this purpose, systems for the manipulation of single-cells, mentioned before, are coupled with techniques that surpass the problem of limited nucleic acid content of individual cells. Multiple strand displacement amplification[14] enables whole-genome amplification with a uniform representation of the genome, facilitating NGS of single-cell genomes or even transcriptomes (Urbieta *et al.* 2015).

Metagenomics is another field that has surpassed the cultivation requirement of bioprospection. Most microbial communities have a high complexity and can embark hundreds of

[14] Multiple strand displacement amplification allows the amplification of minute amounts of DNA by annealing random hexameric primers to the template DNA and letting ϕ29 DNA polymerase (high-processivity enzyme) synthetize DNA with high fidelity and at constant temperature (Dean *et al.* 2002).

different species, few of which are culturable. Even if most of them were amenable to cultivation, studying each individual species would be currently impractical. Metagenomics allows to unlock the vast amounts of genetic information that are contained in such communities. Generalizing, there is essentially two ways to take advantage of metagenomics in enzyme discovery programs. Expression libraries, prepared by cloning environmental DNA in appropriate vectors, can be subjected to function- or sequence-based screening processes. Each clone of the metagenomic library will contain a fragment of the genome of a microorganism belonging to the community and, in theory, the whole genomic information of all members of the community can be represented if sufficient clones are obtained. Alternatively, total environmental DNA can be purified and sequenced, and, analogously to a genome, enzyme-encoding genes are searched by parsing DNA sequences using databases for comparison. Following bioinformatics analysis, cloning and expression of selected ORFs, industrial relevant enzymes can be unveiled and characterized (Cowan *et al.* 2005).

To conclude, since the ultimate goal of enzyme biodiscovery requires the direct observation of the enzyme's activity, functional assays are still a necessary widespread practice. Nevertheless, as whole-genome sequencing evolves and becomes increasingly more accessible and straightforward, it might develop into a standard screening procedure, with the ability to expedite the first stages of the bioprospection process.

## 1.5 Setting the stage for long-read whole-genome sequencing

Since the development of the first sequencing methods, several different sequencing technologies have emerged, leading to an extraordinary decrease in the cost of sequencing per megabase (Figure 1.5.1), and stirring sequencing systems from a novelty position to a routine part of biological research.

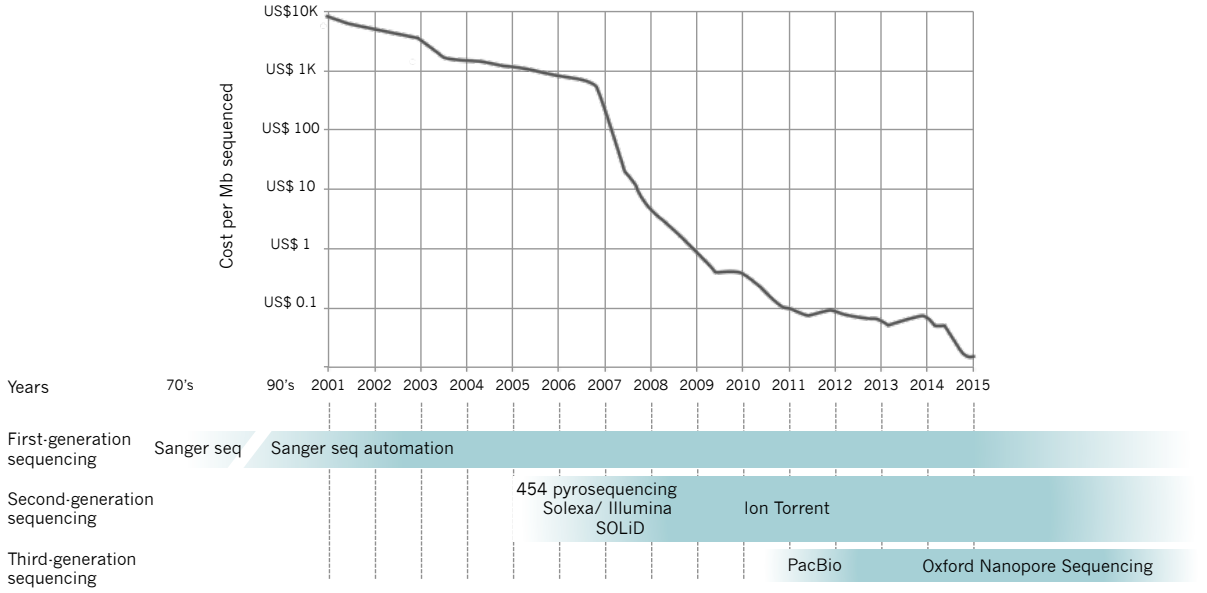


Figure 1.5.1 | **Decrease of sequencing cost through the years and sequencing technologies evolution.** This figure was constructed based on Morey *et al*. 2013 and Wetterstand 2016.

Sanger sequencing was first developed during the 70's and quickly became the sequencing method of choice, prevailing with this status for over 30 years. It enabled the accomplishment of extraordinary feats through the years, the most prominent being the completion of the Human Genome Project in 2003. Several adaptations were made to this technology, including its automation, but overall it is based on a PCR carried out with both deoxynucleotides and chain-terminating dideoxynucleotides. During elongation, some strands eventually incorporate dideoxynucleotides that stop elongation. The collection of different size strands at the end of PCR is then separated in a gel and their terminal base, which is generally labeled, is identified, enabling sequence inference (Morey *et al.* 2013; Goodwin, McPherson & McCombie 2016).

Sanger sequencing can sequence up to 700 bases, which is generally considered a good size read. Yet, it has limited throughput and high cost (Morey *et al.* 2013); indeed, the first human genome sequencing was estimated to cost $0.5–1x10^9$ US dollars (Reuter, Spacek & Snyder 2015). Nevertheless, in the mid-2000's, an all-new generation of sequencing platforms emerged, finally offering true high-throughputs and, with that, the steepest drop in the cost of sequencing ever observed. Next-generation sequencing - or second-generation sequencing -, provided us with enormous quantities of data, but errors that are slightly higher than those of Sanger sequencing and reads that are much shorter (Goodwin, McPherson & McCombie 2016).

Several short-read sequencing technologies have appeared through the years, which can be divided into two broader groups: 'sequencing by ligation' (*e.g.* SOLiD) and 'sequencing by synthesis', the latter being dependent on a polymerase. 'Sequencing by synthesis' technologies can be further subdivided into 'cyclic reversible termination' (*e.g.* Illumina platform) and 'single nucleotide addition' (*e.g.* 454 pyrosequencing platform - discontinued -, and Ion Torrent). 'Cyclic reversible termination' technologies take advantage of labeled 3' blocked nucleotides, which prevent further elongation when incorporated into a strand. Following nucleotide identification, the blocking is reversed and the process continues in a cyclic manner. In technologies of 'single nucleotide addition' there is no need for 3' blocked deoxynucleotides. The four different nucleotides are added sequentially instead of simultaneously, and the incorporation at each addition step is detected by a signal that depends on the technology itself (Delseny, Han & Hsing 2010; Goodwin, McPherson & McCombie 2016).

Different second-generation technologies also use different strategies to generate their sequence libraries. But overall, they depend on the generation of clonal templates of each sequence on a solid surface. Having many thousands of identical copies of a DNA fragment in a defined area ensures a strong enough signal to surpass the lower limits of the detection systems used. A sequencing platform can collect, simultaneously, signals from several million localized reaction centers, thus sequencing many DNA molecules in parallel and leading to the high-throughput characteristic of these technologies. For a comprehensive review on sequencing technologies see Goodwin, McPherson & McCombie 2016.

Currently, Illumina still holds the largest market share of sequencing technologies, offering a very broad range of sequencing instruments, from lower (MiniSeq) to ultra-high-throughputs (HiSeq X). Read lengths go up to 300 bases, with an average accuracy of 99.50%. Since it fits into the 'cyclic

reversible termination' category, it also displays low homopolymer-derived[15] errors - common in other sequencing technologies. Yet, Illumina data has under-representation of AT-rich and GC-rich regions, as well as some substitution errors (Goodwin, McPherson & McCombie 2016).

Whole-genome sequencing has emerged as one of the most widely used applications of NGS, particularly within microbiology fields. Researchers can now have a broader comprehension of the genetic and genomic information and its biological implications. However, short read lengths add much of a challenge in reconstructing *de novo* genomes, making it difficult to resolve the order of some contiguous sequences. The hurdle of assembling short reads brought on a rather fortunate advance in computational biology, with the development of new algorithms that have, in the last years, been the major approach to bypass the short-read problem (Lavezzo *et al.* 2016).

Alternatively, there are synthetic approaches of generating longer reads from short-read sequencing technologies. These approaches fragment template molecules, which are subsequently ligated to barcode sequences, and sequenced in standard NGS instrumentation. Following sequencing, fragments can be split *in silico* by barcode and reassembled with the notion that sharing barcodes means the fragments come from the same original template molecule (Goodwin, McPherson & McCombie 2016).

However, the sequencing scenario is changing again, since technical advances have allowed for the development of the so-called third-generation sequencing technologies. Third-generation sequencing has brought two major improvements for whole-genome sequencing: the ability to sequence single molecules, avoiding the errors and biases introduced during clonal amplification by PCR, and, most importantly, the massive increase of read length. Long reads help to ameliorate the hurdle of sequence assembly. They can span long regions of the genome. This includes complex or repetitive regions that even with high-end computational assembly systems are still an issue for second-generation sequencing (Lavezzo *et al.* 2016). These characteristics make third-generation sequencing not only suitable for projects involving *de novo* assembly of small bacterial (Loman & Quick & Simpson 2015) and viral genomes (Wang J. *et al.* 2015) but also for genome finishing and improvement of reference genomes (English *et al.* 2012). Reconstructing genome structural variation (Norris *et al.* 2016) and isoform usage in transcriptomes (Sharon *et al.* 2013) are also applications where these technologies have advantages over their short-read counterparts.

There are two different third-generation technologies in the market at the time of this dissertation: Pacific Biosciences Sequencing and Nanopore Sequencing.

The Pacific Biosciences (PacBio) sequencing technology takes advantage of a stationary polymerase covalently linked to a well with a transparent bottom. It uses a flow cell with several thousands of these narrow wells where single template molecules are sequenced by progressing through the fixed polymerase. Incorporation of labeled nucleotides is followed continuously with a laser and a camera. They record, in real-time, duration and emitted light as the incorporated nucleotides momentarily pause at the bottom of the transparent well. The polymerase then cleaves the fluorophore, leading to its diffusion away from the sensor before the next labeled nucleotide is

[15] A homopolymer, in this context, refers to regions of DNA consisting of repetitions of the same base.

incorporated. This technology can produce reads longer than 20 kilobases, however the error rates can go up to 20%. This issue can be surpassed to some extent by the circularization of the template. Using circular templates allows for multiple passages of the same molecule through the polymerase and consequently increases accuracy up to 99.90%, by creating a final consensus sequence. However, this process only works for templates no longer than 3 kilobases since it is limited by the polymerase lifetime. The PacBio RS II instrument is large and has a high capital cost. Although the operational cost per sample can be quite low, it is not straightforwardly implemented and still requires extensive infrastructure, making it more suitable for dedicated research/sequencing centers (MacLean, Jones & Studholme 2009; Goodwin, McPherson & McCombie 2016; Lavezzo *et al.* 2016).

Single-molecule sequencing using biological nanopores was proposed nearly 20 years ago, but only in the spring of 2014 did Oxford Nanopore Technologies release the first true nanopore sequencer, the MinION. The MinION was initially released in an early access program, the MinION Access Program (MAP) [16], to a set of participating research groups that were required to test the system. Finally, as of mid-2016, the MinION reached commercial-level.

Like the PacBio system, it fits into the third-generation sequencing category, and it is able to deliver long-read and real-time sequencing of individual molecules. However, the MinION nanopore sequencer has two very distinctive features. Firstly, it is as its name implies nanopore-based, rather than synthesis-based. This means it does not depend on the monitorization of nucleotide incorporation by a secondary signal, but instead directly detects the DNA base composition. Lastly, it is an USB-powered small and portable machine (Figure 1.5.2 C), no larger than a typical smartphone (measuring 10 x 3 x 2 cm). Additionally, and contrary to PacBio, the MinION requires a small initial investment, making it accessible to smaller research groups and bringing ownership of the sequencing process back to the researchers, instead of big sequencing companies (Brown 2015; Brown 2016).

Nanopore sequencing technology has advanced tremendously since its first available version. Nevertheless, most of the work published till now was done with version R7 and R7.3. This dissertation also took advantage of the R7.3 chemistry, the Mk I device, and the software and algorithms developed for these versions, described below.

In nanopore-based sequencing, biological engineered nanopores are embedded in an electrically resistant polymer membrane (Figure 1.5.2 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 1.5.2 B), which in turn can be decoded into sequences.

If left unattended, the DNA would cross the pores at speeds that would not allow appropriate sequence discrimination. For that reason, a motor protein is added during sequencing library preparation, attached to an adapter that will be linked to the DNA (Figure 1.5.2 A). This protein acts as a ratchet break, unwinding the double stranded DNA as it feeds one of the chains through the

[16] Some of the information that is presented in this dissertation regarding the MinION and nanopore sequencing was disclosed to MAP members and it is not publically available.

nanopore, base by base, controlling the speed at which the DNA traverses the pore. Additionally, a tether is also attached to the same adapter (Figure 1.5.2 A), which acts to concentrate the DNA on the surface of the membranes near the pores. Tethering greatly increases the sensitivity of the technology, maximizing the amount of DNA read by pore whilst maintaining input material requirements in the low nanogram range (Brown 2015; Brown 2016).

In the squiggles resulting from a DNA molecule passing through the nanopore, shifts in current are representative of specific *k*-mers, rather than single bases. That is, although the DNA moves one base at a time through the pore, the specific interference in current at each point in time results from (and is characteristic of) a particular oligomer with *k* bases, where length *k* depends on the version of the technology. That means that instead of existing only 4 possible current signatures, one for each base, there are more than 1 000 possibilities for pores that read *k*-mers of length 5, for instance. Thus, basecalling, *i.e.* the process of assigning bases to the squiggle lines, is not straightforward. In fact, the algorithms that model this data are constantly being improved. Particularly, the method still struggles with homopolymeric regions and modified bases; modified bases will typically alter the current shift produced by a given *k*-mer.

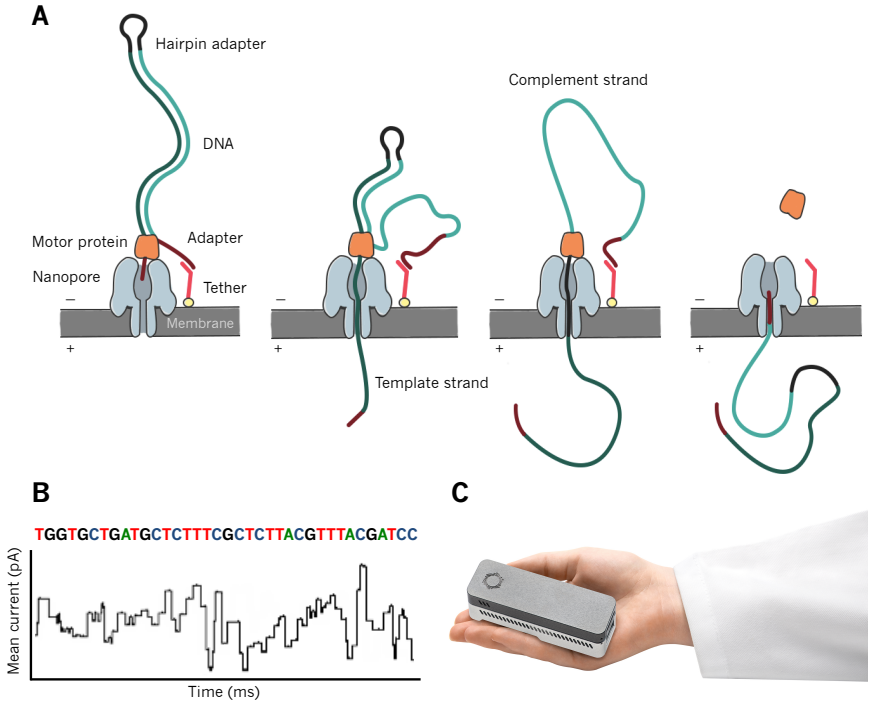


Figure 1.5.2 | **Schematic representation of nanopore-based sequencing of 2D reads (A) and squiggle lines resulting from a DNA molecule passing the nanopore (B); MinION sequencer picture – property of Oxford Nanopore Technologies (C).** For 2D nanopore sequencing (A), a hairpin adapter is added during 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 (B).

Hence, for the R7/R7.3 version of the technology, there is still a high error rate associated with the basecalling process, reaching up to 30%. The reads resulting from R7/R7.3 nanopore sequencing are mostly dominated by indel errors, but since these errors seem to be randomly distributed,

sufficient depth of sequencing can overcome this limitation. Additionally, to further decrease the error rates, similarly to the circular template used by PacBio, nanopore sequencing uses a hairpined sequence library that allows for the contiguous translocation of both forward and reverse strands of a single DNA molecule. Sequencing information from both strands can eventually be conjugated to generate a consensus sequence, which will have a higher quality score.

The hairpin adapter is added during library preparation and links both strands of DNA (Figure 1.5.2 A). When the forward DNA strand passes through the pore, it is followed by the hairpin sequence, and finally the reverse strand. The forward or template strand generates what is called a '1D Template' read (D standing for direction), whilst the reverse or complement strand generates a '1D Complement' read. The consensus sequence obtained from the joint analysis of template and complement reads is called a '2D' read. Note that not all fragments that pass through the pore have success in generating 2D reads; some only generate template reads, others template and complement, and finally only a small minority has sufficient quality to produce a 2D consensus from their template and complement reads (Brown 2015; Brown 2016).

Contrary to other technologies, the read lengths offered by nanopore sequencing have no theoretical instrument-imposed limitation. That is, provided that the DNA is kept intact during library preparation, there is no upper limit for the read lengths obtained (Brown 2015; Brown 2016). Some groups have obtained multiple reads with over 100 kilobases (Urban *et al.* 2015) and the maximum size ever reported for an alignable read is of approximately 255 kilobases 2D (Brown 2016). Sample manipulation and library preparation is key to obtain the desired read length. Fortunately, library preparation is very minimal, involving few pipetting steps to ligate the sequencing adapters.

As mentioned before, long reads have conceivable applications not accessible to short-read sequencing technologies. They can span and resolve large repetitive regions, evidence genome structural variation or reveal complex genomic structures. For instance, Ashton *et al.* (2015) used nanopore-sequencing long reads to identify the position and structure of a bacterial antibiotic resistance island. Additionally, long reads are also ideal for *de novo* genome assembly, for improvement of the contiguity of genome assemblies and for complete-transcript sequencing. Due to the high error rates of the nanopore-sequencing long reads, some of the first trials to assemble small bacterial genomes with these data took use of an hybrid approach, where the long and error-prone reads were used as a scaffold to map accurate Illumina reads (Madoui *et al.* 2015). But quickly it became clear that, despite the error rates, nanopore reads can be used alone to assemble complete bacterial genomes with an accurate reconstruction of gene order, which can be further improved by applying error-correcting algorithms (Loman, Quick & Simpson 2015).

Nevertheless, the most distinctive characteristic of nanopore sequencing, and specifically the MinION device, is its portability and small footprint. Other sequencing technologies are very difficult to employ in remote locations, where availability of infrastructure, laboratory space and trained personnel might be limited. Particularly, sequencing instruments that depend on optical sensing require repeated calibrations by specialized engineers. The MinION, however, runs off a personal computer and can be implemented in a very straightforward manner. This gives the MinION superior convenience for its use

in near-patient clinical tests or environmental monitoring in hard-to-reach field locations. Quick *et al.* (2016) for instance, already put the portability attribute to test by using it for Ebola surveillance in Guinea. They have shown that the device can be established quickly to monitor outbreaks in a resource-limited setting. Yet the most extreme application of the MinION's probably is its deployment to space, into the microgravity environment of the International Space Station.

Although some equipment is still required for library preparation, improvements in the protocol and equipment miniaturization can potentially reduce the space required for a fully-functional sequencing bench to the size of a briefcase. Currently, Oxford Nanopore Technologies is working towards the reduction of library preparation steps and automation of the protocol into lab-on-chip devices such as the Voltrax and Zumbador, that would integrate with the MinION and feed it the sequencing library directly (Brown 2015; Brown 2016).

The MinION Mk I device works with reusable, but ultimately disposable flow cells that have a sensor array over an electrical detection grid. The sensor array contains 2 048 nanopores inserted into a proprietary membrane across a microsupport to provide structure. These are in turn connected to an application-specific integrated circuit chip with 512 sensing channels, capable of individual sequencing at approximately 70 bases per second. That means that a maximum of 512 nanopores can sequence independently at the same time. At the beginning of an experiment, a scan is conducted to determine and choose the best-working nanopores. The remainders that were not selected in the first scan are still available for use later on during the experiment. Throughout the sequencing experiment the number of working pores decreases till it eventually reaches a point where there are no more available pores and the flow cell is no longer usable.

The number of actually working-pores of a R7/R7.3 flow cell at the beginning of an experiment is typically less than the 2 048 total pores, mainly due to the nature of the fabrication process, but also due to the storage/delivery conditions to which the particular flow cell was subjected (Brown 2015; Brown 2016). That means that there is a lot a variability in the number of active nanopores between flow cells, and since this number correlates with the yield of sequencing, different flow cells will generally yield different amounts of data.

Quick, Quinlan & Loman (2014) reported 247.00 megabases of 1D reads, 64.53 megabases of 2D reads, of which only 55.68 megabases had quality scores higher than 9, the so called '2D Pass' reads ('Pass' referring to passing quality filters). However, other groups have reported yields much lower than that, or much higher, reaching up to 2 gigabases of throughput with 50-70% of 2D Pass reads. If higher throughputs are required, one can chose to use PromethION. The PromethION is a different instrument that still takes advantage of nanopore sequencing. However, it has a completely different intended purpose, since it is a larger machine that has some infrastructural requirements. It has 48 individual flow cells that can work simultaneously to generate 2-4 terabases in a 2-day run, placing it as an ultra-high-throughput platform, equivalent to some NGS instruments in the market (Brown 2015; Brown 2016).

Nanopore sequencing also enables real-time analysis, meaning that there is no need to wait till the end of an experiment to get access to sequence information. Since each nanopore sequences

independently, as soon as a strand translocates through a pore, a file can be written with the squiggle lines, giving us immediate access to the data while the experiment keeps progressing. The squiggles are recorded as fast5 files by the MinKNOW software, the controller software of the MinION, one file for each DNA strand. At this point, the files do not contain sequence information. To obtain the sequences, typically, the files are streamed to the cloud-based EPI2ME Metrichor platform for basecalling. The DNA sequences that are generated during the online basecalling are then downloaded into the users computer and made available for further analysis (Brown 2015).

Decreasing the sample-to-response time by doing real-time analysis is a major upgrade over second-generation technologies. It creates opportunities for the development of new time-sensitive applications, the most straightforward being its use for diagnostic purposes or antibiotic resistance profiling. Furthermore, it allows for more flexibility in sequencing experiments, since now we are able to control and determine when sufficient information has been collected and the sequencing run can be stopped.

But in order to take advantage of the real-time potential of nanopore sequencing there is a need to develop streaming bioinformatics algorithms, which continuously update their inference about the sample as each sequence read is generated. WIMP – 'What's in my pot', is the first example of an application that takes advantage of the real-time feature of nanopore sequencing. It is an analysis pipeline that is able to classify and identify microbial species in real-time (Juul *et al.* 2015). Additionally, some open-source software packages have already been developed to facilitate real-time analysis. For instance npReader (Cao *et al*. 2015), continuously scans the folder where the sequencing data is saved, extracts sequences and streams them to downstream pipelines of our liking, allowing for the construction of tailored real-time analysis systems.

The real-time feature of nanopore sequencing can go even further. It is possible to preemptively analyze the first portion of a sequence while the DNA molecule is still being translocated through the pore. This feature has already enabled a primitive version of selective sequencing called 'Read Until'. In Read Until (Loose, Malla & Stout 2016), the sequencing of a specific molecule can be interrupted, based on the information obtained from its already sequenced first portion; the molecule is rejected by reversing the potential across the nanopore. The applications of this methodology are still not even recognized to the fullest. It may allow, for instance, the rejection of a specific organism's DNA from a mixed pool of DNA molecules, or control the representation of barcoded sequences.

Typical bioinformatics algorithms for sequence analysis and assembly are not natively able to handle long reads or error rates of the level of third-generation sequencing, since they were designed with accurate short reads in mind. Thus, standard algorithms used for NGS do not scale up well to properly deal with the types of reads generated by nanopore sequencing. This eventually motivated the development of a new cohort of algorithms specifically constructed to perform well with long and error-prone reads; from tools for *de novo* assembly like Canu (Koren *et al.* 2016), to methods for error correction, such as Nanopolish (Loman, Quck & Simpson 2015), and many other analytical pipelines for very specific applications in different fields (Juul *et al.* 2015). The plethora of options for dealing with nanopore-sequencing data is now bewildering. However, the choice of bioinformatics solutions

should take into consideration the biological problematic or the specific application, which will eventually limit the options.

As previously mentioned, nanopore sequencing technology has been advancing at a steep pace since it first came out, with several different chemistries released, meaning different engineered pores with different signal characteristics, different motor proteins, tethers, membranes or run conditions, that improve yields and decrease error rates. Even during the time of this dissertation, major changes were made to the nanopore technology. The previous R7/R7.3 versions were discontinued and the Mk II device was released, and with it a new flow cell design and an entirely different chemistry, the R9 (now the R9.5). The R9 nanopore - disclosed to be a modified CsgG *E. coli* amyloid secretion pore (Brown 2016) -, brought on a new paradigm for the technology. It generates squiggles that are more spread and allow higher discrimination. Consequently, the error rates of 1D reads decreased, reaching the level of accuracy of 2D reads. That is, 1D sequencing can now replace 2D sequencing, with a major drop on time and complexity of library preparation (10 minutes). This improvement in read accuracy did not originate only from the evolution of the chemistry, but also from the development of new algorithms to perform basecalling. Basecalling algorithms are shifting towards Recurrent Neural Networks, which are able to learn and improve their predictions overtime, ultimately decreasing basecalling errors. Additionally, focused basecalling software (*e.g.* Scrappie) is being developed to *post hoc* correct homopolymers - a weak point of the technology -, and modified DNA bases are already being modeled from the nanopore data, with potential applications in epigenomic studies (McIntyre *et al.* 2017).

Other lines of development are also being followed. For instance, the new system throughput scaled up to 9 gigabases per flow cell with the introduction of the 'fast mode', sequencing 450 bases per second (rather then 70). Additionally, nanopore sequencing is now completely independent of Internet connection - the online basecalling was substituted by Albacore, an algorithm that runs locally on a personal computer -, and sequencing reagents are bound to become free from refrigeration requirements. This adds up to the portability aspect of nanopore sequencing. Moreover, further miniaturization of the system is in progress, with the development of SmidgION. SmidgION will work attached to, and powered by, a mobile phone, and will be able to perform DNA sequencing and eventually other focused nanopore-sensing assays (Brown 2016).

In summary, third-generation sequencing, specifically nanopore sequencing, enables portable, long- and single-read sequencing in real-time. As it continues to improve, and error rates diminish to reach NGS standards, it may quickly become a viable stand-alone option for whole-genome sequencing, with applications on all types of sequencing projects.

## 1.6 Portugal: A privileged place for marine and hydrothermal vent exploration

Due to its fortunate coastal position, Portugal detains exclusive economic rights over a very large area of the North Atlantic Ocean. That includes all natural resources of the water column, the soil

and subsoil of the sea that extends 200 nautical miles from both the Azores and Madeira islands and the mainland coastline (Figure 1.6.1). This zone surrounding Portugal's landmass, *i.e.* the Exclusive Economic Zone (EEZ), delimitates 1.7 million km$^2$. It represents nearly 18 times the country's land area (Rodrigues *et al.* 2011), and confines a variety of aquatic ecosystems of the highest interest, as for instance both the Menez Gwen and Lucky Strike hydrothermal vents (Glowka 2003). Thus, Portugal has privileged access to a relevant compartment of marine biodiversity.

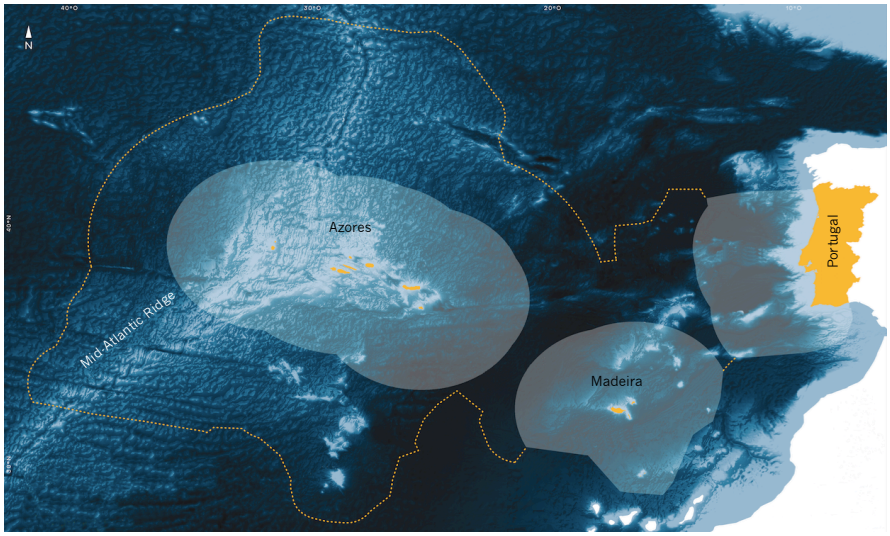


Figure 1.6.1 | **Portugal's EEZ and proposed limits for the continental shelf extension.** Light blue regions represent Portugal's EEZ and orange dashed lines delimit the continental shelf proposed by the EMEPC - Estrutura de Missão para a Extensão da Plataforma Continental. This figure was adapted from the original map created by the EMEPC.

Some efforts have been made to ensure an even more strategic position of Portugal in marine sovereignty. In 2005, the EMEPC - Estrutura de Missão para a Extensão da Plataforma Continental, was created, with the goal to extend Portugal's continental shelf. To clarify, the continental shelf is the submarine continental landmass that stretches from the shoreline of a country. Portugal's continental shelf extends further than the defined 200 nautical miles. It should in fact reach at least 350 nautical miles, equating to an area of approximately 4 million km$^2$ (Figure 1.6.1), 42 times the size of the country's dry land area (Barriga *et al.* 2013). On May 11, 2009, The EMEPC presented a proposal to the Commission on the Limits for the Extension of Continental Shelf created under the United Nation Convention on the Law of the Sea[17]. If accepted[18], Portugal will exert jurisdiction over the resources that are contained in the soil and subsoil but not the water column of the extended continental shelf (Estrutura de Missão para a Extensão da Plataforma Continental 2015). Regardless, Portugal will still have exploration rights not only over gas, metals or minerals that might exist, but also over the seafloor biological resources, that are a major source of biotechnological potential. Indeed, this area would enclose a very extensive portion of the Mid-Atlantic Ridge, with numerous hydrothermal vent fields of enormous potential.

[17] The United Nations Convention on the Law of the Sea of 1982 established guidelines for the rights and restraints of the world nations over the ocean resources.
[18] It is expected that the proposal starts to be evaluated around 2017-2018, where the EMEPC also intends to add a new addend with new information that supports the case. Only in 2020 should the process be concluded (Firmino 2016).

Overall, the sea is a staple of national identity, it is entailed in the country's history, and supports a key set of economical activities[19]. The marine biodiversity under Portugal sovereignty is one of the country's greatest richness, but one that is still immensely underexplored. Over the years, some initiatives have been tiptoeing into the hydrothermalism around the Azores region, one of which was the SEAHMA project, followed by the SEAVENTzymes project, described bellow. Yet, Portugal still falls short in the sector of marine biotechnology (Rodrigues *et al.* 2011). Indeed, few countries have developed structured programs for national research focusing on marine biotechnology, even thought, its global market is estimated to reach a value of *ca.* 4.8 x $10^9$ US dollars in 2020 (Global Industry Analysts, Inc. 2015). Grasping the potential in the Portuguese seas and hydrothermal vents would be a major propeller of the country's economy.

## 1.7 The SEAHMA project

The 2002 SEAHMA (SEAfloor and subseafloor Hydrothermal Modeling in the Azores sea) project was a large proposal which fell into the scope of the discoveries, made at the time, on the Mid-Atlantic Ridge near the Azores islands, namely the Rainbow and the Saldanha hydrothermal vent fields. This project had the aim of broadening our understanding of hydrothermalism in the region and thus, it integrated with a larger initiative, the InterRidge MOMAR (MOnitoring the Mid-Atlantic Ridge). This initiative selected the Mid-Atlantic Ridge as a target for multidisciplinary studies for years to come, focusing particularly on the five main hydrothermal camps of the region (Table 1.7.1): Lucky Strike, Menez Gwen, Menez Hom, Mount Saldanha and Rainbow.

Table 1.7.1 | **General characteristics of the hydrothermal vents visited during the SEAHMA project.** Information was retrieved from Kádár *et al.* 2005, Colaço *et al.* 2006 and Dias & Barriga 2006.

| Hydrothermal vent | Characteristics |
|---|---|
| Lucky Strike (37° 17' N; 32° 16' W) | ◆Depths of 1550-3000 m.<br>◆21 active chimney sites.<br>◆Temperatures between 152-333°C.<br>◆pH between 3.5-4.9.<br>◆Fluids with low sulfur and high methane contents. |
| Menez Gwen (37° 50' N; 31° 31' W) | ◆Depths of 800-1000 m.<br>◆High hydrothermal activity.<br>◆Temperatures between 265-283°C.<br>◆pH between 4.2-4.9.<br>◆Fluids with low content of metals, hydrogen and hydrogen sulfide and high content of methane. |
| Menez Hom (37° 09' N; 32° 26' W) | ◆Poorly understood vent field that is probably still in formation.<br>◆Diffuse vents with no detected chimneys.<br>◆Low temperature.<br>◆Fluids with high methane content. |
| Mount Saldanha (36° 33' N; 33° 25' W) | ◆Depths of 2100-3150 m but elevates up to 800 m.<br>◆Reduced hydrothermal activity with diffuse vents and no chimneys.<br>◆Temperatures only 3-4°C higher that the surrounding water.<br>◆Fluids with high content of methane, metals and sulfur oxides. |
| Rainbow (35° 13' N; 33° 54' W) | ◆Depths of 2230-2500 m.<br>◆Black smoker type chimneys.<br>◆Temperature between 360-400°C.<br>◆pH between 2.8 and 3.1.<br>◆Fluids enriched with metals, hydrogen, methane and carbon dioxide. |

[19] Coastal tourism and fisheries are the major sectors of the country's marine activities (European Commission Directorate-General for Maritime Affairs and Fisheries 2008).

Thus, between the 29th of July and the 14th of August of 2002, these five hydrothermal fields were visited by the Portuguese research cruise SEAHMA-1. On board of the L'Atalante research vessel of IFREMER - Institut Français de Recherche pour l'Exploitation de la MÉR (Figure 1.7.1 A), several surveys were performed along a 270 km line southwest of the Azores archipelago. These surveys consisted not only of geophysical and electromagnetic recordings, but also of extensive observations on the biosphere, taking advantage of the Victor 6000 ROV (Figure 1.7.1 B) to collect a multiplicity of samples (Figure 1.7.1 C).

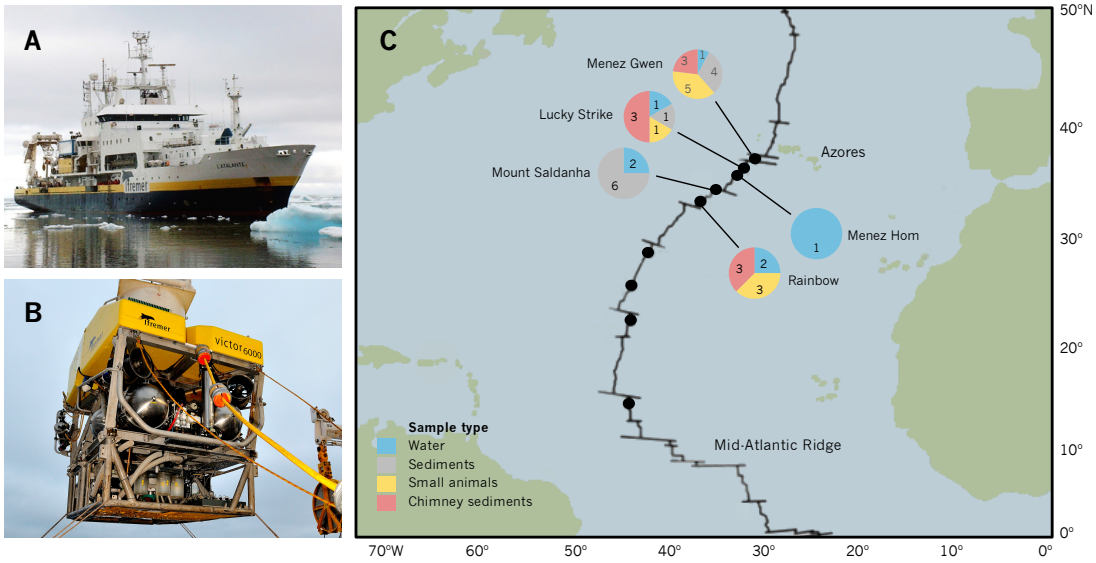


Figure 1.7.1 | **L'Atalante ship (A) and the Victor 6000 ROV (B) used during the SEAHMA project; map of sampled hydrothermal vents during the SEAHMA project (C).** Pictures A and B are property of IFREMER (Institut Français de Recherche pour l'Exploitation de la MÉR). In map C, each hydrothermal sea vent is accompanied by a piechart representing the number and types of samples retrieved from each site.

A total of 36 samples were collected by the Victor 6000 ROV and brought on board for microbial biodiversity analysis. To evaluate a broad range of niches, different types of samples were taken, from water, to seafloor sediments, chimney sediments and small animals such as *Microcaris* sp. and *Rimicaris* sp. shrimp and *Bathymodiolus azoricus* mussels.

Two different approaches were taken to study the prokaryotic diversity enclosed in these samples. The first approach was molecular-based, by PCR-TGGE, with the use of universal primers for *Bacteria* and *Archaea* (Tenreiro 2005). The second approach was culture-based, and took advantage of sea salt culture media to isolate marine aerobic and anaerobic bacteria and archaea. For more information on sample processing and prokaryotic isolation see Appendix B.

At the end of the project, a total of 296 prokaryotes had been isolated in axenic conditions and characterized by a polyphasic approach. This polyphasic approach was done with multiple methods of fingerprinting (Appendix B, Figure B.2) with differential discriminant power. Three different DNA fingerprinting techniques were used, csM13, RAPD PH[20] and RAPD 1281[21], targeting different DNA

[20] RAPD PH uses the primer PH described in Massol-Deya *et al.* 1995, which is directed to the 3' extremity of the 16S rRNA gene. It may allow discrimination at the level of species.
[21] RAPD 1281 uses the primer 1281 described in Akopyanz *et al.* 1992, which is an arbitrary chosen sequence that enables infraspecific discrimination.

regions. Additionally, protein fingerprinting by whole-cell protein profiling using SDS-PAGE – Sodium Dodecyl Sulfate PolyAcrylamide Gel Electrophoresis –, was also performed (Tenreiro 2005). While the DNA fingerprinting methods referred above are very useful to characterize infraspecific diversity, and possible clonal relationships, the whole-cell protein profiling has been shown to be correlated with DNA-DNA hybridization (Vandamme *et al.* 1997), which is known to enable species discrimination. From the composite dendrogram integrating all generated profiles, a dereplication of clonal isolates allowed for partial 16 rRNA gene sequence analysis of a smaller representative set. The selected isolates were identified as belonging to a diverse array of genera (*e.g. Alkaliphilus*, *Azospirillum*, *Caminibacillus*, *Clostridium*, *Desulfovibrio*, *Marinobacter*, *Propionibacterium*, *Pseudomonas*, *Thermococcus*, *Thermotoga* and others), some of which representing novel species.

At this stage, the SEAVENTbugs collection was formed. This collection of deep-sea hydrothermal vent isolates from the Azores region, besides being of immense interest with regards to understanding vent biodiversity, also constitutes a great source for bioprospecting different bioactive compounds. For instance, the aqueous and organic extracts produced by these microorganisms, were tested for pharmaceutical and cosmetic usage by Bioalvo (Rodrigues *et al*. 2011). Furthermore, the potential of these microorganisms to produce industrial relevant enzymes was also assayed, in a project so-called SEAVENTzymes, described bellow.

## 1.8 The SEAVENTzymes project

The SEAVENTzymes project arose as the natural progression of the SEAHMA project. That is, at the time, a privileged collection of 296 hydrothermal vent isolates was available for further exploration, and there was much interest in the application of these extreme-inhabiting organisms, particularly in the use of their extremozymes as biocatalysts. Thus, the SEAVENTzymes project emerged with the purpose of searching for biotechnologically relevant enzymes in the hydrothermal vent prokaryotes of the SEAVENTbugs collection. Specifically, it aimed for the isolation, cloning and heterologous expression of coding genes of:

(I) Novel thermostable hydrolytic enzymes with industrial applications, such as amylases, cellulases, xylanases, mannanases, pectinases, chitinases, proteases and lipases.

(II) Intracellular enzymes with applications in molecular biology, such as DNA polymerases, DNA ligases and restriction endonucleases of type II.

Two different screening approaches were taken to assess the potential of the isolates: a function-based approach by phenotypic assays and a molecular-based approach using PCR degenerate-primers. In a fist stage of the project, the isolates were screened for the enzymes of interest by classic phenotypic techniques. Since the collection had both mesophilic aerobes and anaerobes and thermophilic anaerobes, the isolates were separated into two operational groups. Anaerobes, either thermophilic or mesophilic, require certain conditions that make the phenotypic screening processes unfeasible or extremely complicated. Thus, this set was not subjected to phenotypic screening. Rather, they were subjected to molecular-based screening by the use of

degenerate PCR primers designed to amplify genes coding for the enzymes of interest. Several pairs of primers were designed and used for different families of the target genes; yet, there were virtually no genes of interest amplified, even with multiple stages of PCR optimization.

Nonetheless, the phenotypic screening of the mesophilic aerobes yielded several positive results, particularly for the production of biomass-degrading enzymes. However, attempts to amplify and isolate the genes responsible for the production of those enzymes failed, just as the molecular-based screening of the anaerobic isolates. Thus, in the end, the project had limited success in the exploitation of the SEAVENTbugs collection.

## 1.9 SEAVENTzymes II: Dissertation purpose and outline

After 13 years from the first instance of the SEAVENTzymes project, several technological advances have emerged, some of which were described throughout this introduction. Yet, the interest in extremozymes for industrial purposes still persists. Fortunately, with the conservation and maintenance of the SEAVENTbugs collection, we are now able to revisit the project with a fresh approach. SEAVENTzymes II is the second more modern attempt of searching for biotechnologically relevant enzymes in the hydrothermal vent prokaryotes of the SEAVENTbugs collection.

Sequencing methods of bioprospecting offer great advantages over the screening approaches taken during the first SEAVENTzymes project. They are quite versatile and can be implemented to bioprospect in a culture-dependent or -independent manner, which is an advantage for the study of fastidious hydrothermal vent microorganisms. Moreover, whole-genome sequencing data allows the prospection of multiple enzymes from a single experiment, with no need for a directed test for each enzyme of interest or particular inducing conditions for enzyme expression. Nanopore sequencing, in particular, is a novel sequencing technology that offers an additional set of potential advantages. Long reads may reduce data processing needs and facilitate enzyme mining. Moreover, its unique portability and real-time implementation can eventually be translated into competitive advantages in the biodiscovery process from locations where sampling is reduced to unique visits, *e.g.* hydrothermal vents.

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 dissertation 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.

# Chapter 2. Materials and methods

## 2.1 Reanalysis of the screening results from the SEAVENTzymes project

### 2.1.1 Data analysis

During the SEAVENTzymes project, a subset of 139 isolates of the SEAVENTbugs collection - the mesophilic aerobic isolates -, was subjected to phenotypic screening methods for the detection of biomass-degrading enzymes with industrial potential. Two different screening methods were applied: growth and colorimetric assays. For a summary on the methods used see Appendix C.

All data resulting from the screening assays was revised and compiled. Cramér's V test was applied to measure association between the two different phenotypic screening methods regarding each targeted enzyme. For that purpose, a contingency table was created with the multivariate distribution of positive versus negative results of both screenings. Both the results of either screening method were non-binary, and for the purpose of constructing the contingency table they were transformed into nominal data, that is, either positive or negative. Colorimetric results were considered negative when no change in color was observed, and positive when a taint appeared, regardless of the intensity of the color. Growth assays were considered positive when NAUCr(ES)/NAUCr(BM)>6, *i.e.* when the relative Net Area Under the growth Curve (NAUCr) obtained in base medium plus the enzymes substrate (ES) was at least 6 times higher than the NAUCr in base medium alone (BM). For more information on the NAUCr calculation see Appendix C Figure C.1. This arbitrary operational limit for the definition of positive results was constructed based on a conservative approach. It corresponds to the maximum difference observed between the normalized NAUCr values of replicates (within the 10% replicates), when excluding outliers by the Modified Thompson Tau test. A chi-squared test was used to infer the statistical significance of the Cramér's V association at the significance level of $\alpha=0.05$.

All results from both screening methods were integrated and subjected to a Principal Component Analysis (PCA) in NTSYSpc version 2.21q (Exeter Software). The PCA was performed

based on a matrix of the results normalized by assay, with the calculation of a correlation matrix and its consequent eigenvalue decomposition.

Additionally, fingerprinting profiles of all screened isolates obtained during the SEAHMA project were used to create a composite dendrogram in BioNumerics (Applied Maths) version 6.6, using the Pearson correlation coefficient and the Unweighted Pair Group Method with Arithmetic mean (UPGMA) clustering algorithm. For more information on how these profiles were obtained see Appendix B Figure B.2. The reproducibility of the composite dendrogram was calculated as the average of the reproducibility of each type of fingerprinting, determined during the SEAHMA project (Appendix B Figure B.2).

## 2.2 Isolate recovery and identification

### 2.2.1 Growth conditions

Isolates of the SEAVENTbugs collection were recovered from cryopreserved cultures maintained in Nutrient Broth (BIOKAR Diagnostics) with 4% (w/v) Sea Salts (Sigma) and 20% (v/v) Glycerol at -80ºC. For the selected isolates, 5 µl of the cryopreserved cultures were streaked onto plates with sterile Marine Broth (Difco) plus 1.5% (w/v) of Bacteriologic Agar (BIOKAR Diagnostics). Inoculated plates were incubated at 22ºC for 3 to 5 days until growth was visible.

### 2.2.2 DNA extraction

Genomic DNA was extracted from plate-grown cultures of the recovered isolates using a modified version of the Guanidium Thiocyanate Method (Pitcher, Saunders & Owen*,* 1989). Cells collected from an agar plate were resuspended in 250 µl of lysis buffer (50 mM Tris; 250 mM Sodium chloride; 50 mM EDTA; 0.3% (w/v) Sodium dodecyl sulfate - SDS; pH 8.0) and 100 µl of microspheres. After a homogenization step in a vortex for 2 minutes, the cells were incubated at 65ºC for 30 minutes, followed by another 2 minutes of homogenization. 250 µl of GES (5 M Guanidium thiocyanate; 10 mM EDTA; 0.5% (w/v) Sarkosyl; pH 8.0) was added to the tube, which was mixed by inversion and incubated on ice for 10 minutes. After the incubation step, 125 µl of cold Ammonium acetate 10 M was added, followed by 500 µl of Chloroform:Isoamyl alcohol 24:1. The tube was again mixed by inversion and centrifuged at maximum speed for 10 minutes. The supernatant was recovered into a new tube and an equal volume of cold Isopropanol was added. Following a centrifugation step at maximum speed for 10 minutes, the supernatant was discarded. The DNA pellet was washed with 1 ml of cold 70% (v/v) Ethanol and centrifuged at maximum speed for 10 minutes. The supernatant was pipetted off and the DNA pellet was left to air dry for 2-3 minutes. Finally the DNA pellet was resuspended in 50 µl of TE buffer (10 mM Tris; 1 mM EDTA; pH 8.0) and stored at -20°C.

### 2.2.3 csM13 and RAPD PH fingerprinting

Two different DNA fingerprinting methods were applied to the recovered isolates, namely csM13 and RAPD PH, which vary in the primer and the annealing temperature used during the PCR protocol. csM13 PCR uses the primer 5' GAGGGTGGCGGTTCT' 3 described by Meyer *et al*. (1993), whilst RAPD PH employs the PH primer 5' AAGGAGGTGATCCAGCCGCA' 3 described by Massol-Deya *et al.* (1995). Each amplification reaction was carried out in a total volume of 25 µl, containing 1X PCR reaction buffer, 3 mM of Magnesium chloride, 2 µM of primer, 100 µM of each of the four deoxynucleotides, 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: 5 minutes of initial denaturation at 94ºC, followed by 40 cycles of denaturation at 94ºC for 1 minute, annealing at 50ºC (csM13) or 37ºC (RAPD PH) for 2 minutes and extension at 72ºC for 2 minutes, with a final extension at 72ºC for 10 minutes. PCR products were separated by electrophoresis in a 1.2% (w/v) agarose gel, with 0.5X TBE buffer (40 mM Tris; 45 mM Boric acid; 1 mM EDTA; pH 8.3) and a constant voltage of 2.5 V/cm for 3 hours. The gel was stained with Ethidium bromide and revealed in an Alliance 4.7 UV transiluminator (UVItec).

### 2.2.4 Partial amplification of the 16S rRNA gene and sequence analysis

16S rRNA gene was partially amplified using the universal primers PA 5' AGAG TTTGATCCTGGCTCAG 3' (Massol-Deva *et al*. 1995) and 907r 5' CCGTCAATTCMTTTRAGTTT 3' (Muyzer *et al*. 1998). Each amplification reaction was carried out in a total volume of 50 µl, containing 1X PCR reaction buffer, 2 mM of Magnesium chloride, 1 µM of each primer, 50 µM of each of the four deoxynucleotides, 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 minutes of initial denaturation at 94ºC, followed by 35 cycles of denaturation at 94ºC for 1 minute, annealing at 55ºC for 1 minute and extension at 72ºC for 1 minute, with a final extension at 72ºC for 3 minutes. The amplification products were purified using the JetQuick PCR Product Purification Spin Kit (Genomed), following the manufacturer's instructions. The purified PCR products were then sequenced by Biopremier (Lisbon, Portugal). Sequencing results were subjected to a BLAST search against the NCBI nucleotide collection (nr/nt) database to determine and retrieve the closest known relatives of the isolates based on partial 16S rRNA sequence comparison. A phylogenetic reconstruction with both the isolates' partial 16S rRNA gene sequences, and their top BLAST hits, was generated using the MEGA[22] software version 7.0.16 (Kumar, Stecher & Tamura 2016). Additionally, the partial gene sequences of the type strains of the type species of each represented genus were included. Sequence alignment was performed by the Clustal version embedded into MEGA and clustering was done on a total of 683 positions using the neighbor-joining algorithm accompanied by a bootstrap analysis of 1000-fold.

[22] MEGA 7 is available at http://www.megasoftware.net.

## 2.3 Whole-genome nanopore sequencing

### 2.3.1 2D genomic DNA library preparation

MG SD 082 isolate was streaked onto a plate of Marine Broth (Difco) plus 1.5% (w/v) of Bacteriologic Agar (BIOKAR Diagnostics) and incubated for 72 hours at 22ºC. Cells were harvested and DNA extraction was performed with the Promega Wizard Genomic DNA Purification Kit, following the manufacturer's instructions for Gram-positive bacteria DNA extraction.

The remainder of the protocol for 2D DNA library preparation, described bellow, was developed by Oxford Nanopore Technologies.

DNA was quantified using a Qubit 2.0 fluorometer (Invitrogen) using the dsDNA BR Assay Kit, as described by the manufacturer, and diluted to approximately 33 ng/µl. 45 µl of the diluted DNA solution, amounting to approximately 1.5 µg of DNA, was loaded into a Covaris fragmentation g-tube and centrifuged at 6000 rpm in an Eppendorf 5424 R centrifuge for 1 minute. The Covaris g-tube was positioned in the centrifuge in the inverted position and centrifuged again for 1 minute.

To the 45 µl of fragmented DNA, 5 µl of Control DNA was added, provided in the Oxford Nanopore Genomic Sequencing Kit SQK-MAP-006. The pooled DNA was then end-repaired and dA-tailed using the NEBNext Ultra II End-Repair/dA-Tailing Module (New England Biolabs), according to the manufacturer's instructions. The resulting DNA was cleaned-up using 1X by volume of magnetic Agencourt AMPure XP beads (Beckman Coulter), according to the manufacturer's instructions, and eluted in 31 µl of nuclease-free water. At this stage, 1 µl of the DNA solution was subjected to quantification by Qubit 2.0 fluorometer, as before, to assess if recovery was over 700 ng of DNA.

50 µl Blunt/TA ligase master mix (New England Biolabs) was added to the DNA, plus 10 µl of the Adapter Mix, 2 µl of the Hairpin Adapter - both provided in the sequencing kit -, and 8 µl of nuclease-free water. The tube was mixed by inversion, spinned down and the reaction was left to proceed at room temperature for 10 minutes. At this stage 1 µl of Hairpin Tether provided in the sequencing kit was added. The tube was again mixed by inversion spinned down and incubated at room temperature for another 10 minutes.

The sample was cleaned-up with magnetic Dynabeads MyOne StreptavidinC1 beads (Invitrogen). For that purpose, 50 µl of beads were pelleted in a magnetic rack, and the supernatant was removed. The beads were then washed twice with Bead Binding Buffer, provided in the sequencing kit, vortexing to resuspend, pelleting the beads and removing the supernatant in between washes. The beads were finally resuspended into 100 µl of Bead Binding Buffer, and subsequently added to the prepared DNA. The mixture was left to incubate 5 minutes at room temperature. The tube was placed on the magnetic rack, and the beads were pelleted. The supernatant was pipetted of. The beads-DNA complex was washed twice with 150 µl of Bead Binding Buffer, resuspending by pipetting gently in between washes, pelleting and discarding the supernatant. A final step of pelleting was done to remove any residual buffer. The tube was removed from the magnetic rack and the DNA

was eluted from the beads in 25 µl of Elution Buffer, provided in the sequencing kit, for 10 minutes at 37ºC. The beads were pelleted in the magnetic rack and the supernatant library was retrieved to a new tube. At this stage, 1 µl of the DNA solution was subjected to quantification by Qubit 2.0 fluorometer, as before, to assess DNA recovery.

Two different DNA libraries were performed from two independent cultures of the same isolate. Library 1 was performed exactly as described above. Library 2 was performed as described above, with the exception that the cleaning step with magnetic Agencourt AMPure XP beads (Beckman Coulter) was done with 0.6X by volume instead of 1X.

### 2.3.2 MinION sequencing set-up

Two sequencing experiments were performed with two independently prepared sequencing libraries. For each sequencing run (Run 1 and Run 2) a new R7.3 flow cell was retrieved from storage at 4ºC and left to equilibrate to room temperature. The flow cell was then mounted into the MinION Mk I device (Figure 2.3.2.1), which was connected via USB 3.0 to a PC that met the requirements for running the MinION and associated software (Windows 7; 8 Gb RAM; SSD; i7 processor; USB 3.0). The control software MinKNOW version 0.51.2.40 was initiated and the 'quality control' script was run to determine flow cell quality and to assess number of active pores.

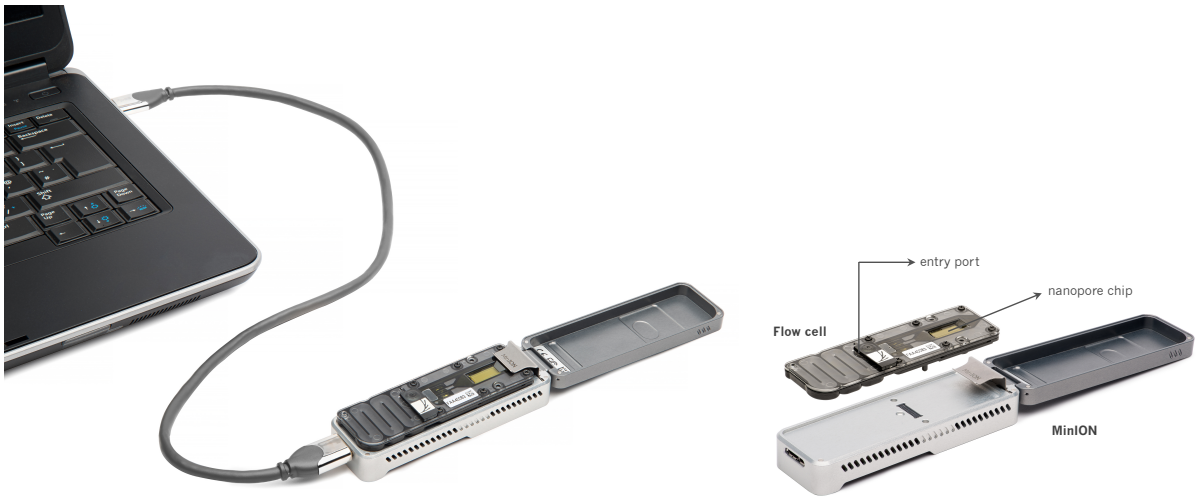


Figure 2.3.2.1 | **MinION set-up.** The disposable flow cell is mounted into the MinION and the device is then connected by a USB 3.0 cable to a computer with the MinKNOW control software. This figure is an adaptation of the original pictures made available by Oxford Nanopore Technologies.

At the end of the quality control protocol the flow cell was primed. Priming Buffer was prepared by adding 500 µl of Running Buffer and 26.6 µl of Fuel Mix, provided in the Oxford Nanopore Genomic Sequencing Kit SQK-MAP-006, to 473.4 µl of nuclease-free water. 500 µl of Priming Buffer were loaded through the flow cell entry port with a P1000 pipette. The solution was left for 10 minutes to prime the flow cell. This step was repeated a second time.

At this stage, the Sequencing Mix was prepared with 6 µl of the DNA library (section 2.3.1), mixed with 75 µl of Running Buffer, 4 µl of Fuel Mix, and 65 µl of nuclease-free water. This 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 without stopping the run.

## 2.4 Sequencing data analysis

### 2.4.1 Basecalling and sequence extraction

Basecalling of the sequencing data was performed in real-time throughout the sequencing runs using the cloud-dependent Metrichor system EPI2ME version 2.39.3. Sequencing analytics were retrieved from the online interface of EPI2ME. Basecalled reads were automatically downloaded in fast5 format and sorted into two folders, the 'pass' folder, which contains 2D Pass reads *i.e.* reads with quality scores equal or above 9, and the 'fail' folder, which in turn contains both 2D reads below the quality threshold and 1D reads.

The sequencing data of both runs was pooled together after the completion of both sequencing experiments and then repartitioned into three separate datasets: all 2D reads, 2D Pass reads and all 1D reads (both template and complement). Basecalled fast5 files of the three datasets were parsed, and sequences were extracted in fasta format using the Poretools version 0.3.0[23] (Loman & Quinlan 2014), as described by the developers. Just as Poretools, all software used locally to further process nanopore-sequencing reads was installed with all its dependencies and ran in a command-line interface on an Ubuntu System 14.04 LTS, unless otherwise specified.

### 2.4.2 Read processing and analysis of datasets

As a first instance of data analysis, RAST[24] online server was used with standard parameters 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][25] (Chen *et al*. 2007), 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, 2D, 2D Pass and 1D reads, was subjected to independent correction, assembly and polish. Canu[26] was used to correct each of the original datasets. Ten iterations of correction were done, as suggested by Canu developers for erroneous nanopore reads. For the first round of correction the command option "–nanopore-raw" was added, and subsequent iterations of the algorithm were done using the "–nanopore-corrected" command option. Since this

[23] Source code for Poretools is available at https://github.com/arq5x/poretools/blob/master/docs/index.rst and usage documentation at https://poretools.readthedocs.io/en/latest/.
[24] RAST server is available at http://rast.nmpdr.org.
[25] Formerly classified as the type strain of *B. amyloliquefaciens* subsp. *plantarum*.
[26] Canu source code is available at https://github.com/marbl/canu and usage documentation at http://canu.readthedocs.io/en/latest/.

pipeline requires information regarding the size of the genome to be assembled, the value inputted was the one corresponding to the genome of the closest neighbor, as determined by RAST, *i.e.* 3.9 megabases (command genomeSize = 3.9m) (Chen *et al*. 2007). The threshold for minimum read length accepted was set to 100 bases (command minReadLength = 100), whilst the threshold for minimal overlap between reads was set to 50 bases (command minOverlapLength = 50). Following correction, canu –trim and canu –assemble commands were run with the same parameters[27] as before to complete the assembly pipeline.

The assembled datasets resulting from Canu were further polished using Nanopolish version 0.2.0[28], which uses information of the squiggle lines recorded in the original fast5 files. For this purpose, and as described by the developers in the usage documentation, Burrows-Wheeler Aligner (BWA)[29] (Li & Durbin 2009) and Sequence Alignment/Map tools (SAMtools)[30] (Li *et al.* 2009) were first used to index and map the original current shift information in fast5 files to the assembled fasta data resulting from Canu. Only then, was the Nanopolish algorithm applied.

At this stage, each original dataset (2D, 2D Pass and 1D), generated a set of three derived datasets from their processing steps, namely correction, assembly and polish. Each of the 12 datasets was subjected to a series of read and mapping quality assessments (Figure 2.4.2.1). Statistical analysis concerning the resulting metrics was performed in RStudio version 1.0.143[31].

*K*(5)-mer composition of our chosen reference and each dataset was determined using the 'kmer' script from Poreminion version 0.0.4[32] (Urban *et al.* 2015). Based on the frequency tables of the *k*-mer counts, Kullback-Leibler divergence (Vinga & Almeida 2003) 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, $R$ represent the reference and $f$ the relative frequency of the *k*-mer $i$ in the total of 1024 possible *k*-mers of length 5.

Read and contig metrics of each dataset were obtained by QUAST[33] (Gurevich *et al.* 2013), taking as a reference the genome of *B. velezensis* strain FZB42 [NC_009725.1]. Mapping potential of the different datasets was assessed based on the mapping of the reads against the same reference, using a local installation of BLAST, BLAST+[34] (Camacho *et al*. 2009), with standard parameters. For

[27] It is usually difficult to predict the optimal parameters for assembly *ab initio* and instead they must be determined empirically. The parameters here defined were the ones that generated the best results in our multiple tests, as evaluated by QUAST.
[28] Source code for Nanopolish is available at https://github.com/jts/nanopolish.
[29] BWA is available at http://bio-bwa.sourceforge.net.
[30] SAMtools is available at http://samtools.sourceforge.net.
[31] RStudio is available at https://www.rstudio.com/products/rstudio/download/.
[32] Source code for Poreminion is available at https://github.com/JohnUrban/poreminion.
[33] QUAST - QUality ASsessment Tool for genome assemblies - is available at http://quast.sourceforge.net/quast.
[34] BLAST+ is available at ftp://ftp.ncbi.nlm.nih.gov/blast/executables/blast+/LATEST/ or as a package from the installation of Blast2GO. Usage information is in https://www.ncbi.nlm.nih.gov/books/NBK279690/.

the purpose of counting mapped reads and evaluate mapping statistics, only the highest scored mapping for each independent read was considered.

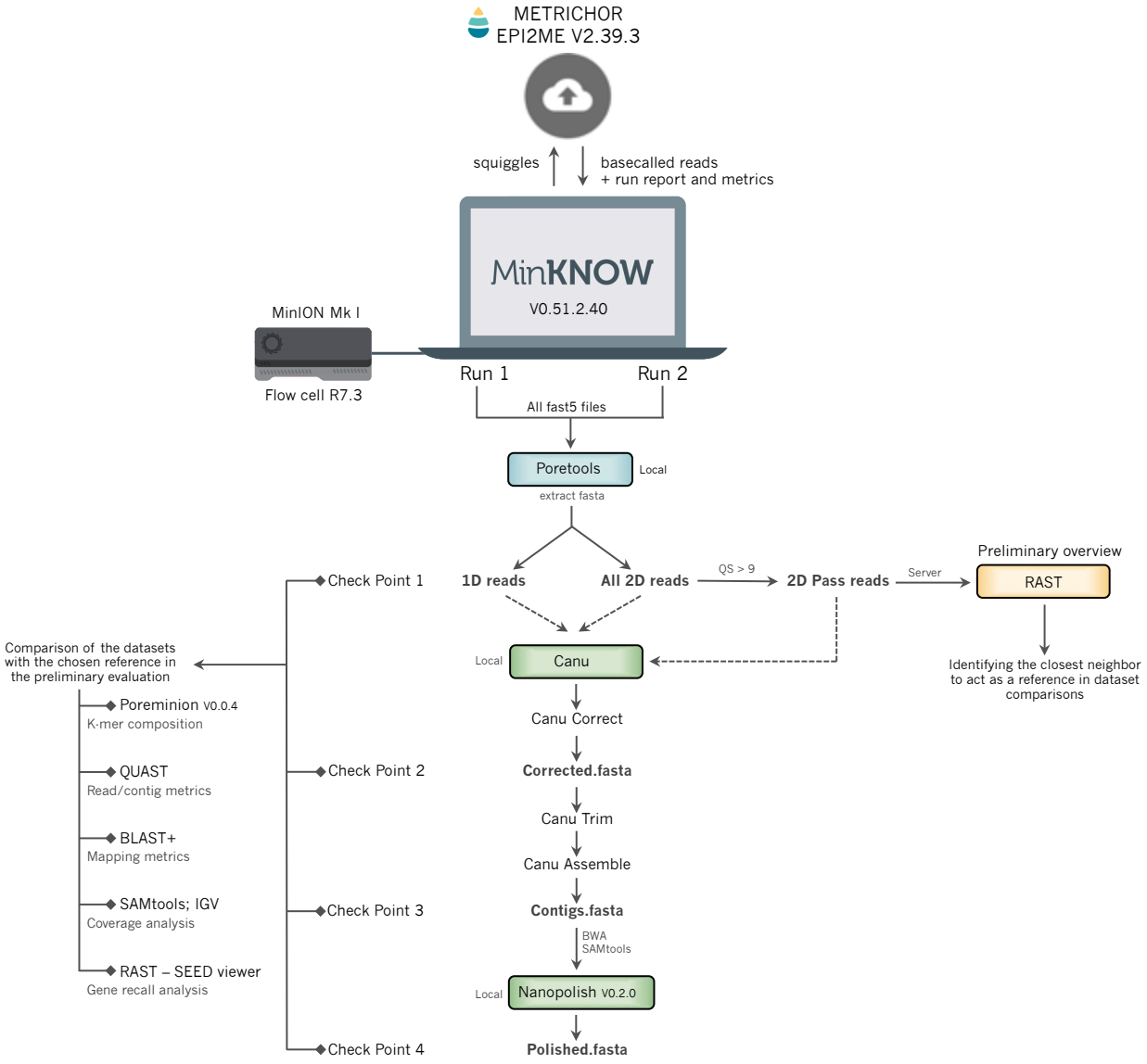


Figure 2.4.2.1 | **Data processing workflow from sequence generation, to dataset partition, read correction, assembly, polish, and quality assessment.**

The datasets were further indexed and mapped against the chosen reference in SAMtools, enabling coverage per base calculation. The resulting sam files were plotted in the Integrative Genomics Viewer (IGV) version 2.3.91[35] software for graphical coverage analysis.

Lastly, all datasets were subjected to RAST with standard parameters for determination of gene recall potential, *i.e.* to determine the number of genes of the chosen reference identified in the particular dataset. For that purpose, a sequence-based comparison with the chosen reference was performed in the SEED viewer, on the RAST server interface.

[35] IGV is available at https://software.broadinstitute.org/software/igv/download.

### 2.4.3 Annotation and enzyme identification

2D reads were submitted to RAST online server for annotation, with standard parameters, taking advantage of the embedded RAST ORF finder. Results from the RAST were exported in xls format. 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. See Appendix A for a more detailed list of biomass-degrading industrial relevant enzymes. The selected annotated sequences were further subjected to PSORTb version 3.0.2[36].

For Blast2GO[37] annotation, the 2D dataset reads were first genecalled using Prodigal[38] gene finder. 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 extracted in xls and then 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[39] database, as well as the CAZy[40] database, to confirm the annotation of peptidases and carbohydrate-active enzymes, respectively. MEROPS BLASTP was performed in the MEROPS database interface whilst CAZy BLASTP was performed through dbCAN web server[41].

Since the 2D dataset has a redundant nature, 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.

[36] PSORTb is available at http://www.psort.org/psortb/.
[37] Blast2GO is available at https://www.blast2go.com.
[38] Prodigal is available at https://github.com/hyattpd/prodigal/releases/.
[39] MEROPS – peptidase database - is available at http://merops.sanger.ac.uk.
[40] CAZy – carbohydrate-active enzymes database – is available at http://www.cazy.org.
[41] dbCAN is available at http://csbl.bmb.uga.edu/dbCAN/.

# Chapter 3. Results and discussion

## 3.1 Reanalysis of the results from the SEAVENTzymes project and isolate selection

### 3.1.1 Growth and colorimetric assays portray different aspects of enzyme production capability

In a first instance of this dissertation, we set ourselves to choose a single isolate to work towards our aim of testing whole-genome nanopore sequencing screening capabilities. For that purpose, we reviewed the overall potential of the SEAVENTbugs collection based on the screening results for industrial enzymes obtained during the SEAVENTzymes project. We focused on the data obtained for the screening of biomass-degrading enzymes, since these represent a major group of economically relevant enzymes and are our chosen working subset of enzymes. The data resulted from two different screening methods: growth screening assays and colorimetric assays, both of which were not performed during this dissertation but are fully described in Appendix C.

Growth assays were performed in basal medium, supplemented with the substrate of the target enzymes as the sole source of a particular required nutrient, either carbon or nitrogen. The organism's ability to grow in a medium where a necessary nutrient is only available by the degradation of the substrate, should be reflective of the organism's ability to produce enzymes acting on the said substrate. The substrates tested were starch, cellulose, xylan, mannan, pectin, chitin, casein and a mixture of 'tweens', aiming to detect starch-, cellulose-, xylan-, mannan-, pectin- and chitin-degrading enzymes, proteases and lipases/esterases, respectively. The growth of each isolate in base medium supplemented with the enzyme substrates (ES) was evaluated based on the calculation of the relative Net Area Under Curve - NAUCr(ES). NAUCr acts as a single parameter to describe the overall growth curve of the organism. To account for residual growth in the base medium (BM) alone, NAUCr(ES) values were normalized to NAUCr(BM), *i.e.* NAUCr(ES)/NAUCr(BM). Attending to the formula (Appendix C, Figure C.1), a value of 1 should indicate that the growth in base medium plus the enzyme substrate is equal to the growth in base media without the substrate. In this case, there is no evidence of preferential growth in the presence of substrate, and consequently no evidence for the

production of the searched enzyme acting on that substrate. In the same manner, values higher than 1 mean that the growth in base medium with the enzyme substrate was higher than the growth in the control situation with base medium only, which should be reflective of the production of enzymes degrading the substrate, and the mobilization of nutrients for growth[42]. The full subset of mesophilic aerobic isolates from the SEAVENTbugs collection was screened with growth assays, yielding a total of 139 tested isolates.

The colorimetric assays were mostly based on a set of commercial chromogenic azurin-dyed cross-linked substrates (AZCL), which were incorporated into solid base medium as sole sources of carbon or nitrogen. These modified substrates are useful to detect the production of endo-hydrolyzing enzymes in a straightforward manner, since the hydrolytic action of the screened enzyme on the substrate leads to the diffusion of a visible blue product in the medium. Just as for growth assays, starch-, cellulose-, xylan-, mannan-, chitin-degrading enzymes and proteases were screened. Pectin-degrading enzymes, for which there was no AZCL-substrate, were not screened by colorimetric assays. In the specific case of lipases/esterases, instead, the colorimetric screening assay depended on the co-addition off a mixture of 'tweens' and calcium chloride. The hydrolysis of 'tweens' by the target enzymes leads to the released of fatty acids which form a yellow precipitate with calcium. Alternatively, if fatty acids are completely degraded, a clear halo appears around the colonies.

In the initial plan of the SEAVENTzymes project, colorimetric assays were intended to be applied only to confirm positive results detected by growth assays. Quickly it became apparent that this was not the best approach. Several of the isolates that were detected as the highest producers of specific enzymes in growth assays did not even show slight coloration in colorimetric assays. Being that most results were not concordant between the two different screening methods, and in an attempt to understand these differences, the colorimetric assays were again applied as a generalized screening procedure, rather than as a confirmation step. At this point, a smaller subset of isolates was chosen, since a polyphasic analysis by fingerprinting methods had allowed for the dereplication of clonal isolates. Thus only 107 of the 139 mesophilic aerobic isolates of the collection were screened with colorimetric assays.

Here, we compiled these results in Figure 3.1.1.1 and tried to assess if there is indeed an association between the two different screening procedures applied. For that purpose, the set of results obtained for the 107 isolates tested in both screenings were transformed into binary data (positive or negative) and a Cramér's V ($\Phi_C$) test was applied.

Most enzyme assays have shown to have no significant (Chi-squared test, $p$>0.05, α=0.05) association between the two screening methods. The highest association found was $\Phi_C$=0.54 for the starch-degrading enzymes screening, and still, although statistically significant (Chi-squared test, $p$=2.94x10$^{-8}$, α=0.05), it represents an underwhelming association. Since the limit for the definition of positive results for the growth assays was arbitrarily defined as 6 to account for data variability, we evaluated the impact of changing this limit in the association measurements. However, shifting the

[42] Although in theory NAUCr(ES)/NAUCr(BM)>1 should represent a positive result for the production of the screened enzyme, to account for some variability encountered in the growth assays, a result was only considered positive when NAUCr(ES)/NAUCr(BM)>6, as described in the methods section.

operational limit below or above 6 did not change the initial association assessment (data not shown)[43].

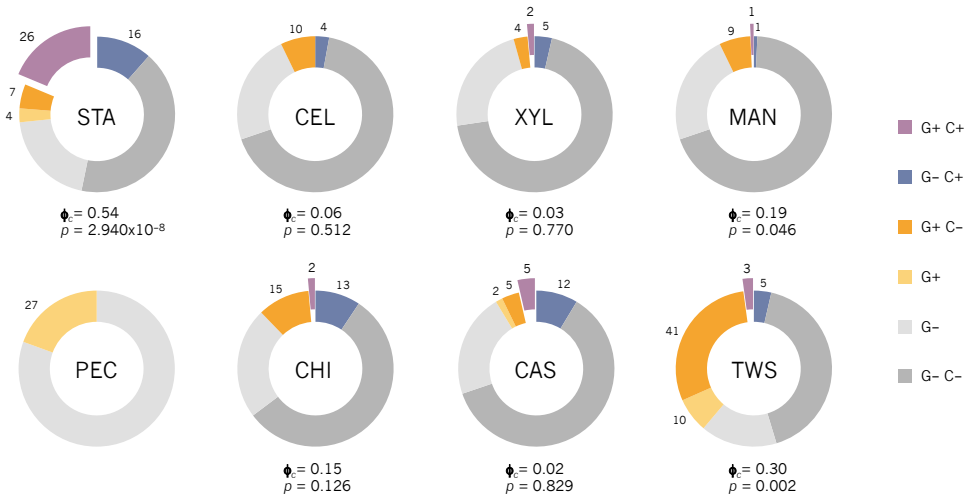


Figure 3.1.1.1 | **Summary of the results obtained from the growth and colorimetric assays performed during the SEAVENTzymes project for the detection of different groups of industrial relevant biomass-degrading enzymes.** Microplate growth assays were done to a total of 139 mesophilic aerobic isolates. From these 139 isolates only 107 were also tested with colorimetric assays. Bellow each graph we present Crámer's V (Φc) values as a measure of association between the two different screening methods, based on the results obtained from the 107 isolates tested in common. *p*-values for this association are also presented and were calculated using a chi-squared test with an α level of 0.05. Each chart represents the number of isolates with positive (+) or negative (–) results in growth (G) or colorimetric (C) assays performed with the particular substrate: STA – starch, CEL – cellulose, XYL – xylan, MAN – mannan, PEC – pectin, CHI – chitin, CAS – casein and TWS – mixture of 'tween' 20 and 'tween' 80. Pectin was only used in growth assays. Results from the colorimetric assays with AZCL-amylose and AZCL-pullulan were pooled together in 'STA', since they both screen for enzymes acting on different components of starch. Thus the values shown represent the number of isolates with amylose– and/or pullulan–degrading capabilities.

These differences in the results of the screening methods were not desired, since the use of two different methods was intended to augment confidence in the screening. However, the lack of association was not completely surprising. The two methods employed are intrinsically different. The growth assays are based on the use of polymeric natural substrates and evaluate the capacity of an isolate to digest the component within the context of complex natural substrates, providing no information about individual enzyme specificities. Contrariwise, AZCL-based colorimetric assays evaluate the activity of very specific enzymes. That is, AZCL substrates are cross-linked, which means they do not have free-ends, and their hydrolysis, and consequent appearance of color, depends on the activity of very specific endo-acting hydrolytic enzymes. Growth assays do not restrict for these types of enzymes and may yield positive results even in the absence of endo-hydrolyzing enzymes. An isolate may be able to grow by the degradation of the substrates by utilizing exo-acting hydrolyzing enzymes and/or other groups of biomass-degrading enzymes. Thus, positive results for growth assays do not necessarily imply a positive result for colorimetric assays.

[43] Spearman correlation coefficient (ρ) was also calculated to assess monotonic correlation. In this analysis the isolates are first ranked from best producers to worst producers for each enzyme assay, taking into consideration the continuous and interval values of the growth and colorimetric results, respectively. This test then allows to evaluate monotonic correlation between the rankings, which should be expected if the assays were ranking the isolates from best to worst producers in a similar manner. However, the results were just as in Cramér's V, with statistical significance only found for a slight association between the assays contemplating starch-degrading enzymes (ρ=0.39, $p$=1.56x10$^{-5}$, a=0.05), meaning there is no evident correlation between the ranking of best to worst producers by the two screening methods.

Note that, for the screening of starch-degrading enzymes, the association value between the two different methods was only reached because two different colorimetric assays were pooled together in the comparison. That is, the growth results in starch-supplemented medium were compared with the pooled results for colorimetric tests in AZCL-amylose and AZCL-pullulan, two different components of starch. Indeed, when association was assessed for each independent assay, with either AZCL-amylose or AZCL-pullulan, no significant association was found (Chi-squared test, $p>0.05$, $\alpha=0.05$). This again shows how the results from the growth assays may be the reflection of the action of multiple enzymes on the substrate.

This might explain to some extent the absence of association between the two different methodologies, but not completely. For instance, the production of lipolytic enzymes was not evaluated based on these AZCL substrates, but rather in the use of the same mixture of 'tweens' as in the growth assays, with the extra addition of calcium chloride. Since the substrate used for both growth and colorimetric screenings was the same, we would expect a higher correlation between these assays. Indeed, after the starch-degrading enzymes screening, the only significant association found was for lipases/esterases screening. Still, it was a very slight association ($\Phi_C=0.30$, $p=0.002$, $\alpha=0.05$). Control assays with calcium chloride had shown that there was no impact of this compound on the normal growth of the isolates, so it should not be the factor leading to the differences found. Most likely, the lack of association in the results may be due to the different conditions applied in the two different assays, specifically the liquid versus solid media conditions and the different substrate concentrations (Appendix C).

Even thought there was a lack of association between the two screening methods, both of the methods had high reproducibility in discerning positive from negative isolates, with 94.6% and 94.8% concordant results for growth and colorimetric assays, respectively. Thus, overall, both methods are most likely portraying different aspects of the enzyme production capability of each isolate. If we intended to select only very specific hydrolytic enzymes, colorimetric assays would be more appropriate since they screen based on specificity. For the purpose of selecting promising isolates with overall biomass-degrading capabilities, there might be advantages in integrating all results. Indeed, the isolates may produce a set of enzymes with potential interest that are not recognized with the colorimetric assays employed.

### 3.1.2 Integrating all results by PCA allows to pinpoint promising isolates

Since each screening method seems to be relevant in its own way of portraying the system, to pinpoint interesting isolates, all non-transformed results from both screening methods were compiled and integrated by PCA. In Figure 3.1.2.1 we show a projection of the isolates on a new space system formed by the three first principal components, which account for 58.3% of the total variability of the system. In this projection most isolates were grouped in a focused cluster. Only a smaller set of isolates was distinguishable from the main group (indicated in Figure 3.1.2.1. in colored circles). This subset most likely represents isolates with unique responses to the screening methods and for that reason, they were further investigated.

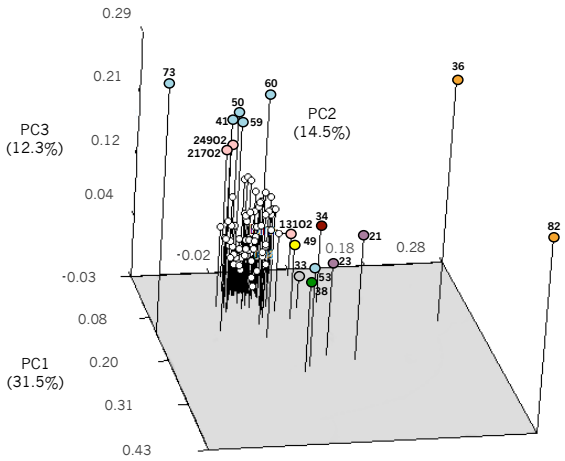


Figure 3.1.2.1 | **Projection of the isolates on the principal component space constructed from the integrated analysis of all results from the phenotypic screening performed during the SEANVENTzymes project.** This projection was obtained from a PCA integrating all data obtained from both growth and colorimetric screening methods. For each of the three first principal components (PC), we indicate the percent variance of the system associated with it. 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 (shown in Figure 3.1.2.2).

The isolates distinguishable by PCA were plotted on the dendrogram obtained by multiple fingerprinting profile analysis of all mesophilic aerobic isolates, to assess their relatedness (more on the polyphasic characterization of the isolates in Appendix B Figure B.2). From the observation of Figure 3.1.2.2, it seems that several of the isolates sub-selected based on PCA form very coherent clusters by fingerprinting profile analysis.

For instance, Operational cluster 1 comprises the isolates MG SD 036[44] and MG SD 082. These isolates originated from the same Menez Gwen (MG) sediments (SD) sample. Furthermore, they are clustered together above the reproducibility threshold, which means they cannot be distinguished based on the integrated fingerprinting analysis, indicating a possible clonal relationship. Indeed, whilst whole-cell protein profiling by SDS-PAGE is known to correlate with DNA-DNA hybridization - the standard for species discrimination -, the RAPD PH, RAPD 1281 and csM13 DNA fingerprinting methods employed are useful for the characterization of infraspecific diversity and identification of possible clonal relationships (Meyer *et al.* 1993; Vandamme *et al.* 1997).

Operational cluster 2 comprises isolate MG CR 023 and MG CR 021, grouped together with two other closely related isolates that were not pinpointed by PCA. Both MG CR 023 and MG CR 021 were recovered from the same crab (CR) sample from Menez Gwen (MG).

Finally, operational cluster 3 comprises the isolates RB BA 059, RB BA 053, LS WA 073, RB RS 041, RB PS 050, RB BA 060 and two other isolates which were not selected based on PCA, namely RB BA 058 and MG CR 186O2. This cluster is the most diverse, with isolates coming from different vents, namely Rainbow (RB), Lucky Strike (LS) and Menez Gwen (MG), and different sample types, such as *Bathymodiolus azoricus* (BA) mussels, water (WA), *Rimicaris* sp. (RS) shrimp, *Pachichara* sp. (PS) fish and crab (CR). Isolates RB BA 059 and RB BA 053 come from the same sample and are

[44] Each one of the isolates is designated by 4 letters and 3 digits. The first two letters codify the vent field from which they were isolated, the third and forth letters refer to the type of sample, and the digits denote the sequential order of isolation. An 'O2' added in front of the code of the isolate indicates that the isolate was obtained from an aerobic reisolation from an anaerobic original culture.

also not distinguishable by fingerprinting analysis, which may indicate clonality.

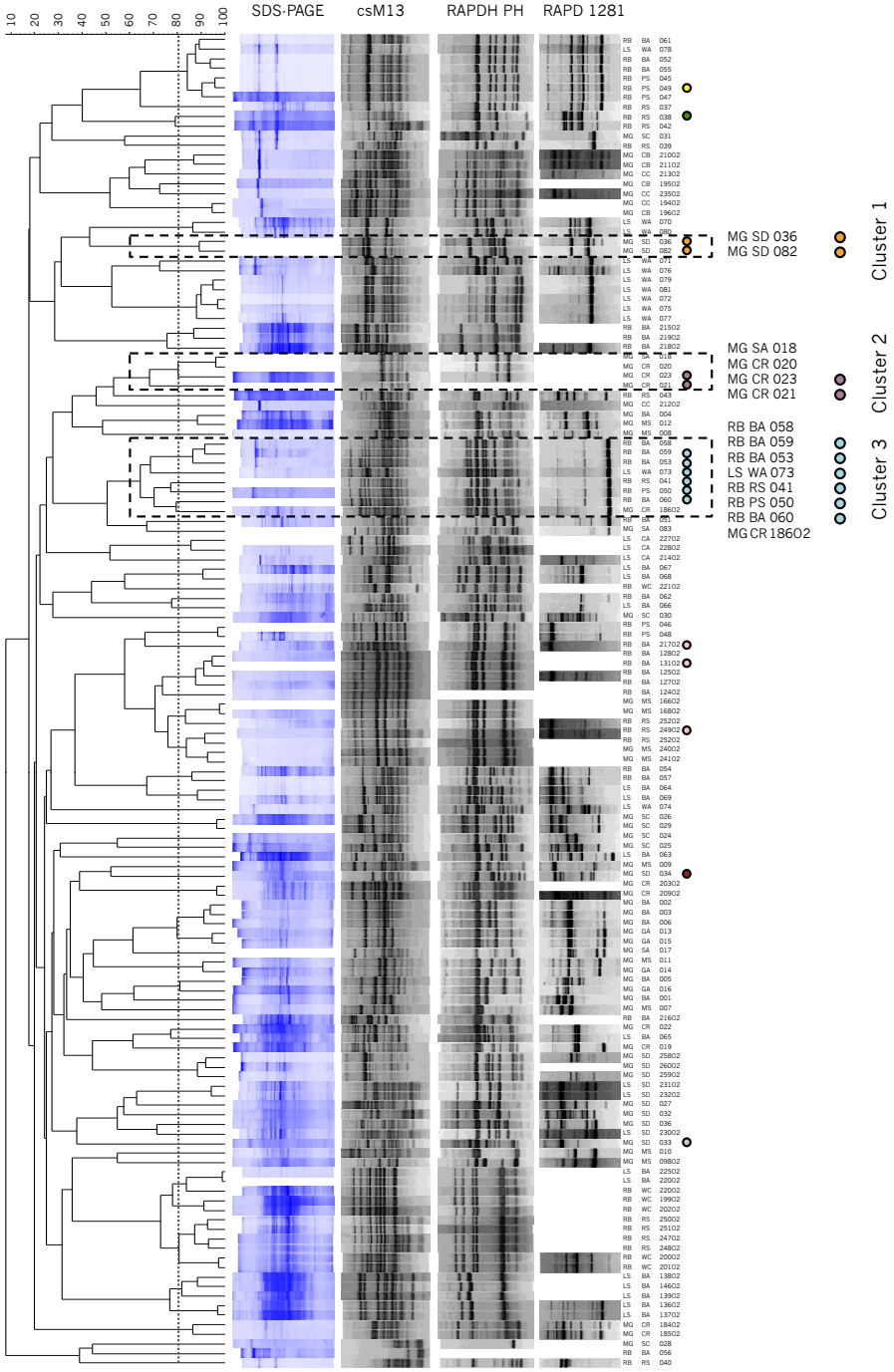


Figure 3.1.2.2 | **Composite dendrogram obtained from the analysis of DNA fingerprinting (csM13, RAPD PH and RAPD 1281) and SDS-PAGE profiles of all 139 mesophilic isolates screened during the SEAVENTzymes project.** The dendrogram was constructed using the Pearson correlation coefficient and the UPGMA clustering method**.** The scale represents percent similarity. The set of isolates that were distinguishable by PCA are identified in the dendrogram with circles. Dashed boxes delimit the clusters of isolates that were selected for further investigation. The vertical line indicates reproducibility level obtained by the average of reproducibility of each fingerprinting method.

A closer observation on the initial projection in Figure 3.1.2.1 reveals that, these specific isolates, besides being clustered together based on fingerprinting analysis, are also grouped at some level on the projection. The isolates represented in orange - operational cluster 1 - are the most further apart from the remaining isolates. Isolates represented in purple – operational cluster 2 - are concentrated in

the middle of the principal component space, together with other isolates belonging to different clusters. Conversely, isolates represented in light blue – operational cluster 3 - are also closely distributed with the exception of the isolates RB BA 053 and LS WA 073, which are further apart. Since the PCA was based on the integrated analysis of the phenotypic results, this observation reflects a coherence in the way closely related isolates respond to phenotypic assays, at least to some extent, and that this coherence can be evidenced by PCA.

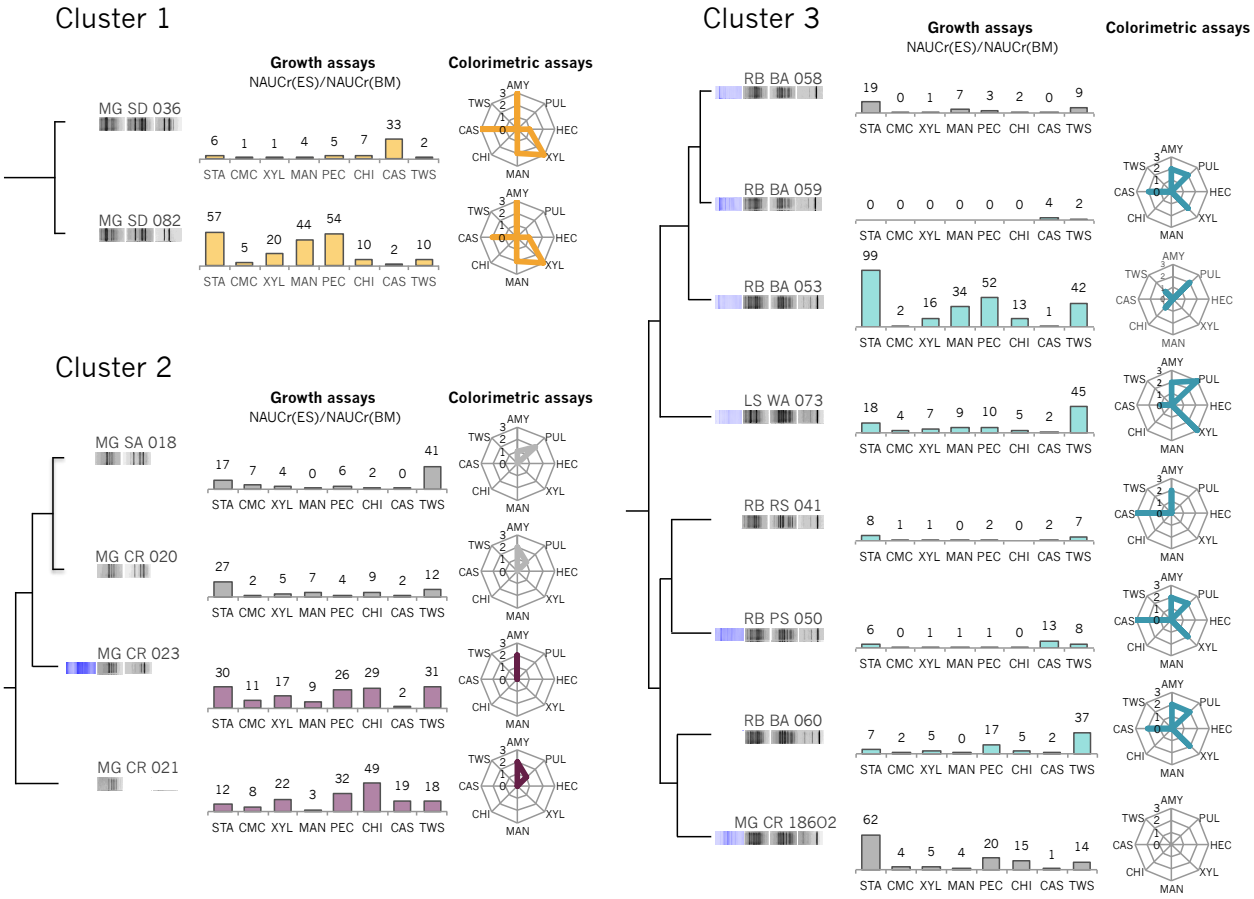


Figure 3.1.2.3 | **Growth and colorimetric screening results of selected isolates based on PCA.** The represented clusters were retrieved from dendrogram in Figure 3.1.2.2 and contain isolates selected based on PCA (colored in orange, purple and blue). Bar graphs represent the results for growth assays as NAUCr(ES)/NAUCr(BM) in media with STA – starch, CMC – carboxymethylcellulose, XYL – xylan, MAN – mannan, PEC – pectin, CHI – chitin, CAS – casein and TWS – 'tween' 20 and 'tween' 80. Radial graphs represent 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 AMY – AZCL-amylose, PUL – AZCL-pullulan, HEC – AZCL-hydroxyethylcellulose, XYL – AZCL-xylan, MAN – AZCL-glucomannan, CHI – chitin-azure, CAS – AZCL-casein and TWS – 'tween' 20 and 'tween' 80 plus calcium chloride. The isolate RB BA 058 was not tested with colorimetric assays.

Indeed, these isolates that were distinguishable by PCA represented some of the overall best producers of enzymes in the collection (Figure 3.1.2.3), in terms of total enzyme number and level of production/activity. All three clusters seem to comprise quite promising candidates for further investigation. However, since the purpose of this dissertation focuses on a single isolate, we must select only one from this subset. For a more informed decision, the isolates of these operational clusters were further analyzed by partial 16S rRNA gene sequencing, in an attempt to taxonomically position them.

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

The SEAVENTbugs collection was kept cryopreserved at -80ºC and all isolates of the pre-selected clusters were subjected to a recuperation protocol as described in the methods. Two of the 14 recuperated isolates, namely MG CR 186O2 and LS WA 073, did not grow from the cryopreserved cultures and were thus removed from the analysis. The remaining isolates were grown and further subjected to csM13 fingerprinting as well as RAPD PH fingerprinting. These DNA fingerprinting methods allowed the confirmation of the isolates identity. That is, comparing the fingerprinting profiles obtained during this work with the profiles obtained during the SEAHMA project enabled us to determine if we were indeed in the presence of the original isolates, and not eventual contaminants. All 12 isolates were confirmed to correspond to the original isolates (data not shown).

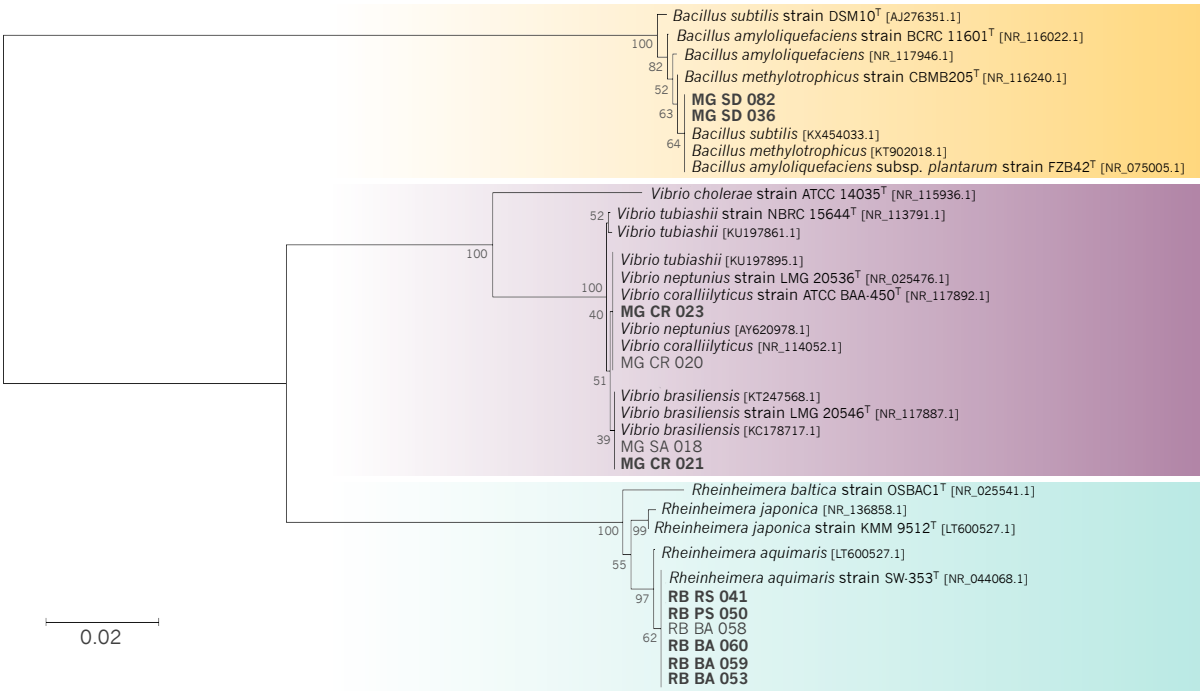


Figure 3.1.3.1 | **Phylogenetic reconstruction of recuperated isolates and their top BLAST hits by neighbor-joining clustering of their 16S rRNA partial gene sequences.** Percent bootstrap values, derived from 1000-fold sampling, are indicated near the respective nodes. Isolates selected based on PCA 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$^T$.

The 16S rRNA gene of these isolates was partially sequenced and a phylogenetic reconstruction is shown in Figure 3.1.3.1. We were able to determine that, just as expected based on fingerprinting analysis, all isolates of each cluster are closely related, at least at the genus level.

The isolates MG SD 082 and MG SD 036 of operational cluster 1 belong to the *Bacillus* genus and seem to be indistinguishable by comparison of partial 16S rRNA gene sequence, just as in the integrated fingerprinting profile analysis. Furthermore, these isolates cannot be discriminated from the *B. amyloliquefaciens* subsp. *plantarum* type strain FZB42$^T$ and two other *Bacillus* strains, one belonging to *B. subtilis* and the other to *B. methylotrophicus*. When carefully observing the *Bacillus* cluster in Figure 3.1.3.1, it seems that strains representing the same species did not necessarily

cluster together by comparison of the analyzed partial 16S rRNA gene sequence.

Ash *et al.* (1991) resolved the basic systematic structure of the *Bacillus* genus by comparison of 16S rRNA gene sequences, dividing the genus into five main clusters. Among them, one cluster consisted of *B. subtilis* and closely related species, termed *Bacillus sensu stricto*. However, discrimination of species within the *Bacillus sensu stricto* group has been proven quite difficult, with 16S rRNA gene sequences showing insufficient discrimination capabilities (Porwal *et al.* 2009). Here, we were also unable to identify these isolates at the level of species.

Albeit not pursued in this dissertation, if we intended to discriminate between species of this *Bacillus sensu stricto* group, sequence comparison of protein encoding genes, such as *gyrA* - coding for DNA gyrase subunit A -, or *rpoB* - coding for RNA polymerase β subunit -, which exhibit much higher genetic variation, could be used as an alternative to the 16S rRNA gene (Porwal *et al.* 2009). Besides these single-gene based approaches, and DNA-DNA hybridization - the golden standard for bacterial species discrimination -, there are many other alternative approaches which have been taken through the years for the discrimination of *Bacillus* species: from the more classic phenotypic or biochemical tests, to fatty acid methyl ester profiling, several DNA fingerprinting methods, MultiLocus Sequence Analysis (MLSA) of housekeeping genes (Liu Y. *et al.* 2013) or even alignment free whole-genome comparisons (Wang A. & Ash 2015).

Nevertheless, 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). Dunlap *et al.* took advantage of genomic data to reevaluate species discrimination in this genus and revealed that the type strains of *B. methylotrophicus* KACC 13015$^T$, *B. velezensis* NRRL B-41580$^T$ and *B. amyloliquefaciens* subsp. *plantarum* FZB42$^T$, have *in silico* DNA-DNA hybridization values greater than 84%, which is well above the standard species definition threshold of 70%. That means that these strains, which were divided into three different species, are most likely later heterotypic synonyms of *B. velezensis*, and should be reclassified as such, since the publication of this specie precedes that of the others. 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,* one being the type strain KACC 13015, may indicate that the isolates belong to, or are closely related to, this restructured *B. velezensis* species.

Operational cluster 2 isolates fit into the *Vibrio* genus. It seems that the four isolates of this cluster most likely belong to at least two different *Vibrio* species, since they were separated into two different clusters. MG CR 021 as well as its counterpart MG SA 018, were clustered in a very cohesive group with strains of *Vibrio brasilliensis*. Contrariwise, MG CR 23 and MG CR 020 were grouped with *Vibrio neptunius* and *Vibrio coraliilyticus*, being indistinguishable from both of these species' type strains. Curiously, although MG SA 018 and MG CR 020 are separated based on 16S rRNA partial gene sequence comparison, they were clustered together above the reproducibility threshold in the composite dendrogram of Figure 3.1.2.2. However, upon further inspection of the composite dendrogram we can see that these two isolates were grouped based on only two of the four

fingerprinting profiles, which ultimately gives us less confidence in their clustering.

Finally, operational cluster 3 belongs to the *Rheinheimera* genus and all isolates seem to be very closely related between them and with the *Rheinheimera aquimaris* type strain, even thought they were retrieved from very diverse samples.

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

At this stage, 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 deep-sea hydrothermal vent field.

As seen in Figure 3.1.4.1, the MG SD 082 isolate seems to produce extracellular endo-hydrolytic enzymes acting on starch, cellulose, xylan, mannan and casein, *i.e.* amylases, cellulases, xylanases, mannanases and proteases, as evidenced by the colorimetric assays. The production of starch-, xylan- and mannan-degrading enzymes is further confirmed by growth assays. Contrariwise, the NAUCr(ES)/NAUCr(BM) 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 was higher that the growth in base media alone, they still fall under the defined threshold for positive results based on replicate analysis NAUCr(ES)/NAUCr(BM)>6. Moreover, growth assays further evidenced the production of chitin-degrading enzymes and lipases, which were however, not observable by colorimetric assays.

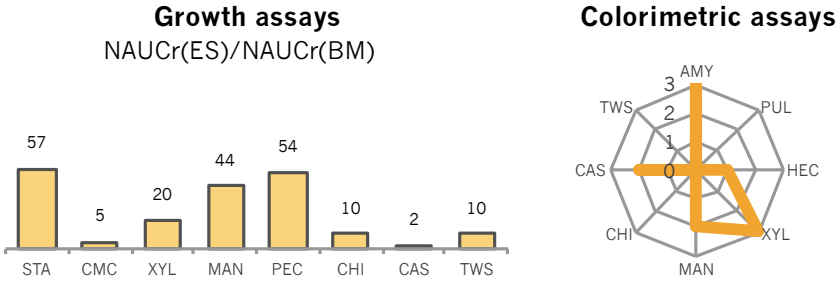


Figure 3.1.4.1 **| Growth and colorimetric screening results obtained during the SEAVENTzymes project for the selected isolate *Bacillus* sp. MG SD 082.** Bar graph represents the results for growth assays as NAUCr(ES)/NAUCr(BM) in media with STA – starch, CMC – carboxymethylcellulose, XYL – xylan, MAN – mannan, PEC – pectin, CHI – chitin, CAS – casein and TWS – 'tween' 20 and 'tween' 80. 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 AMY – AZCL-amylose, PUL – AZCL-pullulan, HEC – AZCL-hydroxyethylcellulose, XYL – AZCL-xylan, MAN – AZCL-glucomannan, CHI – chitin-azure, CAS – AZCL-casein and TWS – 'tween' 20 and 'tween' 80 plus calcium chloride.

*Bacillus* contains one of the most researched organisms as its type species, *i.e. B. subtilis*, the model for Gram-positive organisms. Members of this genus have long been described as aerobic Gram-positive bacteria, with rod-shaped and spore-forming capabilities (Bhandari *et al.* 2013). It is amongst the most diverse and prolific prokaryotic genera, with more than 220 recognized species distributed widely across terrestrial and aquatic habitats, including marine sediments, where they seem to be a common isolated taxon (Jørgensen & Boetius 2007). This ubiquity is attributed not only

by their ability to form resilient spores, which can be easily transported and subsist in diverse environmental settings, but also by their great metabolic versatility and ability to grow under physico-chemical extremes (Sass *et al.* 2008; Ettoumi *et al.* 2009; Ettoumi *et al*. 2013).

Due to the resistant nature of spores to harsh conditions, and the fact that marine *Bacillus* isolates do not display characteristic marine traits, it still remains unclear whether such spore-forming bacteria can be indigenous to extreme marine habitats or were simply transported from other environments. Nevertheless, some authors have shown evidences to support that certain *Bacillus* spp. have the potential to participate in several oceanic biogeochemical cycles near hydrothermal sediments and plumes, even in their spore form (Ettoumi *et al*. 2013). Furthermore, it seems that some species may also be active *in situ*, since they are able to grow in artificial similar extreme conditions (Sass *et al.* 2008). If *Bacillus* spp. are active in these extreme settings, bioprospecting for their enzymes might unveil some catalyst with extreme resistant features of industrial interest.

Indeed, this genus comprises several biotechnological important species, which are typically inserted in the *Bacillus sensu stricto* group. Species such as *B. licheniformis* and *B. amyloliquefaciens* are very well know for their versatile metabolic capabilities, and among other things, are useful in the production of antibiotic or probiotic components, antagonistic substances or surfactants, and several industrial important enzymes (Dunlap *et al.* 2016). They are common sources of thermostable and halotolerant enzymes such as amylases (Asoodeh, Chamani & Lagzian 2010), pullulanases, cellulases (Trivedi *et al.* 2011), xylanases (Khandeparker, Verma & Deobagkar 2011), mannanases (Cheng *et al.* 2016), pectinases (Joshi *et al.* 2012), chitosanases (Chulhong *et al.* 2011), proteases (Zhou *et al.* 2013), lipases (Lailaja & Chandrasekaran 2013) and esterases (Karpushova *et al.* 2005).

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 seems to be recurrent in seafloor sediments and well known for its biotechnological utility and production of industrial relevant enzymes - the object of study of this dissertation. Furthermore, a quick visit to the NCBI genome database revealed that there is genomic data available for over 150 different *Bacillus* species, which will most likely facilitate our evaluation of the nanopore-sequencing data by offering a large collection of reference data.

## 3.2 Whole-genome nanopore sequencing of the *Bacillus* sp. MG SD 082 isolate

### 3.2.1 Independent sequencing runs differ in yields and read length distributions

Two nanopore-sequencing runs were performed using two separate flow cells and two independently prepared 2D-sequencing libraries of the *Bacillus* sp. MG SD 082 DNA. Although 2D nanopore sequencing ultimately aims to generate 2D reads, the resulting sequencing data is still partitioned into three different datasets: 1D data composed by the direct sequencing of either template

or complement DNA strands, 2D data that results from the consensus calling of paired 1D template and 1D complement reads and finally 2D Pass data that corresponds to a subset of 2D reads which have mean quality scores[45] (QScore) equal or higher than 9. Overviewing the sequencing metrics revealed that the two runs generated considerably different data yields and read length distributions.

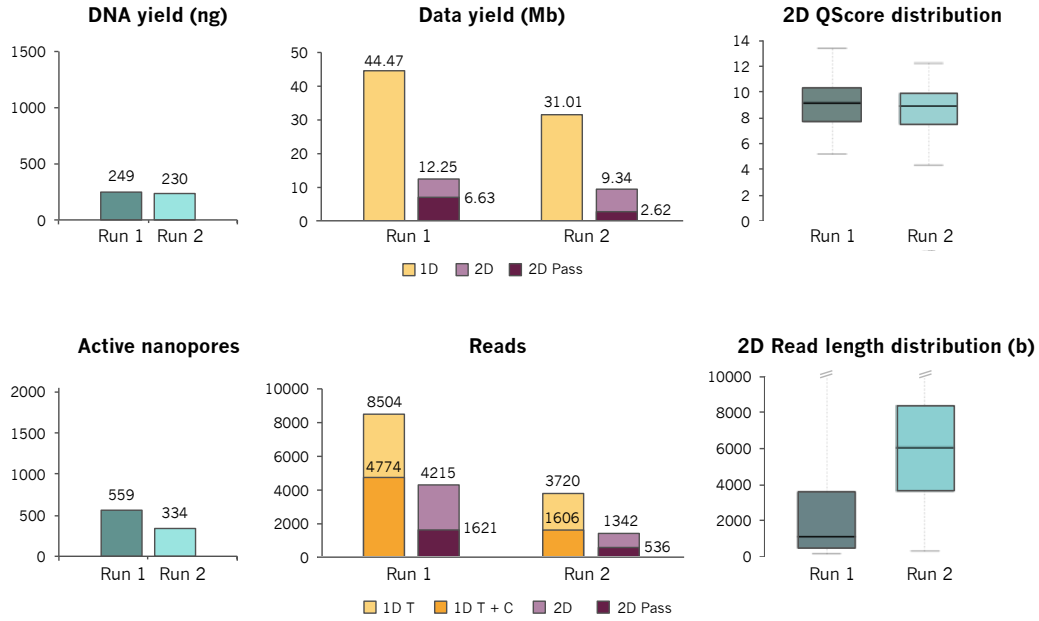


Figure 3.2.1.1 | **Sequencing metrics of the *Bacillus* sp. MG SD 082 two nanopore-sequencing runs.** 'DNA yield' refers to the total DNA amount obtained at the end of the sequencing library preparation. 'Active nanopores' refers to the number of available working pores at the beginning of the sequencing run. 'Data yield' in megabases and number of 'Reads' are partitioned into the three generated data types: 1D data - composed by 1D template reads (1D T) and 1D complement reads (1D C) -, 2D data generated by the consensus of 1D template and complement and finally 2D Pass data which is a subset of 2D reads with quality scores equal or above 9. Boxplots are represented with Spear whiskers that extend to minimum and maximum values. For 2D read length distributions the maximum values are omitted for the purpose of clarity – see table 3.2.1.1. QScore of a read refers to the per base quality score mean.

Table 3.2.1.1 | **Read length and quality metrics of the two independent nanopore-sequencing runs.**

| | Length mean (b) | Length mode (b) | Length median (b) | Longest read (b) | QScore mean | QScore mode | QScore median |
|---|---|---|---|---|---|---|---|
| **Run 1** | | | | | | | |
| 1D Template | 3.50 K | 701 | 2.81 K | 130.00 K | 4.6 | 5.3 | 4.6 |
| 1D Complement | 2.84 K | 392 | 1.31 K | 93.81 K | 4.6 | 4.7 | 4.6 |
| 2D | 2.97 K | 413 | 1.38 K | 106.00 K | 8.8 | 7.3 | 8.8 |
| 2D Pass | 2.64 K | 413 | 2.20 K | 21.36 K | 10.2 | 9.6 | 10.1 |
| **Run 2** | | | | | | | |
| 1D Template | 7.08 K | 3.22 K | 6.43 K | 405.85 K | 4.3 | 4.3 | 4.7 |
| 1D Complement | 6.24 K | 3.27 K | 5.49 K | 503.42 K | 4.4 | 4.3 | 4.3 |
| 2D | 6.60 K | 3.53 K | 5.94 K | 83.95 K | 8.7 | 9.3 | 8.9 |
| 2D Pass | 6.92 K | 3.53 K | 6.31 K | 25.88 K | 9.9 | 9.3 | 9.8 |

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 (Figure 3.2.1.1). Furthermore,

[45] The per base quality scores of other sequencing technologies correspond to the Phred scale, where scores indicate a specific likelihood of error for that base. The nanopore calculated quality scores are an indication of how well the current squiggles fit into the basecalling model, but do not follow Phred expected error rates.

for Run 1, only 6.63 Mb of the 2D data had quality scores equal or above 9, representing the 2D Pass data, whilst Run 2 yielded 2.62 Mb of 2D Pass data.

The differences in total data yield between the two runs do not seem to be related with the amount of DNA resulting from the library preparation itself. DNA yields (Figure 3.2.1.1) were both close to the expected 250 ng reported by the developers of the technology (249 ng for Library 1 and 230 ng for Library 2). However, the amount of available pores of the flow cells at the beginning of the experiments was low and somewhat different. Heterogeneity in throughput of nanopore-sequencing experiments has been associated with the number of working pores of the flow cells used (Brown 2015), which can vary greatly, as a result of the manufacture of the flow cell itself and the storage conditions to which it was subjected. 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 in the literature using the same R7.3 chemistry. For instance, Quick, Quinlan & Loman (2014) reached 247 Mb of 1D data, almost 6 times more data than the highest throughput we achieved. However, there is no mention on the number of active nanopores of the flow cells used.

Throughout the period we worked with nanopore sequencing, we witness an evolution in pore availability. The first flow cells showed as little as 100 active nanopores, greatly conditioning the amount of data that could be retrieved from it. It seems that flow cells were being damaged in transit, with the formation of air bubbles over the nanopore chip. The air-liquid interface was reported to be mechanically erasing the nanopore chip. Nevertheless, improvements were substantial in the latter flow cells received. Currently, the new flow cell design and chemistry of the nanopore technology exhibit a significant progress in throughput, with the latest reports showing as much as 2 Gb of sequences per flow cell (Brown 2016).

It has been reported that the proportion of 2D reads can also vary between experiments, with as much as 70% 2D reads being the best described till now. Here we have seen (Figure 3.2.1.1) that from the total reads passing the pore (8 504 for Run 1 and 3 720 for Run 2), approximately 50% (4 215 reads) originated 2D consensus in Run 1, and 36% in Run 2 (1 342 reads). From these 2D reads, only 38% (Run 1) and 40% (Run 2) were 2D Pass reads.

The proportion of reads that generated a 2D consensus was quite low compared to the best-case scenario where each template read would have a paired complement and generate a consensus 2D read. It seems the limiting factor was the number of molecules that did not had the complement strand sequenced. From those who indeed had both template and complement strands sequenced (1D T + C in Figure 3.2.1.1), a high proportion of reads (88% for Run 1 and 84% for Run 2) were able to generate 2D data. This means that the low proportion of 2D data most likely derives from the 2D library preparation. A possible explanation might reside in a low yield of properly hairpined DNA molecules. However, since there is an enrichment step in hairpined sequences during library preparation, it is more likely that the low proportion of 2D reads results from low DNA quality. For

instance, nicks in the template strands of the DNA will interrupt the 2D continuous sequencing of the two strands of a molecule, which will then only generate a 1D template read.

While Run1 had generally higher yields in terms of total bases and reads, it also demonstrated a 2D read length distribution much more skewed towards smaller read lengths that Run 2 (Figure 3.2.1.1). Indeed, the mean 2D read length for Run 1 was 2.97 Kb (Table 3.2.1.1), statistically different from the mean of Run 2, which was 6.60 Kb (Mann-Whitney test, $p$=2.91x10$^{-81}$, α=0.05).

Note that, the sequencing library was prepared targeting a mean 2D read length of approximately 6 Kb, by using an appropriate Covaris g-tube fragmentation protocol. The fragmentation protocol was applied not because it was necessary, but because it has been designed to optimize both data yield and read length. Nanopore sequencing can sequence very large fragments, but working with extremely long molecules during library preparation can eventually lead to a lower yield and quality of the sequencing libraries. As library preparation is performed, pipetting steps add up, and the likelihood of breaking long fragments increases; if it happens after the addition of the adapters and hairpin, it renders impossible 2D sequencing of the molecule. Thus, here we implemented the suggested protocol, which had been already tested in-house in preliminary experiments, generating a profile of 2D read lengths with a mean of approximately 6 Kb and a mode of 3 Kb.

The skewed distribution of 2D read lengths of Run 1 is reflected in an abnormally low 2D read length mode of 413 bases and median of 1.38 Kb (Table 3.2.1.1), even though the maximum length achieved was of 106 Kb. This implies a high concentration of low molecular weight 2D sequences in the sequencing library, which might have been a result of unwanted fragmentation of the DNA prior to library preparation - either during DNA extraction or fragmentation with the Covaris g-tube. That is, short molecules were subjected to library preparation, with hairpin adapter linkage, and sequenced by the device, leading to the generation of 2D short reads. If fragmentation had happened post adapter linkage, it would not allow for 2D sequencing, but rather the sequencing of small 1D template reads.

Run 2 however, had the desired average 2D read length of 6.60 Kb and a mode of 3.53 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 magnetic 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. Library preparation is still the read length limiting stage. Fortunately, sequencing kits were already optimized in latter versions of the technology to improve the quality of longer DNA libraries.

In terms of 2D mean quality scores, the two runs generated distributions (Figure 3.2.1.1) with means that, although statistically different (Mann-Whitney test, $p$=2.05x10$^{-7}$, α=0.05), are still very close in value, with Run 1 generating a slightly higher mean of 8.8 versus 8.7 for Run 2 (Table 3.2.1.1). This comes to show that quality depends more on the chemistry and basecalling of the technology and less on library or flow cell variability. Furthermore, in our preliminary experiments (data not shown) we verified that read quality is also independent of read length or even time of sequencing. This means that there is no evidence indicating that sequencing quality decreases with the

augmentation of sequence length, or throughout the sequencing experiment. Thus, the technology seems to enable the sequencing of long reads with no systematic decrement in the quality of the data.

## 3.3 Comparison of datasets and evaluation of read processing needs

### 3.3.1 2D reads represent a smaller but higher-quality fraction of the nanopore-sequencing data

2D nanopore sequencing generates three sets of data, one of which, 2D Pass, is automatically filtered with the intention of constituting the usable higher-quality dataset. Yet, in our preliminary tests (data not shown), we have found that using all 2D data, rather than just 2D Pass, can increase several fold the coverage of the dataset. Additionally, we have seen in the previous section, 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 use 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.

Overall, all datasets offer possible advantages and for that reason, we intended to better characterize each dataset to understand their usefulness for our ultimate goal of mining industrial enzymes. We anticipate that there are 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.

Since the overall yield of either run was lower than expected, firstly, the data from both experiments were pooled together to create a richer dataset, closer to what has been reported for the chemistry here used. Figure 3.3.1.1 presents summary metrics for the pooled data.

The genome that was used as a reference to assess dataset quality was chosen by submitting high quality nanopore-sequencing data (2D Pass reads) to RAST. The genome of *B. velezensis* strain FZB42 (formerly known as the type strain of *B. amyloliquefaciens* subs. *plantarum*) was the highest scored neighbor genome of our data, present in the RAST curated database. In parallel, and although not explored in depth in the body of this dissertation, the isolate was further subject to a nanopore-sequencing based real-time identification system, which allowed us to further corroborate the identity of the MG SD 082 isolate as *Bacillus velezensis* (Appendix D). These evidences are in accordance with what was expected based on the previous 16S rRNA partial gene sequence analysis. Thus, from this stage on, the genome of *B. velezensis* strain FZB42 was consistently used whenever a reference was necessary.

Pooling all 1D reads amounts to 75.48 Mb of data, with a mean read length of 4.3 Kb (Figure 3.3.1.1). This data represents a maximum theoretical coverage of 19.26-fold, using as reference the genome size of *B. velezensis* strain FZB42.

Pooling all 2D reads creates a 21.59 Mb 2D dataset with a read length mean of 3.8 Kb, which represents a 5.51-fold theoretical coverage of the expected genome (Figure 3.3.1.1). Furthermore, the

subset of 2D data with quality scores equal or above 9, that is, the 2D Pass data, accounts for a maximum theoretical coverage of 2.36-fold, with a mean read length of 3.7 Kb. Thus overall, even when pooling data from two independent sequencing experiments, we are still working with low-coverage 2D datasets.

As expected, quality scores are lower for 1D data than for the 2D consensus derived reads. 1D reads wonder around mean quality scores of 4.5 and 2D consensus data around 8.8. Subsetting 2D Pass data, which excludes 2D data with quality scores below 9, generates a dataset with mean quality scores of 10.1 (Figure 3.3.1.1).

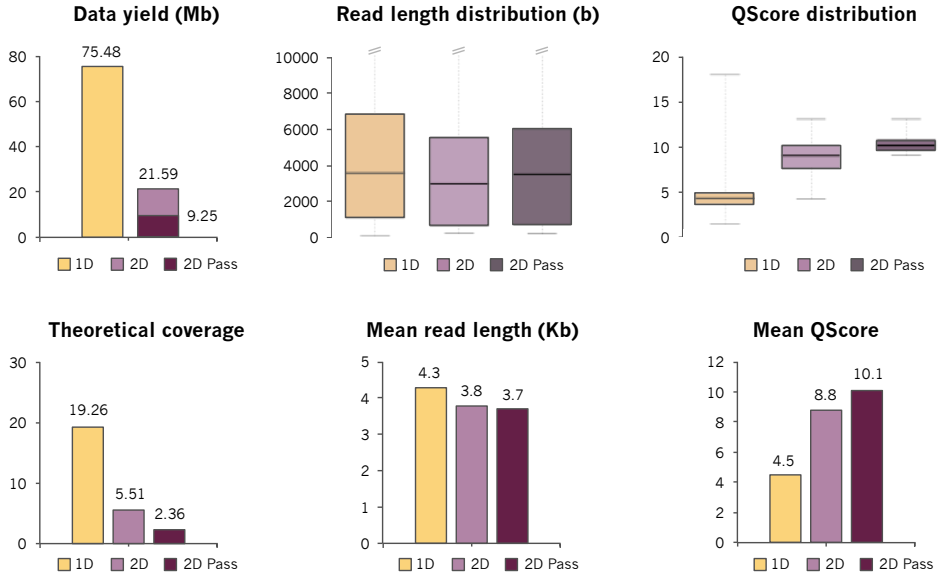


Figure 3.3.1.1 | **Yield, read and quality metrics of the repartitioned datasets 1D, 2D and 2D Pass.** Theoretical coverage was calculated as the ratio of data yield by the size of the genome of *B. velezensis* strain FZB42. Boxplots are represented with Spear whiskers that extend to minimum and maximum values. For 2D read length distributions the maximum values are omitted for clarity purposes – see Table 3.3.1.1. QScore of a read refers to the per base quality score mean. Mean read lengths (Kruskal-Wallis test, $p$=3.7x10$^{-31}$, $\alpha$=0.05) and mean quality scores (Kruskal-Wallis test, $p$=4.40x10$^{-232}$, $\alpha$=0.05) are statistically different between the datasets.

Each of the three possible datasets has a different profile, not so much in terms of read length, but mostly in terms of data yield and mean quality scores, meaning that the expected coverage and error rate distribution of the datasets is quite different. Indeed, we have already established that either dataset has mean read lengths sufficient to span entire bacterial genes - assuming an average gene size of 1 000-1 200 bases -, that the highest theoretical coverage is achieved with 1D data, and that 2D Pass data offers the highest quality data. These differences between the datasets are most likely going to lead to very dissimilar responses to downstream enzyme mining systems.

Note that, even though 2D reads have higher quality scores, for the R7.3 nanopore-sequencing chemistry, reports still average 2D error rates to 15% (Brown 2015). This is still a significant amount of errors, which will most likely impact the ability of algorithms to correctly identify ORFs from the raw data. Most groups exploring nanopore sequencing perform high-coverage sequencing and employ several iterations of corrections before and after assembling the data, with algorithms specifically prepared to *post hoc* process long error-prone reads, reducing the impact of

errors on the downstream analysis. But this post-experiment processing of the data adds a new level of complexity to the analysis, besides deeming the real-time character of nanopore sequencing useless. Thus 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 metrics, which are typically employed for draft assembly quality assessment. Here, these metrics were employed in an analogous manner to describe the read/contig length distributions of each of our datasets. N50 is the length of the largest contig that reaches half of the assembly total length in an array of all contigs ordered by size. In other words, 50% of the entire assembly is contained in contigs equal or greater than the N50 value. L50 is the smallest number of contigs whose lengths sum up to half of the assembly size, *i.e.* the number of contigs with lengths equal or greater than N50. Thus, the best datasets are those that have the highest N50 lengths and the lowest L50. NG50 and LG50 represent modified versions of this metrics that enable the comparison between different sized datasets (Bradnam *et al.* 2013), normalizing them to the size of the expected genome, rather than of the assembly. In this case, it took into consideration the size of the genome of the chosen reference *B. velezensis* strain FZB42.

Table 3.3.1.1 | **Read/contig metrics of the 1D, 2D and 2D Pass datasets and their corrected, assembled and polished versions.**

| Dataset | Total (b) | Number of reads/contigs | Number of reads/ contigs > 1 Kb | NG50 (b) | 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.
*1D template and 1D complement reads count as independent reads.

After processing the data, 1D non-processed reads still comprise the highest amount of sequencing data in terms of total bases and reads, representing the highest theoretical coverage we could achieve with the data generated (Table 3.3.1.1). Furthermore it offers the most promising NG50 (85 087 bases) and LG50 (8) metrics. Additionally, it comprises 12 408 reads with more than 1 Kb in length, providing a large amount of sequences, which could, in theory, span entire bacterial genes. For the purpose of finding genes, it is generally considered that a useful dataset, normally referring to a draft genome, is one that has a high number of scaffolds greater than the length of the average organism's gene (Bradnam *et al.* 2013). In the case of nanopore sequencing, reads can generally

surpass gene lengths, even with no assembly. Moreover, nanopore-sequencing non-processed data can surpass in length that of the contigs or scaffolds produced by assembling similar coverage data from second-generation technologies.

1D reads correction decreased greatly the amount of 1D data to 14% (10.22 Mb). The correction algorithm by Canu is a lossy process that tends to eliminate low quality regions of the data. Being that these data have low quality scores, as already established, this decrease, although major, it was not unexpected. However, it ended up overpowering the foreseen advantage of 1D data having higher coverage. Furthermore, the elimination of low quality data eventually led to a decrease in NG50 and an increase in LG50, which indicates an elimination of a part of the larger reads in the original dataset (Table 3.3.1.1).

The effect of correction in the 2D and 2D Pass datasets also had the same general impact. Yet, the decrease in data amount was less 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.

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, thus the lack of NG50 and LG50 metrics. One could assume that the assembly reduced the amount of data by compiling the 10.22 Mb of corrected data into non-redundant 260 contigs covering 618.65 Kb of the genome (Table 3.3.1.1). However, it seems quite unlikely. This dataset comprises only 154 reads with over 1000 bases, which is a farfetched assembly result taking into consideration the corrected dataset used, which had a promising NG50 of 74 291 bases and LG50 of 10 (Table 3.3.1.1) The results are most likely the reflection of aggressive trimming applied on the error-rich 1D reads by the assembly pipeline.

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. Yet again, it seems that the aggressive trimming, combined with the generally low coverage of the data (5.51 for 2D data and 2.36 for 2D Pass), trumped the benefits of assembling the reads, and decreased the NG50 length while increasing LG50 (Table 3.3.1.1).

Further polishing the assembled data, that is, applying post-assembly correction, did not seem to have any impact in the yield or read/contig length distribution of either dataset.

It seems there might not be an obvious advantage in assembling the data to increase sequence length and contiguity. The algorithms that are applied in error-prone reads depend on high coverage datasets to compensate for the aggressive trimming of lower quality data. Since we are working with low-coverage data, the algorithms underperformed. Nonetheless, assembling the data can bring several other benefits which are still worth exploiting. For instance, assembling reduces the redundant nature of a dataset by compiling repetitive information into consensus sequences, which not only augments accuracy but also reduces the downstream computational effort of analyzing repetitive information.

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 3.3.1.2). *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 (Dubinkina *et a*l. 2016). This method is thus a more suitable manner of comparing complete datasets, rather than just depending on the mapped/aligned regions. In the case where a dataset is compared with itself, the graphic representation of relative *k*-mer frequencies is seen as y=x. Comparisons straying from that line reflect differences in *k*-mer frequencies, and therefore, differences in sequence. Additionally, the entropy of the comparison can be calculated as a Kullback-Leibler divergence ($d^{KL}$).

Although 1D data had the highest potential for gene mining in terms of read length distribution and theoretical coverage, the data seems to be highly divergent from the reference ($d^{KL}$=0.176), as seen in Figure 3.3.1.2. Moreover, 1D data processing worsens the overall quality of the data ($d^{KL}$>>0.176). This indicates that the type and distribution of errors of these 1D reads does not fit well within the models used to correct and assemble the data.

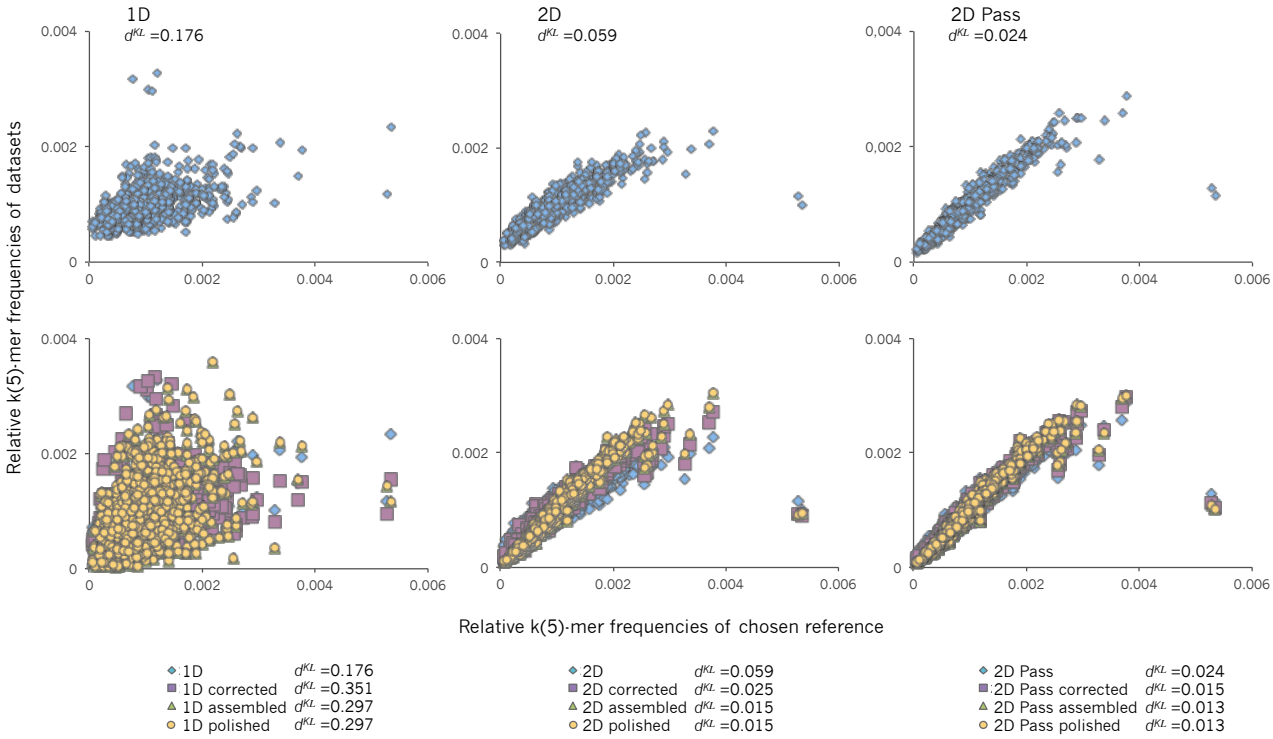


Figure 3.3.1.2 | ***K*(5)-mer relative frequencies comparison between each dataset and the chosen reference *B. velezensis* FZB42.** Each point in the graphs represents one of the 1024 possible *k*(5)-mers traced as its relative frequency in the specific dataset (y-axis) versus its relative frequency 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.

As expected by the quality score 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.

2D data is, as expected, more divergent than 2D Pass data, but when corrected, they reach 2D Pass levels ($d^{KL}$=0.025). This set of 2D corrected reads may bring an advantage over the 2D Pass data, since it comprises a larger dataset (15.87 Mb rather than 9.25 Mb), offering a 2-fold increase in theoretical coverage

Again, polishing the assembled data did not seem to have any obvious impact on the accuracy of the dataset. Just as the other algorithms, polishing depends on coverage, by aligning the squiggle information of the original reads against the assembled data. Its correction is based on the probability of the new assembled consensus sequences corresponding to the original squiggles. Yet, since we were working with low coverage original data and low quality assemblies, the changes in the polished dataset were very minimal, affecting only some scarce bases along the sequences (data not shown).

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. It could probably be less useful for the purpose of gene identification, affecting mapping and eventual genecalling. 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.

### 3.3.2 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, using as reference the genome of *B. velezensis* strain FZB42 (Figure 3.3.2.1 and Figure 3.3.2.2).

Although 1D data has demonstrated, in the previous sections, consistently lower quality scores and accuracy, it still had a high number of reads mapped (2546), only surpassed by the 2D dataset and its corrected version (Figure 3.3.2.1). However, mapped reads only represent 13.7% of the total reads of the 1D dataset. Moreover, on average, only 31.9% of the extension of a particular read is mapped, showing mean percent identities of 76.9% (Figure 3.3.2.2). 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 (Brown 2015; Loman, Quick & Simpson 2015), can create frameshifts that hinder the genecalling of the data. Thus, even though 1D data had a large potential in terms of maximum theoretical coverage, this did not revealed itself as a major contributor for the purpose of mining enzymes, due to the high error rates of the data.

Again, we can see how processing 1D reads was detrimental for their usefulness, reducing greatly the number of mapped reads from 2 456 to only 403. More importantly, the number of genes of the reference that were recalled decreased from 47 to only 1 after correction and 0 after assembly.

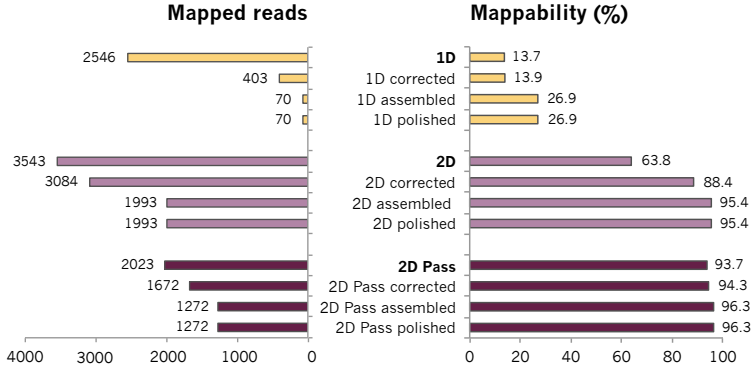


Figure 3.3.2.1 | **Mappability of 1D, 2D and 2D Pass datasets and their corrected, assembled and polished versions.** This figure shows both the total number of mapped reads as well as the percentage of reads of a particular dataset that was able to map against the genome of the chosen reference *B. velezensis* strain FZB42.

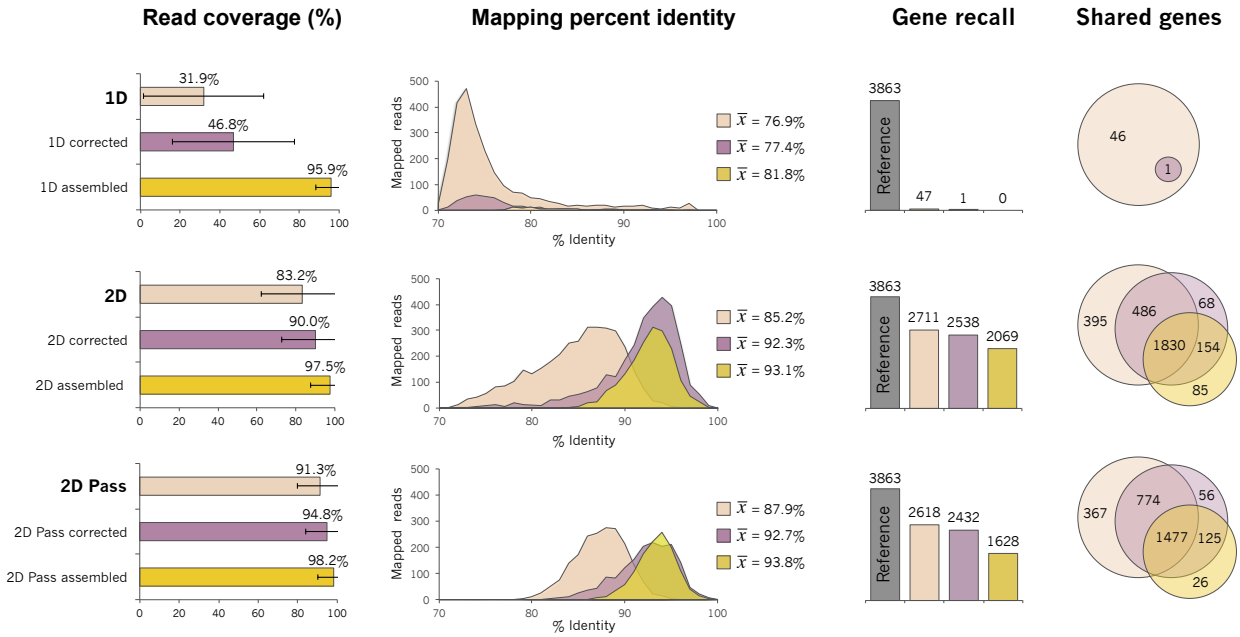


Figure 3.3.2.2| **Read mapping coverage, distribution of mapping percent identity and gene recall of each dataset.** These metrics were calculated using as reference the genome of *B. velezensis* strain FZB42. Results from the polished assembly were omitted since values were equal to the assembled datasets. '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.

Only 63.8% of the 2D data mapped against the reference (Figure 3.3.2.1). Still, it 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% (Figure 3.3.2.2). Regardless, 2D reads have 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 1152 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 (Brown 2015; Loman, Quick & Simpson 2015).

Correction of 2D reads shifted the mapping identity and read coverage up, reaching a mean of 92.3% and 90.0%, respectively (Figure 3.3.2.2). Yet, the loss of data in the correction process, discussed before, led to a decrease in the number of mapped reads (3 084), as well as 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.

The 2D Pass subset of 2D reads showed matches that span nearly the entire length of the reads (91.3%) with an average identity of 87.9%. Comparing the percent identity distribution of 2D reads with the 2D Pass subset reveals that they mostly differ in the subtraction of the lower tail of the 2D distribution, which reached 70% (Figure 3.3.2.2). With 2D Pass reads, mappings do not reach lower than 80% identity. The removal of this lower identity mapping was predictable, since 2D Pass reads are a subset created by the rejection of lower quality 2D reads. Nevertheless, even when subtracting a large part of the data, gene recall was still very high, with a total of 2 618 genes of the reference identified. Moreover, true depth of sequencing was verified to be as low as 2.3-fold (data not shown).

Just as for the other datasets, processing the reads led to an increase in read coverage and average percent identity but a decrease in gene recall. The corrected version of 2D Pass reads, although it had lower recall in general, enabled the recall of 181 new genes of the reference that were not pinpointed in the 2D Pass dataset. Further assembling the data disclosed 26 new genes.

Polishing the assembled datasets with Nanopolish did not alter the mapping or gene recalling metrics of either dataset as it was expected by the results obtained in the previous sections.

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 equal or 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 (in this case 21.59 Mb rather than 9.25 Mb), which not only requires more computational effort, but also more analysis time.

Curiously, although the 2D corrected and 2D Pass datasets had revealed before that they shared the same general similarity with the reference genome in terms of *k*-mer frequencies, the

distribution of their mapping percent identities is quite different. 2D corrected reads reach a much higher mean percent identity of 92.3% while 2D Pass reads average to 87.9% (Mann-Whitney test, $p$=1.40x10$^{-231}$, α=0.05). Indeed, percent identity only considers the fraction of reads that were mapped, while *k*-mer comparison takes into consideration the totality of the data, including reads that were unmapped. The unmapped reads of the 2D corrected dataset are most likely very erroneous and increase the entropy of *k*-mer comparisons, balancing out the higher accuracy of the remaining reads.

Either way, being that the 2D corrected dataset had a 2-fold higher coverage, we would expect an also higher gene recall. Yet, this did not happen. It might be that the higher number of mapped reads of the 2D corrected data were mainly redundant information, in combination with a loss of some information due to the lossy correction process. Nevertheless, upon further inspection we found that the two different datasets were able to grasp different genes. The 2D corrected dataset allowed to pinpoint 285 genes that were not unveiled with the 2D Pass data, and the 2D Pass data revealed 365 genes that the 2D corrected dataset failed to disclose.

2D reads are a particularly interesting dataset, since its use does not depend on any sort of processing. Furthermore, since 2D reads are the standard output of nanopore sequencing, they could potentially be analyzed in real-time. That is, in theory a read could be mined for coding sequences as soon as it is sequenced. Contrariwise, to use 2D corrected reads, the sequencing experiment has to be over, since the correction depends on sequence overlap determination and consensus creation. This would eliminate the possibility of real-time analysis, one of the most unique characteristics of the nanopore technology. Indeed, 2D corrected reads seem to have a much higher accuracy, which can be important for annotating and mining the sequencing data for enzymes. But 2D reads alone, with no correction or any sort of processing, should be able to give us enough accuracy to allow database comparison and good gene recall.

Moreover, 2D Pass is a very good alternative to 2D full data. It is in this higher-quality, lower-coverage data, that most of the gene mining potential is harbored. One can chose to take advantage of this dataset if worried with the extra computational effort involved with the full 2D dataset. Nevertheless, the 2D dataset still offers a higher coverage of the genome and higher gene recall in the case here portrayed. Since the increase in computational effort was not limiting in the specific context of this dissertation, the 2D dataset was chosen to be subjected to mining for industrial relevant enzymes.

## 3.4 Mining sequencing data for industrial relevant enzymes

### 3.4.1 Blast2GO - in combination with Prodigal – and RAST annotation systems generate different sets of annotations from the same 2D nanopore-sequencing data

All 2D non-processed data, consisting of a total of 5 557 reads amounting to 21.59 Mb, was

used for enzyme mining taking advantage of both RAST and Blast2GO annotation systems. For that purpose, the sequence data was first subjected to gene prediction algorithms. RAST uses an internal genecalling algorithm in its pipeline, contrary to Blast2GO, which, at the time it was used, still depended on external genecalling. Thus, 2D reads were genecalled with Prodigal before submission to the Blast2GO annotation pipeline.

As already discussed, nanopore-sequencing data generated by the R7.3 chemistry, is highly erroneous, even in its 2D consensus form. At this stage, there was no particular genecalling algorithm that was able to purposely surpass this limitation of the nanopore-sequencing data. We chose Prodigal because it is a very fast gene recognition tool, that is easily implemented locally on a computer, and performs well with a wide range of GC content genomes (Hyatt *et al*. 2010). Furthermore, Blast2GO developers had shown that Prodigal has the overall best performance when compared with Glimmer or GeneMarkS, offering a compromise between precision and recall[46].

As seen in Figure 3.4.1.1, the ORFs called by Prodigal, and submitted to Blast2GO reached a total of 21 348. Note that the original data used consisted of all 2D non-processed reads, which entails a redundant nature, and hence the high number of ORFs obtained. 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.

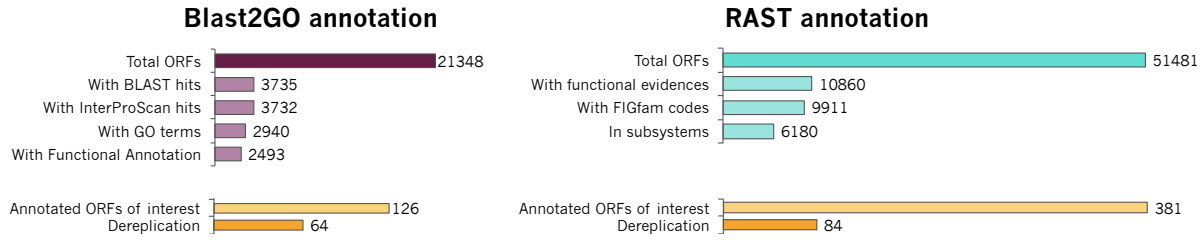


Figure 3.4.1.1| **Annotation metrics from the analysis of the *Bacillus velezensis* MG SD 082 2D-nanopore-sequencing data using Blast2GO and RAST.** The original dataset consisted of 5 557 non-processed 2D reads, amounting to 21.59 Mb. Blast2GO was used in combination with Prodigal genecaller.

At the end of the annotation process, Blast2GO had attributed functional annotations to 2 493 ORFs and RAST to 10 860, from which 9 911 were associated with FIGfams, and 6 180 were in subsystems – the highest level of annotation by RAST (Figure 3.4.1.1). In a first glance both systems were able to assign putative biological functions to an extensive set of ORFs; in between the annotations we were able to find enzymes involved in primary and secondary metabolism, carbohydrate-active enzymes, sporulation proteins, proteins involved in drug and metal resistance, membrane transporters for nutrient uptake, non-ribosomal peptide synthetases and others. The full set of annotations was then manually parsed to retrieve only those with industrial potential, specifically, starch-, cellulose-, xylan-, mannan-, pectin- and chitin-degrading enzymes, proteases and

[46] The tests by Blast2GO developers were done using the genome of *Streptococcus thermophilus* and can be found on https://www.blast2go.com/blast2go-pro/request-free-pro-trial/23-unpublished/116-genefinding.

lipases/esterases. In the set of RAST annotations we found 381 entries, which may represent proteins of interest. Blast2GO only generated 126.

Taking into consideration that the 2D data has the redundant nature of a non-assembled dataset, we were expecting some replicated annotations. Thus, at this stage we dereplicated the selected ORFs by eliminating repetitive annotations in different reads. This was done with the help of BLAST, to verify that equally annotated ORFs were indeed in reads spanning the same genome region and genes of the reference.

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. The genecallers applied seem to have fail to call entire genes, which leads to the appearance of several smaller ORFs, equally annotated, spanning different consecutive regions of the same read. Indeed, in the subset of the selected entries of interest we found that a gene could be annotated in as much as 4 fragments. 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, which were eventually compared and integrated. Blast2GO in combination with Prodigal generated a total of 64 annotations which fit into the industrial enzymes category, whereas RAST revealed 84 (Figure 3.4.1.1). Both annotation systems shared a total of 37 annotations and the remaining were uniquely identified by each system.

When evaluating the subset of 37 annotations that both systems generated in common, it became apparent that, as a result of the fragmented annotation noted before, the length of the predicted proteins was shorter than the reference's corresponding proteins (data not shown). Moreover, the mean length expected between the two annotation systems was also different (Mann-Whitney, $p=2.25\text{x}10^{-29}$, $\alpha=0.05$). Mean length for the reference set of proteins was 412 amino acids (±157)[47]. Contrariwise, for the same set of proteins, the mean length for Prodigal plus Blast2GO derived annotations was 92 amino acids (±37), and for RAST derived annotations it was 215 amino acids (±110). This differences mostly likely originated at the beginning by the application of different genecallers. Overall, it seems RAST over-performed when compared with Blast2GO, not only by generating more ORFs of interest, but also longer ORFs spanning a higher extension of the actual protein. Nevertheless, Blast2GO was still able to identify genes that were not pinpointed by RAST.

Either way, our intention was never to extensively compare the two annotation systems, which are intrinsically different and were bound to generate different results. Rather, we intended to expand the perspective over the data by applying different systems to analyze the same data.

[47] Values in brackets represent standard deviation.

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

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 111 putative industrial relevant enzymes capable of acting on the degradation of starch, cellulose, xylan, mannan, pectin, chitin, proteinaceous compounds and lipids (Figure 3.4.2.1). See Appendix E Table E.1 and Table E.2 for an extended list. Furthermore, the mining results seem to be in accordance with the phenotypic assays performed during the SEAVENTzymes project showed in Figure 3.1.4.1.

For instance, mining the whole-genome nanopore-sequencing data unveiled a putative extracellular α-amylase (EC 3.2.1.1). 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 (See Appendix A for a detailed description of the enzymes). 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.

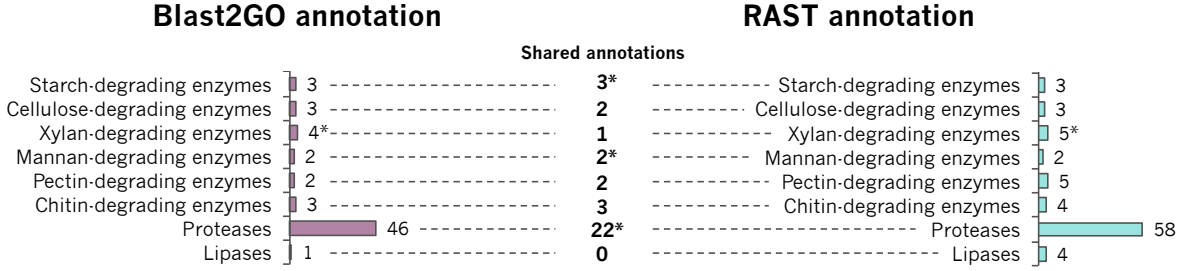


Figure 3.4.2.1| **Industrial relevant enzymes annotation overview from the analysis of the *Bacillus velezensis* MG SD 082 2D-nanopore-sequencing data using Blast2GO and RAST.** Blast2GO was used in combination with Prodigal genecaller. '*' indicates the presence of extracellular endo-hydrolytic enzymes.

In a similar manner, we were able to find several extracellular endo-acting 1,4-β-xylanases (EC 3.2.1.8), as well as several enzymes acting on the removal of xylan side-groups, such as an acetylxylan esterase (EC 3.1.1.72) and two α-*L*-arabinofuranosidases (EC 3.2.1.55). A mannan endo-1,4-β-mannosidase (EC 3.2.1.8) precursor, and at least 6 different extracellular proteases (EC 3.4.-.-) were also found. These enzymes could represent the underlying activities responsible for the positive results in the colorimetric assays with AZCL-xylan, AZCL-mannan and AZCL-casein, and in the growth assays with their respective natural substrates. In the particular case of the growth assays with casein, the NAUCr(ES)/NAUCr(BM) value observed was higher than 1, but still lower than the threshold for the definition of a positive result. Now, attending to the results of both colorimetric assays and whole-

genome data mining, we can consider that the negative result in growth assays is most likely erroneous.

Enzymes acting on pectin-degradation were also found in the set of selected annotations, including two pectin lyases (EC 4.2.2.10) and one pectase lyase (EC 4.2.2.2), as well as enzymes acting on pectin side groups such as an arabinogalactan endo-1,4-β-galactanase (EC 3.2.1.89) and an arabinan endo-1,5-α-*L*-arabinase (EC 3.2.1.99). This collection of predicted enzymes can potentially reflect the capability of the isolate to growth with pectin as the source of carbon, explaining the positive results obtained for the growth assays.

The same applies for the case of chitin-degrading enzymes. The data unveiled an endo-acting chitosanase (EC 3.2.1.132) which acts on the deacetylated form of chitin, *i.e.* chitosan, a β-hexosaminidase (EC 3.2.1.52), which acts on the external linkages of chitin/chitosan, as well as two chitooligosaccharides-acting deacetylases (EC 3.5.1.-). Although the isolate does not seem to have a chitinase (EC 3.2.1.14), which is, to a limited extent, corroborated by the absence of that enzyme in the closest neighbor genome, it shows a full set of enzymes that would allow the use of chitin, by deacetylating it into chitosan and subsequently degrade it. Although this is likely to happen with the natural chitin polymer used in the growth assays, it might not stand true for the modified chitin substrate used for the colorimetric assays. It could be that these enzymes are unable to act on the modified chitin and thus the negative discrepant result observed in the colorimetric assays.

Lipases/esterases production was detected in growth assays and may be a response of either one of the three carboxylesterases or two lipases found by mining the data. However neither of these enzymes had any indication of being extracellular. Thus, it remains the doubt regarding the positive results observed for the growth assays with “tweens”, since the use of the same substrate for colorimetric assays did not yield concordant results.

There was a specific result in the colorimetric assays for which we could not detect the responsible enzyme. That is, there was no evidence of an extracellular endo-acting cellulase in the data that could account for the positive results in the colorimetric assays. Rather, we identified some enzymes with potential to act on the external ends of cellulose, none of which had any indication of being extracellular. 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, by observing the coverage of the 2D data in the Integrative Genomics Viewer software, we found that neither of the two genes coding for putative endo-glucanases, *i.e.* 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.

Note that, although we were able to identify a series of genetic determinants that may be in the basis of the phenotypic results observed, we cannot by this approach obtain unquestionable associations between the predicted enzymes and the phenotypic screening results.

Furthermore, as sequencing-based bioprospecting goes, there is still much to be done. At this stage we only screened for enzymes of interest, and it is mostly in this step were sequencing can be an asset. But to truly evaluate the potential of an enzyme one must be able to grasp its activity in a non-abstract manner. Identifying the enzymes is just a first step. It has to be followed by heterologous expression of the predicted gene. Until then, its product remains a possibility rather that a certainty.

### 3.4.3 Whole-genome nanopore sequencing can be a valuable approach for the bioprospection of deep-sea hydrothermal vent prokaryotes

Overall, even with the erroneous nature of 2D data, and the consequent fragmentation of the protein sequences into smaller annotated chunks, we were still able to use whole-genome nanopore-sequencing long-read data to pinpoint industrial relevant enzymes from *Bacillus velezensis* MG SD 082, several of which may be responsible for the phenotypic results observed.

Although not directly explored during this dissertation, we were also able to find multiple other biotechnological interesting elements in the sequencing data, which transcended the realm of industrial biomass-degrading enzymes. For instance, several biosynthetic clusters of secondary metabolites were identified. Microbial secondary metabolism represents a rich source of high-value chemicals with potential therapeutic applications; genome mining of these gene clusters has become a trending approach for novel compound discovery (Martínez-Núñez & López 2016). Nanopore sequencing has the capability to improve this approaches by producing long reads that are able to span these complex biosynthetic gene clusters, which are usually repetitive and modular and are very hard to assemble with short-read data.

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 with a single non-directed experiment. Consequently, here, we were able to identify several different groups of enzymes 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.

Note that, some industry-associated biodiscovery projects still prefer target-based phenotypic screening approaches, since they usually have very specific goals and applications in mind. However, if one aims to grasp the overall biotechnological potential of an isolate or sample, sequencing approaches can be advantageous and unveil previously unfathomable potential.

As discussed before, the use of 2D long reads, that have the potential to span entire genes, allows to directly annotate the data with no need for data assembly or processing. However, erroneous 2D-nanopore-sequencing data may result in several genes being uncalled, or called in such a way that does not allow for proper annotation - falling under the quality threshold applied by the annotation systems. To counteract the erroneous nature of the 2D data, one could use datasets resulting from *post hoc* correction and assembly. But for the case here portrayed, where the original

data had very low-coverage, the processing of the reads proved to be limiting in the sense that it led to more loss of information than gain. Besides, the processing algorithms depend on the conclusion of the sequencing experiment to evaluate the pooled data and create a new consensus. Thus, even if we were working with higher coverage data, and processing would generate better quality datasets, the application of these algorithms would still render the real-time capability of the nanopore sequencing useless, which is an unfortunate disadvantage.

Besides the generation of long reads, and albeit not explored directly in the work pertained in this dissertation, it is the MinION real-time and portability aspects, accompanied by the possibility of sequencing sample metagenomes, 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 fixed 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 more of a 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 dissertation 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.

Annotation systems are becoming much faster and straightforward, but there is still no implementation of an annotation system in real-time, since this did not constitute a possibility till very recently. Even if real-time annotation is not implemented, and analysis is performed at the end of the sequencing experiment, the process from nanopore sequencing to results can take less than 24 hours with the latest improvements of the technology. Indeed, the run time of the sequencing experiment is as flexible as one may want, and it can be stopped as desired, depending solely on the goal. Furthermore, we have found that half of the data produced in a nanopore-sequencing run is obtained in the first 3 hours (data not shown).

Regardless, in the particular case here portrayed, the high error rate still persists as a major problem for the purpose of mining enzymes, and there are some approaches that can ameliorate the use of this data for our intended purpose. For instance, although the annotations span only small portions of the protein sequence, the original reads span the entire genes. That means that, in a first instance, this data can be used to quickly screen the potential of an isolate by annotating the called

ORFs, but if desired, *post hoc* data correction and curation can further augment the confidence on the genes of interest and their products. This curation is most likely necessary to guide the eventual protocols for heterologous expression of the genes for further physico-chemical studies.

One could alternatively argue in favor of the development of algorithms that better model gene finding and annotation with error-prone reads, but at the stage of the technology in the end of this dissertation, error rates have already decreased to levels close to second-generation technologies and soon, this may not represent a problem.

In the particular case here portrayed, there was also a fraction of the genes that were not identified due to the low coverage of the sequencing data, rather than the error rates of 2D data. Yet, taking into consideration the shallow resolution of the data, the results were still quite promising, delivering a large collection of annotations.

There is always a cost-benefit assessment that should be done, and one could chose to purposely sequence with lower coverage depending on the goal, since it represents a quicker, cheaper and less computational heavy approach. For instance, for a quick assessment of the biotechnological potential of an isolate, or even a sample, low-coverage sequencing, followed by direct annotation of long reads, can generate a collection of annotations that represent a broad overview of the activities and capabilities of the tested subject, even if failing to uncover certain genes. This is not completely farfetched since low-resolution data has already been used to make inferences about microbial community functionality (Fierer *et al.* 2012). After a superficial and quick assessment with low-coverage sequencing, further sequencing could be deployed if the isolate or sample in question revealed any characteristics of interest for the biodiscovery project. As previously mentioned, nanopore technology, and particularly the MinION, offers an easy implementation and high flexibility of usage, with sequencing runs being as long or as short as one may want. One could stop the sequencing when the desired depth has been reached.

But the low coverage in this dissertation did not result from a deliberate plan, but from an underwhelming low sequencing yield. The low data yields that were seen in this dissertation do not reflect the general yields reported for the nanopore technology and resulted from low quality flow cells. Either way, the problem of low coverage could be solved in a very natural manner, by increasing the sequencing yields with further sequencing experiments. That would, however, entail a consequent increase in the price and time of the experiment. With the current developments in the flow cells and the sequencing chemistry, as well as the introduction of fast-mode sequencing, yields are reaching much higher values for a single experiment, than those we were able to achieve. Either way, one could still choose to sequence with any depth desired.

Note that, by mining the sequencing data we were able to identify genes and enzymes with a high level of activity specificity, something that did not happen in the growth assays, for instance. Albeit this only happened because we were focusing on a set of enzymes for which there is already an extensive knowledge base.

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. The main approaches to predict function of a called ORF are based on the recognition of sequence similarity with already described genes, from which functional homology is inferred with variable levels of confidence. Yet, the vast majority of predicted ORFs in databases still lack any association with functionality - the so-called hypothetical genes. We can easily find cases where 20% to 50% of the ORFs of a genome are still of unknown function, even if many of them are conserved among several organisms (Eisenstein *et al.* 2000). Many of the ORFs in the nanopore-sequencing data of the MG SD 082 isolate were annotated as hypothetical proteins. Some may represent false called genes, but others may actually code for enzymes, which could be of interest in the context of the investigation and which we will fail to detect. In the case of this dissertation, this was not a major concern, since the aim was to identify novel variants of conserved and known protein groups, rather than completely new functions.

Nonetheless, understanding the physiological function of these gene products is the major challenge that still limits the use of sequencing technologies for bioprospection.

Some strategies have emerged to surpass this problem. Advances in comparative genomics have recently inspired several initiatives that aim to annotate genomes via gene context. This approaches go beyond the simple recognition of sequence similarity. For instance, the functional association of proteins is sometimes reflected in their cohabitation into operons or cluster of genes, or in their evolution in a correlated manner and their fusion as a single sequence in another organism (Enault, Suhre & Clevarie 2005). The application of nanopore sequencing, and other third-generation sequencing technologies, can have a positive impact on this approach by generating long reads and consequently improving contiguity of genomes and enabling a better overview of gene context.

Other tactics are based in the coupling of all ‘omics’ technologies, such as genomics, transcriptomics, proteomics and metabolomics into intricate systems biology approaches. The democratization of ‘omics’ technologies, in terms of pricing and analytical tools, will almost certainly convert this strategy from a novelty status to a standard procedure. Nanopore sequencing already enables direct RNA sequencing with the same portable device by implementing a different library sequencing kit. The long-term prospection is to eventually enable the sequencing of proteins, which is not an implausible next stage of the technology, knowing that the new R9 technology uses a CsgG nanopore whose function in nature is to secrete peptides. Thus, nanopore sequencing has the potential to become an integrated interface for coupling omic’s technologies into a single device.

Fortunately, even though some genes may escape us under our current understanding of sequence information, the sequence data obtained has permanent character. This means it can be revisited time and time again as our knowledge base augments. Sequencing data increases its value with time, as knew methods of studying and understanding these sequences develop and disclose new opportunities and potential in “old” data.

Overall, this leads us to propose that whole-genome nanopore sequencing has the potential to become a relevant system for the biotechnological assessment of prokaryotic isolates or samples from deep-sea hydrothermal vents or other environments (Appendix F).

# Chapter 4. Conclusions and future perspectives

The oceans cover over 70% of Earth's surface. This is probably the most common phrase in marine sciences literature. The intent of this phrase is not so much to inform, but rather to evoke in the reader a promise of what the oceans still have to offer. This was also the case in the context of this dissertation. Indeed, the oceans' potential is immeasurable and marine biotechnology has been able to grasp what we can only imagine to be a very small fraction of it. Exploring microorganisms in unexamined habitats employing novel screening methodologies has been suggested as a way to further propel the field. The broader theme of this dissertation fits into the suggested guidelines, by aiming to explore the biotechnological potential of deep-sea hydrothermal vent microorganisms, using novel sequencing technologies.

This dissertation was the continuation of the work that began with the visit to the hydrothermal vents surronding the Azores during the SEAHMA project, from which a collection of isolates was obtained and explored in the SEAVENTzymes project. This project emerged to assess the potential of the collection for the production of industrial relevant enzymes, which it was able to do, to some extent, by taking advantage of phenotypic screening assays.

Classical phenotypic assays are still a major part of any bioprospection pipeline. To truly evaluate the potential of an enzyme one must be able to evaluate its function in a non-abstract manner. However, as a screening method for new enzymes in hydrothermal vent microorganisms it has some disadvantages. For instance, it is not a suitable procedure for the study of fastidious hydrothermal vent microorganisms, which require growth conditions that are not compatible with many of the assays. Indeed, the phenotypic screening during the SEAVENTzymes project was only implemented on a minor subset of isolates of the collection, the mesophilic aerobes. This approach ended up excluding the major source of potential of thermo-resisting enzymes, the thermophilic isolates.

Sequencing methods of screening offer great advantages over the screening approaches taken during the first SEAVENTzymes project. They are quite versatile and can be implemented to bioprospect in a culture-dependent or -independent manner, which makes more sense for the study of

hydrothermal microorganisms. Moreover, whole-genome sequencing screening acts as a window to the full genomic potential of an isolate or a sample, deeming the screening independent of multiple focused tests, enzyme expression conditions, and even growth requirements of the microorganism.

Nanopore sequencing, in particular, brings an additional set of novel characteristics to the sequencing territory. It produces single-molecule long reads, from a straightforward library preparation, and enables real-time portable sequencing. From the nature of this technology, we were anticipating some competitive advantages for the bioprospecting of enzymes from hydrothermal vent microorganisms. For instance, long reads, spanning entire gene clusters, would enable direct annotation with no need for prior read assembly. Consequently, the reads could be directly streamed to data mining pipelines as soon as they are sequenced, taking advantage of the real-time capability of this technology. Moreover, the portability of the MinION nanopore device would eventually enable the implementation of in-field real-time sequencing.

As the long-term intention is to implement nanopore sequencing to reassess the biotechnological potential of the SEAVENTbugs collection, we first had to evaluate if indeed this technology has the potential for it. In a primary stage, we aimed to evaluate the use of nanopore sequencing to screen a small set of industrial relevant enzymes from a single isolate of the SEAVENTbugs collection. We set ourselves to choose a promising isolate by reevaluating the results of the SEAVENTzymes project. By integrating all the results, we were able to select a good candidate, even though there were disparities between data resulting from different phenotypic approaches. This isolate, a *Bacillus* sp. from the Menez Gwen sediments, was successfully subjected to whole-genome nanopore sequencing.

The nanopore-sequencing device was simple to implement, yet, the heterogeneity in throughput between flow cells was underwhelming. Nevertheless, with the new R9 chemistries released for this technology, reports are showing much more consistent results (Brown 2016). That means that at this point, one could expect to obtain more and better quality data from this technology. Independently of this current advance, we were still able to obtain results in accordance with what we already knew about the isolate's production of industrial relevant enzymes, despite the lower-throughput and less-than-optimum error rates of the used R7.3 version of the technology.

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.

Thus, in this dissertation, we indeed proof-of-concept the use of whole-genome nanopore sequencing, in combination with automatic annotation systems, to evaluate the biotechnological potential of a *Bacillus* sp. isolate from hydrothermal vent sediments, with regard to industrial relevant enzymes.

Although not explored in depth in this dissertation, ultimately, the same whole-genome

nanopore-sequencing data also allowed to unveil several other ORFs of biotechnological interest that transcended our initial set of industrial relevant enzymes. Moreover, it enabled the identification of the isolate at the species level – *Bacillus velezensis* -, with only minor additional effort, which comes to show the versatility of the data.

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.

That is, nanopore-sequencing long reads enable direct annotation with no need for data processing, reducing the analysis time and complexity. In turn, this lack of processing requirements can eventually support the real-time and in-field implementation of sequencing-based screening. Consequently, this technology can transform the bioprospecting of remote locations into a more proactive activity, guiding the sampling strategy itself whilst on the field. Moreover, being a sequencing-based strategy, it can enable culture-dependent and -independent analysis. Thus, overall it has the competency to overview the biotechnological potential of a sample in a quick and straightforward manner.

As for the SEAVENTbugs collection, we have still not grasped all its potential. We are now in a position were we can implement this technology to bioprospect all isolates of interest, maybe by barcoding them and sequencing them in batches, or in a more broad approach by sequencing the metagenome of the preserved samples.

Yet, as bioprospecting goes, there is still much to be done. Identifying the enzymes is just a first step. It has to be followed by heterologous expression of the predicted gene and biochemical characterization of its product. 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. Although here, nanopore-sequencing reads shown high error rates, generating fragmented annotations, they still span the entirety of the genes. *Post hoc* correction of the data can improve its quality and facilitate the isolation of the genes of interest for cloning by guiding, for instance, the proper design of primers for the amplification of the selected genes, the design of targeting sequences for CRISPR-Cas9[48] mediated gene isolation and enrichment, or alternatively, the construction of bait sequences for target enrichment and capture systems.

Overall, even with the current limitations of sequence-based methods, the MinION revealed itself as being a useful and accessible sequencing platform. Its portability and real-time potential was not explored directly in the work pertained in this dissertation, nor its implementation with metagenomic samples but, 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.

[48] CRISPR-Cas9 is a novel genome editing technology that takes advantage of a bacterial defense system where a RNA-guided Cas nuclease cuts foreign genetic elements. For more information on this system see Fujita, Yuno & Fujii 2015.

# References


Adams, M.W., Perler, F.B. & Kelly, R.M., 1995. Extremozymes: expanding the limits of biocatalysis. *Biotechnology,* 13(7), pp.662-668.

Akopyanz, N. *et al.,* 1992. DNA diversity among clinical isolates of *Helicobacter pylori* detected by PCR-based RAPD fingerprinting. *Nucleic Acids Research*, 20(19), pp.5137-5142.

Allsopp, M., *et al.,* 2009. State of the World's Oceans, *Amsterdam, The Netherlands: Springer,* pp.1-10.

Altschul, S.F. *et al.*, 1990. Basic local alignment search tool. *Journal of Molecular Biology*, 215(3), pp.403-410.

Anderson, E.R., Sogin, M.L & Baross, J.A., 2015. Biogeography and ecology of the rare and abundant microbial lineages in deep-sea hydrothermal vents. *FEMS Microbiology Ecology,* 91(1), pp.1-11.

Appeltans, W. *et al.*, 2012. The magnitude of global marine species diversity. *Current Biology*, 22(23), pp.2189-2202.

Arezi, B. *et al.*, 2003. Amplification efficiency of thermostable DNA polymerases. *Analytical Biochemistry*, 321(2), pp.226-235.

Ash, C. *et al.,* 1991. Phylogenetic heterogeneity of the genus *Bacillus* revealed by comparative analysis of small-subunit-ribosomal RNA sequences. *Letters in Applied Microbiology*, 13(4), pp.202-206.

Ashton, P.M. *et al.,* 2015. MinION nanopore sequencing identifies the position and structure of a bacterial antibiotic resistance island. *Nature Biotechnology*, 33(3), pp.296-300.

Ashwini, K. *et al*., 2013. Purification and activity of amylases of Marine *Halobacillus* sp. *amylus* HM454199. *Indian Journal of Geo-Marine Sciences*, 42(6), pp.781-785.

Asoodeh, A., Chamani, J. & Lagzian, M., 2010. A novel thermostable, acidophilic-amylase from a new thermophilic "*Bacillus* sp. Ferdowsicous" isolated from Ferdows hot mineral spring in Iran: Purification and biochemical characterization. *International Journal of Biological Macromolecules*, 46(3), pp.289-297.

Ayadi, M.A. *et al.*, 2009. Influence of carrageenan addition on turkey meat sausages properties. *Journal of Food Engineering*, 93(3), pp.278-283.

Barbieri, E. *et al.,* 1999. Occurrence, diversity, and pathogenicity of halophilic *Vibrio* spp. and non-O1 *Vibrio cholerae* from estuarine waters along the Italian Adriatic coast. *Applied and Environmental Microbiology*, 65(6), pp.2748-2753.

Barriga, F.J.A.S. *et al*., 2013. Estimating and finding seafloor and subseafloor sulfide mineralization : optimists versus pessimists. Paper presented at the *42nd Conference of the Underwater Mining Institute: Recent Developments in Atlantic Seabed Minerals Exploration and Other Topics*, Rio de Janeiro & Ouro Preto & Porto de Galinhas, Brazil, October 21st-29th, 2013.

BCC Research, 2014. Global Markets for Enzymes in Industrial Applications. Available at *http://www.bccresearch.com/market-research/biotechnology/enzymes-industrial-applications-bio030h.html.* Accessed September 2nd, 2016.

Beaumont, N.J. *et al*., 2008. Economic valuation for the conservation of marine biodiversity. *Marine Pollution Bulletin*, 56(3), pp.386-396.

Bell, E.M. & Heuer, V.B., 2012. The Deep Biosphere: Deep Subterranean and Subseafloor Habitats. In Bell, E.M., ed., Life at Extremes: Environments, Organisms and Strategies for Survival, *Wallingford, United Kingdom: CABI*, pp.345-363.

Besemer, J. & Borodovsky, M., 2005. GeneMark: Web software for gene finding in prokaryotes, eukaryotes and viruses. *Nucleic Acids Research*, 33, pp.451-454.

Bhandari, V. *et al*., 2013. Molecular signatures for *Bacillus* species: Demarcation of the *Bacillus subtilis* and *Bacillus cereus* clades in molecular terms and proposal to limit the placement of new species into the genus *Bacillus*. *International Journal of Systematic and Evolutionary Microbiology*, 63(7), pp.2712-2726.

Boers, S.A., Hays, J.P. & Jansen, R., 2015. Micelle PCR reduces chimera formation in 16S rRNA profiling of complex microbial DNA mixtures. *Scientific reports*, 5, p.14181.

Boeuf, G., 2011. Marine biodiversity characteristics. *Comptes Rendus - Biologies*, 334(5-6), pp.435-440.

Bouchet, P., 2006. The magnitude of marine biodiversity. In Duarte, C.M., ed., The Exploration of Marine Biodiversity: Scientific and Technological Challenges, *Bilbao, Spain: Fundación BBVA*, pp.31-62.

Bradnam, K.R. *et a*l., 2013. Assemblathon 2: evaluating *de novo* methods of genome assembly in three vertebrate species. *GigaScience*, 2, p.10.

Brandt, A., 2008. Deep-Sea Ecology: Infectious Impact on Ecosystem Function. *Current Biology*, 18(23), pp.1104-1106.

Briggs, J.C., 1987. Biogeography and plate tectonics, *Amsterdam, The Netherlands: Elsevier*, pp.157-167.

Brown, C., 2015. Owl Stretching with examples. Plenary lecture in *London Calling 2015 by Oxford Nanopore Technologies*, London, United Kingdom, May 14th-15th, 2016. Available at *https://vimeo.com/128281064.* Accessed September 3rd, 2016.

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.

Burgess, J.G., 2012. New and emerging analytical techniques for marine biotechnology. *Current Opinion in Biotechnology*, 23(1), pp.29-33.

Camacho, C. *et al.,* 2009. BLAST plus: architecture and applications. *BMC Bioinformatics*, 10(1), p.421.

Canals, M. *et al.,* 2006. Flushing submarine canyons. *Nature*, 444(7117), pp.354-357.

Cao, M.D. *et al.,* 2015. Real-time analysis and visualization of MinION sequencing data with npReader. *Bioinformatics*, 32(5), pp.764-766.

Cardigos, F. *et al.*, 2005. Shallow water hydrothermal vent field fluids and communities of the D. João de Castro Seamount (Azores). *Chemical Geology*, 224(1-3), pp.153-168.

Carrasco, M. *et al*., 2016. Screening and characterization of amylase and cellulase activities in psychrotolerant yeasts. *BMC Microbiology*, 16, p.21.

Cerqueira, T. *et al.,* 2015. Microbial diversity in deep-sea sediments from the Menez Gwen hydrothermal vent system of the Mid-Atlantic Ridge. *Marine Genomics*, 24, pp.343-355.

Chalfie, M. *et al.*, 1994. Green fluorescent protein as a marker for gene expression. *Science* 263(5148), pp.802-805.

Chen, X.H. *et al.,* 2007. Comparative analysis of the complete genome sequence of the plant growth-promoting bacterium *Bacillus amyloliquefaciens* FZB42. *Nature Biotechnology*, 25(9), pp.1007-1014.

Cheng, L. *et al.*, 2016. Purification and characterization of a thermostable $\beta$-Mannanase from *Bacillus subtilis* BE-91 : Potential application in inflammatory diseases. *BioMed Research International,* 2016, p.6389147.

Chulhong, O. *et al.*, 2011. Isolation, purification, and enzymatic characterization of extracellular chitosanase from marine bacterium *Bacillus subtilis* CH2. *Journal of Microbiology and Biotechnology*, 21(10), pp.1021-1025.

Colaço, A. *et al*., 2006. Bioaccumulation of Hg, Cu, and Zn in the Azores triple junction hydrothermal vent fields food web. *Chemosphere*, 65(11), pp.2260-2267.

Conesa, A. *et al*., 2005. Blast2GO: A universal tool for annotation, visualization and analysis in functional genomics research. *Bioinformatics*, 21(18), pp.3674-3676.

Connon, S.A. & Giovannoni, S.J., 2002. High-throughput methods for culturing microorganisms in very-low-nutrient media yield diverse new marine isolates. *Applied and Environmental Microbiology*, 68(8), pp.3878-3885.

Corliss, J.B. *et al*., 1979. Submarine Thermal Springs on the Galápagos Rift. *American Association for the Advancement of Science*, 203(4385), pp.1073-1083.

Costa, M., 2015. What, if anything, is an extremophile? Plenary lecture in *MicroBiotec'15*, Évora, Portugal, December 10$^{th}$-12$^{th}$, 2015.

Cottrell, M.T. *et al.,* 2000. Selected chitinase genes in cultured and uncultured marine bacteria in the alpha and gama-subclasses of the *Proteobacteria. Applied and Environmental Microbiology,* 66(3), pp.1195-1201.

Cowan, D. *et al.*, 2005. Metagenomic gene discovery: Past, present and future. *Trends in Biotechnology*, 23(6), pp.321-329.

Cui, H.-L. *et al.,* 2006. *Halorubrum lipolyticum* sp. nov. and *Halorubrum aidingense* sp. nov., isolated from two salt lakes in Xin-Jiang, China. *International Journal of Systematic and Evolutionary Microbiology*, 56(7), pp.1631-1634.

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.

Danovaro, R., Snelgrove, P.V.R. & Tyler, P., 2014. Challenging the paradigms of deep-sea ecology. *Trends in Ecology and Evolution*, 29(8), pp.465-475.

DasSarma, S., Coker, J.A & DasSarma, P., 2009. *Archaea* (Overview). In Schaechter, M., ed., Encyclopedia of Microbiology Third Edition, *Oxford, United Kingdom: Elsevier*, Volume 2, pp.1-26.

Dean, F.B. *et al.,* 2002. Comprehensive human genome amplification using multiple displacement amplification. *Proceedings of the National Academy of Sciences*, 99(8), pp.5261-5266.

Delcher, A.L. *et al.,* 1999. Improved microbial gene identification with GLIMMER. *Nucleic acids research*, 27(23), pp.4636-4641.

Delseny, M., Han, B. & Hsing, Y.I., 2010. High-throughput DNA sequencing: The new sequencing revolution. *Plant Science*, 179(5), pp.407-422.

Demirjian, D.C., Morís-Varas, F. & Cassidy, C.S., 2001. Enzymes from extremophiles. *Current Opinion in Chemical Biology*, 5(2), pp.144-151.

Dias, Á.S. & Barriga, F.J.A.S., 2006. Mineralogy and geochemistry of hydrothermal sediments from the serpentinite-hosted Saldanha hydrothermal field (36°34′N; 33°26′W) at MAR. *Marine Geology*, 225(1-4), pp.157-175.

Dick, G.J. & Tebo, B.M., 2010. Microbial diversity and biogeochemistry of the Guaymas Basin deep-sea hydrothermal plume. *Environmental Microbiology*, 12(5), pp.1334-1347.

Dionisi, H.M., Lozada, M. & Olivera, N.L., 2012. Bioprospection of marine microorganisms: biotechnological applications and methods. *Revista Argentina de Microbiología*, 44(1), pp.49-60.

Dubinkina, V.B. *et al*., 2016. Assessment of *k*-mer spectrum applicability for metagenomic dissimilarity analysis. *BMC Bioinformatics*, 17, p.38.

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.

Dupont, C.L. *et al.*, 2015. Genomes and gene expression across light and productivity gradients in eastern subtropical Pacific microbial communities. *The International Society for Microbial Ecology Journal*, 9(5), pp.1076-1092.

Edwards, K.J., Wheat, C.G. & Sylvan, J.B., 2011. Under the sea: microbial life in volcanic oceanic crust. *Nature Reviews Microbiology*, 9(10), pp.703-712.

Edwards, R. *et al.*, 2006. Using pyrosequencing to shed light on deep-mine microbial ecology. *BMC Genomics*, 7, p.57.

Egan, S., Thomas, T. & Kjelleberg, S., 2008. Unlocking the diversity and biotechnological potential of marine surface associated microbial communities. *Current Opinion in Microbiology*, 11(3), pp.219-225.

Egorova, K. & Antranikian, G., 2005. Industrial relevance of thermophilic *Archaea*. *Current Opinion in Microbiology*, 8(6), pp.649-655.

Eisenstein, E. *et al.,* 2000. Biological function made crystal clear---annotation of hypothetical proteins via structural genomics. *Current Opinion in Biotechnology*, 11(1), pp.25-30.

Elleuche, S. *et al.*, 2014. Extremozymes-biocatalysts with unique properties from extremophilic microorganisms. *Current Opinion in Biotechnology*, 29, pp.116-123.

Elleuche, S. *et al*., 2015. Exploration of extremophiles for high temperature biotechnological processes. *Current Opinion in Microbiology*, 25, pp.113-119.

Enault, F., Suhre, K. & Claverie, J.M., 2005. Phydbac "Gene Function Predictor": a gene annotation tool based on genomic context analysis. *BMC Bioinformatics*, 6, p.247.

English, A.C. *et al.*, 2012. Mind the Gap: Upgrading genomes with Pacific Biosciences RS long-read sequencing technology. *PLOS ONE*, 7(11), p.47768.

Estrutura de Missão para a Extensão da Plataforma Continental, 2015. Projeto de Extensão da Plataforma Continental. Available at *http://www.emepc.pt*. Accessed August 30th, 2016.

Etter, W. & Hess, H., 2015. Reviews and syntheses: The first records of deep-sea fauna - A correction and discussion. *Biogeosciences*, 12(21), pp.6453-6462.

Ettoumi, B. *et al.,* 2009. Diversity and phylogeny of culturable spore-forming *Bacilli* isolated from marine sediments. *Journal of Basic Microbiology*, 49(1), pp.13-23.

Ettoumi, B. *et al.,* 2013. Microdiversity of deep-sea *Bacillales* isolated from Tyrrhenian sea sediments as revealed by ARISA, 16S rRNA gene sequencing and BOX-PCR fingerprinting. *Microbes and Environments*, 28(3), pp.361-369.

European Comission Directorate-General for Maritime Affairs and Fisheries, 2008. EU Maritime Policy: Facts and Figures – Portugal. Available at *https://bookshop.europa.eu/en/eu-maritime-policy-pbKL7807407/.* Accessed August 30th, 2016.

Farrant, G.K. *et al.*, 2016. Delineating ecologically significant taxonomic units from global patterns of marine picocyanobacteria. *Proceedings of the National Academy of Sciences*, 113(24), pp.3365-3374.

Fautin, D. *et al.*, 2010. An overview of marine biodiversity in United States waters. *PLOS ONE*, 5(8), p.11914.

Ferreira, A.C.R. *et al.*, 2015. DGGE and multivariate analysis of a yeast community in spontaneous cocoa fermentation process. *Genetics and Molecular Research*, 14(4), pp.18465-18470.

Ferrera, I., Banta, A.B. & Reysenbach, A.-L., 2014. Spatial patterns of *Aquificales* in deep-sea vents along the Eastern Lau Spreading Center (SW Pacific). *Systematic and Applied Microbiology*, 37(6), pp.442-448.

Fiedler, H.-P. *et al.*, 2008. Proximicin A, B and C, novel aminofuran antibiotic and anticancer compounds isolated from marine strains of the actinomycete *Verrucosispora*. *The Journal of Antibiotics*, 61(3), pp.158-163.

Fierer, N. *et al.,* 2012. Comparative metagenomic, phylogenetic and physiological analyses of soil microbial communities across nitrogen gradients. *The International Society for Microbial Ecology Journal*, 6(5), pp.1007-1017.

Firmino, T., 2016. A plataforma continental, a missão que aí vem e a novidade de uma adenda. Available at *https://www.publico.pt/2016/08/27/ciencia/noticia/a-plataforma-continental-a-missao-que-ai-vem-e-a-novidade-de-uma-adenda-1742414.* Accessed August 30th, 2016.

Fischer, S.G. & Lerman, L.S., 1979. Length-independent separation of DNA restriction fragments in two-dimensional gel electrophoresis. *Cell*, 16(1), pp.191-200.

Flores, G.E. *et al.,* 2012. Distribution, abundance, and diversity patterns of the thermoacidophilic "deep-sea hydrothermal vent euryarchaeota 2". *Frontiers in microbiology*, 3(1):47.

Fredriksson, N.J., Hermansson, M. & Wilén, B., 2013. The choice of PCR primers has great impact on assessments of bacterial community diversity and dynamics in a wastewater treatment plant. *PLOS ONE*, 8(10), p.76431.

Fujita, T., Yuno, M. & Fujii, H., 2015. Efficient sequence-specific isolation of DNA fragments and chromatin by *in vitro* enChIP technology using recombinant CRISPR ribonucleoproteins. *bioRxiv*, p.033241.

Fung, A., Hamid, N. & Lu, J., 2013. Fucoxanthin content and antioxidant properties of *Undaria pinnatifida*. *Food Chemistry*, 136(2), pp.1055-1062.

Global Industry Analysts, Inc., 2015. The Global Marine Biotechnology Market Trends, Drivers & Projections. Available at *http://www.strategyr.com/MarketResearch/Marine_Biotechnology_ Market_Trends.asp.* Accessed September 1st, 2016.

Glowka, L., 2003. Putting marine scientific research on a sustainable footing at hydrothermal vents. *Marine Policy*, 27(4), pp.303-312.

Gobalakrishnan, R. & Sivakumar K., 2016. Systematic characterization of potential cellulolytic marine actinobacteria *Actinoalloteichus* sp. MHA15. *Biotechnology Reports*, 12, pp30-36.

Gonzalez, J.M. *et al*., 2012. Amplification by PCR artificially reduces the proportion of the rare biosphere in microbial communities. *PLOS ONE*, 7(1), p.29973.

Goodfellow, M. & Fiedler, H.P., 2010. A guide to successful bioprospecting: Informed by actinobacterial systematics. *Antonie van Leeuwenhoek,* 98(2), pp.119-142.

Goodwin, S., McPherson, J.D. & McCombie, W.R., 2016. Coming of age: ten years of next-generation sequencing technologies. *Nature Reviews Genetics*, 17(6), pp.333-351.

Guckert, J.B. *et a*l., 1996. Community analysis by Biolog: curve integration for statistical analysis of activated sludge microbial habitats. *Journal of Microbiological Methods*, 27, pp.183-197.

Gurevich, A. *et al.,* 2013. QUAST: Quality assessment tool for genome assemblies. *Bioinformatics*, 29(8), pp.1072-1075.

Haas, B.J. *et al*., 2011. Chimeric 16S rRNA sequence formation and detection in Sanger and 454-pyrosequenced PCR amplicons. *Genome Research*, 21(3), pp.494-504.

Hay, M.E., 2009. Marine Chemical Ecology: Chemical signals and cues structure marine populations, communities, and ecosystems. *Annual Review of Marine Science*, 1(1), pp.193-212.

Hernández-González, I.L. & Olmedo-Álvarez, G., 2016. Draft whole-genome sequence of the type strain *Bacillus aquimaris* TF12$^{T}$. *Genome announcements*, 4(4), p.640.

Hill, D.P. *et al.,* 2008. Gene Ontology annotations: what they mean and where they come from. *BMC Bioinformatics*, 9(5), p.2.

Hough, D.W. & Danson, M.J., 1999. Extremozymes. *Current Opinion in Chemical Biology*, 3(1), pp.39-46.

Hunter, S. *et al*., 2009. InterPro: the integrative protein signature database. *Nucleic Acids Research*, 37, pp.211–215.

Hyatt, D. *et al.,* 2010. Prodigal: prokaryotic gene recognition and translation initiation site identification. *BMC Bioinformatics*, 11(1), p.119.

Ijaq, J. *et al*., 2015. Annotation and curation of uncharacterized protein – challenges. *Frontiers in Genetics,* 6, p.119.

Ínceošlu, Ö. *et al*., 2010. Effect of DNA extraction method on the apparent microbial diversity of soil. *Applied and Environmental Microbiology*, 76(10), pp.3378-3382.

Jeanthon, C., 2000. Molecular ecology of hydrothermal vent microbial communities. *Antonie van Leeuwenhoek*, 77(2), pp.117-133.

Jebbar, M. *et al.,* 2015. Microbial diversity and adaptation to high hydrostatic pressure in deep-sea hydrothermal vents prokaryotes. *Extremophiles*, 19(4), pp.721-740.

Johnson, D.B., 1998. Biodiversity and ecology of acidophilic microorganisms. *FEMS Microbiology Ecology*, 27(4), pp.307-317.

Joint, I., Mühling, M. & Querellou, J., 2010. Culturing marine bacteria - an essential prerequisite for biodiscovery. *Microbial Biotechnology*, 3(5), pp.564-575.

Jørgensen, B.B. & Boetius, A., 2007. Feast and famine - microbial life in the deep-sea bed. *Nature Reviews Microbiology*, 5(10), pp.770-781.

Joshi, M.M. *et al.,* 2012. Pectinase from marine *Bacillus subtilis*: An efficient bioscouring agent. *Journal of Biotechnology and Biomaterials*, 2(6), p.228.

Juul, S. *et al.,* 2015. What's in my pot? Real-time species identification on the MinION. *bioRxiv*, p.030742.

Kádár, E. *et al.,* 2005. Enrichment in trace metals (Al, Mn, Co, Cu, Mo, Cd, Fe, Zn, Pb and Hg) of macro-invertebrate habitats at hydrothermal vents along the Mid-Atlantic Ridge. *Hydrobiologia*, 548(1), pp.191-205.

Kanehisa, M. & Goro, S., 2000. KEGG:Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Research*, 20(1), pp.27-30.

Karl, D.M., 2007. Microbial oceanography: paradigms, processes and promise. *Nature Reviews Microbiology*, 5(10), pp.759-769.

Karpushova, A. *et al.,* 2005. Cloning, recombinant expression and biochemical characterization of novel esterases from *Bacillus* sp. associated with the marine sponge *Aplysina aerophoba*. *Applied Microbiology and Biotechnology*, 67(1), pp.59-69.

Kashefi, K. & Lovley, D.R., 2003. Extending the upper temperature limit for life. *Sicence*, 301(5635), p.934.

Khandeparker, R., Verma, P. & Deobagkar, D., 2011. A novel halotolerant xylanase from marine isolate *Bacillus subtilis* cho40: Gene cloning and sequencing. *New Biotechnology*, 28(6), pp.814-821.

Kim, S.-K. & Venkatesa, J., 2015. Introduction to Marine Biotechnology. In Kim, S.-K., ed., Springer Handbook of Marine Biotechnology, *Berlin & Heidelberg, Germany: Springer*, pp.1–10.

Klippel, B. *et al.,* 2014. Carbohydrate-active enzymes identified by metagenomic analysis of deep-sea sediment bacteria. *Extremophiles*, 18, pp.853-863.

Kopf, A. *et al.*, 2015. The ocean sampling day consortium. *GigaScience*, 4(1), pp.27-31.

Koren, S. *et al.*, 2016. Canu: scalable and accurate long-read assembly via adaptive *k*-mer weighting and repeat separation. *bioRxiv*, p.71282.

Kumar, S., Stecher, G. & Tamura, K., 2016. MEGA7: Molecular Evolutionary Genetics Analysis version 7.0 for bigger datasets. *Molecular Biology and Evolution*, 33(7), pp.1870-1874.

Lailaja, V.P. & Chandrasekaran, M., 2013. Detergent compatible alkaline lipase produced by marine *Bacillus smithii* BTMS 11. *World Journal of Microbiology and Biotechnology*, 29(8), pp.1349-1360.

Lavezzo, E. *et al.*, 2016. Third-generation sequencing technologies applied to diagnostic microbiology: benefits and challenges in applications and data analysis. *Expert Review of Molecular Diagnostics*, 16(9), pp.1011-1023.

Lee, Z.P. *et al.*, 2007. Euphotic zone depth: Its derivation and implication to ocean-color remote

sensing. *Journal of Geophysical Research: Oceans*, 112(3), pp.1-11.

Li, H. & Durbin, R., 2009. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics*, 25(14), pp.1754-1760.

Li, H. *et al.*, 2009. The Sequence Alignment/Map format and SAMtools. *Bioinformatics*, 25(16), pp.2078-2079.

Liu, J. *et al.*, 2015. Comparison of ITS and 18S rDNA for estimating fungal diversity using PCR-DGGE. *World Journal of Microbiology and Biotechnology*, 31(9), pp.1387-1395.

Liu, Y. *et al*., 2015. Phylogenetic diversity of the *Bacillus pumilus* group and the marine ecotype revealed by Multilocus Sequence Analysis. *PLOS ONE,* 8(11), p.80097.

Loman, N.J., Quick, J. & Simpson, J.T., 2015. A complete bacterial genome assembled *de novo* using only nanopore-sequencing data. *Nature Methods*, 12(8), pp.733-735.

Loman, N.J. & Quinlan, A.R., 2014. Poretools: A toolkit for analyzing nanopore-sequence data. *Bioinformatics*, 30(23), pp.3399-3401.

Loose, M., Malla, S. & Stout, M., 2016. Real time selective sequencing using nanopore technology. *Nature Methods*, 13(9), pp.751-754.

Lundberg, K.S. *et al.*, 1991. High-fidelity amplification using a thermostable DNA polymerase isolated from *Pyrococcus furiosus*. *Gene*, 108(1), pp.1-6.

MacLean, D., Jones, J.D.G. & Studholme, D.J., 2009. Application of 'next-generation' sequencing technologies to microbial genetics. *Nature Reviews Microbiology*, 7(4), pp.287-296

Madoui, M.-A. *et al.*, 2015. Genome assembly using Nanopore-guided long and error-free DNA reads. *BMC Genomics*, 16(1), p.327.

Martin, W. *et al.*, 2008. Hydrothermal vents and the origin of life. *Naure Reviews Microbiology*, 6(11), pp.805-814.

Martínez-Núñez, M.A. & López, V.E.L, 2016. Nonribosomal peptides synthetases and their applications in industry. *Sustainable Chemical Process*, 4, p.13.

Massol-Deya, A.A. *et al.*, 1995. Bacterial community fingerprinting of amplified 16S and 16-23S ribosomal DNA gene sequences and restriction endonuclease analysis (ARDRA). In Akkermans, A.D.L, Van Elsas, J.D. & De Brujin, F.J., ed., Molecular Microbial Ecology Manual, *Amsterdam, The Netherlands: Springer,* pp.289–296.

McDonald, A.G. & Tipton, K.F., 2013. Fifty-five years of enzyme classification: Advances and difficulties. *FEBS Journal*, 281(2), pp.583-592.

McIntyre, A.B.R. *et al.*, 2017. Nanopore detection of bacterial DNA base modifications. *bioRxiv*, p.127100.

Médigue, C. & Moszer, I., 2007. Annotation, comparison and databases for hundreds of bacterial genomes. *Research in Microbiology*, 158(10), pp.724-736.

Meyer, W. *et al.*, 1993. Hybridization probes for conventional DNA-fingerprinting used as single primers in the Polymerase Chain-Reaction to distinguish strains of *Cryptococcus neoformans*. *Journal of Clinical Microbiology*, 31(9), pp.2274-2280.

Mills, E. L., 1983. Problems of deep-sea biology: an historical perspective. In Rowe, G.T., ed., Deep-sea Biology, *New York, USA: Wiley*, pp.1-79.

Moeseneder, M.M. *et al.*, 1999. Optimization of terminal-restriction fragment length polymorphism analysis for complex marine bacterioplankton communities and comparison with denaturing

gradient gel electrophoresis. *Applied And Environmental Microbiology*, 65(8), pp.3518-3525.

Morey, M. *et al.*, 2013. A glimpse into past, present, and future DNA sequencing. *Molecular Genetics and Metabolism*, 110(1-2), pp.3-24.

Morgan, R., Xiao, J.-P. & Xu, S.-Y., 1998. Characterization of an extremely thermostable restriction enzyme, PspGI, from a *Pyrococcus* strain and cloning of the PspGI restriction-modification system in Escherichia coli. *Applied and Environmental Microbiology*, 64(10), pp.3669-3673.

Muyzer, G., De Waal, E.C. & Uitterlinden, A.G., 1993. Profiling of complex microbial populations by denaturing gradient gel electrophoresis analysis of polymerase chain reaction-amplified genes coding for 16S rRNA. *Applied and Environmental Microbiology*, 59(3), pp.695-700.

Muyzer, G. *et al.*, 1998. Denaturing gradient gel electrophoresis (DGGE) in microbial ecology. In Akkermans, A.D.L, Van Elsas, J.D. & De Brujin, F.J., ed., Molecular Microbial Ecology Manual, *Amsterdam, The Netherlands: Springer,* pp. 2645-2671.

Nakagawa, S. & Takai, K., 2008. Deep-sea vent chemoautotrophs: Diversity, biochemistry and ecological significance. *FEMS Microbiology Ecology*, 65(1), pp.1-14.

National Research Council US, 2002. Marine Biotechnology in the Twenty-First Century: Problems, Promise, and Products, *Washington, DC: The National Academies Press,* pp. 3-28.

Nichols, D. *et al.*, 2010. Use of iChip for high-throughput in situ cultivation of uncultivable microbial species. *Applied and Environmental Microbiology*, 76(8), pp.2445-2450.

Norris, A.L. *et al.*, 2016. Nanopore sequencing detects structural variants in cancer. *Cancer Biology and Therapy*, 17(3), pp.246-253.

Nyyssönen, M. *et al.*, 2013. Coupled high-throughput functional screening and next-generation sequencing for identification of plant polymer decomposing enzymes in metagenomic libraries. *Frontiers in Microbiology*, 4, p.282.

Ochman, H., Gerber, A.S. & Hartl, D.L., 1988. Genetic applications of an inverse polymerase chain reaction. *Genetics*, 120(3), pp.621-623.

Olive, D.M. & Bean, P., 1999. Principles and applications of methods for DNA-based typing of microbial organisms. *Journal of Clinical Microbiogly*, 37(6), pp.1661-1669.

OSPAR Commission, 2010. Background Document for Oceanic ridges with hydrothermal vents/fields. Available at *http://www.ospar.org/documents?v=7220*. Accessed on August 16th, 2016.

Oulas, A. *et al.*, 2015. Metagenomics: Tools and insights for analyzing next-generation sequencing data derived from biodiversity studies. *Bioinformatics and Biology Insights*, 9, pp.75-88.

Overbeek, R. *et al.*, 2014. The SEED and the Rapid Annotation of microbial genomes using Subsystems Technology (RAST). *Nucleic Acids Research*, 42(1), pp.206-214.

Parkes, R.J., Cragg, B.A. & Wellsbury, P., 2000. Recent studies on bacterial populations and processes in subseafloor sediments: A review. *Hydrogeology Journal*, 8, pp.11-28.

Parkes, R.J. *et al.*, 2014. A review of prokaryotic populations and processes in subseafloor sediments, including biosphere: Geosphere interactions. *Marine Geology*, 352, pp.409-425.

Piel, J. *et al.*, 2004. Antitumor polyketide biosynthesis by an uncultivated bacterial symbiont of the marine sponge *Theonella swinhoei*. *Proceedings of the National Academy of Sciences of the United States of America*, 101(46), pp.16222-16227.

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.

Podar, M. & Reysenbach, A.-L., 2006. New opportunities revealed by biotechnological explorations of extremophiles. *Current Opinion in Biotechnology*, 17(3), pp.250-255.

Poli, A. *et al.*, 2012. *Geobacillus subterraneus* subsp. *aromaticivorans* subsp. nov., a novel thermophilic and alkaliphilic bacterium isolated from a hot spring in Sırnak, Turkey. *The Journal of General and Applied Microbiology*, 58(6), pp.437-446.

Porwal, S. *et al*., 2009. Phylogeny in aid of the present and novel microbial lineages: Diversity in *Bacillus*. *PLOS ONE*, 4(2), p.4438.

Preiss, L. *et al.*, 2015. Alkaliphilic bacteria with impact on industrial applications, concepts of early Life forms, and bioenergetics of ATP synthesis. *Frontiers in Bioengineering and Biotechnology*, 3, p.75.

Quick, J. *et al.*, 2016. Real-time, portable genome sequencing for Ebola surveillance. *Nature*, 530(7589), pp.228-232.

Quick, J., Quinlan, A.R. & Loman, N.J., 2014. A reference bacterial genome dataset generated on the MinION™ portable single-molecule nanopore sequencer. *GigaScience*, 3(1), p.22.

Ramirez-Llodra, E., Shank, T. & German, C., 2007. Biodiversity and Biogeography of Hydrothermal Vent Species: Thirty years of discovery and investigations. *Oceanography*, 20(1), pp.30-41.

Rani, A., Souche, Y.S. & Goel, R., 2009. Comparative assessment of *in situ* bioremediation potential of cadmium resistant acidophilic *Pseudomonas putida* 62BN and alkalophilic *Pseudomonas monteilli* 97AN strains on soybean. *International Biodeterioration and Biodegradation*, 63(1), pp.62-66.

Ravaux, J. *et al.*, 2003. Heat-shock response and temperature resistance in the deep-sea vent shrimp *Rimicaris exoculata*. *The Journal of Experimental Biology*, 206(14), pp.2345-2354.

Reuter, J.A., Spacek, D.V. & Snyder, M.P., 2015. High-Throughput sequencing technologies. *Molecular Cell*, 58(4), pp.586-597.

Rice, A.L. *et al.,* 1986. Seasonal deposition of phytodetritus to the deep-sea floor. *Proceedings of the Royal Society of Edinburgh*, 88, pp.265-279.

Rodrigues, C. *et al*., 2011. Exploring the industrial value of the portuguese deep-sea microorganisms. *Spatial and Organization Dynamics Discussion Papers,* 8, pp.107-116.

Romano, I. *et al.,* 2006. *Halomonas alkaliphila* sp. nov., a novel halotolerant alkaliphilic bacterium isolated from a salt pool in Campania (Italy). *The Journal of General and Applied Microbiology*, 52(6), pp.339-348.

Rosenbaum, V. & Riesner, D., 1987. Temperature-gradient gel electrophoresis. Thermodynamic analysis of nucleic acids and proteins in purified form and in cellular extracts. *Biophysical Chemistry*, 26(3), pp.235-246.

Rusch, D.B. *et al.*, 2007. The Sorcerer II Global Ocean Sampling expedition: Northwest Atlantic Through eastern tropical Pacific. *PLOS Biology*, 5(3), p.77.

Sakharkar, K.R. & Chow, V.T.K., 2008. Microbial genomics: Rhetoric or reality? *Indian Journal of Microbiology*, 48(2), pp.156-162.

Sarethy, I.P. *et al.,* 2011. Alkaliphilic bacteria: applications in industrial biotechnology. *Journal of Industrial Microbiology & Biotechnology*, 38(7), pp.769-790.

Sass, A.M. *et al.,* 2008. Diversity of *Bacillus*-like organisms isolated from deep-sea hypersaline anoxic sediments. *Saline Systems*, 4, p.8.

Sharon, D. *et al.,* 2013. A single-molecule long-read survey of the human transcriptome. *Nature*

*Biotechnology*, 31(11), pp.1009-1014.

Sjostrom, S.L. *et al.,* 2014. High-throughput screening for industrial enzyme production hosts by droplet microfluidics. *Lab On a Chip*, 14(4), pp.806-813.

Soni, R., Soni, S.K. & Goyal, N., 2007. Biotechnology for Developing Novel Microbes. In Soni, S.K., ed., Microbes: A Source of Energy for the 21st Century, *New Delhi, India*: *New India Publishing Agency,* pp.496-536.

Srivastava, S., Ghosh, N. & Pal, G*.,* 2013. Metagenomics: Mining Environmental Genomes. In Kuhad, R.C. & Singh, A., ed., Biotechnology for Environmental Management and Resource Recovery, *New Delhi, India: Springer*, pp.161-190.

Staley, J.T. & Konopka, A., 1985. Measurement of *in situ* activities of nonphotosynthetic microorganisms in aquatic and terrestrial habitats. *Annual Review of Microbiology*, 39(1), pp.321-346.

Stothard, P. & Wishart, D.S., 2006. Automated bacterial genome analysis and annotation. *Current Opinion in Microbiology*, 9(5), pp.505-510.

Sucharita, K. *et al.*, 2009. *Shewanella chilikensis* sp. nov., a moderately alkaliphilic gammaproteobacterium isolated from a lagoon. *International Journal of Systematic and Evolutionary Microbiology*, 59(12), pp.3111-3115.

Sunagawa, S. *et al.*, 2015. Structure and function of the global ocean microbiome. *Science*, 348(6237), pp.1261359-1261359.

Synnes, M., 2007. Bioprospecting of organisms from the deep-sea: Scientific and environmental aspects. *Clean Technologies and Environmental Policy*, 9(1), pp.53-59.

Tenreiro, T., 2005. Diversidade procariota em fontes hidrotermais. *Bachelor Dissertation at the Faculty of Sciences*, University of Lisbon.

Teske, A. & Sørensen, K.B., 2008. Uncultured archaea in deep marine subsurface sediments: have we caught them all? *The International Society for Microbial Ecology Journal*, 2(1), pp.3-18.

Thakur, N.L. *et al.*, 2008. Marine molecular biology: An emerging field of biological sciences. *Biotechnology Advances*, 26(3), pp.233-245.

Trivedi, N. *et al.*, 2011. Solvent tolerant marine bacterium *Bacillus aquimaris* secreting organic solvent stable alkaline cellulase. *Chemosphere*, 83(5), pp.706-712.

Tzeneva, V.A. *et al.*, 2009. Effect of soil sample preservation, compared to the effect of other environmental variables, on bacterial and eukaryotic diversity. *Research in Microbiology*, 160(2), pp.89-98.

Urban, J.M. *et al.*, 2015. Sequencing ultra-long DNA molecules with the Oxford Nanopore MinION. *bioRxiv,* p.019281.

Urbieta, M.S. *et al.*, 2015. Thermophiles in the genomic era: Biodiversity, science, and applications. *Biotechnology Advances*, 33(6), pp.633-647.

Vandamme, P. *et al*., 1997. *Streptococcus difficile* is a nonhenolytic Group B, Type Ib *Streptococcus*. *International Journal of Systematic Bacteriology,* 47(1), pp.81-85.

Van den Burg, B., 2003. Extremophiles as a source for novel enzymes. *Current Opinion in Microbiology*, 6(3), pp.213-218.

Van Dover, C.L., 2011. Tighten regulations on deep-sea mining. *Nature*, 470(7332), pp.31-33.

Van Dover, C.L. & Lutz, R.A., 2004. Experimental ecology at deep-sea hydrothermal vents: a

perspective. *Journal of Experimental Marine Biology and Ecology*, 300(1-2), pp.273-307.

Venter, J.C. *et al.*, 2004. Environmental genome shotgun sequencing of the Sargasso Sea. *Science*, 304(66), pp.66-74.

Vermelho, A.B. *et al.,* 2013, Diversity and Biotechnological Applications of Prokaryotic Enzymes. In Rosenberg, E. *et al.*, ed., The prokaryotes: Applied Bacteriology and Biotechnology Fourth Edition, *Berlin & Heidelberg, Germany: Springer*, pp.213-240.

Vester, J.K., Glaring, M.A. & Stougaard, P., 2014. Discovery of novel enzymes with industrial potential from a cold and alkaline environment by a combination of functional metagenomics and culturing. *Microbial Cell Factories*, 13(1), p.72.

Vieille, C. & Zeikus, G.J., 2001. Hyperthermophilic Enzymes : sources, uses, and molecular mechanisms for thermostability. *Microbiology and Molecular Biology Reviews*, 65(1), pp.1-43.

Vinga, S. & Almeida, J., 2003. Alignment-free sequence comparison - A review. *Bioinformatics*, 19(4), pp.513-523.

Waite, J.H. & Tanzer, M.L., 1981. Polyphenolic Substance of *Mytilus edulis*: novel adhesive containing L-Dopa and Hydroxyproline. *Science*, 212(4498), pp.1038-1040.

Wang, A. & Ash, G.J., 2015. Whole genome phylogeny of *Bacillus* by Feature Frequency Profiles (FFP). *Scientific Reports*, 5, p.13644.

Wang, J., *et al.,* 2015. MinION nanopore sequencing of an influenza genome. *Frontiers in Microbiology*, 6, p.766.

Webb, T.J., 2009. Biodiversity research sets sail: showcasing the diversity of marine life. *Biology Letters*, 5(2), pp.145-147.

Wetterstrand K.A., 2016. DNA Sequencing Costs: Data from the NHGRI Genome Sequencing Program (GSP). Available at *www.genome.gov/sequencingcostsdata*. Accessed September 4th, 2016.

Wilson E.O., 1997 Introduction. In Reaka-Kudla, M.L., Wilson, D.E. & Wilson, E.O., ed., Biodiversity II Understanding and Protecting our Biological Resources, *Washington: Joseph Henry Press*, pp.1-3.

Yamauchi, Y. *et al.,* 2013. *Halarchaeum salinum* sp. nov., a moderately acidophilic haloarchaeon isolated from commercial sea salt. *International Journal of Systematic and Evolutionary Microbiology*, 63(3), pp.1138-1142.

Yu, N.Y. *et al.,* 2010. PSORTb 3.0: Improved protein subcellular localization prediction with refined localization subcategories and predictive capabilities for all prokaryotes. *Bioinformatics,* 26(13), pp.1608–1615.

Zhang, C. & Kim, S.-K., 2010. Research and application of marine microbial enzymes: status and prospects. *Marine drugs*, 8(6), pp.1920-1934.

Zhang, J. *et al.*, 2015. Microbial diversity in the deep-sea sediments of Iheya North and Iheya Ridge, Okinawa Trough. *Microbiological Research*, 177, pp.43-52.

Zhou, M.-Y. *et al.,* 2013. Diversity of both the cultivable protease-producing bacteria and bacterial extracellular proteases in the coastal sediments of King George Island, Antarctica. *PLOS ONE*, 8(11), p.79668.

Zinger, L., Gobet, A. & Pommier, T., 2012. Two decades of describing the unseen majority of aquatic microbial diversity. *Molecular Ecology*, 21(8), pp.1878-1896.

# Appendix A. Classification of industrial relevant biomass-degrading enzymes

Bellow we present summary lists regarding the major economically relevant groups of biomass-degrading enzymes, their most significant applications and their classification[49].

Table A.1 | **List of major economically relevant groups of biomass-degrading enzymes and their most significant application areas.**

| Enzyme group | Substrate description | Applications |
|---|---|---|
| Starch-degrading enzymes | Starch is a common reserve molecule in plants. It is a polysaccharide composed by (I) amylose - linear insoluble polymer of glucose units linked by 1,4-α-glycosidic bonds -, and (II) amylopectin - branched soluble component that links glucose linearly via 1,4-α-glycosidic bonds but branches out via 1,6-α-glycosidic bonds. | Starch-degrading enzymes represent 25% of the global enzyme market and are used in the liquefaction and saccharification of starch granules in industrial processes of the food industry (*e.g.* brewing, baking, fruit juice processing). Saccharification is typically performed at high temperatures and heat-stable enzymes would offer advantages to the process. |
| Cellulose-degrading enzymes | Cellulose is the major component of plant cell walls and the most abundant polysaccharide on Earth. It is composed by long chains of glucose units linearly linked by 1,4-β-glycosidic bonds. | Cellulose-degrading enzymes are used in the paper recycling processes by acting on paper fibers, leading to the gentle dislodgment of ink and washing of the paper. For this process enzymes should be stable at high-temperatures and alkaline environments. They are also used in the textile and food industries, as well as in the production of 2nd generation bioethanol. |
| Xylan-degrading enzymes | Hemicelluloses are non-cellulosic components of plant cell walls. They are highly branched polysaccharides composed of 1,4-β-linked backbones of xylose (in xylan) or mannose (in mannan), and several side-groups of different pentoses or hexoses. | These enzymes have industrial applications in food and feed processing, production of 2nd generation bioethanol and biobleaching. Bleaching of alkaline wood pulp, *i.e.* the whitening of the pulp by exclusion of lignin, is typically done using chlorine. However, thermostable and alkali-stable xylanases and mannanases can be useful in the development of ecologically friendly alternative processes. |
| Mannan-degrading enzymes | | |
| Pectin-degrading enzymes | Pectin is another component of plant cell walls. Typically it has a backbone of 1,4-α-linked *D*-galacturonic acid residues, or of repetitions of the disaccharide 1,4-α-*D*-galacturonic acid-1,2-α-*L*-rhamnose, with side chains of *D*-galactose, *L*-arabinose or other sugars. It can also be acetylated or *O*-methyl-esterified in different proportions. | Pectin-degrading enzyme, particularly thermostable pectinases, are enzymes of high interest in the beverage industry, since they can be used to extract fruit juice, having applications in, for instance, the production of wine and other beverages. |
| Chitin-degrading enzymes | Chitin is a linear molecule of 1,4-β-linked *N*-acetylglucosamine residues that appears as the major structural component of fungi cell walls and exoskeletons of insects and crustaceans. | The use of these enzymes is an impending area of development. They offer great potential for the exploitation of marine chitinous waste from the seafood industry or as possible agents for the biocontrol of fungi and insect pests. |
| Peptidases | Proteases have a large spectrum of substrate specificities. They catalyze the degradation of proteinaceous material by hydrolyzing more or less specific peptide bonds on proteins or peptides. | Proteases global sales represent 60% of the total enzyme market. They are extremely useful for very specific pharmaceutical or chemical applications, but also for more broad applications in the food, feed, cosmetic and detergent industries. To be used in detergent formulations they should withstand a broad range of pH, and be stable in the presence of surfactants and other additives. |
| Lipases/Esterases | Lipases are carboxylic ester hydrolases that catalyze the cleavage of ester bonds (or the reverse esterification reaction) in lipidic substrates. Esterases prefer short-chain acyl esters with less than 10 carbon atoms whilst lipases act on long chain fatty acid esters. | These enzymes are used in pharmaceutical and fine chemical processes, largely because of their ability to produce optically pure compounds. They are also used in the food, dairy, cosmetic, agrochemical, biosurfactant and paper industries. Lipases and esterases that are alkali-stable also have a long tradition as supplements in the laundry industry. |

[49] For polysaccharide-degrading enzymes, there is a different classification system besides EC codes, which groups carbohydrate-active enzymes into families based on sequence similarity. This system is compiled into the CAZy - carbohydrate-active enzymes database (http://www.cazy.org).

Table A.2 | **Classification of polysaccharide-degrading enzymes.** Information was retrieved from Dalmaso, Ferreira & Vermelho 2015; Elleuche *et al.* 2015 and curated with ExplorEnz database (http://www.enzyme-database.org).

| Enzyme accepted name | Cleavage site | Released product | Classification | Comment |
|---|---|---|---|---|
| **Starch-degrading** | | | | |
| α-amylase | Internal 1,4-α linkages | Dextrins[1] | EC 3.2.1.1 | Endoamylase |
| β-amylase | Second 1,4-α linkage of non-reducing ends[2] | Maltose | EC 3.2.1.2 | Exoamylase |
| Glucan 1,4-α-glucosidase | External 1,4-α linkage of non-reducing ends | Glucose | EC 3.2.1.3 | Exoamylase |
| α-glucosidase | External 1,4-α linkage of non-reducing ends | Glucose | EC 3.2.1.20 | Exoamylase |
| Pullulanase | 1,6-α linkages of pullulan[3] | Maltotriose | EC 3.2.1.41 | Debranching |
| Isoamylase | 1,6-α linkages | Malto-oligosaccharides | EC 3.2.1.68 | Debranching |
| Limit dextrinase | 1,6-α linkages | Maltose | EC 3.2.1.142 | Debranching |
| Oligo-1,6-glucosidase | 1,6-α linkages | Glucose | EC 3.2.1.10 | Debranching |

[1]Mixture of low-molecular-weight polysaccharides of *D*-glucose. [2]The end without a reducing aldehyde group.
[3]Polymer of units of maltotriose (three maltose residues linked by 1,4-α-glycosidic bonds) linked by 1,6-α glycosidic bonds.

| | | | | |
|---|---|---|---|---|
| **Cellulose-degrading** | | | | |
| Cellulase | 1,4-β linkages | Cellulose oligosaccharides | EC 3.2.1.4 | Cellulase Endoglucanase |
| β-glucosidase[1] | External 1,4-β linkage of non-reducing ends | Glucose | EC 3.2.1.21 | Cellulase Exoglucanase |
| Cellulose 1,4-β-cellobiosidase | Second 1,4-β linkage of non-reducing ends | Cellobiose | EC 3.2.1.91 | Cellulase Exoglucanase |

[1]β-glucosidase can also act on glucosyl side-groups of mannans.

| | | | | |
|---|---|---|---|---|
| **Xylan-degrading** | | | | |
| Endo-1,4-β-xylanase | Internal 1,4-β linkages | Xylo-oligosaccharides | EC 3.2.1.8 | Endoxylanase |
| α-*D*-xyloside xylohydrolase | External 1,4-β linkage of non-reducing ends | Xylose | EC 3.2.1.177 | Exoxylanase |
| Xylan 1,4-β-xylosidase | External 1,4-β linkage of non-reducing ends | Xylose | EC 3.2.1.37 | Exoxylanase |
| α-*L*-arabinofuranosidase | Terminal non-reducing α-*L*-arabinofuranoside linkage to xylose residues | *L*-arabinofuranose and debranched xylan | EC 3.2.1.55 | Removal of xylan side-groups |
| α-glucuronidases | Glucuronosyl linkage to xylose residues | Glucuronate and debranched xylan | EC 3.2.1.139 | Removal of xylan side-groups |
| Acetylxylan esterase | Acetyl linkage to xylose residues | Acetate and deacetyladed xylan | EC 3.1.1.72 | Removal of xylan side-groups |
| Feruloyl esterase | Ferulic acid linkage to xylose residues | Ferulate and debranched xylan | EC 3.1.1.73 | Removal of xylan side-groups |
| **Mannan-degrading** | | | | |
| Mannan endo-1,4-β-mannosidase | Internal 1,4-β linkages | Mannan oligosaccharides | EC 3.2.1.78 | Endomannanase |
| β-mannosidase | External 1,4-β linkage of non-reducing ends | Mannose | EC 3.2.1.25 | Exomannanase |
| α-galactosidase | Terminal non-reducing galactoside linkage to mannose residues | Galactose and debranched mannan | EC 3.2.1.22 | Removal of mannan side-groups |
| **Pectin-degrading** | | | | |
| Polygalacturonase | 1,4-α linkages to galacturonic acid residues | Pectin oligosaccharides and galacturonic acid | EC 3.2.1.15 | Pectinase |
| Rhamnogalacturonan hydrolase | 1,2-α linkage between rhamnogalacturonans disaccharide units | Rhamnogalacturonan oligosaccharides | EC 3.2.1.171 | Endopectinase |
| Arabinan endo-1,5-α-*L*-arabinase | 1,5-α linkages between arabinose residues | Arabinose | EC 3.2.1.99 | Degradation of pectin side-chains |
| Arabinogalactan endo-1,4-β-galactanase | 1,4-β-galactosidic linkages in arabinogalactan | Galactose | EC 3.2.1.89 | Degradation of pectin side-chains |

| Enzyme accepted name | Cleavage site | Released product | Classification | Comment |
|---|---|---|---|---|
| Pectin esterase | Methyl linkage to galacturonic acid residues | Methanol and pectate[1] | EC 3.1.1.11 | Removal of pectin side-groups |
| Pectate lyase | 1,4–α linkages to galacturonic acid residues | Oligosaccharides with 4–deoxy–α–galacto–4–enuronosyl groups at the non–reducing end | EC 4.2.2.2 | Removal of pectin side-groups |
| Pectin lyase | 1,4–α linkages to galactoran methyl ester | Oligosaccharides with 4–deoxy–6–*O*–methil–α–galacto–4–enuronosyl groups at the non–reducing end | EC 4.2.2.10 | Removal of pectin side-groups |

[1]A demethylated version of pectin.

**Chitin-degrading**

| | | | | |
|---|---|---|---|---|
| Chitinase | 1,4–β linkages | Chitodextrins and *N*–acetylglucosamine | EC 3.2.1.14 | Endo/Exo chitinase |
| β–*N*–acetylhexosaminidase | External 1,4–β linkage of non–reducing ends | *N*–acetylglucosamine | EC 3.2.1.52 | Exochitinase |
| Chitin deacetylase | *N*–acetamido linkage to *N*–acetylglucosamine residues | Acetate and chitosan[1] | EC 3.5.1.41 | Removal of chitin side-groups |
| Chitosanases | Internal 1,4–β linkages between glucosamine residues of chitosan | Chitosan oligosaccharides | EC 3.2.1.132 | Endochitosanase |

[1]A deacetyladed version of chitin forming a linear polysaccharide composed by 1,4–β linked glucosamine (deacetylated) units and *N*–acetylglucosamine (acetylated) units.

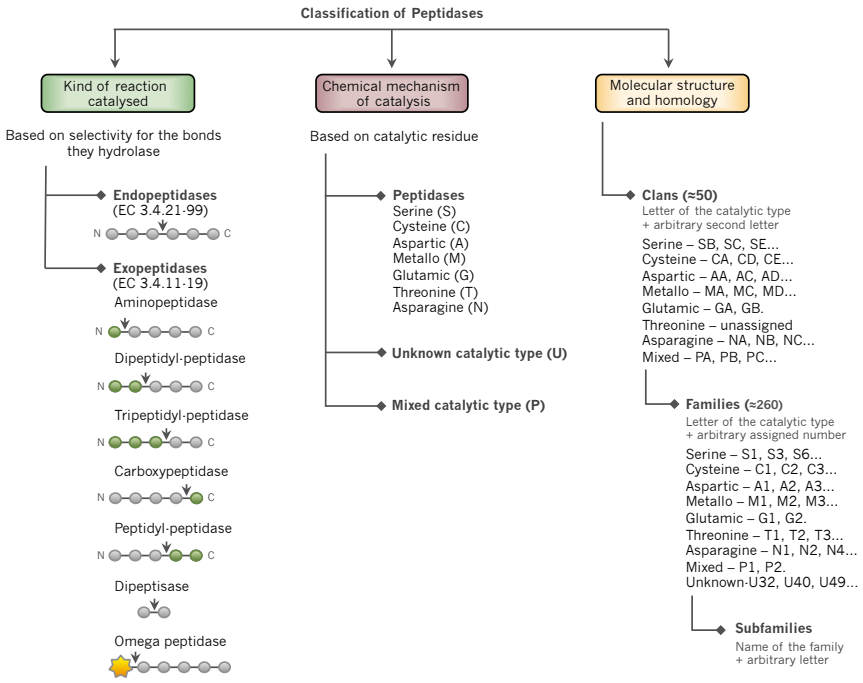


Figure A.1 | **Classification of peptidases.** Peptidases have three different possible classification systems based on (I) the kind of reaction catalyzed (type of bonds hydrolyzed), (II) the chemical mechanism of catalysis (depending on the amino acid or metallic ion at the active site) and (III) the molecular structure and homology (by comparison of amino-acid sequence and three-dimensional structures of different peptidases). The homology-based classification is compiled into the MEROPS database (http://merops.sanger.ac.uk/index.shtml). Adapted from Vermelho *et al.* 2013 and curated using MEROPS database.

Table A.3 | **Classification of lipolytic enzymes.** Adapted from Dalmaso, Ferreira & Vermelho 2015 and curated with ExplorEnz database (http://www.enzyme-database.org).

| Enzyme accepted name | Reaction catalyzed | Classification | Comment |
|---|---|---|---|
| Triacylglycerol lipase | Triacylglycerol + $H_2O$ ⇌ Diacylglycerol + Carboxylate | EC 3.1.1.3 | Lipase |
| Carboxylesterase | Carboxylic ester + $H_2O$ ⇌ Alcohol + Carboxylate | EC 3.1.1.1 | Esterase |

# Appendix B. The SEAHMA project: More on the isolation and polyphasic characterization of the isolates

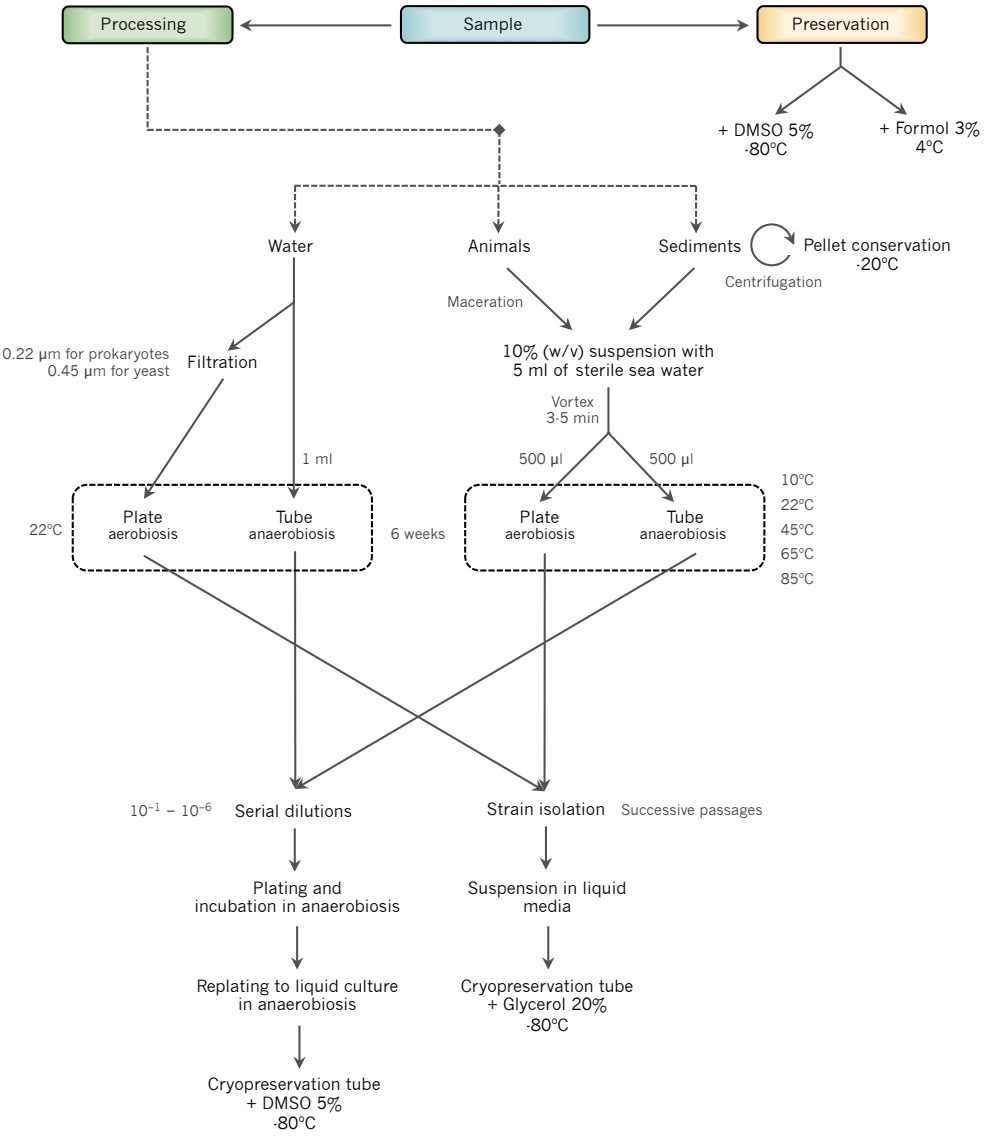


Figure B.1 | **Schematic representation of the isolation workflow during the SEAHMA project.**

Table B.1 | **Composition of general culture media and supplements used for the isolation of prokaryotes during the SEAHMA project.**

| Medium 1 | Medium 2 | Supplement |
|---|---|---|
| BHI[1] 0.9% (w/v) | Peptone 0.1% (w/v) | Sodium nitrate 0.170% (w/v) |
| | Cellobiose 0.5% (w/v) | Iron(III) sulfate 0.200% (w/v) |
| | Yeast extract 0.05% (w/v) | Manganese sulfate 0.085% (w/v) |
| PIPES[2] 0.6% (w/v) | PIPES[2] 0.6% (w/v) | |
| Sea Salts (Sigma) 3.0% (w/v) | Sea Salts (Sigma) 3.0% (w/v) | |
| Sulfur 1.0% (w/v) | Sulfur 1.0% (w/v) | |

[1]BHI – Brain Heart Infusion commercial medium. [2]PIPES – Piperazine-N,N'-bis(2-ethanesulfonic acid).

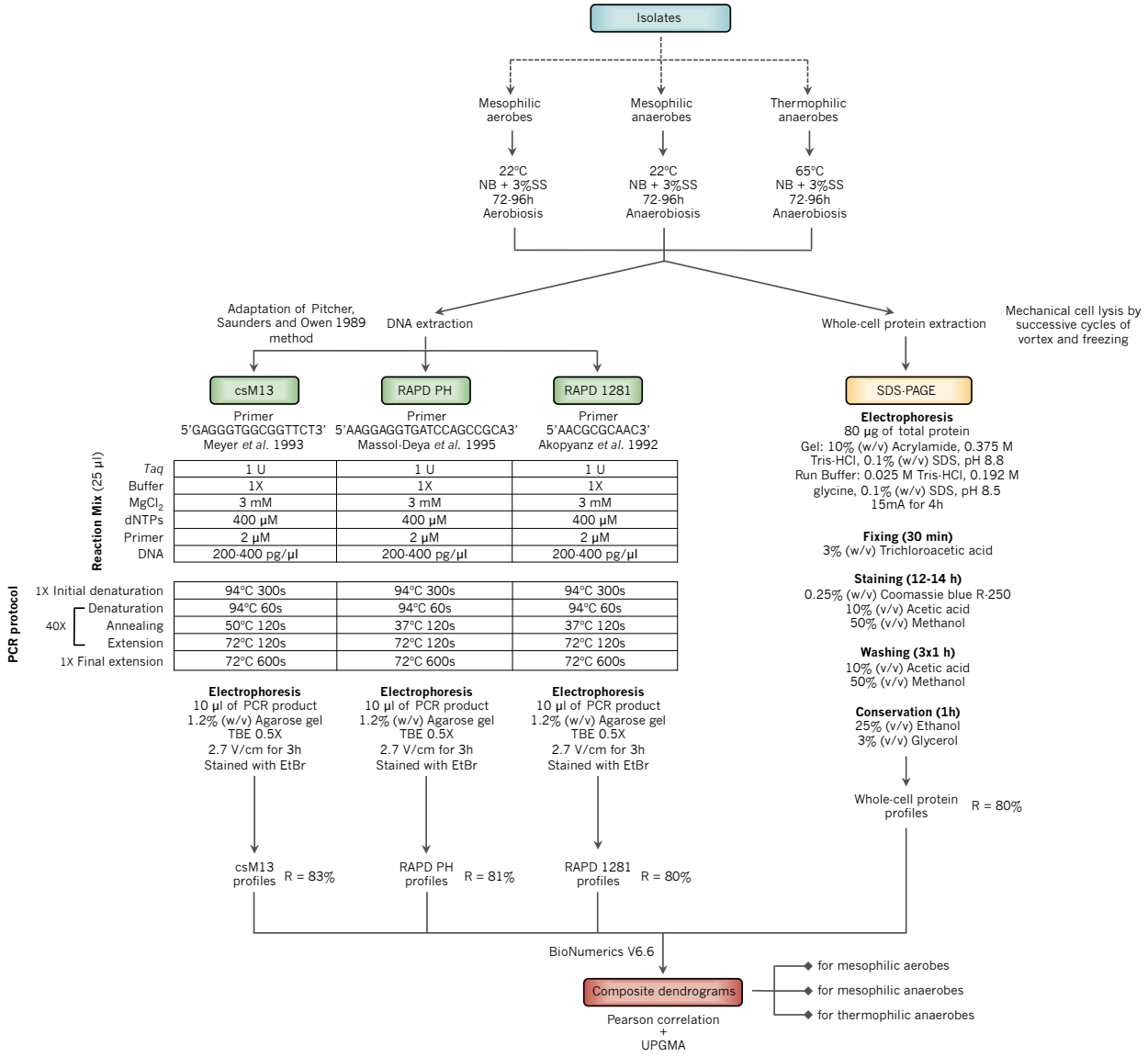


Figure B.2 | **Schematic representation of the polyphasic characterization of the isolates during the SEAHMA project.** R (%) represents the reproducibility of each method based on 10% replicates. Reproducibility was calculated as the average distance between each isolate and its replicate in the dendrogram constructed with the profiles resulting from the specific method. Abbreviations: dNTPs – deoxyribonucleotides; EtBr – ethidium bromide; NB – nutrient broth; RAPD – random amplified polymorphic DNA; SDS – sodium dodecyl sulfate; SDS-PAGE – sodium dodecyl sulfate polyacrylamide gel electrophoresis; SS – sea salts (Sigma); TBE – 40 mM Tris, 45 mM Boric acid, 1 mM EDTA, pH 8.3; UPGMA – unweighted pair group method with arithmetic mean clustering.

# Appendix C. The SEAVENTzymes project: Summary of the phenotypic screening methods

The SEAVENTbugs collection is constituted by 296 prokaryotic isolates, which resulted from the isolation procedures implemented during the SEAHMA project. During the SEAVENTzymes project, 139 of those isolates - the mesophilic aerobic isolates-, were subjected to phenotypic screening methods for the detection of biomass-degrading enzymes with industrial potential. Bellow we present a summary of these methods.

## Growth assays

Microplate growth assays were performed to assess the ability of each isolate to grow in base medium supplemented with a target-enzyme's substrate as the only source of a required nutrient. Growth in such conditions should indicate the production of enzymes that degrade the specific substrate, enabling nutrient mobilization for growth.

Each isolate was subject to 8 different assays with different substrates incorporated into the base medium: starch, carboxymethylcellulose, xylan, mannan, pectin, chitin, casein and finally a mixture of 'tween' 20 and 'tween' 80. These assays screen for starch-, cellulose-, xylan-, mannan-, pectin-, chitin-degrading enzymes, proteases and lipases, respectively. In the assays for the screening of glycoside-acting enzymes, YNB (Yeast Nitrogen Base, DIFCO) supplemented with 3% (w/v) of Sea Salts (Sigma) was used as base medium, and the specific substrate was added in a final concentration of 0.5% (w/v) as the only source of carbon. For the screening of proteases, YCB (Yeast Carbon Base, DIFCO) was used, supplemented with 3% (w/v) of Sea Salts (Sigma) and 0.5% (w/v) casein as the sole source of nitrogen. For the assays screening lipases/esterases, YNB (Yeast Nitrogen Base, DIFCO) supplemented with 3% (w/v) of Sea Salts (Sigma) was used as base medium, and a mixture of 'tween' 20 and 'tween' 80 was added in a final concentration of 0.5% (w/v) as the sole source of carbon. Additionally, each isolate was grown in YNB and YCB supplemented with 3% (w/v) of Sea Salts (Sigma) to determine residual growing in base medium alone.

Growth was monitored using the automatic monitoring system for optical density reading BIOSCREEN C (LabSystems), which allows simultaneous growth assessment of 200 microwells. Each well of the BIOSCREEN microplate had a total volume of 300 µl of medium inoculated with 1 µl-loop of plate-grown cells of the tested isolate. The plate was incubated with constant agitation at 22ºC and optical readings were taken every 30 minutes for 120 hours, at a wavelength of 600 nm. The matrix of total optical readings, that constitute the growth curves, was subjected to numerical transformations, as described in Figure C.1, to give NAUCr (relative Net Areas Under the Curve). To account for residual growth in base medium, NAUCr obtained in the medium plus substrate was

normalized to the NAUCr obtained in base medium alone (Figure C.1). This method was adapted from Guckert *et al.* 1996.

As in large screening experiments, to avoid the experimental effort of creating duplicates or triplicates of each single test, a set of 10% of isolates was randomly chosen to be tested and replicated, in a way it would be possible to infer about the reproducibility of the method.

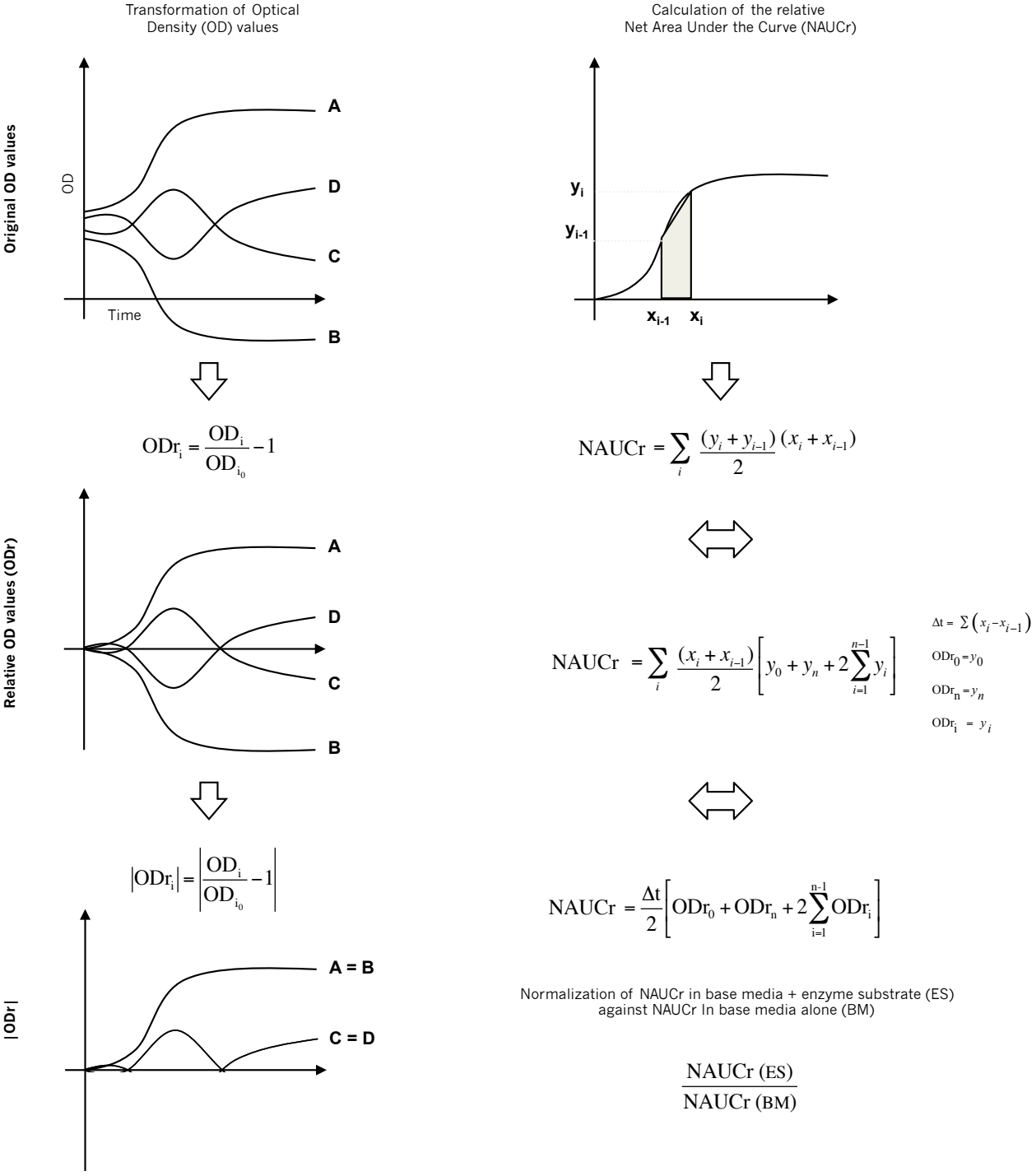


Figure C.1 | **Transformation of the isolates' growth curves into relative and normalized NAUCs.** This figure is an adaptation of the original scheme presented in the reports of the SEAVENTzymes project.

## Colorimetric assays

Colorimetric assays were performed in 24-well microplates where each well was filled with solid base medium incorporated with the target-enzyme's chromogenic substrate as the sole source of a required nutrient. To ensure a higher accessibility of the chromogenic substrates, 750 µl of base medium was added first to each well, and only after solidification, was 750 µl of substrate suspension added, with constant agitation for an homogeneous distribution. After inoculation of each well with 1 µl-loop of cultured cells of the tested isolates, plates were incubated at 22ºC for 7 days.

In the assays for the screening of glycoside-acting enzymes, YNB (Yeast Nitrogen Base, DIFCO) supplemented with 1.5% (w/v) of bacteriologic agar and 3% (w/v) of Sea Salts (Sigma) was used as base medium. The specific substrate was added as a sterile suspension equating to a final concentration of 0.1% (w/v). The chromogenic substrates used were AZCL-modified (AZurin-dyed Cross-Linked) polymers: AZCL-amylose for the detection of α-amylase; AZCL-pullulan for the detection of debranching enzymes such as pullulanases and dextrinases; AZCL-hydroxyethyl cellulose for the detection of endo-cellulase; AZCL-xylan for the detection of endo-1,4-β-D-xylanase; AZCL-glucomannan for the detection of endo-1,4-β-D-mannanase; chitin-azure for the detection of chitinases. For the screening of proteases, YCB (Yeast Carbon Base, DIFCO) was used as base medium, supplemented with 1.5% (w/v) bacteriologic agar and 3% (w/v) of Sea Salts (Sigma). AZCL-casein was added as a sterile suspension in a final concentration of 0.1% (w/v). Hydrolysis of these insoluble AZCL-modified substrates by specific enzymes results in the diffusion of blue-dyed oligosaccharides, enabling direct identification of enzyme production based on the change of color of the medium to blue. Results were considered positive when a blue taint appeared. Additionally, results were further subdivided into slightly positive (1), clearly positive (2) and strongly positive (3) depending on the intensity of the appearing color. For the assays screening lipases/esterases, YNB (Yeast Nitrogen Base, DIFCO) supplemented with 1.5% (w/v) bacteriologic agar and 3% (w/v) of Sea Salts (Sigma) was used as base medium. A mixture of 0.1% (w/v) 'tween' 20 and 'tween' 80 and 0.05% (w/v) of calcium chloride was added as a sterile suspension. The release of fatty acids from the action of lipolytic enzymes on 'tweens', in the presence of calcium chloride, leads to the formation of calcium salts of fatty acids, which appear as a yellow precipitate. Eventually, if complete degradation of these fatty acids occurs, a clear halo surrounding colonies should appear. Thus, results were considered positive for the production of lipases/esterases when a yellow precipitate or a clear halo was evident.

As in large screening experiments, to avoid the experimental effort of creating duplicates or triplicates of each single test, a set of 10% of isolates were randomly chosen to be tested and replicated, in a way it would be possible to infer about the reproducibility of the method.

# Appendix D. Real-time isolate identification using whole-genome nanopore-sequencing data

WIMP –'What's in my Pot' (Juul *et al.* 2015) is an analysis pipeline that takes advantage of the real-time capability of nanopore sequencing and implements sequence-based real-time identification of bacterial, viral and fungal species in what may be a complex metagenomic sample. During a nanopore-sequencing run, as soon as a DNA strand translocates a pore, a file is created with the sequencing data and streamed to the WIMP online pipeline. Each read is basecalled and classified by determining its most likely placement in the NCBI Taxonomy tree, giving it a classification score. For that purpose, the pipeline uses bioinformatics tools that map *k*-mers of length 24 of the sequencing data to nodes in the reference NCBI Taxonomy tree, which is a pre-built reference database enclosing all bacteria, viral and fungal genomes available in RefSeq. A report is created and automatically updates with new classified reads, at regular intervals, or by refreshing the browser throughout the run, providing a straightforward and interactive interpretation of the results.

Here we used whole-genome nanopore-sequencing data from Run 2, described in the methods section, and subjected it to the WIMP online pipeline with the aim of further identifying the sequenced *Bacillus* sp. MG SD 082 isolate (Figure D.1). Most of the reads (254) were identified as belonging to *Bacillus velezensis* strain AS43.3 (formerly classified as *B. methylotrophicus*), followed by a large number of reads (167) identified as *Bacillus velezensis*, with no strain specification. 94 reads did not allow for species discrimination, being positioned only at the level of *Bacillus subtilis* group. The remaining reads were identified as belonging to different strains of the *Bacillus velezensis* specie. Overall, the data seems to indicate that the isolate belongs to the specie *Bacillus velezensis*.

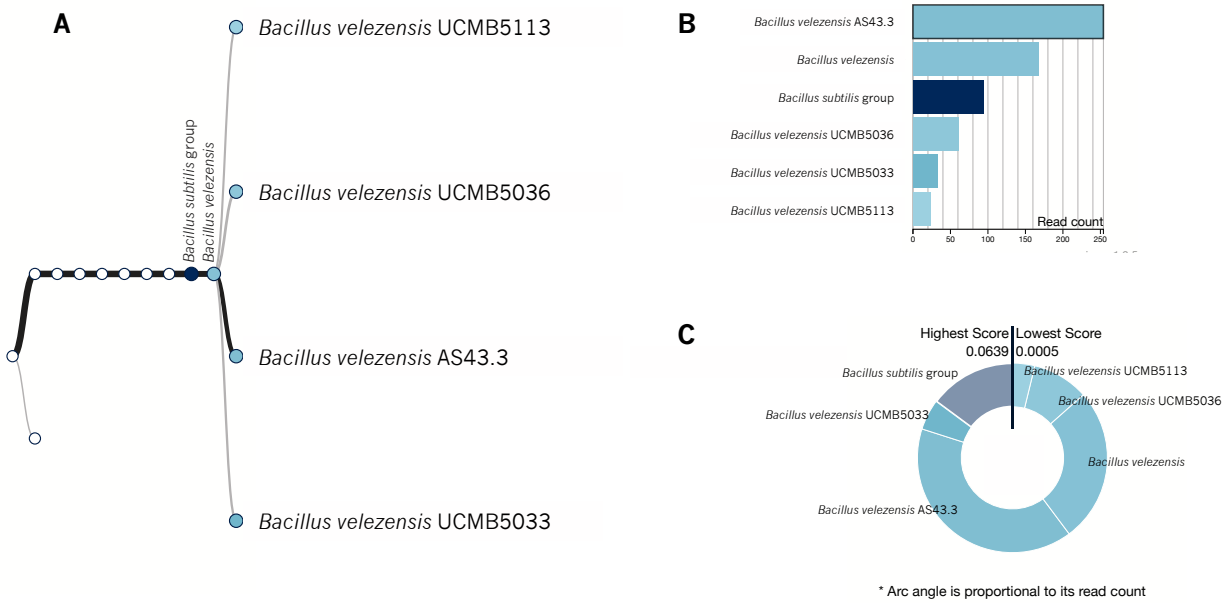


Figure D.1 | **WIMP – 'What's in my pot' real-time identification of *Bacillus* sp. MG SD 082 using whole-genome nanopore-sequencing data.** Figure A depicts the NCBI tree nodes represented in the analyzed sequencing data. Bar chart B shows the number of reads with each different classification. Donut chart C shows relative proportion of reads for each identification ordered by confidence level.

# Appendix E. Selected annotation results of the *Bacillus velezensis* MG SD 082 sequencing data

Table E.1 | **Putative polysaccharide-degrading enzymes with industrial potential identified in the *Bacillus velezensis* MG SD 082 nanopore-sequencing data.** 'Annotation description' refers to the description attributed by the annotation systems. For RAST derived annotations the FIGfam code is shown. Correspondingly, for Blast2GO derived annotations, the GI number (NCBI GenInfo Identifier) of the Top Blast Hit is presented. For PSORTb only results with scores above 7 are reported. For CAZy families only results with *E*-values lower that $5x10^{-4}$ are reported.

| Annotation description | Annotation system | FIGfam | Top Hit GI number | CAZy Family | PSORTb |
|---|---|---|---|---|---|
| **Starch-degrading** | | | | | |
| α-amylase (EC 3.2.1.1) | RAST; Blast2GO | FIG00004763 | GI:700308105 | GH13 | E |
| α-glucosidase (EC 3.2.1.20) | RAST; Blast2GO | FIG00745599 | GI:328551956 | GH13 | C |
| Oligo-1,6-glucosidase (EC 3.2.1.10) | RAST; Blast2GO | FIG00086220 | GI:919440866 | GH13 | C |
| **Cellulose-degrading** | | | | | |
| 6-phospho-β-glucosidase | Blast2GO | na | GI:597504640 | GH4 | U |
| β-glucosidase (EC 3.2.1.21) | RAST; Blast2GO | FIG00001469 | GI:1074988413 | GH1 | C |
| β-glucosidase (EC 3.2.1.21); 6-phospho-β-glucosidase (EC 3.2.1.86) | RAST; Blast2GO | nd | GI:406858681 | GH1 | U |
| β-glucosidase (EC 3.2.1.21); 6-phospho-β-glucosidase (EC 3.2.1.86) | RAST | nd | na | GH1 | U |
| **Xylan-degrading** | | | | | |
| 1,4-β-xylanase | Blast2GO | na | GI:57338944 | GH11 | E |
| Acetylxylan esterase related enzyme | RAST | FIG01376833 | na | CE6 | U |
| α-N-arabinofuranosidase (EC 3.2.1.55) | RAST; Blast2GO | FIG00007157 | GI:1004876411 | GH51 | U |
| α-N-arabinofuranosidase 2 (EC 3.2.1.55) | RAST | nd | na | GH51 | U |
| β-xylosidase (EC 3.2.1.37) | RAST | FIG00003086 | na | GH43 | U |
| Endo-1,4-β-xylanase A precursor (EC 3.2.1.8) | RAST | FIG00475203 | na | GH43 | E |
| Glucuronoxylanase [Bacillus pumilus] | Blast2GO | na | GI:647227048 | GH30 | E |
| Xylanase chitin deacetylase | Blast2GO | na | GI:387170786 | CE4 | U |
| **Mannan-degrading** | | | | | |
| α -galactosidase (EC 3.2.1.22) | RAST; Blast2GO | FIG00002020 | GI:549062795 | GH4 | C |
| Mannan endo-1,4-β-mannosidase precursor (EC 3.2.1.78) | RAST; Blast2GO | nd | GI:505053743 | GH26 | U |
| **Pectin-degrading** | | | | | |
| Arabinan endo-1,5-α-L-arabinosidase (EC 3.2.1.99) | RAST; Blast2GO | FIG00036772 | GI:799135232 | GH43 | C |
| Arabinogalactan endo-1,4-β-galactosidase | RAST | FIG01550130 | na | GH53 | U |
| Pectin lyase like protein | RAST | FIG01451709 | na | PL9 | E |
| Pectin lyase like protein | RAST | FIG01451709 | na | PL9 | U |
| Pectin lyase precursor (EC 4.2.2.2); Pectate lyase | RAST; Blast2GO | FIG00905834 | GI:158198564 | PL1 | E |
| **Chitin-degrading** | | | | | |
| β-hexosaminidase (EC 3.2.1.52) | RAST; Blast2GO | FIG00001088 | GI:328551872 | GH3 | U |
| Chitosanase | RAST | FIG01371742 | na | GH46 | U |
| N-acetylglucosamine-6-phosphate deacetylase (EC 3.5.1.25) | RAST; Blast2GO | FIG00076542 | GI:1016513308 | CE9 | U |
| Peptidoglycan N-acetylglucosamine deacetylase (EC 3.5.1.-); Chitooligosaccharide deacetylase | RAST; Blast2GO | nd | GI:1074987444 | CE4 | U |

In CAZy families classification GH stands for Glycoside Hydrolase, CE for Carbohydrate Esterase and PL for Polysaccharide Lyase.
PSORTb results are represented as E – Extracellular, C – Cytosolic and U – Unknown.
na – not applicable.
nd – not defined.

Table E.2 | **Selection of putative potentially relevant proteases and lipases identified from the *Bacillus velezensis* MG SD 082 nanopore-sequencing data.** 'Annotation description' refers to the description attributed by the annotation systems. For RAST derived annotations the FIGfam code is shown. Correspondingly, for Blast2GO derived annotations, the GI number (NCBI GenInfo Identifier) of the Top Blast Hit is presented. For PSORTb only results with scores above 7 are reported. The peptidases presented are only a subset of all peptidases disclosed by the annotation and represent only those that were either identified as being extracellular or belong to peptidase families of commonly extracellular peptidases. For MEROPS families only results with *E*-values lower that $7x10^{-7}$ are reported.

| Annotation description | Annotation system | FIGfam | Top Hit GI number | MEROPS Family | PSORTb |
|---|---|---|---|---|---|
| **Peptidases** | | | | | |
| Extracellular serine protease | RAST | FIG01386572 | na | S8A | E |
| Glutamyl endopeptidase precursor (EC 3.4.21.19), blaSE | RAST; Blast2GO | FIG00019562 | GI:1045807747 | S1D | E |
| Peptidase M20 | Blast2GO | na | GI:1052480118 | M20D | U |
| Peptidase M4 | Blast2GO | na | GI:115304415 | M4 | E |
| Serine alkaline protease (subtilisin E) ( EC:3.4.21.62 ) | RAST | FIG01230769 | na | S8A | E |
| Serine protease | Blast2GO | na | GI: 504071785 | S1E | U |
| Serine protease | Blast2GO | na | GI: 597507379 | S8A | U |
| Serine protease, DegP/HtrA, do-like (EC 3.4.21.-) | RAST; Blast2GO | FIG00083017 | GI:983384791 | S1C | U |
| Serine protease, DegP/HtrA, do-like (EC 3.4.21.-) | RAST; Blast2GO | FIG00083017 | GI:914788432 | S1C | E |
| Serine protease, DegP/HtrA, do-like (EC 3.4.21.-) | RAST | FIG00083017 | na | S1C | U |
| Thermostable carboxypeptidase 1 (EC 3.4.17.19) | RAST | FIG00229345 | na | M32 | U |
| Zinc metalloproteinase precursor (EC 3.4.24.29) / aureolysin | RAST | nd | na | M4 | E |
| **Lipases/Esterases** | | | | | |
| Carboxylesterase (EC 3.1.1.1) | RAST | FIG01225679 | na | na | U |
| Carboxylesterase (EC 3.1.1.1) | RAST | FIG01225679 | na | na | U |
| Carboxylic ester hydrolase | RAST | FIG01343121 | na | na | U |
| FIG006988 Lipase/Acylhydrolase with GDSL-like motif | RAST | FIG00006988 | na | na | CM |
| Lipase | Blast2GO | na | GI:1004872375 | na | U |

PSORTb results are represented as CM – Cytoplasmic Membrane, E – Extracellular and U – Unknown.
na – not applicable.
nd – not defined.

# Appendix F. Proposed pipeline for biotechnological potential assessment using nanopore sequencing

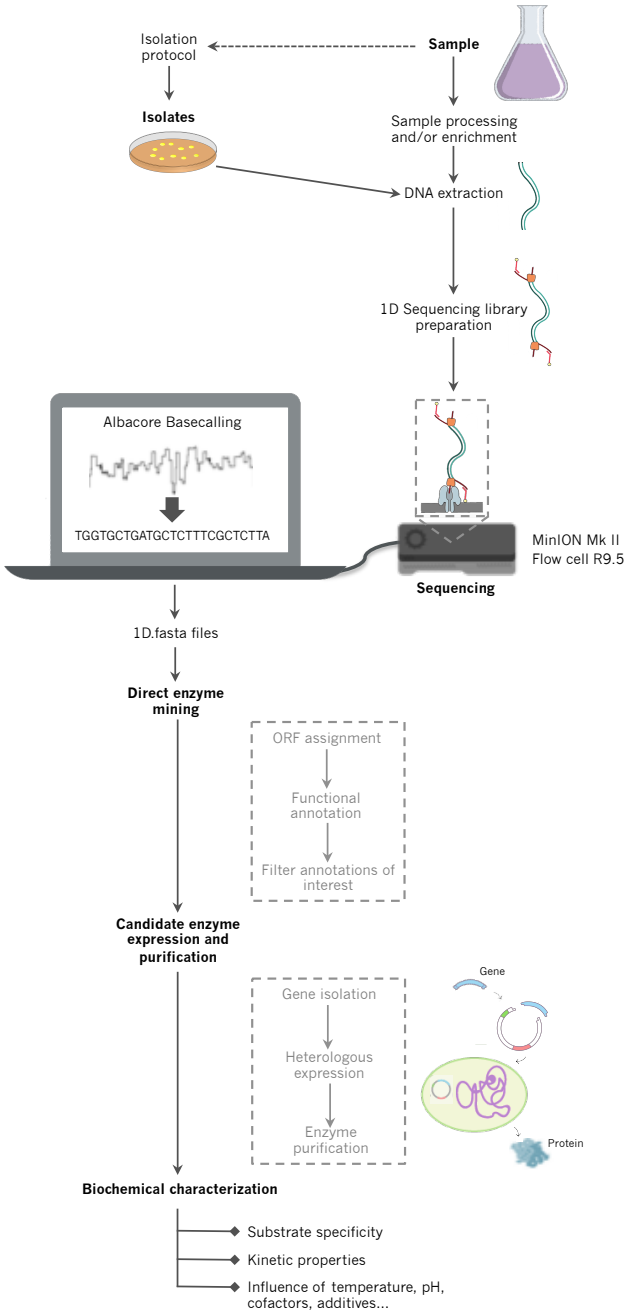


Figure F.1 | **Proposed workflow for nanopore-sequencing based biotechnological potential assessment.** This figure does not represent all the possible pathways of utilizing nanopore-sequencing data. Rather, it represents the basic skeleton of a pipeline aiming to identify the enzyme-encoding potential of an isolate or sample by taking advantage of the latest version R9.5 of nanopore sequencing. The post-sequencing stages can be more or less automated and more or less high-throughput depending on the specific technologies and protocols applied. Annotation can also be general, or alternatively focalized for a particular set of proteins by using specific databases or tools. At the current stage of the technology, in less than 24 hours we can go from sample to annotation results, which allows for a quick initial screening of the potential of the isolate/sample. Then, the results can be further explored by expressing the candidate genes in a more time-investing manner. Nanopore-sequencing data can be useful for the quick screening of other biotechnological interesting characteristics, surpassing the enzyme-encoding potential, but for this purpose different downstream analysis pipelines and functional assays have to be employed.