UNIVERSITE BRETAGNE LOIRE / SCIENCES DE LA MER ET DU LITTORAL

Thèse préparée à l'Université de Bretagne Occidentale

pour obtenir le diplôme de DOCTEUR délivré de façon partagée par L'Université de Bretagne Occidentale et l'Université de Bretagne Loire

Spécialité : Microbiologie

École Doctorale Sciences de la Mer et du Littoral

présentée par

Damien Courtine

Préparée au Laboratoire de Microbiologie des Environnements Extrêmes (LM2E)
UMR6197, UBO – Ifremer – CNRS

Génomique comparative d'isolats phylogénétiquement proches appartenant au genre *Thermococcus*, une archée hyperthermophile

Thèse soutenue le 19 Décembre 2017
devant le jury composé de :

Anna-Louise REYSENBACH
Professeur, Université de Portland (USA) / *Rapporteur*

Simonetta GRIBALDO
Directrice de Recherche, Institut Pasteur / *Rapporteur*

Dominique LAVENIER
Directeur de Recherche, CNRS Rennes / Examinateur

Karine ALAIN
Chargée de recherche, CNRS Brest / *Directrice de thèse*

Loïs MAIGNIEN
Maître de Conférence, Université de Bretagne Occidentale / *Co-encadrant*

A. Murat EREN
Maître de conférence, Université de Chicago (USA) / *Co-encadrant*

Yann MOALIC
Enseignant Chercheur, Université de Bretagne Occidentale / *Invité*

UBO Université de Bretagne Occidentale

Acknowledgements

This thesis work was carried out at the Laboratoire de Microbiologie des Environnements Extrêmes (LM2E) at the Institut Universitaire Européen de la Mer (IUEM). It was financed by the Région Bretagne (Brittany Region) and Labex MER. I thank Mohamed Jebbar, Anne Godfroy and Didier Flament for welcoming me in this laboratory.

First of all, I would like to sincerely thank the members of the jury who agreed to evaluate this work. I would like to thank Anna-Louise Reysenbach, Professor at the Portland University, and Simoneta Gribaldo, Research director at the Institut Pasteur, for agreeing to review this thesis. I would also like to thank Dominique Lavenier, Research Director at the CNRS, and Yann Moalic, Researcher at the Université de Bretagne Occidentale, for reviewing this work.

I would like to thank Karine Alain for her support, availability and kindness. Many thanks to Lois Maignien, for his guidance; this work could not have taken place without his advices and its availability. I would also like to thank A. Murat Eren (aka Meren) for its advices, for allowing me to sequence more genomes in this project, and also for hosted me for 1 month in his lab at the University of Chicago. This visit was a very enriching experience.

Un merci particulier à Patrick Forterre, Jacques Oberto et Violette Da Cunha de m'avoir donner accès à des données génomiques supplémentaires, et également d'avoir finalisé le séquençage de génomes *via* l'ERC EVOMOBIL.

Je remercie Myriam Georges pour toute son aide apportée lors des manips de bio-mol et pour tous les autres moments.
Un merci également à Nadège Bienvenu pour l'utilisation de la souchothèque, et Stéphane L'Haridon pour ses nombreux conseils en culture des *Thermococcales*.
Merci Stéphanie, pour ton aide et ta bienveillance au quotidien.
Merci à toutes les personnes ayant participé de près ou de loin à ce travail.
Merci à l'ensemble des membres du LM2E pour leur disponibilité et leur gentillesse.

Un remerciement particulier à la team EcoGenomics, Florian, Clarisse et la nouvelle recrue Blandine, pour toutes les discussions, les réunions du lundi, merci pour ces moments.
Également à l'ensemble des thésards, passés, présents, futurs, pour toutes les discussions, soutiens, et bons moments passés ensemble, en commençant par les anciens, Gaëlle, Sandrine, Matthieu, Simon, Coraline; la génération 2014: Mélanie, Charlène, Vincent, Gwenn et Tiphaine; Les prochains sur la liste: Sébastien, Yang,

Caroline, Clarisse, et Florian (encore toi?!); les « jeunes »: Sarah, Jordan, Pierre et les derniers arrivés, Marc, Blandine et David.

Je souhaite adresser un sincère merci au FEIRI, et plus particulièrement à Sandy, Frédéric et Emmanuel, pour toute l'aide informatique apportée : les debugs de Penduick, quand le mac ne veut plus démarrer… Merci.

Margane, bien que tu sois partie du labo peu après mon arrivée, un grand merci pour tous les bons moments !

Je voudrais également adresser un merci à Ivan, qui m'a longuement aidé et poussé à faire cette thèse.

Enfin je voudrais remercier ma famille, Maman, Papa, Steph, Brigitte, Mathis, Mémé, d'avoir cru en moi, je n'en serais pas là sans vous.
À ma moitié, pour avoir enduré mes crises et être à mes côtés tous les jours, dans les bons et les autres moments, je ne te remercierai jamais assez.

Table of contents

List of figures

List of tables

List of abbreviations

AAI: Average Amino acid Identity
aLRT: approximate Likelihood-Ratio Test
AFLP: Amplified Fragment Length Polymorphism
AMPHORA: AutoMated PHylogenOmic infeRence
ANI: Average Nucleotide Identity
anvi'o: an ANalysis and Vizualisation platform for 'Omics data
BAM: Binary Alignment/Map format
BBH: Best BLAST Hit
BLAST: Basic Local Alignment Search Tool
BIONJ: Improved version of Neighbor Joining
BMGE: Block Mapping and Gathering with Entropy
bp / kbp / Mbp / Gbp: Base pair / kilo bp / Mega bp / Giga bp
COG: Cluster of Orthologous Gene
DDH: DNA-DNA Hybridization
ddNTP: DideoxyNucleotide-5'-TriphosPhate
DHVE2: Deep-sea Hydrothermal Vent Euryarchaeota 2
DNA: DeoxyriboNucleic Acid
dNTP: DeoxyNucleotide-5'-TriphosPhate
DSHV: Deep-sea Hydrothermal Vent
DSMZ: Deutsche Sammlung von Mikroorganismen und Zellkulturen
EDTA: EthyleneDiamineTetraacetic Acid
EMBOSS: European Molecular Biology Open Software Suite
EPR: East Pacific Rise
GGDC: Genome-to-Genome Distance Calculator
HGT/LGT: Horizontal Gene Transfer / Lateral Gene Transfer
HMM: Hidden Markov Model
HR: Homologous Recombination
iToL: Interactive Tree Of Life
ITS: Internal Transcribed Spacer
KAAS: KEGG Automatic Annotation Server
KEGG: Kyoto Encyclopedia of Gene and Genome
LG: Le and Gascuel
LM2E: Laboratoire de Microbiologie des Environnements Extrêmes (Lab of Microbiology of Extremes Environments)
MAFFT: Multiple Alignment using Fast Fourier Transform
MAR: Mid-Atlantic Ridge
MBGE: Molecular Biology of Gene in Extremophiles
MBL: Marine Biological Laboratory
MCL: Markov Clustering Algorithm
MPa: Mega Pascal

MUMmer: Maximal Unique Match software
MUSCLE: MUltiple Sequence Comparison by Log-Expectation
NCBI: National Center for Biotechnology information
NGS: Next Generation Sequencing
NNI: Nearest Neighbor Interchange
OTU: Operational Taxonomic Unit
PC: Protein Cluster
PCR: Polymerase Chain Reaction
qPCR: Quantitative PCR
RDP: Ribosomal Data Project
RFLP: Restriction Fragment Length Polymorphism
RNA: RiboNucleic Acid
rRNA: Ribosomal RNA
SAM: Sequence Alignment/Map format
SCG: Single Copy core-Gene
SINA: Silva INcremental Aligner
SMS: Smart Model Selection in PhyML
SNP: Single-Nucleotide Polymorphism
SPC: Specific Protein Cluster
SRA: Sequence Read Archive
SSU: Small Sub-Unit
TACK: *Thaumarchaeota*, *Aigarchaeota*, *Crenarchaeota*, *Korarchaeota*
PhyML: Phylogenetic Maximum Likelihood
Tm: Temperature of Melting
TRM: *Thermococcales* Rich Medium
UBOCC: Université de Bretagne Occidentale (University of Western Brittany) Culture Collection
UV: Ultra-Violet
WGS: Whole Genome Shotgun

Introduction

Chapter I: Introduction

I) Introduction of deep-sea hydrothermal vents

For a long time, no one suspected that life could thrive in the abyssal depths of the oceans. The year 1977 was a turning point in the field of biology: a dive of the submersible Alvin in Galápagos rift at 2500 m below sea level has shed light on organisms living near deep-sea thermal hot spring (Corliss and Ballard, 1977; Corliss et al., 1979). These hot springs are also known as deep-sea hydrothermal vent. Light can only diffuse in the upper layers of oceans, so these ecosystems are based on chemosynthetic microorganisms, which turn inorganic carbon into organic molecules (Corliss et al., 1979) using energy from inorganic molecules instead of light.

1) Generalities about deep-sea hydrothermal vents

Since their discovery in 1977, deep-sea hydrothermal vents (DSHV) are subject to intense scientific investigations as their ecology and functioning is in sharp contrast with surface ecosystems. They are present all around oceans in active volcanic areas (Figure 1), and they are distributed from sea level to 4964 m. The Beebe vent field at the Cayman rise is the actual deepest known vent field (Connelly et al., 2012). In these active areas, the oceanic crust is fissured due to the thermal activities and this leads to increase in its permeability. Cold seawater can percolate through these cracks, the heat source beneath the ocean crust. On its way, water reacts with warm rocks and washes them out. The resulting fluid is hot (up to 400°C) and chemically reduced. It then rises through the ocean crust. When the hydrothermal fluid mixes with cold seawater, 2-3°C, minerals carried by the fluid precipitate (Figure 2). This phenomenon forms a growing vertical structure called hydrothermal chimney or “Black smoker” (Edmond et al., 1982).

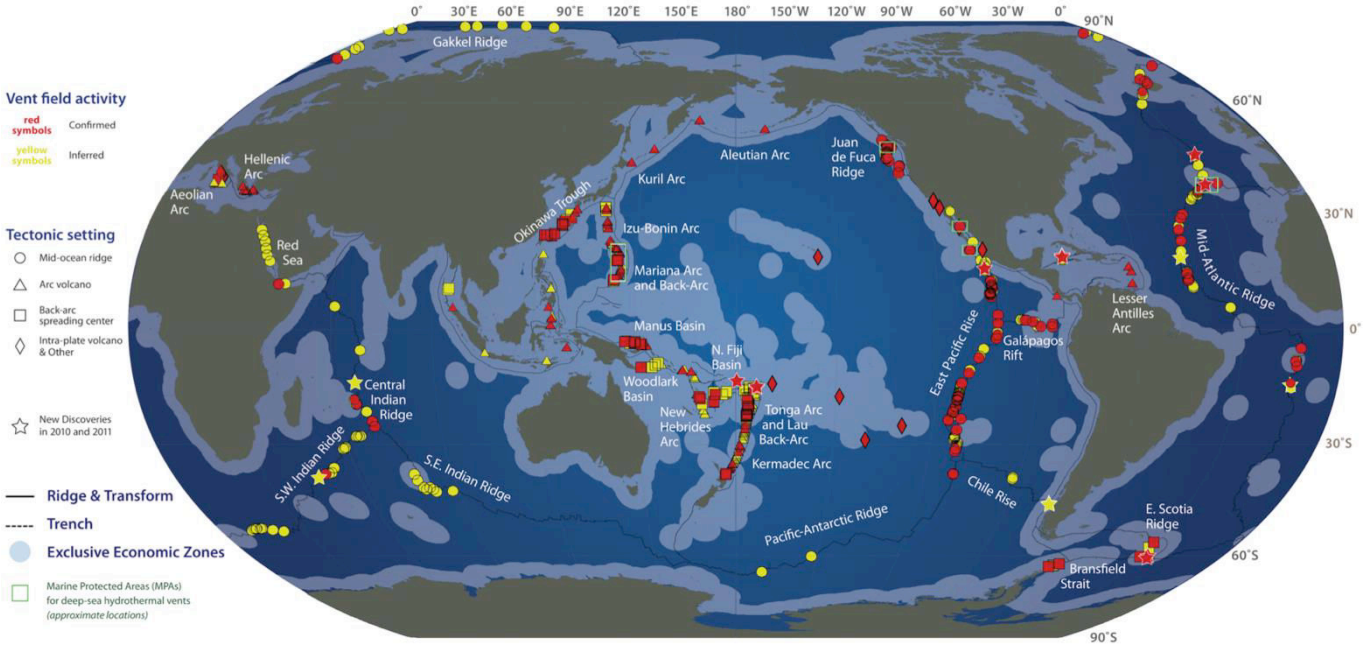


Figure 1: Localization of known hydrothermal vents

From http://www.whoi.edu/page.do?pid=83497&tid=7842&cid=71854

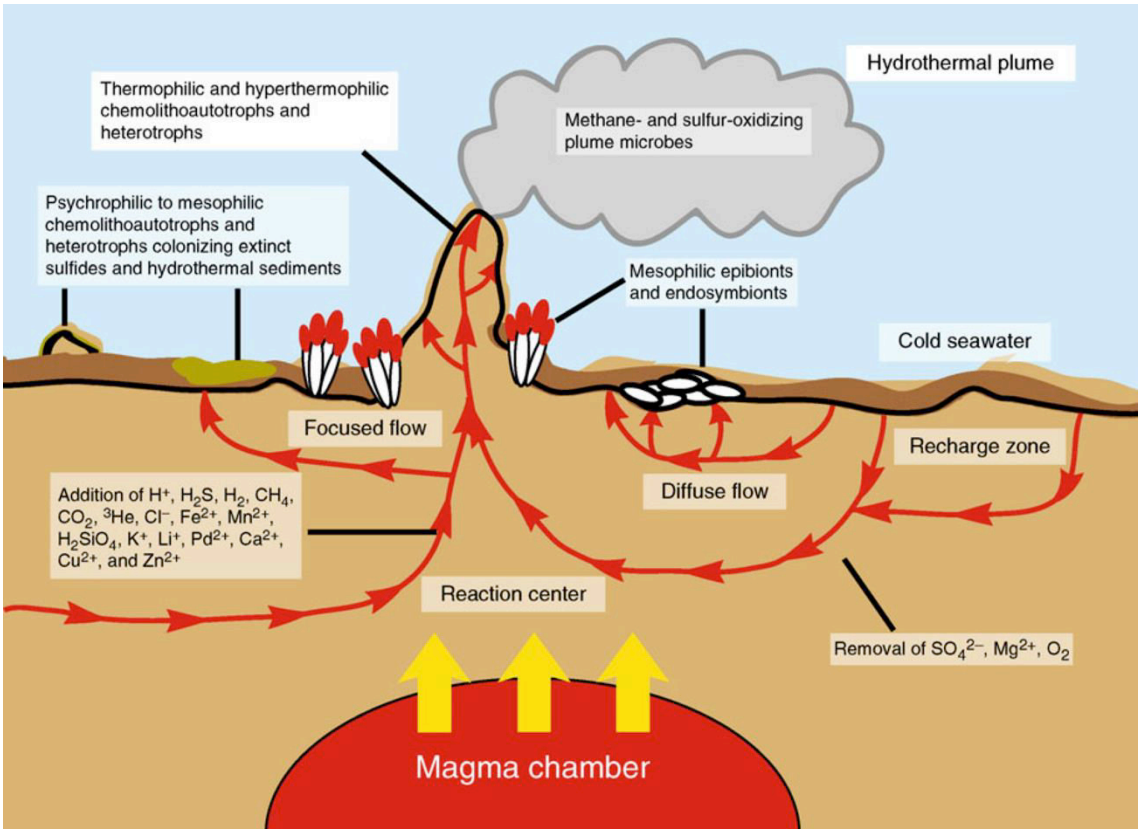


Figure 2: Formation of a deep-sea hydrothermal vent

From Flores and Reysenbach, 2011.

Each hydrothermal vent field is different in terms of chemical composition. This is due to multiple parameters: dilution rate of fluid in seawater, seabed mineral composition, and the geophysical context (pressure and temperature used to influence chemical reactions) (Wetzel and Shock, 2000). When seawater penetrates the ocean crust, it starts to loose chemical elements such as dissolved 0_2, SO_4^{2-}, PO_4^{2-}, NO_3^-, Mg^{2+}, and on the other hand, leaches metallic compounds (Zn^{2+}, Mn^{2+}, Fe^{2+}, Cu^{2+}) and dissolved gases like H_2, H_2S, CH_4, CO and CO_2 (Kelley et al., 2002; Von Damm, 1995). At this stage, it is no longer seawater but hot and anoxic hydrothermal fluid. Mainly the rocks of seabed, basalts or peridiotites, influence its composition (Wetzel and Shock, 2000). This leads into two categories of vent fields: on the one hand ultramafic vent systems with a peridotite basement, and on the other hand, basalt-hosted system with predominance of basalts in seabed. In general, hydrothermal fluid in ultramafic context is characterized by high dihydrogen and methane concentrations, whereas fluid from basalt-hosted hydrothermal systems is enriched in dihydrogen sulfide and is deprived of dihydrogen and methane (McCollom, 2007; Wetzel and Shock, 2000).

Hydrothermal vents are located on average at 2000 m water depth, well below the light penetration limit of the ocean surface. In addition, it is estimated that only 1% of the organic material produced in the upper layers of the ocean can sink to the seabed (Jannasch and Taylor, 1984). Consequently, chemosynthetic lithotrophic *Bacteria* and *Archaea* that can produce complex organic molecules from the fixation of CO_2 are at the basis of trophic web. In abysses, there is a strong hydrostatic pressure. It increases of 0.1 MPa every 10 m of water column, either the value of the atmospheric at the sea level. Further, the hydrothermal fluid can be very hot, 350-400°C, and acid, pH 2-4 (Martin et al., 2008; Von Damm, 1995), while surrounding water is cold, 2-3°C and nearly neutral. When hydrothermal fluid and water mix together, it creates intense

physical-chemical gradients. All these parameters mean that DSHV are qualified as an extreme environment.

2) Deep-sea hydrothermal vent ecology

Hydrothermal vents are often compared to oases of life within abyssal deserts. They indeed harbor a striking number of metazoans: shrimps (*Rimicaris exoculata*), crabs (*Bythograea thermydron*), mollusks (*Bathymodiolus thermophilus*), giant tube worm (*Alvinella pompejana*), etc. (Lutz and Kennish, 1993).

Also, microbial diversity is studied since the discovery of DSHV. The first way to do this is classical in microbiology: isolating microorganisms in pure culture and studying their phenotypes. But as many bacterial and archaeal lineages are still eluding cultivation, additional approaches based on direct environmental DNA sequencing are often used to explore taxonomic and functional diversity. Since the advent of high throughput sequencing technologies, the number of candidate microbial species/genera associated to DSVH is constantly increasing. Of course, viruses and *Eukarya* should not be forgotten, but this work is only focused on microorganisms belonging to an archaeal order.

Both *Archaea* and *Bacteria* are found in different ecological niches: in the hydrothermal plume, on the walls of chimney, in sediments and associated with fauna.

Concerning *Archaea*, most of the lineages belong to the *Euryarchaeota* phylum. It includes *Thermococcales*, *Methanococcales*, *Methanopyrales*, *Archaeoglobales*, *Halobacteriales*, and the Deep-sea Hydrothermal Vent Euryarchaeota group 2 (DHVE2) (Flores et al., 2012; Fortunato and Huber, 2016; Nakagawa and Takai, 2008; Reveillaud et al., 2016; Roussel et al., 2011; Takai and Nakamura, 2011). Some other lineages belong to the TACK (*Thaumarchaeota*, *Aigarchaeota, Crenarchaeota*, and *Korarcheota*) superphylum (Guy and Ettema, 2011). These are mainly *Desulfurococcales* and

Nitrosupumilales (Fortunato and Huber, 2016). Phylogenetic placement of some newly discovered archaeal clades is still controversial, like the *Lokiarchaeota* phylum (Da Cunha et al., 2017; Spang et al., 2015). Spang and collaborators identified these new organism from metagenomics data, collected at Loki's Castle vent site (Arctic Ocean).

Bacterial communities are found mostly in less warm niches of the hydrothermal vents. Mainly, *Proteobacteria* (by abundance: *Epsilon*-, *Gamma*-, *Alpha*- and *Delta*-) represent the vast majority of bacterial lineages in DSHV. There are also *Aquificales*, *Bacteroidetes*, *Firmicutes* and *Thermotogales* (Figure 3) (Flores et al., 2011; Huber et al., 2006; Nakagawa and Takai, 2008; Orcutt et al., 2011; Reveillaud et al., 2016; Teske et al., 2002). The deep-sea shrimp, *Rimicaris exoculata* is known to harbor a complex symbiosis with microorganisms. The community associated with *R. exoculata* is dominated by *Epsilonproteobacteria* (Guri et al., 2012).

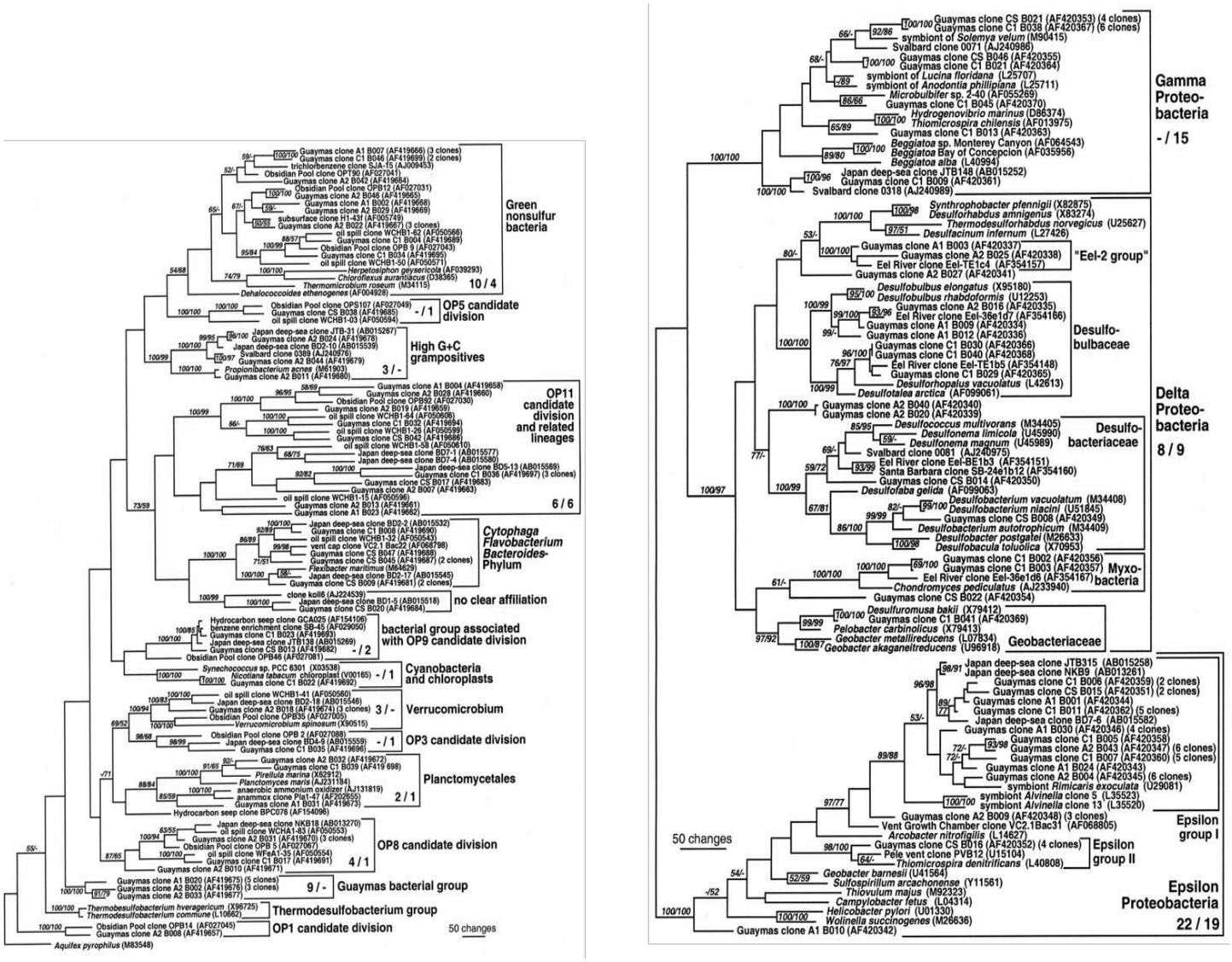


Figure 3: Bacterial diversity of hydrothermal active sediments of the Guaymas Basin

These distance trees are based on full size 16S rRNA gene sequences. Left tree: *Bacteria* without phylum of *Proteobacteria*. Right tree: *Proteobacteria* only. From Teske et al., 2002.

3) Extremophiles

Extremophile, from the Latin "*extremus*" and the Greek "*philiā*", means literally "love the extreme". The word Extremophile was proposed in the 1974 to facilitate the communication about "*organisms able to populate environments hostile to mesophiles, or organisms which grow only in intermediate environments*" (Macelroy, 1974). To date, an accepted definition of extremophile is an organism that thrives under chemical or physical extreme conditions, such as high or low pH or temperature, high salinity or pressure (Horikoshi and Bull, 2011; Madigan et al., 2012). This definition is anthropocentric thought. What seems extreme for a human, for example temperature above 80°C, is absolutely normal and contribute to the proper development of hyperthermophilic *Archaea*. Although the word extremophile often refers to unicellular microorganisms members of *Archaea* or *Bacteria*, some fungi can live in extreme acid environments. And other *Eukarya* like the well-known micro-animal Tardigrade that can bear with multiple extreme conditions (Rothschild and Mancinelli, 2001; Seki and Toyoshima, 1998). Even though 14 different extremophilies are listed in the *Extremophile Handbook* (Horikoshi and Bull, 2011), it refers mostly to the following physicochemical parameters: temperature (psychrophile and (hyper-)thermophile), pH (acidophile and alkaliphile), salinity (halophile) and pressure (piezophile).

Thermophile means every organism that needs a sufficiently hot environment to thrive. The thermophilic feature is delimited by the value of 40°C as the lower threshold for an organism to develop. This group of organisms is split into two categories: a thermophile organism has an optimum growth temperature upper or equal to 45°C, and a hyperthermophile has an optimum growth temperature above 80°C (Figure 4) (Madigan et al., 2012). Within thermophiles, are present some *Eukarya*, among which fungi have the upper temperature limit. This value is between 60°C and 65°C (Madigan

et al., 2012; Maheshwari et al., 2000; Tansey and Brock, 1972). Beyond this limit of 65°C, only *Bacteria* and *Archaea* can live. If we increase the temperature to 80°C and above, almost all microorganisms able to grow are anaerobic *Archaea* (Stetter, 2006).

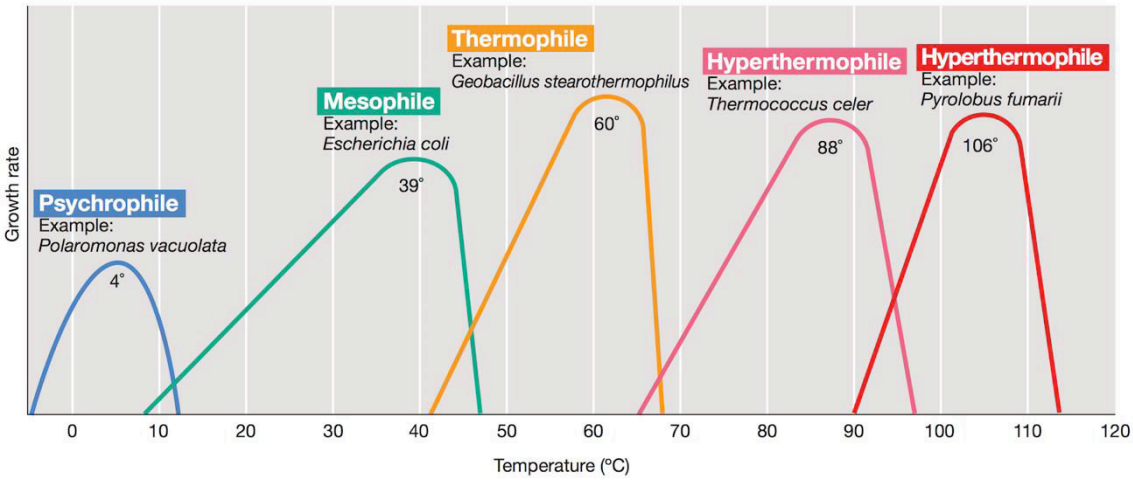


Figure 4: Optimal growth temperature of microorganisms
From Madigan et al., 2012

The story of thermophilic microorganisms started in 1965 when a microbiologist, Thomas Brock, discovered in thermal springs of Yellowstone national park and California (USA), a bacterium that can optimally grow at 70°C (Brock and Freeze, 1969). Consequences of this discovery are multiple: it opened a new field of research on the discovery of new organisms able to grow at high temperature; study the physiology and ecology of such organisms; find new molecule of biotechnological interest. As an example of this last topic, the DNA polymerase (an enzyme that catalyze the synthesis of DNA during the replication step) from *Thermus aquaticus*, the *bacterium* isolated by T. Brock, has allowed to develop a technic widely use in molecular biology: the polymerase chain reaction (PCR) (Chien et al., 1976; Guyer and Koshland, 1989; Saiki et al., 1985).

II) The third domain of life: *Archaea*

1) Generalities

This work has focused on *Thermococcus*, a hyperthermophilic *Archaea* found primarily in deep-sea hydrothermal environments. *Archaea* is the 3rd domain of life. It is by working on the composition of ribosomal RNA sequences that Woese and Fox emitted the hypothesis of a third domain of life (Woese and Fox, 1977). Indeed, they found that life is split into 3 domains: the first contains all organisms that are affiliated to *Bacteria*. The second gather all eukaryotic organisms together (further *Eukarya*). And the third is composed of only "methanogenic *bacteria*" (Woese and Fox, 1977). For the authors, this new group is more than a subdivision of the *Bacteria* kingdom, like "Gram-positive" and "Gram-negative". So they proposed a name for this group, the *Archaebacteria*. The Greek root of *Archaea* (*ἀρχάίος*) means ancient. This name refers to the methanogenic phenotype of these organisms that is suspected to have appeared earcly on during life development on the Earth (Woese and Fox, 1977). Few years later, within this new group and based on 16S rRNA genes, a study proposed to divide *Archaeabacteria* into two phyla. The first comprised methanogens and halophiles, while the second is composed of *Sulfolobus*, a thermoacidophile (Fox et al., 1980). These two groups have subsequently been confirmed, and named *Euryarchaeota* and *Crenarchaeota* respectively. Ten years later, Woese and collaborators proposed the concept of domain: a taxonomic rank above the kingdom, because their result did not match with the Whittaker's taxonomy based on the five kingdoms (*Animalia*, *Plantae*, *Fungi*, *Protista* and *Monera*) and the dichotomy *Eukaryote-Prokaryote* (Whittaker and Margulis, 1978; Woese et al., 1990). These 3 domains are: *Archaea*, *Bacteria* and *Eukarya* (previously named *Eukaryota*). Woese and colleagues proposed to keep the actual kingdom ranks,

and create two new kingdoms for the *Archaea* domain: the *Euryarchaeota* and the *Crenarchaeota*.

Phylogeny of *Archaea* is a broadly discussed topic within the scientific community. Metagenomics and single-cell genomics, two major technics that allow the discovery of new lineages in a culture-independent way, is now widely used to populate the archaeal tree. Even with this massive amount of data, the *Archaea* domain root and the structure of archaeal phyla/superphyla/kingdom is still under debate (Da Cunha et al., 2017; Hug et al., 2016; Petitjean et al., 2015; Raymann et al., 2015; Rinke et al., 2013; Williams et al., 2017; Zaremba-Niedzwiedzka et al., 2017), and even the existence of an archaeal domain (Forterre, 2015; Spang et al., 2015).

According to recent phylogenies, the *Archaea* domain is divided in 4. The first part is the phylum *Euryarchaeota*, which include, among other orders, *Thermococcales*. The second is the superphylum TACK (Guy and Ettema, 2011), witch includes the four initials phyla: *Thaumarchaeota*, *Aigarchaeota*, *Crenarchaeota* and *Korarchaeota*; and the three new phyla: *Bathyarchaeota* (Meng et al., 2014), *Vestraetearchaeota* (Vanwonterghem et al., 2016) and *Geoarchaeota* (Kozubal et al., 2013). The third is the superphylum Asgard, where the phyla *Heimdallarchaeota* (Zaremba-Niedzwiedzka et al., 2017), *Thorarchaeota* (Seitz et al., 2016), *Odinarchaeota* (Zaremba-Niedzwiedzka et al., 2017) and *Lokiarchaeota* (Spang et al., 2015) are present. The last major cluster is the superphylum DPANN (Castelle et al., 2015; Rinke et al., 2013), which includes the following phyla: *Woesarchaeota*, *Pacearchaeota*, *Mamarchaeota*, *Pavarchaeota*, *Nanoarchaeota*, *Nanohaloarchaeota*, *Aenigmarchaeota*, *Micrarchaeota*, *Diapherotrites* and *Altiarchaeales* (Figure 5).

In contrast to TACK and *Euryarchaeota*, the two superphyla DPANN and Asgard are only composed of genomes extracted from metagenomes, and none of them are cultivated (Adam et al., 2017, Castelle and Banfield, 2018).

All works of this thesis will focus on hyperthermophilic and anaerobic microorganisms that belong to the *Thermococcales* order.

2) The order *Thermococcales*

Thermococcales is an archaeal order that belongs to the domain of *Archaea*, phyla of *Euryarchaeota* and class of *Thermococci*. This order is composed of three genera: *Thermococcus* (Zillig et al., 1983), *Pyrococcus* (Fiala and Stetter, 1986) and *Palaeococcus* (Takai et al., 2000). Members of *Thermococcales* are cocci of about 1 µm in diameter. They are strict anaerobes, with some exceptions (Amend et al., 2003; Thorgersen et al., 2012), hyperthermophiles and heterotrophs. Additionally, few members of *Thermococcus* have the ability to grow lithotrophically by anaerobic oxidation of CO (Lee et al., 2008; Sokolova et al., 2004). They can use a broad range of molecules as carbon source (proteins, peptides, carbohydrates) by fermentation. For certain isolates, S^0 stimulates their growth, while others require S^0 to grow indicating a mixed fermentative / anaerobic respiration energy conservation strategy. *Thermococcales* are often isolated from both deep-sea and shallow hydrothermal environments (*Pyrococcus* is only found in deep-sea).

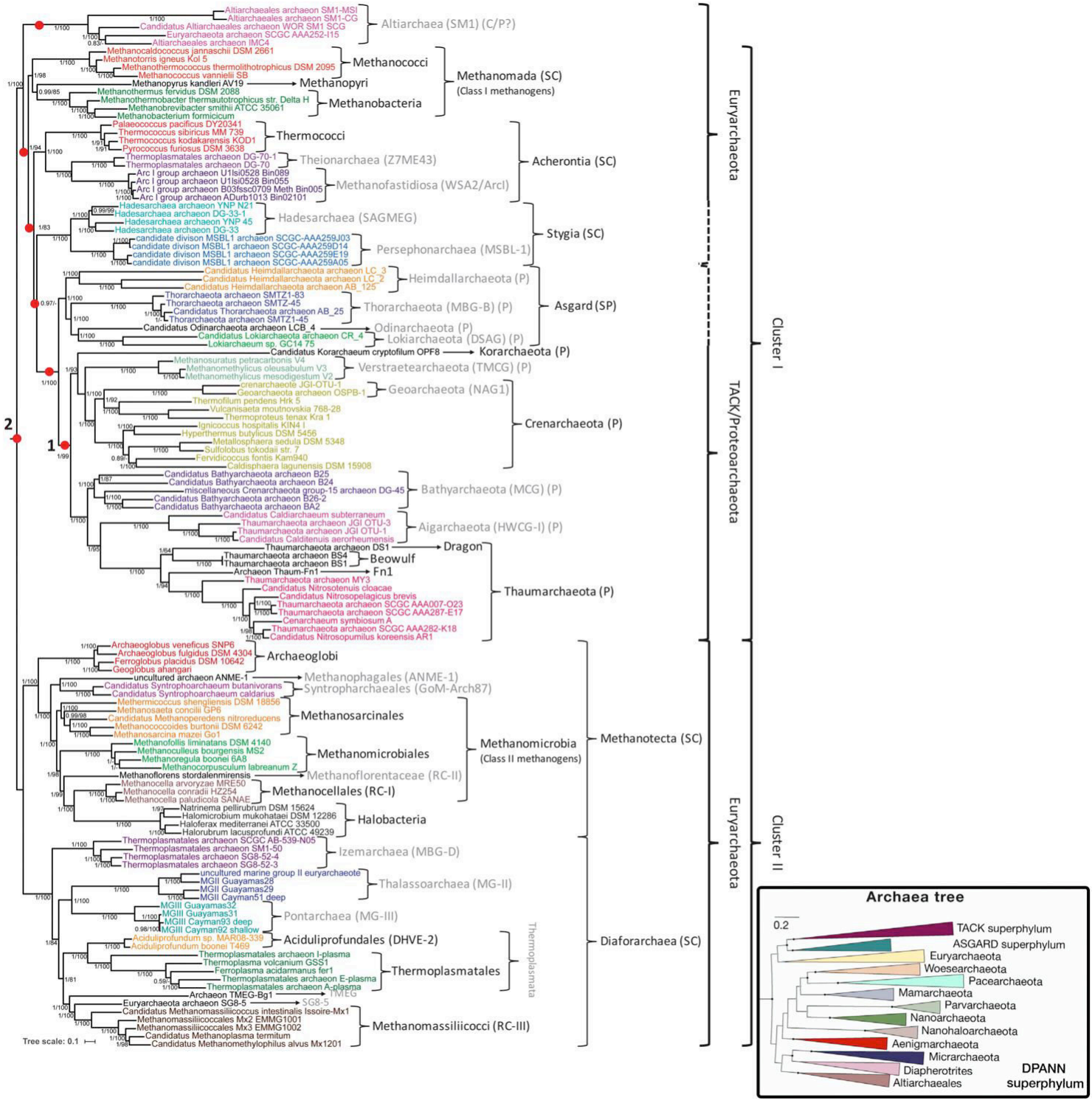


Figure 5: Phylogenetic tree relating the archaeal domain

Phylogenetic tree from Adam et al. (2017). Red dots represent different root for this tree. The #1 represent the traditional root (Woese et al., 1990), between *Euryarchaeota* and *Crenarchaeota* (now TACK). The #2 represents a new well-supported root proposed by Raymann et al (2015). Grey names represent clades for which no isolates are available. Bottom right panel shows a simplified archaeal phylogeny, from Castelle and Banfield 2018. SC: Super Class; P: Phylum; SP: Super Phylum.

There is however exceptions, for example *Thermococcus sibiricus*, which was isolated from a siberian high-temperature oil reservoir (Miroshnichenko et al., 2001). Their optimal growth temperature range between 75 and 100°C (Kobayashi, 2015; Stetter and Huber, 2015; Takai, 2015; Zillig and Reysenbach, 2015). To date, 45 species of *Thermococcales* are characterized (Table 1). In molecular studies, *Thermococcales* are always found in deep-sea hydrothermal ecosystems, all around the globe. From these studies, *Thermococcales* form a minor part of the microbial community at vent sites. However, they are relatively easy to isolate, as evidenced by the large number of isolates available in 'Taxonomy' database (NCBI) and in culture collections. Unsurprisingly, *Thermococcales* thus constitute the order with the largest number of isolates in the LM2E culture collection (UBOCC), with around 300 isolates available.

From the biotechnology point of view, the most striking application is the DNA polymerase of *Pyrococcus furiosus*. This enzyme is known as '*Pfu* polymerase'. It is widely used in PCR thanks to its high fidelity compared to the traditional *Taq* polymerase. In addition of its polymerase activity, *Pfu* polymerase has a 3'-5' exonuclease activity that enhances its fidelity (Lundberg et al., 1991). This enzyme was patented for the first time in 1996 under patents 'US5489523' and 'US5545552'.

On the genomics side, 38 completes genomes are currently available on NCBI (*Genomes* database). They distribute as follow: 1 *Palaeococcus*, 10 *Pyrococcus* and 27 *Thermococcus* (Table 2). An additional set of 12 genomes is available in public databases, but they are assembled at scaffold or contig level. *Thermococcales* have a single chromosome, with a mean genome length around 2 000 000 base pairs (bp). The largest genome to date is *Thermococcus barophilus* CH5 (2.39 Mbp) and the smallest one is *Thermococcus* sp. P6 (1.52 Mbp). The percentage of G+C is on average 42% for *Pyrococcus* species, with the exception of *Pyrococcus yayanosii* (51.6 %). Concerning

Palaeococcus, the 2 available genomes have different G+C content: 43% for *P. pacificus* and 54% for *P. ferrophilus*. Members of *Thermococcus* are split into two groups: the main with a mean G+C value of 54% [51.1 – 56.4], and the second group, composed of 6 genomes, has a mean GC% value of 42% [40.3 – 44.6] (Table 2). Although these microorganisms are hyperthermophile, their G+C content is not as high as expected. Indeed, *in vitro*, the temperature of DNA denaturation is function of the G+C content of the sequence. But *in vivo*, DNA stability may be enhanced by ions, proteins or metabolites (Jaenicke and Sterner, 2006). No correlation has been established between the optimal growth temperature of a microorganism and the G+C content of its DNA molecule (Galtier and Lobry, 1997). Nevertheless, according to Galtier and Lobry, this correlation exists between sequences of RNA (ribosomal and transfer) and the optimal growth temperature. *Thermococcales* harbor one chromosome, however some members of *Pyrococcus* and *Thermococcus* possess also mobile genetic elements. They are split into two classes: viruses and plasmids. As example of viruses, we can cite the two currently characterized ones: *Pyrococcus abyssi* virus 1 (Geslin et al., 2003) and *Thermococcus prieurii* virus 1 (Gorlas et al., 2012). Concerning plasmids, nearly 20 are described within the genera *Thermococcus* and *Pyrococcus*. The first one was identified in 1992, from a *Pyrococcus abyssi* isolate (Erauso et al., 1996). These small molecules are known as vector for exchanging genetic information between cells. Thus this can increase genetic diversity within their hosts. *Thermococcus nautili* harbors a remarkable plasmid, pTN3, which encode an integrase. This enzyme drives large-scale genomic inversions in few generations. According to Cossu and colleagues (Cossu et al., 2017), this shuffling mechanism could occur in case of rapid environmental changes, by providing alternate patterns of gene expression.

Table 1: List of all *Thermococcales* species characterized to date

According to: http://www.bacterio.net/ and NCBI Taxonomy, https://www.ncbi.nlm.nih.gov/taxonomy

Genus	Species	Reference
Thermococcus	*acidaminovorans*	(Dirmeier et al., 1998)
Thermococcus	*aegaeus*	(Arab et al., 2000)
Thermococcus	*aggregans*	(Canganella et al., 1998)
Thermococcus	*alcaliphilus*	(Keller et al., 1995)
Thermococcus	*altanticus*	(Cambon-Bonavita et al., 2003)
Thermococcus	*barophilus*	(Marteinsson et al., 1999)
Thermococcus	*barossii*	(Duffaud et al., 1998)
Thermococcus	*celer*	(Zillig et al., 1983)
Thermococcus	*celericrescens*	(Kuwabara et al., 2007)
Thermococcus	*cleftensis*	(Hensley et al., 2014)
Thermococcus	*coalescens*	(Kuwabara et al., 2005)
Thermococcus	*eurythermalis*	(Zhao et al., 2015)
Thermococcus	*fumicolans*	(Godfroy et al., 1996)
Thermococcus	*gammatolerans*	(Jolivet et al., 2003)
Thermococcus	*gorgonarius*	(Miroshnichenko et al., 1998)
Thermococcus	*guaymasensis*	(Canganella et al., 1998)
Thermococcus	*hydrothermalis*	(Godfroy et al., 1997)
Thermococcus	*kodakaraensis*	(Atomi et al., 2004)
Thermococcus	*litoralis*	(Neuner et al., 1990)
Thermococcus	*marinus*	(Jolivet et al., 2004)
Thermococcus	*nautili*	(Gorlas et al., 2014)
Thermococcus	*pacificus*	(Miroshnichenko et al., 1998)
Thermococcus	*paralvinellae*	(Hensley et al., 2014)
Thermococcus	*peptonophilus*	(González et al., 1995)
Thermococcus	*piezophilus*	(Dalmasso et al., 2016a)
Thermococcus	*prieurii*	(Gorlas et al., 2013b)
Thermococcus	*profondus*	(Kobayashi et al., 1994)
Thermococcus	*radiotolerans*	(Jolivet et al., 2004)
Thermococcus	*sibiricus*	(Miroshnichenko et al., 2001)
Thermococcus	*siculi*	(Grote et al., 1999)
Thermococcus	*stetteri*	(Miroshnichenko et al., 1989)
Thermococcus	*thioreducens*	(Pikuta et al., 2007)
Thermococcus	*waiotapuensis*	(González et al., 1999)
Thermococcus	*zilligii*	(Ronimus et al., 1997)
Pyrococcus	*chitonophagus*	(Huber et al., 1995; Lepage et al., 2004)
Pyrococcus	*abyssi*	(Erauso et al., 1993)
Pyrococcus	*furiosus*	(Fiala and Stetter, 1986)
Pyrococcus	*glycovorans*	(Barbier et al., 1999)
Pyrococcus	*horikoshii*	(González et al., 1998)
Pyrococcus	*kukulkanii*	(Callac et al., 2016)
Pyrococcus	*wosei*	(Zillig et al., 1987)
Pyrococcus	*yayanosii*	(Birrien et al., 2011)
Palaeococcus	*ferrophilus*	(Takai et al., 2000)
Palaeococcus	*helgesonii*	(Amend et al., 2003)
Palaeococcus	*pacificus*	(Zeng et al., 2013)

Table 2: Genomic characteristics of publicly available complete *Thermococcales* genomes

Species	Strain	Size (Mb)	GC (%)	Plasmid	Accession	Ref
Palaeococcus pacificus	DY20341	1.86	43		CP006019	(Zeng et al., 2015)
Pyrococcus abyssi	GE5	1.77	44.7	1	AL096836	(Cohen et al., 2003)
Pyrococcus chitonophagus	GC74	1.96	44.9		CP015193	(Oger, 2018)
Pyrococcus chitonophagus	GC74	1.97	44.9		LN999010	(Papadimitriou et al., 2016)
Pyrococcus furiosus	DSM 3638	1.91	40.8		AE009950	(Maeder et al., 1999)
Pyrococcus furiosus	COM1	1.91	40.8		CP003685	(Bridger et al., 2012)
Pyrococcus horikoshii	OT3	1.74	41.9		BA000001	(Kawarabayasi et al., 1998)
Pyrococcus kukulkanii	NCB100	1.98	44.6		CP010835	(Oger et al., 2017)
Pyrococcus sp. NA2	NA2	1.86	42.7		CP002670	(Lee et al., 2011)
Pyrococcus sp. ST04	ST04	1.74	42.3		CP003534	(Jung et al., 2012a)
Pyrococcus yayanosii	CH1	1.72	51.6		CP002779	(Jun et al., 2011)
Thermococcus barophilus	MP	2.06	41.7	1	CP002372	(Vannier et al., 2011)
Thermococcus barophilus	CH5	2.39	41.8		CP013050	(Oger et al., 2016)
Thermococcus barossii	SHCK-94	1.92	54.7		CP015101	(Oger, 2018)
Thermococcus celer	Vu 13	1.87	56.4		CP014854	(Oger, 2018)
Thermococcus cleftensis	CL1	1.95	55.8		CP003651	(Jung et al., 2012b)
Thermococcus eurythermalis	A501	2.13	53.5	1	CP008887	(Zhao et al., 2015)
Thermococcus gammatolerans	EJ3	2.05	53.6		CP001398	(Zivanovic et al., 2009)
Thermococcus gorgonarius	W-12	1.67	51.7		CP014855	(Oger, 2018)
Thermococcus guaymasensis	DSM 11113	1.92	52.9		CP007140	(Zhang X. et al., 2014)
Thermococcus kodakaraensis	KOD1	2.09	52		AP006878	(Fukui et al., 2005)
Thermococcus litoralis	DSM 5473	2.22	43.1		CP006670	(Gardner et al., 2012)
Thermococcus nautili	30-1	1.98	54.8	3	CP007264	(Oberto et al., 2014)
Thermococcus onnurineus	NA1	1.85	51.3		CP000855	(Lee et al., 2008)
Thermococcus pacificus	P-4	1.79	54.2		CP015102	(Oger, 2018)
Thermococcus paralvinellae	ES1	1.96	40.3		CP006965	Jung et al., 2013
Thermococcus peptonophilus	OG-1	1.90	51.7	1	CP014750	(Oger, 2018)
Thermococcus piezophilus	CDGS	1.93	51.1		CP015520	(Dalmasso et al., 2016b)
Thermococcus profundus	DT 5432	2.04	53.1	1	CP014862	(Oger, 2018)
Thermococcus radiotolerans	EJ2	1.87	55.6		CP015106	(Oger, 2018)
Thermococcus sibiricus	MM 739	1.85	40.2		CP001463	(Mardanov et al., 2009)
Thermococcus siculi	RG-20	2.03	55		CP015103	(Oger, 2018)
Thermococcus sp. 2319x1	2319x1	1.96	44.6		CP012200	(Gavrilov et al., 2016)
Thermococcus sp. 4557	4557	2.01	56.1		CP002920	(Wang et al., 2011)
Thermococcus sp. 5-4	5-4	1.85	55.7		CP021848	(Cossu et al., 2017)
Thermococcus sp. AM4	AM4	2.09	54.8		CP002952	(Oger et al., 2011)
Thermococcus sp. P6	P6	1.52	54.9		CP015104	(Oger, 2018)
Thermococcus thioreducens	OGL-20P	2.07	53.5		CP015105	(Oger, 2018)

III) Ordering microbial diversity in cohesive units

1) The necessity to classify microorganisms

a) Historical motivations

Classification is an *a priori* condition for rational thought, so we could not perform any reasoning without them. This behavior can be applied to living beings and is motivated by the fact that our environment is so diverse that it is easier to apprehend a summarized and organized version of its entirety. But this brings approximations because we often elude the uniqueness of all individual components of such units. Ordering things into ontological units is a way to reach this goal. In the 18th century, a scientist undertook to classify living things. Carl von Linné, a Swedish botanist and zoologist, undertook to classify the living beings (Animals and Plants). In his main book, *Systema Naturæ* first published in 1735, he classified Animals into 4 400 species and Plants into 7 700 species. As in our modern classification system, Carl von Linné uses nested boxes to order all these living beings. After many attempts to classify living beings, the last accepted classification system is the one published by Woese and colleagues in 1990, with *Bacteria*, *Archaea* and *Eukarya* as the mains upper level units.

b) Classification of living beings

The classification of organisms (living or fossil) followed progresses of science. First, classifications were built based on the morphological features of compared species. Since the advent of DNA sequencing, classification of organisms is based on molecular markers, such as SSU rRNA. Modern classification is generally based on 9 ranks (Figure 6). Living organisms are called by a binomial name (inherited from Linnaeus classification) that is: the genus and species names. The last rank, species, is the most discussed among microbiologists. For *Eukarya*, more precisely Animals and Plants, a common definition of the species is "*a group of organisms with the potential to*

interbreed and produce viable and fertile offspring" (Cohan, 2002; Mayr, 1942). This "biological species concept" has the major disadvantage of being inaplicable to organisms without sexual reproduction such as *Bacteria* and *Archaea*.

Owing to the huge number of microbial cells present on Earth, microbiologists had to find a way to classify them, to make this diversity more apprehensible. It appeared necessary to create units. This can be done by operational criterions, like in the current microbial species definition. In metagenomics, and more generally in microbial community studies, the notion of Operational Taxonomic Unit (OTU) is widely used. It defines groups of sequences, and therefore microorganisms, that share enough similarity.

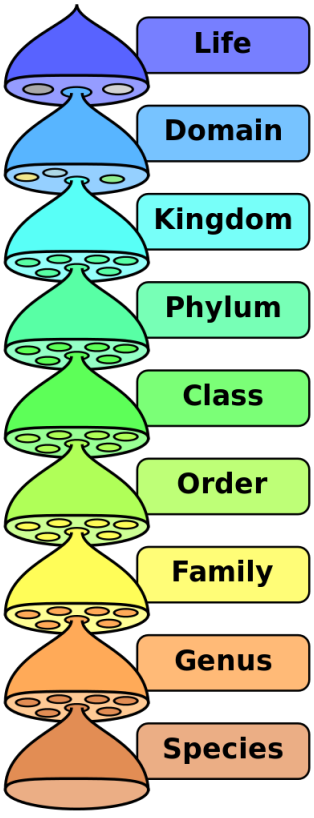


Figure 6: Representation of the ranks in the modern classification

2) The actual definition of a microbial species

a) The definition

Characterizations of new microbial species are published in the *International Journal of Systematic and Evolutionary Microbiology* (IJSEM) and is supported by the deposition of the viable organism in within two international public collections. In 2002, the international committee on systematics of prokaryotes re-evaluated the species definition in microbiology. The current definition of microbial species is the following: "*a species is a category that circumscribes a (preferably) genomically coherent group of individual isolates/strains sharing a high degree of similarity in (many) independent features, comparatively tested under highly standardized conditions"* (Rosselló-Mora and Amann, 2001; Stackebrandt et al., 2002). In other words, a species is a group of individuals that share a high similarity of their DNA sequence (DNA-DNA relatedness, 16S rRNA and/or housekeeping gene sequence similarity), sharing numerous phenotypic resemblances and few differences.

b) How are species delineated?

i) Sequence similarity

- DNA-DNA sequence similarity

To ensure that a newly discovered microbial isolate is a new species or not, the molecular gold-standard method is to determine the DNA similarity between two or more isolates. The method uses the following principle: DNA of a reference and query strains are denatured by heating them up to 95°C. Then temperature is gradually decreased to measure the percentage of re-association (Figure 7). Two ways are available to express DNA-DNA hybridization (DDH) (or pairing, or relatedness): the first consist in measuring the ΔT_m. In molecular biology, the T_m corresponds to the temperature at which 50% of DNA strands are already denatured. It is the thermal stability/denaturation midpoint. Therefore, ΔT_m is the difference between T_m of

homoduplex DNA and T_m of heteroduplex DNA T_m. Homo- and heteroduplex correspond to reference strain DNA and mixture of both reference and tested DNA, respectively. The second way is the Relative Binding Ratio (RBR), which is expressed in percentage of similarity. It is this value that is most often determined. Briefly, it is the relative proportion of DNA heteroduplex in comparison to the DNA homoduplex. The latter is supposed to be 100% (Rosselló-Mora and Amann, 2001). To determine whether two strains belong to the same species, it is recommended to have a DNA-DNA relatedness greater than 70% and/or a ΔT_m lower than 5°C (Wayne et al., 1987). Currently, DDH has been largely replaced by an *in silico* version. For instance, the genome-to-genome distance calculator (GGDC) hosted on the DSMZ website (Auch et al., 2010a, 2010b; Meier-Kolthoff et al., 2014). This new way to access DDH is better, because it is easier to compare a new isolate with a lot of already sequenced species. And thus, avoid positioning errors in the classification while alleviating wetlab work required for *in vivo* studies.

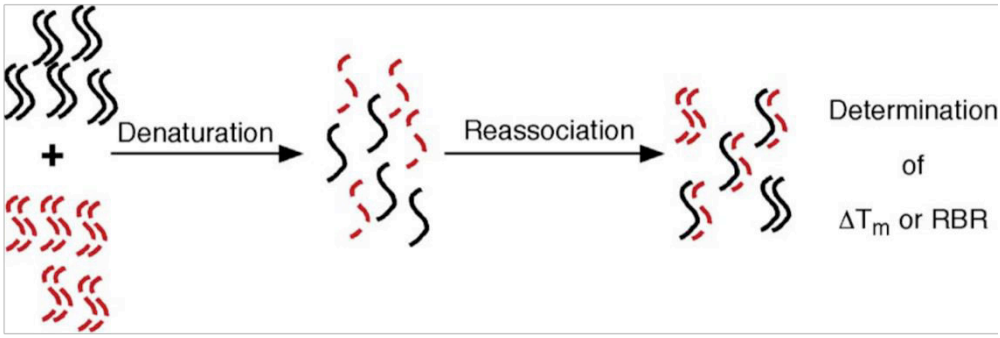


Figure 7: General principle of DNA-DNA hybridization.
Black strands represent DNA of the reference strain. Red strands represent DNA of the tested isolate. In the last step, homoduplex are displayed as two strands of the same color (black or red), while heteroduplex are displayed as a mixture of black and red strands. (From Rosselló-Mora and Amann, 2001).

DDH is a gold standard for delineation of microbial species, but other markers can be used, like the 16S rRNA gene sequence. Firstly, thanks to its ubiquity and the presence of slow and fast evolving regions. Secondly, it is easier and cheaper to obtain the nucleotide sequence of 16S rRNA gene than that of the whole genome. This marker can be a criterion for species delineation, but it should be used in complement of DDH. In general, species having more than 70% DNA relatedness show more than 97% similarity on the 16S rRNA sequence (Stackebrandt and Goebel, 1994). It should be noted that this threshold is still used nowadays for clustering 16S rRNA gene sequences originating from amplicons library next-generation sequencing techniques (NGS). This value has been re-evaluated in 2006. It is now recommended to use the threshold 98.7–99% similarity on the entire 16S rRNA gene sequence for testing the DNA-DNA relatedness (Stackebrandt and Ebers, 2006). Indeed, there is a non-linear relation between 16S rRNA gene sequence similarity and DNA-DNA re-association (Figure 8) (Fox et al., 1992). Authors show that above 70% DNA-DNA similarity, the 16S rRNA is 98.7% or greater, indicating that two strains belong to different species, while the opposite is not true (Figure 8). Above 98.7% 16S rRNA gene sequence similarity between two strains, one cannot conclude if these two strains belong to the same species or not based solely on this marker. DDH is mandatorily to conclude. That is why both values are mandatory for the delineation of new species above 98.7% 16S rRNA gene sequence similarity.

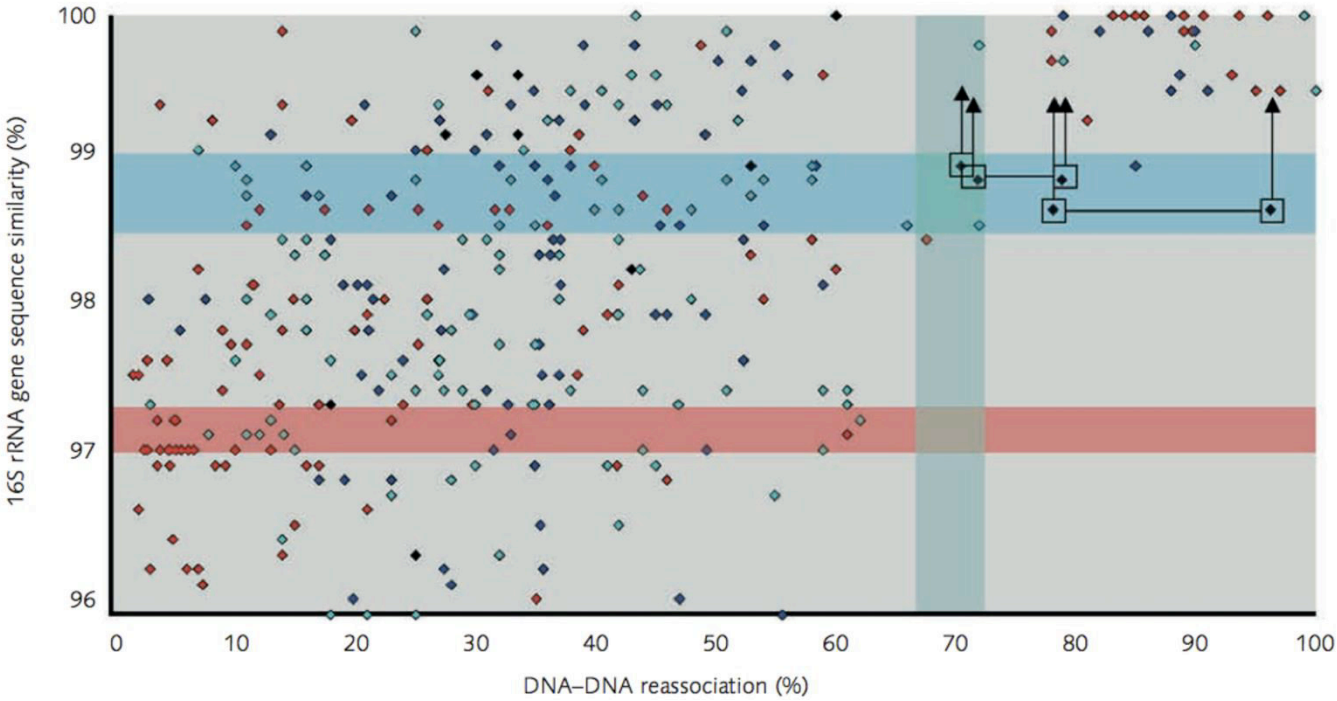


Figure 8: Relation between 16S rRNA gene sequence similarity and DNA-DNA relatedness.
The red horizontal bar indicates the old 97% similarity threshold. The blue horizontal bar relates the new threshold (98.7-99% similarity). And the green vertical bar figure the 70% DNA-DNA reassociation threshold. From Stackebrandt and Ebers, 2006.

- Average Nucleotide Identity

Average Nucleotide Identity (ANI) can be taken into account for defining the microbial species. Even if it is not present in the accepted microbial species definition, this parameter is widely used in characterization of newly isolated strains because it can accurately replace DDH values when genomes of strains are available (Goris et al., 2007; Richter and Rosselló-Móra, 2009). Indeed, on shared genes between strains, an ANI of 94-95% or above correspond to a 70% or greater DDH (Goris et al., 2007; Konstantinidis and Tiedje, 2005). Slightly different values are found when ANI is based on whole genome sequence: 95-96% (Richter and Rosselló-Móra, 2009). The ANI can be used either on shared genes or on whole genome sequence because it is an *in silico* analysis (Auch et al., 2010a; Richter and Rosselló-Móra, 2009).

ii) Phenotype

A new species must share resemblances with its closest relatives and have distinct phenotypic features, like for example carbon sources that the strain uses (Stackebrandt et al., 2002). Table 3 reports examples of phenotypes, determined by cultivating the isolate on lab bench, used to differentiate a new species of the genera *Thermococcus*.

Table 3: Comparative table of phenotypic and genotypic differential characteristics.

Here, *T. piezophilus* was compared to 4 closed species of *Thermococcus*. ND: Not Determined. Adapted from Dalmasso et al., 2016a

Characteristic	*T. piezophilus*	*T. onnurineus*	*T. barosii*	*T. profundus*	*T. coalescens*
Motility	+	+	−	+	+
Temperature (°C; optimum)	60–95 °C (75 °C)	63–90 °C (80 °C)	60–92 °C (82.5 °C)	50–90 °C (80 °C)	57–90 °C (87 °C)
pH (optimum)	5.5–9.0 (pH 6.0)	5.0–9.0 (pH 8.5)	4.0–9.0 (pH 6.5–7.5)	4.5–8.5 (Opt. ND)	5.2–8.7 (pH 6.5)
NaCl (optimum) (%)	2.0–6.0 (3.0)	1.0–5.0 (3.5)	ND	1.0–6.0 (Opt. ND)	1.5–4.5 (2.5)
Growth on:					
Pyruvate	−	−	−	+	−
Maltose	−	−	+	+	−
Starch	−	+	ND	+	−
Yeast extract	+	+	+	+	+
Peptone	+	+	+	+	ND
DNA G+C content (mol %)	51.1	52	54.7	50	53.9

iii) Additional methods

In 2002, the international committee on systematics of prokaryotes proposed new methods to strengthen the integrity of a microbial species. For instance, to sequence a minimum of 5 protein-encoding genes present in the majority of microorganisms (housekeeping genes). The G+C content of DNA can also be taken into account. Other techniques are based on patterns of DNA on electrophoresis gels, like amplified fragment length polymorphism (known as AFLP) or Restriction Fragment Length Polymorphism (RFLP). In the latter technique, DNA samples are cleaved into smaller fragments by template-specific restriction enzymes. It is a kind of enzyme that cut DNA based on a specific template. The frequency of cuts depends on the length of this template. Then, for a sample, restrictions fragments are separated by length on

electrophoresis gel. This gives a profile that can be compared across different samples. And the last technique is DNA array, which is another way to apprehend the similarity of DNA sequences between 2 tested strains (Stackebrandt et al., 2002).

c) The lacking parameters: Ecology

i) Species as a group of ecologically coherent isolates

The aim of the microbial species definition is to provide a frame to order microorganisms for systematics issues. But as it contains thresholds, there may be exceptions. To differentiate species, phenotypes of each isolate are used. But with this criterion, it would be expected that all isolates of one species would have the same ecological niche, or overlapping niches. But it is not the case. For instance, the most studied microbial species, *Escherichia coli*, can be found in the environment, in animal's gut, pathogen or not. Moreover, there are multiple strains that can cause different diseases, like the enterohemorrhagic *E. coli* or the uropathogenic *E. coli* (Konstantinidis and Tiedje, 2005).

Also, it is difficult to assess cell's ecology on the lab bench. Cultivate a microorganism in pure culture under ideal conditions does not reflect its action in the environment. For instance, in the environment, microorganisms can migrate to find better growth conditions. They undergo multiple selection pressures: nutrients sources, predation by microorganisms or viruses, etc. They can interact with other organisms (micro and/or macro) in different ways: in a relationship of symbiosis, or parasitism. Besides, they can incorporate foreign DNA, plasmids, and thus acquire new capacities.

d) Why species definition is still the same?

With all technological advances that have emerged since the last 15 years (NGS, computing resources, databases), we are entitled to expect an update of the microbial species definition released in 2002. But it is not yet formally the case, despite

considerable research in this direction. For Konstantinidis and Tiedje, an update of the thresholds for DDH and/or ANI should better group microorganisms in species, as currently described. But "*it would, however, be impractical to implement because it would instantaneously increase the number of existing species* […]*, and cause considerable confusion in the diagnostic and regulatory (legal) fields*" (Konstantinidis and Tiedje, 2005). Other authors proposed to revisit the actual taxonomy based on the full genome sequence and minimal phenotypic tests (Vandamme and Peeters, 2014). Consequently, each newly characterized strains as to be published with its genome available in public databases, and a hard work has to be done to sequence all type and reference strains without published genome. Other microbiologists share this point of view, but "*it will take time to develop a new coherent species concept. A rush for a new species concept is not needed and would be counterproductive*" (Thompson et al., 2015).

3) What about a new species definition in the genomic era?

a) Currently, definition does not integrate ecology

The current definition does not really encompass the ecology of species, what is the relationship between microorganisms and with their environment. In descriptions of new strain, phenotypes of the isolate are described, but there are no universal rules (Tindall et al., 2010). They can refer to the pathogenicity, the ability to live in water, soil, animal's gut, to play a role in nitrogen fixation, and so on. Here is the issue: how many phenotypes are required? That is why the definition has to move, or to propose a concept that integrates the ecology of strains, in order to understand global mechanisms underlying differentiation processes.

b) How incorporate ecology and genomic data in a more cohesive model?

Even is these two aspects seems to be strongly linked, the ecology do not clearly appear as a major criterion in the definition. For instance, within *E. coli*, there are pathogenic

strains that infect gut, and other that are able colonize the urinary tract. These two strains have a little less than 42% of shared genes (Welch et al., 2002).

A workflow has been proposed in 2014. It consists in sequencing genomes that occur in the same ecological niche, for instance a single rock sample from a black smoker. Then, we have to cluster genomes into genomic units and make hypothesis about their role in the environment, or what distinguish these units (Shapiro and Polz, 2014). These hypotheses can be answered with the predicted functions of genes. For instance, we can find predicted functions that are present only within a genomic cluster, and make an assumption about their role. If it is possible to assess, the information provided by gene flows can help to determine the state of divergence between two or more genomic units. This framework is called "reverse ecology" (Shapiro and Polz, 2014), and has the advantage of being without *a priori* conjectures.

4) Instead of thinking species, think population

a) Definition of a population

With the framework proposed above, it is easier to see all closely related genomes as a population. This population is defined as "*a group of individuals sharing genetic and ecological similarity, and coexisting in a sympatric setting*" (Hunt et al., 2008; Shapiro and Polz, 2014). The sympatric setting refers to microorganisms present in the same place, where barriers to gene flow are low or inexistent. Several mechanisms are involved in the formation of new populations: genetic exchange, for instance homologous recombination (HR), gene frequencies, or mobile genetic elements (Anderson et al., 2017a; Cordero et al., 2012; Shapiro et al., 2012).

b) Genomic units: how to cluster genomes?

i) Sequence similarity

A first and easy way to cluster microorganisms is based on the similarity between sequences, as it is the case with DDH, ANI, 16S rRNA or marker gene sequence similarity

or average amino acid identity (AAI). Here, as in ecological units, the problem is the use of thresholds, because it is not clear where to put borders. The advantage with these criteria is their convenience, but they group strains/sequences within an operational way (Shapiro et al., 2012). Another way to organize genomes is to build phylogenetic trees, which will represent genotypic clusters (Doolittle and Papke, 2006; Thompson et al., 2015). To be strongly supported, phylogenetic tree has to be build using concatenation of multiple markers, a set of genes universally distributed (Petitjean et al., 2015), or even all shared genes, *i.e.* core-genes. Here too, it is not clear where to draw barriers between species.

ii) Genetic units

As all methods based on sequence similarity rely on operational units, other ways have to be used to represent units present in natural ecosystems. Communities of microorganism are not static. They share DNA, they migrate in new niches, they undergo predation by bacteriophages, etc. So clustering genetic units should be done based on migration, mutation rate, recombination and selection (Shapiro et al., 2012). The first, migration, is difficult to assess because speciation arising from geographic isolation is carried out in an allopatric setting. The last two, recombination and selection, are described as two major elements that drive the speciation in natural sympatric population (Shapiro and Polz, 2014).

c) A population is not static

Like all living organisms, genomes of members of a population are dynamic. They can accumulate point mutation, capture free DNA and integrate it in their own genome even if this fragment comes from distant microorganism. This phenomenon is called lateral gene transfer (LGT). In the recipient genome, such transfer can give them new abilities or an allelic variant of a gene. The gene flow, *i.e.* the transfer of genetic information

between microorganisms, is mediated by other mechanisms: transfer mediated by mobile genetic elements, sich as vesicles as in *Thermococcales* (Choi et al., 2015; Soler et al., 2008), or by virus as in the tree domains of life, or by conjugation when a plasmid can be transferred to another cell (Wagner et al., 2017). Gene flow and point mutation can serve as trigger for separation of population and leads to speciation, unless homologous recombination occurs at a very high rate with population because it homogenizes the genotype of population (Shapiro et al., 2012). Thanks to these two levers, two models have been proposed, the first that can leads to separation or collapsing of population, and the second that is better known under the "Ecotype" model.

d) Different models exist

i) Based on recombination/selection ratio

The recombination/selection ratio depends primarily on the number of homologous recombination, which has the role to homogenize loci or even entire genome, across the population. The second is selection, which favor a particular variant and confer an advantage. This advantage is translated by an increase of its fitness, and thus an increase of its frequency in the population. This can lead to the extinction of the other variants. The balance between these two elements defines two speciation models.

The first is the "stable ecotype model", when rate of recombination is well below the rate of selection (Table 4). In this model, an adaptive mutant arises in a population, this can lead to ecological separation in a new niche, or this mutation can purge the diversity by outcompeting its close neighbors in a periodic selection event. The diversity is purged only in the ecotype where the mutant arises, the other ecotypes are not affected by this selection event because they are in independent niches (Cohan and Perry, 2007). This model states that adaptive and neutral loci in the genome cannot be unlinked

through observed rates of homologous recombination because it is too low. So it is a genome-wide selective sweep (Figure 9).

Table 4: Speciation stages under different conditions of selections and homologous recombination

From Shapiro and Polz, 2014

	(A) r/s >>1	(B) r/s << 1
Stage 1	New niche-specifying variant(s) acquired by mutation, homologous recombination, or HGT	
Stage 2	Ecological separation: new variant spreads in new niche by recombination	Ecological separation: new variant spreads in new niche by clonal expansion
Stage 3	Genetic separation driven by genome-wide depression in recombination between new and ancestral niches	Genetic separation driven by periodic selection and drift
Stage 4	Genetic separation maintained by genetic barriers to recombination, including sequence divergence and epistasis; otherwise lineages may merge back together	Genetic separation maintained by further periodic selection and drift events; lineages are permanently separate
Stage 5	Lineages remain ecologically and genetically distinct (at both adaptive and neutral loci, genome-wide) until extinction	

The second case is where homologous recombination is greater than the selection rate. Here, barriers to gene flow are low or non-existent, and the rate of homologous recombination is high enough to avoid incorporation of polymorphism due to genetic drift. This polishing of polymorphism is only valid for adaptive loci. But there could be a mutation in a part of the population that brings it the ability to switch in a new niche (Table 4). This adaptive variant will spread in the new population by gene-sweep, *i.e.* homologous recombination. At this time, if gene flow decreases, it will result in two distinct populations as in figure 10, else the newly divergent groups could merge back together and purge polymorphisms in adaptive loci.

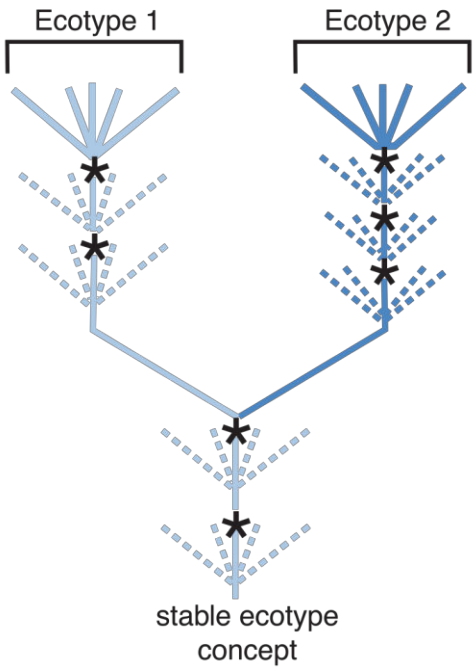


Figure 9: Illustration of the stable ecotype model
This model depicts differentiation of bacterial/archaeal population under the stable ecotype model. Asterisks represent selection events (genome-wide selective sweep) that purge almost all diversity (shown as dashed lines) that has arisen since the last selection event. from Cohan and Perry, 2007; Fraser et al., 2009.

It is important to see these speciation models as dynamic processes, if "*these stages of speciation can be defined, it does not mean that all populations that start at Stage 1 will make it to Stage 5*" (Shapiro and Polz, 2014). Therefore, some lineages can either undergo extinction or new differentiation event at any stage of this process.

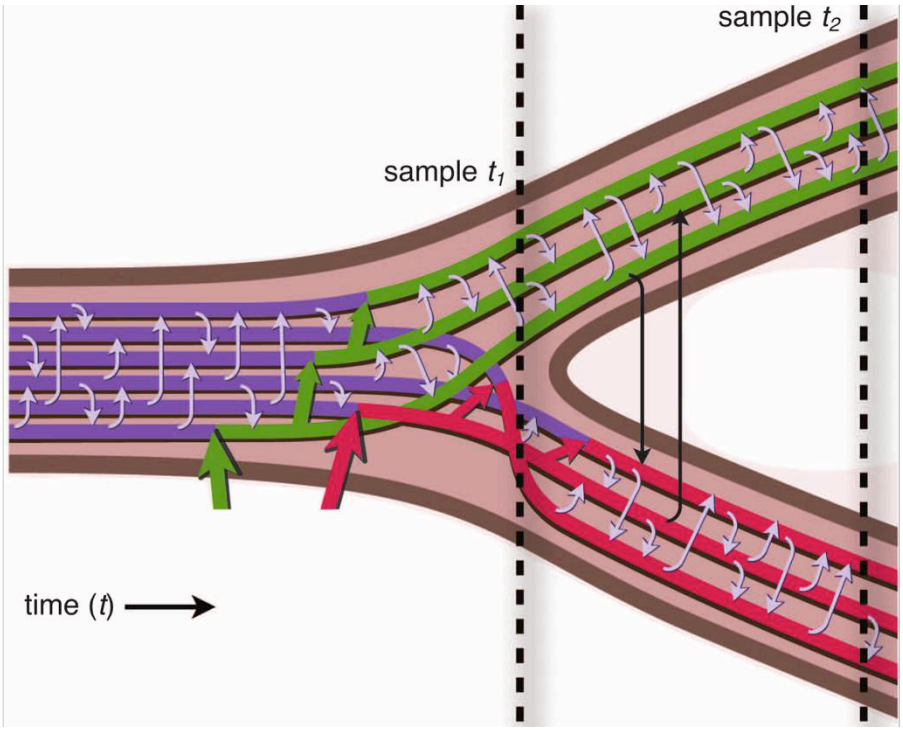


Figure 10: Emergence of 2 populations

Thin purple and black arrows represent genes exchange between members. Big green and red arrows represent the acquisition of an adaptive allele (Adapted from Shapiro et al., 2012)

5) Examples of natural populations

These models of speciation are based on recent observations of natural populations. They were made on marine microbe, because it is probably easier to observe large-scale genetic information exchange and new niches colonization in water than in soil owing to the increases mixing and dispersion in the former environment.

a) *Vibrio*

A study of 20 genomes affiliated to *Vibrio cyclitrophicus* occurring in the same geographical area and sampled in 2006 and 2009, showed that two populations are present and they are at an early stage of the speciation process (Shapiro et al., 2012). Indeed, these authors choose the 20 isolates based on a phenotype: free-living or particle-associated (Hunt et al., 2008). Between these genomes, they identified around 28 000 single-nucleotide polymorphisms (SNP), distributed over all genomes. Other SNPs were identified: 725, localized in 11 regions of which 3 include more than 80% of

these SNPs. They are called “EcoSNPs” because they follow the niche separation of the 20 isolates. This ecological specialization is mediated by the size of the particles to which the cells are attached and this may be the result of small variations in genes of critical pathways, like adhesion to surface, virulence factor or biofilm formation (Shapiro et al., 2012). For the authors, these populations are in an early stage of differentiation (Figure 10, sample t_1), where neutral loci do not allow us to pinpoint the two emerging populations. If the speciation postulate is verified, homologous recombination will be more likely to occur within than between population, and thus reach the complete separation, *i.e.* the formation of 2 distinct genotypes.

b) *Sulfolobus*

Sulfolobus is an archaeal genus belonging to the *Creniarcheota* phylum. *Sulfolobus* species are aerobic acidophiles and thermophiles, mostly found terrestrial hot springs (Madigan et al., 2012). Within a group a 12 isolates affiliated to *Sulfolobus islandicus*, which come from the same hot spring, patterns of gene flow (homologous recombination) are in favor with the presence of well separated populations (Cadillo-Quiroz et al., 2012; Krause et al., 2014). For the authors, this phenomenon supports an ecological differentiation that can prevent competition between the two populations. In addition, with isolates from the same geographic area, it has been shown that recombination shapes the population structure, according to the speciation framework presented above (Whitaker et al., 2005).

IV) Objectives of the thesis

In this work, we aim at better understanding the early stages of diversification at the genomic level. During these 3 years of PhD, the common thread was thus the study of closely related *Thermococcus* isolates genomic diversity, through the lens of comparative genomics studies. Here we expected to observe at least two drivers of diversification: geographical isolation, with which we can assume the presence of migration barriers, and sympatric speciation. For this purpose, a long first part consisted in selecting groups of strains meeting several criteria: number of isolates, geographical origins and group support in marker genes (16S rRNA-ITS) phylogenetic tree.

From these criteria, two groups were selected. They represented 45 genomes, with the first group close to *Thermococcus* sp. 4557 and *T. celericrescens*, and the second group is close to *T. nautili*. Within each group, the aim was to identify functional drivers that group isolates in phylogenetic coherent clusters and thus pinpoint specific genes and functions involved – or resulting – in these diversification processes.

At the end, I have started an exploratory study on the dissemination and distribution of *Thermococcales* cells between deep-sea and shallow hydrothermal vents and also terrestrial hot springs. For this, multiple public metagenomes were mapped on a large dataset of *Thermococcales's* genomes and we observed the percentage of detection for each genome within each metagenome.

Materials and Methods

Chapter II: Materials and Methods

This chapter gathers all the methods used in this work. It aims to provide an overview of all the steps carried out and to provide details about the techniques used. Specific and more concise sections on materials and methods are available in each subsection of the results.

I) The LM2E culture collection: UBOCC

Strains used in this study are stored in an international culture collection. This collection is established within 2 laboratories: the Laboratory of Microbiology of the Extreme Environments (LM2E, Plouzané, France) for the marine sub-collection, and the Laboratory of Biodiversity and Microbial Ecology (LUBEM, Plouzané, France) for the agri-food sub-collection. In total, around 1 300 isolates are present within this collection.

This study focused on *Thermococcales* stored in the marine culture collection. Among the 1300 isolates, about 300 strains belonged to this archaeal order and thus constituted an ideal starting material to address the central questions of this study (Appendix 1).

1) Culture of hyperthermophilic and anaerobic isolates

Given that isolates are anaerobic, they were grown in 50 mL serum bottle (Ref: W012488A, Wheaton®) sealed with a rubber stopper. The dioxygen (O_2) present in the flask and medium was removed by flushing the flask between 5 to 7 times with dinitrogen (N_2). Residual traces of O_2 were removed by adding 1% (v/v) of a 5% (w/v) sodium sulfide solution ($Na_2S.\ 9H_20$). The aim of this step was to get a completely reduced medium. The oxidation state of the medium was visualized with a dye, the resazurin. Few drops of a 1% (w/v) solution are sufficient for 1L of medium. When the medium is totally oxidized (before sterilization notably), resazurin is blue. After

sterilization, if the media was reduced, resazurin was yellow to colorless. Otherwise resazurin was pink due to oxic conditions. To keep the flask anoxic and sterile during sampling or inoculation, we used sterile needles and syringes.

a) Preparation of culture media

Thermococcales isolates grow on rich media, like *Thermococcus* Rich Medium (TRM) or Ravot medium. TRM was composed of (per liter of distilled water): 23 g NaCl; 5 g $MgCl_2.6H_20$; 3.46 g Piperazine-*N,N'*-bis (2-ethanesulfonic acid) (PIPES buffer); 4 g Tryptone; 1 g Yeast extract; 0.7 g KCl; 0.5 g $(NH_4)_2SO_4$; 0.05 g NaBr; 0.01 g $SrCl_2.6H_20$; 0.001 g resazurin. Adjust the pH to 6.8 before sterilization. After sterilization, following solutions were added aseptically: 1mL of 5% (w/v) K_2HPO_4; 1 mL of 5% (w/v) KH_2PO_4; 1 mL of 2%(w/v) $CaCl_2.2H_2O$; 1 mL of 10 mM $Na_2O_4W.2H_20$; 1 mL of 25 mM of $FeCl_3.6H_20$. To finish, the medium was complemented with 0.5 mL of Balch vitamins solution (Balch et al., 1979; Zeng et al., 2009).

Ravot medium was composed of (per liter of distilled water): 1 g NH_4Cl; 0.83 g Sodium acetate trihydrate; 0.2 g $MgCl_2.6H_20$; 3.45 g Piperazine-*N,N'*-bis (2-ethanesulfonic acid) (PIPES buffer); 5 g Tryptone; 5 g Yeast extract; 0.1 g KCl; 0.1 g $CaCl_2.2H_20$; 20 g NaCl; 0.001 g resazurin. Adjust the pH to 7 before sterilization. After sterilization, multiple solutions were added: 5 mL of 7% (w/v) K_2HPO_4, 5 mL of 7% (w/v) KH_2PO_4. To finish, the medium was complemented with 0.5 mL of Balch vitamins solution (Gorlas et al., 2013b).

b) Sterilization steps

Media were sterilized at 121°C for 20 min in a wet atmosphere and under a pressure of 2 bars. For compounds that are thermolabile, like vitamins, they were sterilized by filtration onto 0.22 µm pore-size membranes.

Elemental sulfur (S^0) was sterilized in a specific way, the tyndallization. It consisted of a sterilization cycle in wet atmosphere at 100°C during 30 min. This step was repeated 3 times, 1 time per day.

c) Culture conditions

Uncharacterized putative *Thermococcales* isolates present in the UBOCC grow at 85°C, except 4 isolates that grow at 80°C. In general, isolates belonging to this order grow rather fast, reaching a stationary phase in less than 24 h. In addition to fermentation, these isolates are known to use S^0 as an electron acceptor during respiration. Hence, tyndallized S^0 was added within each vial, at 2 $g.L^{-1}$.

All cultures of microorganisms in this study were performed into 50 mL serum bottle (hereafter named vial). For each vial, 20 mL of complemented medium was added (Figure 11A). Before sealing the flask, S^0 was added under sterile conditions. Then, each vial was flushed with N_2. To finish, the reducing agent was added to remove the last traces of O_2.

All isolates were stored in the culture collection in 1.8 mL Nunc® CryoTubes® at −80°C. To start a culture from a cryovial, the first step was to put isolates back into culture: thaw the sample and transfer 100 to 200 µL aseptically and anaerobically into a vial with 20 mL of reduced medium. Incubate for at least 16 h at the recommended temperature for the isolate. Subcultures were performed by transferring 10% (v/v) of the original culture into a new one, with fresh sterile medium.

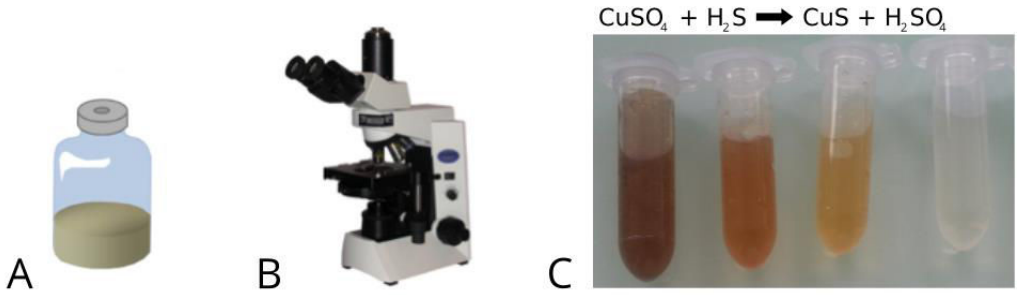


Figure 11: General tools used for *Thermococcales* cultivation
A: a vial used for anaerobic culture; B: Optic microscope for magnificence x400 and x1000; C: Cord –Ruwisch test, illustration of the precipitation of H_2S into CuS. From left to right: Culture producing decreased quantity of H_2S, and negative control (fresh medium).

d) Cells viability

After 16 to 20 h of incubation, each culture was checked under light microscope (Olympus® CX41), at x400 and x1000 magnification (Figure 11B). The cell density was appreciated for each culture. In complement to observation, a Cord-Ruwisch test was performed (Figure 11C). The principle is as follow: all these isolates produce hydrogen sulfide (H_2S) during their growth. The aim was to precipitate this H_2S with a solution of copper sulfate ($CuSO_4$) (Cord-Ruwisch, 1985). To do this, 0.1 µL of the culture was added into 1 mL of the solution of acid $CuSO_4$ (HCl 50 mM + $CuSO_4$ 5 mM). If H_2S was present, a black precipitate of CuS formed immediately. The negative control was performed with 0.1 mL of sterile reduced culture medium.

II) DNA extraction

1) Genomic DNA extraction

DNA of isolates was extracted for two different purposes. The first one was to sequence two taxonomic markers for all *Thermococcales* isolates of UBOCC. These markers are the ribosomal RNA small subunit or 16S ribosomal RNA (16S rRNA), and the Internal Transcribed Spacer (ITS) that is located between the 16S and the 23S rRNA on the genome. The second purpose was the sequencing of the whole genome of 48 isolates that were selected within the UBOCC.

All DNA extractions were performed with the same phenol-chloroform protocol, based on the hydrophobicity of DNA and the amphiphilic nature of proteins. In order to extract a sufficient quantity of DNA for each sample, 2 vials per isolate were used. They were pooled together in 50 mL Falcon® tube before the first step of the protocol. The procedure consisted of the following steps:

- Centrifuge culture at 8 000 rotations per minute (rpm) for 5 min, at 4°C
- Discard supernatant under fume hood
- Homogenize the cell pellet with 1mL of TNE 1X buffer (50 mL Tris-HCl 1 M; 50 mL EDTA 0.5 M pH 8; 10 mL NaCl 5M; q.s.p 500 mL)
- Transfer this solution in a 2 mL microtube
- For lysis, add
    - 100 µL of *N*-Lauroylsarcosine sodium salt (Sarkozyl) 10%
    - 100 µL of Sodium Dodecyl Sulfate (SDS) 10%
    - 50 µL of Proteinase K at 20 $mg.mL^{-1}$ (ref: V3021, Promega®)
- Incubate at 55°C for 1 h in a water bath. Mix slowly from time to time
- Add 20 µL of RNase A at 10 $mg.mL^{-1}$ (Ref: 02101076, MP Biomedicals)
- Incubate 20 to 30 min at 37°C
- Add 1 mL of Phenol-Chloroform-Isoamyl alcohol (PCI) (25:24:1)
- Shake by turning for 45 sec
- Centrifuge at 14 000 rpm for 15 min at 4°C
- Recover the aqueous phase (upper phase)
- If the interface contains a lot of cell fragments, repeat the operation 1 time (1 mL PCI, mix 45 sec, centrifuge, recover supernatant)
- Add 1 mL of chloroform
- Shake by turning for 45 sec
- Centrifuge at 14 000 rpm for 15 min at 4°C
- Recover the aqueous phase (upper phase)
- Add 0.7 volume of frozen isopropanol (−20°C)
- Put the tube at −20°C for at least 1 h. (Possible to leave overnight)
- Centrifuge at 14 000 rpm during 20 min at 4°C
- Discard the supernatant and resuspend the DNA pellet with 0.5 mL of frozen 75% (v/v) ethanol
- Centrifuge at 14 000 rpm during 20 min at 4°C
- Discard the supernatant and dry the DNA pellet at room temperature, tubes returned, or 10 min in SpeedVac (DNA 120, Savant™) at room temperature
- Resuspend the DNA pellet by adding 50 to 100 µL of low EDTA buffer, for example buffer EB (Ref: 19086, Qiagen)

2) Assessing the quality and quantity of DNA

The quality of DNA after extraction is crucial for the downstream analysis. Various ways are available to check the quality and the quantity. The first one is spectrophotometry. This technique scans the sample and measures the absorbance at specifics wavelength directly from the sample. Another technique of spectrometry uses a fluorescent dye that inserts between the DNA strands. The second method is the qPCR quantification. The methods used are described below:

a) With the spectrophotometer NanoDrop®

The device used was the NanoDrop® ND1000 (Figure 12). This first method was rapid and easy to use. The principle was to load 1-2 µL of sample directly on the pedestal and close the sampling arm.

It measured the absorbance of nucleic acids (DNA and RNA), proteins and aromatic organic compounds. Each molecule absorbs at a specific wavelength. Nucleic acids have a strong absorbance at 260 nm. Aromatic amino acids are responsible for the proteins absorbance at a wavelength of 280 nm. And 230 nm is used to evaluate the quantity of organic compounds such as phenol. Moreover, at this wavelength, the peptide bond within proteins absorbs too. The quality evaluation is based on 2 ratios: the A_{260}/A_{280} ratio and the A_{260}/A_{230} ratio. The first ratio determines the contamination of a sample by proteins. A DNA sample is considered pure when the ratio is between 1.8 and 2. Below 1.8, the quantity of contaminant can interfere with downstream analysis. The second ratio indicates the degree of contamination of DNA by organic compounds such as phenol. A sample with a ratio between 2 and 2.2 is considered pure.

The NanoDrop® ND1000 was connected to a computer. Result for each sample was immediately displayed on the screen. Before the first use, the pedestal was loaded with 1-2 µL of sterile MilliQ® water. Then, a blank was made with 1-2 µL of elution buffer.

The samples could then be loaded one after another. Before each measure, wipe carefully the pedestal with non-abrasive wipes. Result for each sample were given in $ng.\mu L^{-1}$. This value should be taken carefully, because this method quantifies both double- and single-stranded DNA (dsDNA - ssDNA), and also RNA.

Figure 12: NanoDrop® ND1000 and Quantus® fluorometers
NanoDrop® ND1000 is presented on the left side, and Quantus® on the right side

b) With a fluorometer Quantus®

This device (Figure 12) has to be used in association with NanoDrop®. The reason is that it can only quantify the targeted molecule (depending on the kit used) within a sample, but do not assess its quality. It is however more accurate because we used a kit designed for dsDNA quantification (Ref. E2670, Promega). This kit contains a fluorescent dye that interleaves within dsDNA, so ssDNA and RNA were not taken into account.

This quantification method depends on a standard curve, prepared with 2 points: 0 $ng.\mu L^{-1}$, and 100 $ng.\mu L^{-1}$, following the manufacturer instructions. The drawback of this method is that the DNA concentration of samples has to fall within this interval. Here, results obtained with the NanoDrop® are a good indicator of the potential need for a dilution.

The protocol proceeds as follow:

- Prepare the buffer TE 1X

- Dilute the dye with buffer at 1:200
- In 500 µL PCR tubes, add 100 µL of TE 1X for each sample and the 2 points of the standard curve
- Add 1 µL of sample or negative or positive control
- Add 199 µL of dye diluted in TE 1X and mix thoroughly
- Incubate 5 min at room temperature, protected from light
- Calibrate the Quantus® with controls
- Measure all samples

c) Illumina® library quantification by qPCR

Illumina sequencing technic allows multiplexing samples on a single flowcell, *i.e.* sequence more than 1 sequencing library at the same time. To avoid biases, each library should be at the same concentration. The manufacturer recommend to use a qPCR protocol, KAPA Library Quantification Kit Illumina® platforms (Kapa Biosystems).

According to the manufacturer protocol, each library was diluted at 1:10 000 and 1:20 000 in triplicate. So, 6 wells of the qPCR were dedicated to quantify 1 library. Then reagents (primers and qPCR Master Mix) were added into each well of the qPCR plate.

The thermal cycler used was a StepOne™ apparatus (Applied Biosystems®). The qPCR cycling program was:

- Initial denaturation: 95°C, 5 min
- 35 cycles of:
  - Denaturation: 95°C, 1 min
  - Annealing / Extension / Data acquisition: 60°C, 45 sec

Given that, qPCR is very sensitive especially to pipetting errors. That is the reason why all was done in triplicate: negative control, standard curve and each dilution of libraries.

3) Preparation of DNA

a) 16S and ITS

For this project, 16S rRNA and ITS markers were sequenced for all isolates using the Sanger technique, this in order to affiliate each isolate within a phylogenetic taxon, *i.e.* find the appropriate genus and identify a set of closely related strains suitable for our comparative genomics study. In a second time, 48 isolates were selected for whole

genome sequencing. For this part, a multiplexed Illumina sequencing strategy was employed to recover all genomes.

i) PCR protocol

16S rRNA gene sequencing is the gold-standard method to assign a taxonomic position to a microbial isolate owing to the large taxonomic databases available. Here we choose to add the ITS in the sequencing to reconstruct a more robust phylogenetic tree. Indeed, ITS can accumulate mutations at a higher rate than 16S rRNA genes because it is a non-coding fragment.

Three primers were employed for the whole process: A4F (TCC GGT TGA TCC TGC CRG) with a Tm of 60°C (Reysenbach et al., 2000a), A1492R (GGC TAC CTT GTT ACG ACT T) with a Tm of 56°C (Teske et al., 2002), and A71R (TCG GYG CCC GAG CCG AGC CAT CC) with a Tm of 62.6°C (Casamayor et al., 2002). The 2 degenerated nucleotides correspond to: “A“ or “G“ (R), and “C“ or “T“ (Y).

Among these primers, A4F and A71R were used for the PCR. The first primer matches at the beginning of the 16S rRNA gene, in forward way, while the second matches the beginning of the 23S rRNA gene in reverse. Thanks to this primer pair, the whole 16S-ITS sequence could be recovered at the same time. The PCR template employed is presented in figure 13.

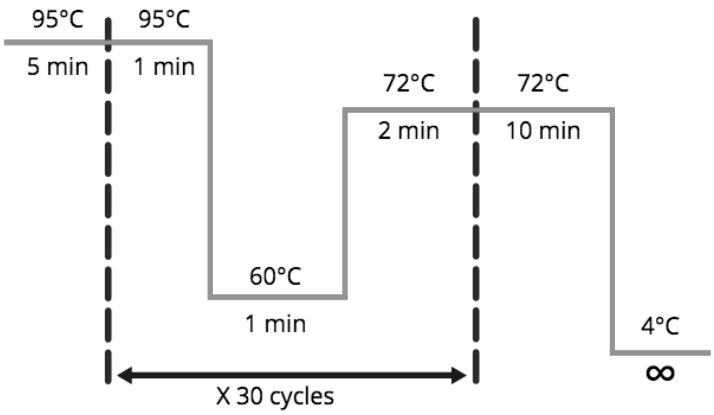


Figure 13: PCR template used to amplified the 16S-ITS DNA sequence

The PCR mixture was prepared using the following final concentrations, for 1 reaction:

- GoTaq® G2 DNA Polymerase (Ref: M7845, Promega): 0.015 $U.\mu L^{-1}$
- Forward and Reverse primer: 0.2 µM
- 5X Green GoTaq® Reaction Buffer (Ref: M791A, Promega): 1X
- Deoxynucleotides triphosphate (dNTP): 0.8 mM
- 1 µL of: DNA sample or negative control or positive control
- Sterile MiliQ® water q.s.p 30 µL

PCR products migrated on a 0.8% (w/v) agarose gel, at 100 V during 45 min. This gel was made with a Tris-Acatate-EDTA (TAE) buffer at concentration 1X. TAE and agarose were boiled and then ethidium bromide was added (5 µL for 100 mL). This molecule has the property of being interleaved between the two strands of DNA, and to fluoresce when exposed to ultra-violet (UV) radiation. For each sample, 5 µL of PCR product was loaded within a well. The ladder used was BenchTop 1 kb DNA ladder (Figure 14) (Ref: G7541, Promega). The expected size of the amplification fragment was around 1800 bp. A transilluminator was used to expose the gel and verify amplification results.

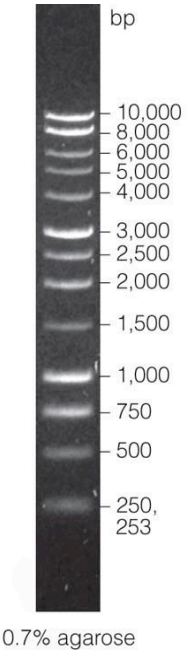


Figure 14: Migration profile of the DNA ladder employed during electrophoresis
This repartition of fragment's size of the BenchTop 1kb DNA ladder loaded within a 0.7% agarose electrophoresis gel.

III) DNA Sequencing

1) Sanger sequencing

Beckman Coulter Genomics (Takeley, UK) sequenced all PCR amplicons, with capillary sequencing machines. Briefly, the technique consists of a classic PCR (one primer, DNA template and dNTP), in which labeled dideoxynucleotides triphosphates (ddNTP) are added. The 4 ddNTP are each labeled with a different fluorescent dye. When the reaction occurs, the polymerase will randomly incorporate a ddNTP (Figure 15), which will result in stopping the reaction. This stop is due to the absence of a hydroxyl group on the 3' extremity of the ddNTP, so the polymerase cannot link the next nucleotide. Then, sample is loaded in long and thin acrylic-fiber capillary instead of classic electrophoresis gel. Under electrical field, fragments migrate and are separated by their length. Shorter fragments migrate quicker. At the end of the capillary, a laser stimulates dyes and a camera records the emitted light. To finish, the result is displayed as an electropherogram, with colored peaks corresponding to the DNA sequence (Figure 15).

The 3 primers cited above were used (A4F, A1492R, A71R). So, for each sample, 3 sequences were generated. We choose this combination of primers to have a sufficiently large overlapping area between sequences. That is, sequences generated from A4F and A1492R have to overlap together. Similarly, primers A1492R and A71R have to overlap.

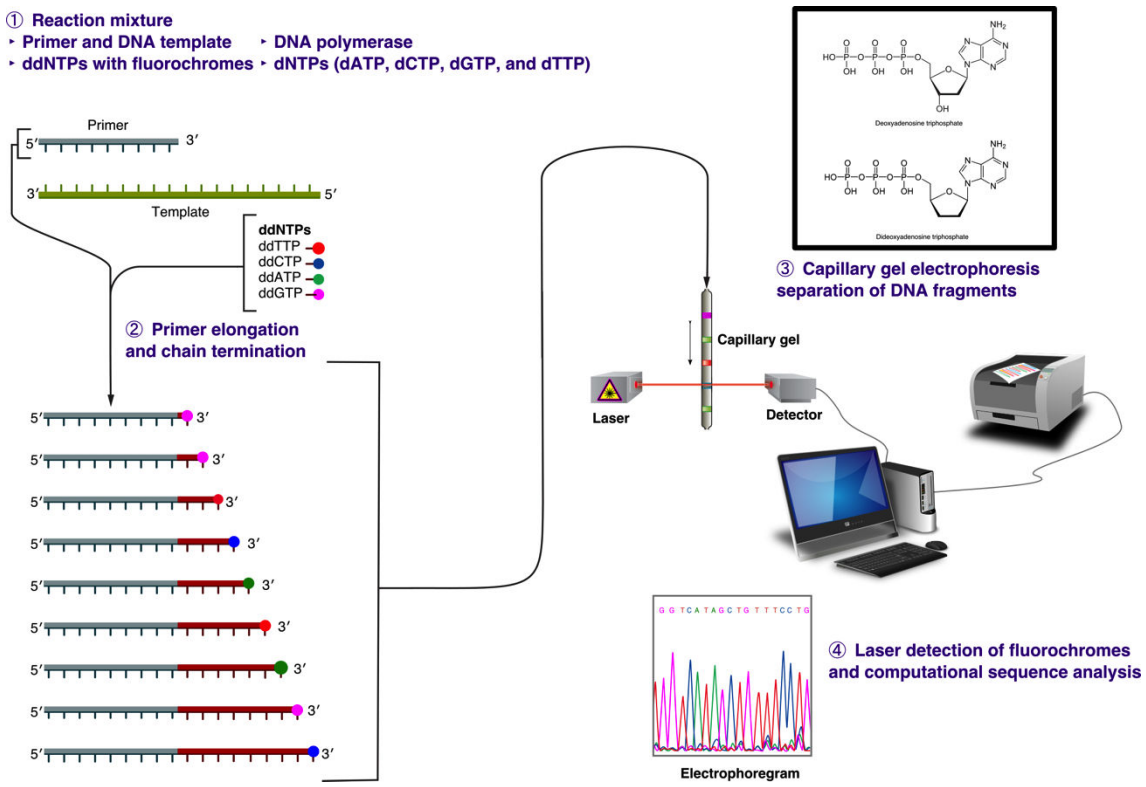


Figure 15: Overview of the Sanger sequencing framework

When the DNA polymerase incorporates a labeled dideoxynucleotide triphosphate (ddNTP), the reaction stops. Then, all fragments are loaded within a capillary gel. This separates fragments according to their length. Then a laser detects each ddNTP and writes the result within a file: the nucleotide detected and its quality score. Adapted from https://commons.wikimedia.org/wiki/User:Estevezj#/media/File:Sanger-sequencing.svg

2) Illumina® sequencing

Illumina is part of the Next-Generation Sequencing (NGS) techniques. The throughput exceeds millions of short sequences per sequencing run. Here, the sequencing of whole genome (WGS) was done on a MiSeq v3 system, in paired-end reads. This consists of sequencing a short fragment of DNA, about 550 bp (also called “insert”), by both ends. And 300 bp are read from each side.

We choose 48 isolates of *Thermococcus* for WGS. They were sequenced in 2 runs, with 24 multiplexed genomes per run. The multiplexing allows sequencing multiple samples in one time. To avoid confusing, each library is flanked with a sequencing tag (or index),

which is specific to one sample. During a single run, the Illumina MiSeq v3 system can produce up to 15 Gbp in paired sequences of 2x300 bp. So for each genome, around 1 million of pairs is sequenced. The length of a *Thermococcus* genome is on average 2 Mbp. We hence targeted a theorical average coverage of about 300X per nucleotide position.

a) Libraries preparation

The preparation of libraries was performed following the manufacturer protocol. The kit used is TruSeq® DNA PCR-Free Library Prep (Ref: 15036187, revision D, Illumina®), with large insert length, *i.e.* 550 bp. For each sample, 2 µg of DNA was required. The first step was to fragment this genomic DNA. The DNA shearing was done with a M220 Focused-ultrasonicator™ (Covaris®) (Figure 16). The sample was loaded in microTUBE AFA Fiber Snap-Cap 6x16 mm (Ref: 520045, Covaris®). The lysis was performed with the following settings: duty factor 20%; Displayed power 50 W; Duration 55 sec; Temperature 20°C.

Then, all steps were performed according to the manufacturer protocol. Briefly, after the shearing, ends of DNA fragments were repaired with both 3'-5' exonuclease and 5'-3' polymerase. Next, DNA fragments that have the right size were selected with small magnetic beads. The next step consists of adding an "A" nucleotide in 3' ends of each fragment, with a ligase. Then adapters were ligated to both sides of each insert. During this step, a specific barcode was allocated to each sample and this will allow multiplexing samples on the same flowcell. Before pooling libraries together, they must go through a validation step. First, the quality was checked with a Bioanalyzer 2100 (Agilent Technologies) (Figure 16). Libraries were loaded on high sensitivity DNA chips (Ref: 5067-4626, Agilent Technologies). In a second time, libraries were quantified by a

qPCR protocol (see below). To finish, libraries were normalized to 2 nM, and they were pooled (5 µL of each normalized library).

Figure 16: Covaris® and Bioanalyzer instruments
On the left side, the Covaris® M220 Focused-ultrasonicator™ used to shear DNA, and on the right, the Bioanalyzer 2100 (Agilent Technologies) used to assess quality of Illimuna® genomic libraries.

b) Sequencing

Half of the libraries were sent for sequencing to the Marine Biological Laboratory (MBL), Woods Hole, MA, USA. The device used was a MiSeq, with the Reagent Kit V3 (Figure 17). It can produce up to 25 M of reads. For the second set of 24 genomes, genomic DNA was sent and libraries were prepared and sequenced directly at the MBL, using the same protocol as above.

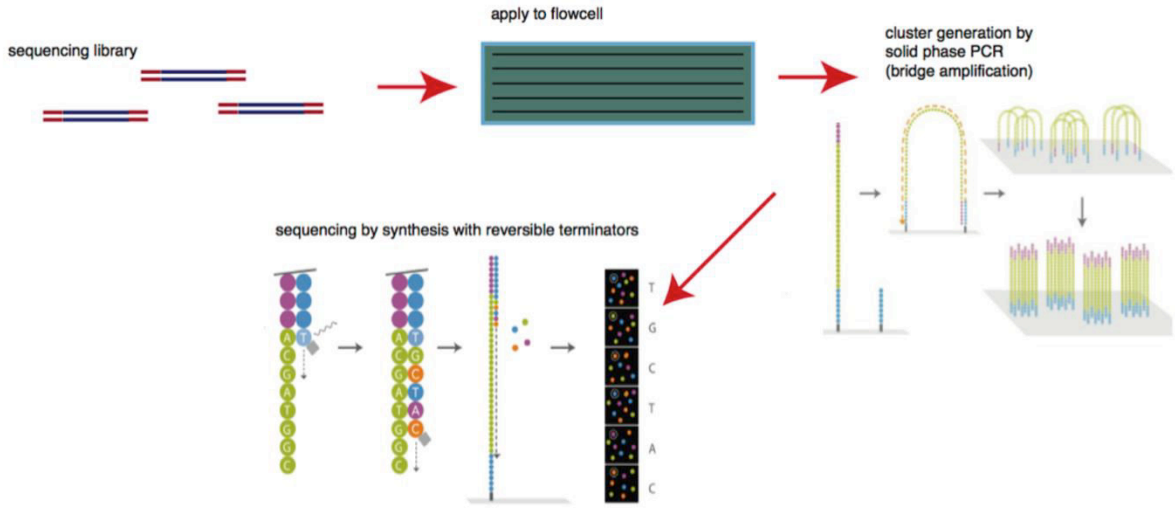


Figure 17: General principle of Illumina sequencing
Adapted from https://bitesizebio.com/13546/sequencing-by-synthesis-explaining-the-illumina-sequencing-technology/

IV) Sequence assembly

In bioinformatics, sequence assembly is the step where reads are merged together in order to reconstruct a longer sequence.

1) Workflow to recover full 16S-ITS sequences

a) 16S-ITS quality check

Before assembling, it is mandatory to check the sequences quality. For the Sanger sequencing technique, Beckman Coulter Genomics provide several reports, in particular whether the sequence passed the sequencer quality filter or not. Then, each sequence came with a file containing the nucleotide sequence and the quality score associated to each position. The sequencer calculates this score following the PHRED format and recommendations (Ewing and Green, 1998; Ewing et al., 1998). Most of the time, the quality of the first 20 to 30 nucleotides of reads is below the threshold of Q30 (99.9% of chance that a nucleotide is assigned correctly). This was reflected in the presence of abnormal peaks (Figure 18), so they were trimmed. The same thing was observed for the end of each sequence. Around position 700 to 800, the quality began to decrease, and it is increasingly difficult to discern the peaks on the electropherogram (Figure 18).

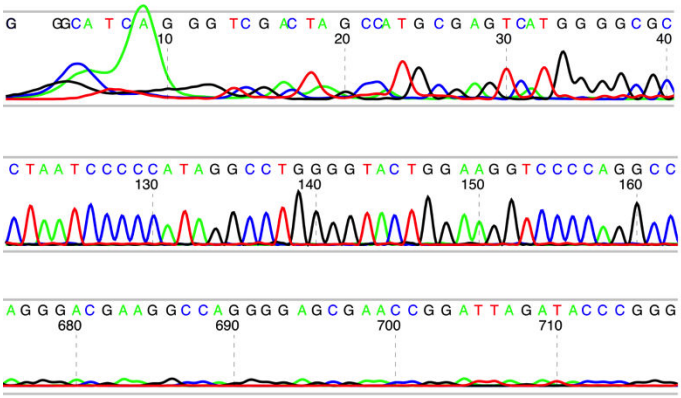


Figure 18: Example of an electropherogram

From top to bottom: Beginning, middle and end of the electropherogram. The beginning is chaotic and stabilizes at position 20-30. Then peaks are clear. At the end, it is no longer possible to distinguish between peaks.

b) 16S-ITS assembly

Here, for each *Thermococcales* isolate, 3 reads were merged. Assembly step was done with in-house Perl and BASH scripts. First, sequences were trimmed at both ends: from position 1 to 25, and from position 800 to the end. These thresholds were chosen because they were suitable for the vast majority of sequences. Next, sequences arising from primers A1492R and A71R were reversed and complemented with *revseq* (package EMBOSS v6.5.7.0), because they were in reverse direction. Then, sequences were aligned with each other using MUSCLE v3.8.31 (Edgar, 2004). Some public 16S-ITS sequences of *Thermococcales* were added in the alignment to guide MUSCLE. After that, overlapping areas were analyzed one by one in order resolve conflicts. When a conflict was found, we referred to the electropherogram, and choose between the two possibilities. The software 4Peaks v1.8 (http://nucleobytes.com/4peaks/) was used to visualize electropherograms.

2) Whole genome

The way to reconstruct a whole genome sequenced with short read differs from assembling Sanger reads. The method used is based on the resolution of de Bruijn graphs (Compeau et al., 2011). The idea is as follow: short reads are decomposed into small pieces of length *k*, or *k*-mer. Usually, the value of *k* is odd and is comprised between 21 and 63 within CLC Genomics Workbench. Then these pieces are aligned. The prefix (length *k*-1) of the suffix of the k-mer *n* has to match exactly with the suffix (length *k*-1) of the k-mer *n+1*. Raw reads were sent by the sequencing center. The only processing they did was the de-multiplexing step, which consisted of sorting reads according to the samples and removed traces of Illumina primer and adapters. Each sample came with 2 files: the first containing the read1 of a pair and the second containing the read2 of the same pair, in FASTQ file format, sorted in the same order.

a) Apply quality filter

Like all sequencing methods, Illumina technologies generate some errors during sequencing, at a rate of around 0.1% (Glenn, 2011). In addition, the quality of reads decreases as its length increases (Figure 19). So, in order to filter sequencing errors out, empirical algorithm has been released to discard low quality reads (Minoche et al., 2011). This algorithm was transcribed in Python, a very common programming language in bioinformatics. This tool, illumina-utils v1.4.2 (Eren et al., 2013), was employed with the command *iu-filter-quality-minoche* and default parameters.

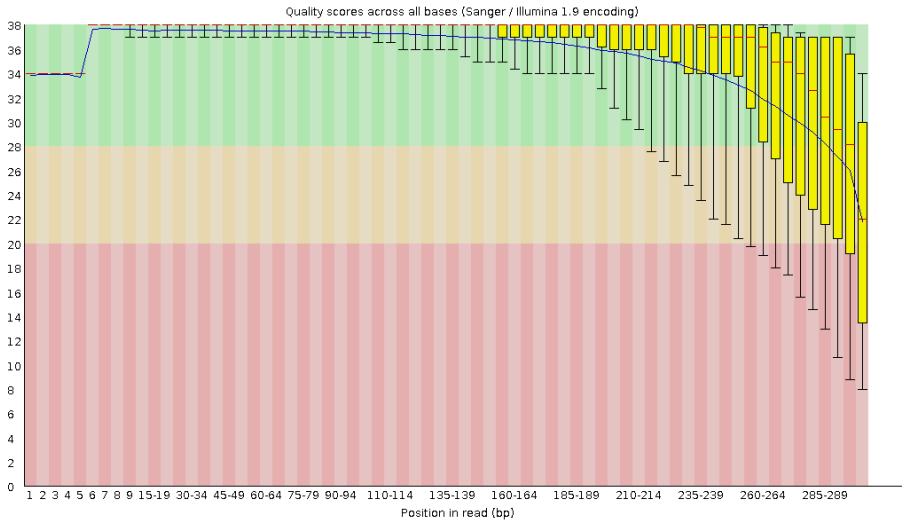


Figure 19: Example of Illumina® read quality

This figure shows the average quality of each read1 from the genome of *Thermococcus* sp. AMTc19 after quality control. Obtained with FastQC v0.11.5 and default parameters (http://www.bioinformatics.babraham.ac.uk/projects/fastqc/)

b) Assemblies of reads

Next, each genome assembly was done with CLC Genomics Workbench v8.5.1 (Qiagen®)(CLC). Default parameters were used for the first run, *i.e.*: *k*-mer set to 21. Furthermore, CLC mapped the reads back on the assembly. Thanks to this, the software calculated the coverage value for every position. If the assembly produced multiple contigs, the "Join Contigs" tool of the Genome Finishing Module of CLC was used. It

consists of linking ends of 2 contigs using linkage information from paired-end reads and/or BLAST. Sometimes, this method was not conclusive, so other assemblies were undertaken, using different values of *k*-mer, from 23 to 63 (limit of the software), until reaching the smallest number of final contigs.

V) Phylogenetic tree

A phylogenetic tree is a way to represent evolutionary relationships between individuals. Since the advent of sequencing technics, phylogenetic are built using molecular markers, both DNA and proteins.

1) Assign taxonomy through 16S-ITS phylogenetic tree

From the UBOCC, 273 16S rRNA – ITS sequences of *Thermococcales* were produced. A way to assign taxonomy to each sequence is to build a phylogenetic tree. First, 16S-ITS sequences of fully sequenced *Thermococcales* genomes were downloaded from RefSeq database (NCBI), either 20 sequences when the tree was built.

The alignment of sequences was done first with MUSCLE, but this was not conclusive. We choose to switch to SINA, an aligner designed for 16S rRNA gene sequences, which use the 16S rRNA gene variable and conserved regions (Pruesse et al., 2012). But this tool does not take into account the ITS. So, sequences were cut at the last nucleotide of 16S rRNA gene sequence and thes sequences were aligned with SINA v1.2.11, with the parameter "profile" set to *Archaea*. All ITS were aligned with MUSCLE v3.8.31, default parameters. Then both parts of each sequence were concatenated.

The next step was the building of the tree. The method employed was Bayesian phylogeny, with MrBayes v3.2.6 (Ronquist and Huelsenbeck, 2003). The evolutionary model was HKY (*nst=2*) (Hasegawa et al., 1985), with an invariable gamma rate (*rates=invgamma*). The run was composed of 60 000 000 occurrences (*ngen=60000000*) that were sampled every 2000 times (*samplefreq=2000*). The first 25% of occurrences

were discarded, during the burn-in phase. Then, the tree was visualized with ARB (Ludwig et al., 2004) and FigTree v1.4.2 (http://tree.bio.ed.ac.uk/software/figtree).

2) Second way to assign taxonomy

Taxonomy of each isolate was assigned with another method that is more convenient: the RDP classifier v2.11, with default parameters (Wang et al., 2007). This tool works by comparing the 16S rRNA gene sequence to a database and return the result with confident index for each taxonomic rank.

3) Phylogenomic tree from rich set of genes

Phylogenomics is the junction of phylogeny and genomics. One aim is to build phylogeny with genomics data, like whole genome DNA or different sets of concatenated genes. Throughout this study, we used Single Copy core-Genes (SCGs). That is to say, for a given set of genomes, all single-copy genes shared by all genomes.

a) Set of genomes

In this work, the dataset was composed by 114 *Thermococcales* genomes. Among them, 40 came from public databases, 46 were sequenced during this project. The laboratory MBGE (Pasteur Institute, Paris, France) sequenced 27 additional genomes. The last genome belongs to a new *Thermococcus* strain isolated in the laboratory, *Thermococcus* sp. MF15 (Le Guellec et al., *In Prep*).

b) Single copy core-genes

The set of single copy genes for the 114 genomes was extracted thanks to the anvi'o visualization tool (Eren et al., 2015) and its pangenomics workflow. The latter is available online (http://merenlab.org/2016/11/08/pangenomics-v2/). Briefly, the pipeline was as follow: Prodigal (Hyatt et al., 2010) found the Coding DNA Sequences (CDS) for each genome. Each CDS was annotated thanks to the Cluster of Orthologous Gene (COG) database (Galperin et al., 2014). Then, anvi'o ran a BLASTP of all proteins

against themselves in order to create a similarity graph (Altschul et al., 1990). This graph was then parsed using the MCL algorithm (Van Dongen, 2000; Van Dongen and Abreu-Goodger, 2012), with inflation (*--mcl-inflation*) parameter set to 6 for the global pangenome, in order to create protein clusters (PC). These latter could be regarded as clusters of orthologous genes, depite not being true COGs. If a PC contained at least 1 protein for each genome, it thus belonged to the core genome. In the particular case where a core genome PC contained a single protein from each genome, this PC was considered part of the Single-copy Core-Gene collection (SCGs)

c) Phylogenomic tree

Phylogenomic tree was built with the concatenation of identified SCGs. First, protein sequences of SCGs were extracted and sorted by PC, one file per PC. So each files contained 114 sequences, 1 for each genome. Then, proteins sequences were aligned with MAFFT v7.055b, in accurate mode (*mafft-linsi*) (Katoh and Standley, 2013). The trimming of sequences was performed with BMGE v1.12, with default parameters (Criscuolo and Gribaldo, 2010). The 2 previous steps were carried out for all files independently. To finish the preparation of data, proteins sequences were concatenated with an in-house Perl script. The resulting file was then visually inspected; invariable positions were discarded and the result was saved in PHYLIP format (Felsenstein, 1981) with SeaView 4.6.1 (Gouy et al., 2010).

The phylogeny was built with PhyML v3.20120412 (Guindon and Gascuel, 2003; Guindon et al., 2010) on a dedicated webserver (http://www.atgc-montpellier.fr/phyml/). The selection of model of evolution was done with SMS, a plug-in of PhyML that tested different models and selected the one that best fit the dataset (Lefort et al., 2017). The parameters selected were: Akaike Information Criterion as selection criterion; BioNJ as starting tree (Gascuel, 1997); NNI as tree improvement;

aLRT SH-like for fast likelihood-based branch support calculation method (Anisimova and Gascuel, 2006).

The resulting tree was visualized with iToL, a web-based platform (Letunic and Bork, 2016).

VI) Species definition

During this work, two groups were chosen for studying the inter-specific genomic diversity of archaeal populations. This implied to highlight species present in our dataset of 114 genomes.

1) Average Nucleotide Identity

A wildly used method to perform this task relies on the ANI between pairs of genomes. This implies that close relative microorganisms have less divergent ANI values than two more distant microorganisms. A value of ANI greater than 94-96% implies that both microorganisms belong to the same species. This value was computed for all pairs within the dataset of 105 genomes, either 11 025 pairs. OrthoANI v1.2 was used to compute all pairs (Lee et al., 2015). This software employed the BLAST-based method to calculate ANI instead of the MUMmer-based method.

All results were compiled in a square matrix. The heatmap was built with R v3.2.2 (R Core Team, 2015) and the package *gplots* v3.0.1 (function *heatmap.2*), using RStudio (https://www.rstudio.com/).

2) *In silico* DNA-DNA hybridization

The *in silico* DNA-DNA Hybridization (DDH) is another way to define microbial species. The value of 70% is accepted as the threshold at which 2 microorganisms are defined as belonging to the same species. In general, this threshold corresponds to the ANI 94-96%. But two combined metrics are more accurate than one.

In silico DDH was computed with the Genome to Genome Distance Calculator (GGDC) v2.1 (Auch et al., 2010a, 2010b, Meier-Kolthoff et al., 2013, 2014). The web service was used to achieve this step, with default parameters (http://ggdc.dsmz.de/ggdc.php).

VII) Searching specific PC in both groups of close genomes

In order to find genomic markers for closely related clades, we defined specific proteins clusters (SPC): these are PC that belongs only to a monophyletic subset of genomes. SPCs were indentified by carrying a new pangenomics analysis for both groups of isolates. The difference with the pipeline applied previously for pangenomics was the parameter --*mcl-inflation*, set to 8.

1) COG and KEGG annotations

We annotated each gene by comparison to the COG database using anvi'o (*anvi-run-ncbi-cogs*). Then, for a given sub-group, all genes present in SPCs were annotated with the KEGG Automatic Annotation Server (KAAS) (Moriya et al., 2007). The aim of this annotation was to identify metabolic reactions only present within a cluster of genomes. KAAS was used with the Best BLAST Hit (BBH) method. The parameter "*gene"* was modified to compare all SPCs to genes belonging to the following list of organisms: *hsa, dme, ath, sce, pfa, eco, sty, hin, pae, nme, hpy, rpr, mlo, bsu, sau, lla, spn, cac, mge, mtu, ctr, bbu, syn, aae, mja, afu, pho, ape, ton, tko, tga, tsi, tba, pab, pfu*.

After computation, identifiers (*ko ids)* were available for each gene, unless KAAS did not found an annotation.

Then each SPC with a *ko id* was compared to all genomes from the same pangenome, at different levels. First, we searched for the presence of the same *ko id* within all PC from the pangenome. Then COG annotations were used in the same way as the *ko ids*: we searched for SPCs that have a unique COG annotation through the entire pangenome.

VIII) Metapangenomics

Metapangenomics is a strategy to study pangenomes conjunction with metagenomes to simultaneously identify gene clusters across closely related genomes and characterize their occurrence in the environment through metagenomic data. During this step, we reviewed the bibliography for the presence of deep-sea hydrothermal vents metagenomes. The aim of this part was to estimate the distributions of *Thermococcales* genomes in this kind of extreme environment.

1) Recovering metagenomes

Metagenomes used in this work were retrieved from the database Sequences Reads Archive (SRA), hosted by the NCBI. The whole database was browsed with keywords such as “hydrothermal”, “hydrothermal vent” or “deep-sea metagenome”. In total, 354 metagenomes sequenced by Illumina technologies were recovered (Appendix 2). Most of them came from deep-sea hydrothermal environments. Eighteen were hot-spring metagenomes, from Yellowstone national park, South Africa or Taiwan.

2) Processing of metagenomes

a) Download

An in-house BASH script was designed to automatize the whole process. First, metagenomes were downloaded in sra file format. This format had to be converted before use. This was done with the SRA toolkit v2.8.2. The command *prefetch* downloaded the file from the NCBI SRA database (with the parameter *–v <SRA-id>*). Then, the file was decompressed in FASTQ format, with the command *fastq-dump*. The argument *--split-files* allowed to split the dataset in two files containing read 1 and read 2 respectively.

b) Quality

On each pair of file, the Minoche quality filter was applied using illumina-utils, as for genome assemblies (IV.2.a).

c) Mapping of reads to genomes

After the quality filtering, reads were mapped on genomes, with Bowtie v2.3.1 (Langmead and Salzberg, 2012). Before running bowtie, all data on which we wanted to recruit reads were grouped within a single file. Then the command *bowtie2-build* was applied to this file. This generated the index, a list of files used by bowtie to map reads. Then, each pair of metagenome reads was mapped on the target genomes. It produced an alignment file in *sam* format (Sequence Alignment/Map). For downstream analysis, it was preferable to have a sorted and indexed alignment file. This was done with SAMtools, v1.3.1 (Li et al., 2009). First, the *sam* alignment file was converted in bam file (Binary Alignment/Map) with *samtools view*. Then it was sorted with the command *samtools sort* and indexed with *samtools index*. At the end, we have one sorted and indexed file for each metagenome. This file was then processed with anvi'o, following profiling and merging steps presented in the following workflow described online (http://merenlab.org/2016/06/22/anvio-tutorial-v2/).

First with *anvi-profile* we synthetized the information for each mapping result on each genome. That is to say, for a given position, it recovered which read mapped and which nucleotide. This was done for the 354 metagenomes separately. Then, all profiles were merged with *anvi-merge* function. This allowed the visualization of all metagenomes mapped on the same genomes simultaneously.

3) Visualization of data

For all metagenomes mapped on genomes, anvi'o calculated multiple metrics. The one that interested us was the "detection". It is the ratio of position on a given genome that

was mapped divided by the total length of the genome. For example, 1% of detection on a 2 Mbp genome corresponds to 20 000 distinct positions of this genome were found within the metagenome.

Results

Chapter III: Results

I) Screening culture collection for *Thermococcus*

1) Abstract

Investigating genomic markers of early diversification requires a collection of closely related microorganisms in order to highlight clade-specific genes and pathways. The LM2E owns a collection of marine microorganisms, sampled during multiple oceanographic cruises. Among the 1300 isolates available in this collection, 305 are classified as *Thermococcales* based on phenotypes features and microscopic observations. In this study, we built a phylogenetic tree based on 16S rRNA genes and 16S-23S spacer (=ITS). Two groups of isolates were then selected based on their multiple geographic origins and phylogenetic placement. In total, this represents 48 isolates suitable for genome sequencing in order to investigate genomic diversity between closely related *Thermococcus* isolates.

2) Introduction

The genomic diversity of microorganisms is nearly infinite. *Bacteria* and *Archaea* are present in nearly all environments on Earth, including the deep-sea hydrothermal vents, and are adapted to these habitats. These adaptations imply a large gene catalogue, spread in both *Bacteria* and *Archaea*. Indeed, pangenomics studies show a continuously increasing number of new genes when the number of compared genome increases, at the genus taxonomic level (Tettelin et al., 2005, 2008). This study will focus on the identification of closely related microorganisms from an extreme environment, the deep-sea hydrothermal vents (DSHV), with the perspective to study genes and pathways associated with early differentiation and / or speciation stages.

DSHV are complex and dynamic structures from geochemical and biological aspects (Wang et al., 2009). They are formed by precipitation of minerals and metals contained

in a hot and reduced fluid, which results from leaching of rocks from the oceanic crust. DSHV are mainly present in two types of geochemical contexts: ultramafic and basaltic (Wetzel and Shock, 2000). The first is characterized by high concentration of H_2 and CH_4 and low H_2S, while it is the opposite for basaltic DSHV (McCollom, 2007). This particular environment, hot and reduced fluid opposed to cold and oxygenated seawater, offers many ecological niches due to the presence of sharp physico-chemical gradients. Microbial communities associated to DSVH are composed of chemolithoautotrophs that act as primary producers and allow the presence of a wide diversity of bacterial and archaeal heterotrophs (Jannasch and Wirsen, 1979; Roussel et al., 2011). DSHV are widespread in oceans, but microbial communities associated with them show high degree of similarities (Anderson et al., 2015). Even if it is possible for *Thermococcales* to migrate between sites and colonize newly formed chimneys (Pagé et al., 2008), they arrive to a new site by chance, possibly following currents (Wirth, 2017). In the case of geographic distant isolates, this should have an impact on the gene flow, with apparition of dispersal barriers. The result is the separation and diversification of populations (Shapiro et al., 2012).

At the LM2E, there are continuous efforts to build a culture collection that can be used to address fundamental questions that we are interested in going after in this dissertation. The UBOCC culture collection comprises hundreds of isolates of marine microorganisms, originating from different places around the world and sampled during different oceanographic cruises. From this culture collection, around 300 isolates are classified as *Thermococcales*, a hyperthermophilic *Archaea* living primarily within hot marine ecosystems like deep-sea hydrothermal vents (Zillig et al., 1983, 1987). The order of *Thermococcales* comprises three genera: *Pyrococcus*, *Palaeococcus* and *Themococcus* (Fiala and Stetter, 1986; Takai et al., 2000; Zillig et al., 1983). To date,

45 characterized species are documented within the *Thermococcales*, of which 34 belong to *Thermococcus*. In this study, we will focus on *Thermococcus*, because they are well studied, easy to grow and isolate. Moreover, genetic tools are available (Thiel et al., 2014), they harbor mobile genetic elements (Cossu et al., 2017; Gaudin et al., 2014; Gorlas et al., 2012), and they are more abundant than the two other genera (Huber et al., 2006; Lepage et al., 2004).

Here we report the framework for taxonomic assignment of *Thermococcales* isolates present in the UBOCC culture collection. The strategy employed was to sequence the 16S rRNA gene marker and also the 16S-23S rRNA internal transcribed spacer (ITS). The 16S rRNA gene allowed the taxonomic affiliation of each isolate, while the ITS was used to discriminate between closely related isolates (García-Martínez et al., 2002; Huber et al., 2006). Indeed, thanks to its broad sequence and length polymorphism, the ITS locus is widely employed in marine bacterial and archaeal population studies (García-Martínez and Rodríguez-Valera, 2000; García-Martínez et al., 2002). With the sequencing of the ITS, we expect to increase phylogenetic tree robustness. The purpose of this tree is to act as a support for selecting groups of isolates for genome sequencing to inspect the genomic diversity of closely related *Thermococcus* isolates.

3) Results

a) Origins of isolates

All strains used in this work were stored at -80°C in the culture collection hosted in the laboratory. Among thousands of isolates, 305 strains were predicted to belong to the order of *Thermococcales* based on microscopic observations and growth culture condition. We cultivated each isolates, extracted the DNA and we sequenced their 16S rRNA gene and ITS sequences. Among these, 95 sequences of 16S rRNA gene were already available in the lab. On the 305 isolates available, 32 sequences of 16S-ITS could

not be obtained, due to a failure at the cultivation step, or presence of contamination in the culture (multiple peaks on the electropherograms). Therefore, we have retained 273 full sequences of 16S-ITS (Appendix 1). These isolates originated respectively from: the Pacific Ocean, with 77 isolates from the East Pacific Rise (EPR) 13°N and 121 from the EPR 9°N; the Atlantic Ocean, with 40 isolates from the hydrothermal field Rainbow and 8 from the hydrothermal site Menez-Gwen; 12 isolates from the Indian Ocean; 4 from the Antarctic island Saint-Paul (South Indian Ocean); and 11 without geographic origin (Appendix 1). After the 16S rRNA gene sequences taxonomic assignment, 14 isolates were assigned to the order *Pyrococcus*, and the other 259 isolates were all assigned to the *Thermococcus* order (Appendix 1).

In order to verify this taxonomic assignment and choose isolates for genome sequencing, we built a phylogenetic tree. As specified above, sequences used were: the 16S rRNA gene sequences and the ITS sequences of the isolates sequenced during this study, and also the 16S rRNA gene and ITS sequences available in the public databases, *i.e.* from sequenced genomes. Both alignments were merged and the figure 20 shows the resulting tree.

On this phylogenetic tree, 2 groups of closely related clades were chosen for whole genome sequencing, based on the following criteria: (i) geographic origin of the isolates. We decided to select in priority groups composed of isolates from multiple geographical origins (hydrothermal fields) in order to address the question of impact of dispersion on the gene flow between sites. (ii) There must be several isolates for each geographical origin. The group should appear monophyletic, meaning with a putative common ancestor to all isolates. Ideally, a group should be composed approximately of 24 genomes (for a total of 48 genomes sequenced). We also considered the presence of potential mobile genetic elements (plasmids, viruses) by targeting isolates close to

known strains with plasmids such as *T. nautili*, because these elements can act as drivers of evolution

The first group included 21 isolates from the Pacific Ocean – namely the EPR 13°N –, the Atlantic Ocean – namely the Rainbow and Menez Gwen sites – and the South Indian Ocean – namely the Saint-Paul island – (Figure 20; Figure 21). This group was close to *Thermococcus* sp. 4557. All isolates come from deep-sea hydrothermal vents, except the 4 that originate from Saint-Paul, which were sampled from a shallow hot spring.

The second group was composed of 27 genomes (Figure 20, Red isolates). Here, isolates come from the 2 sites in the Pacific Ocean, the EPR 9°N and the EPR 13°N (Figure 21). This group was also chosen because isolates are closely related to *T. nautili*. Indeed, this strain harbors 3 plasmids that can potentially serve as a vector of genetic information between isolates (Oberto et al., 2014).

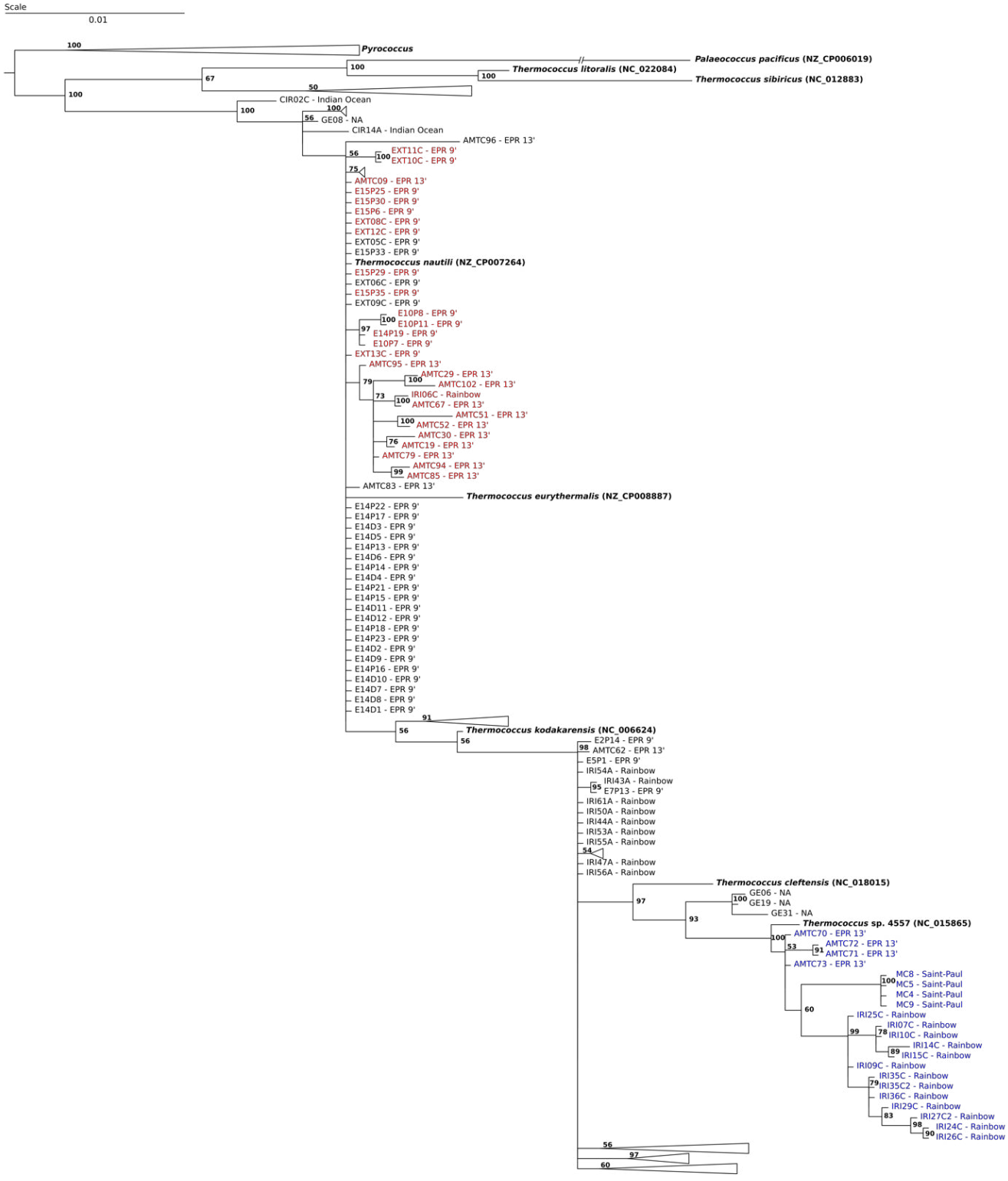


Figure 20: Phylogenetic tree based on 16S rRNA gene-ITS sequences of UBOCC isolates

This phylogenetic tree built by bayesian inference shows a summary of relationships between *Thermococcales* isolates present in UBOCC and genomes of *Thermococcales* available in public databases (Bold leaves). Red and Blue leaves represent isolates chosen for genome sequencing. Bootstrap values greater than 50 are shown next to nodes. Bar: 1 substitution per 100 nucleotides.

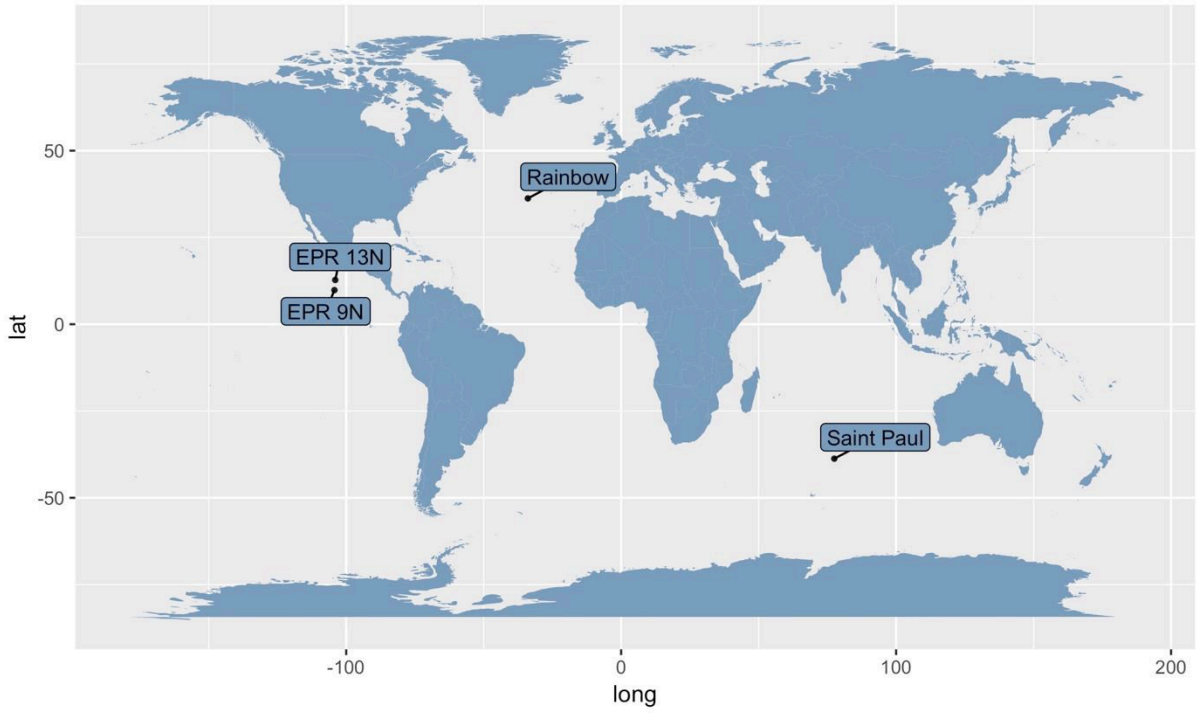


Figure 21: Geographic origins of isolates selected for whole genome sequencing
Isolates of the group I come from the East Pacific Rise (EPR) 13°N, the Rainbow deep-sea vent site and the Saint-Paul hot spring. Isolates of the group II come the EPR 9°N and the EPR 13°N (R Core Team, 2015; Wickham, 2009).

4) Discussion

In this study, our purpose was to select groups of closely related isolates for later investigation of their genomic diversity. We thus choose to sequence 16S rRNA gene sequences and 16S-23 rRNA gene spacers (ITS) to place nearly 300 strains in a phylogenetic context. The ITS spacer is not under a selection pressure because it is a non-coding sequence and usually exhibits a higher variability and can improve phylogenetic resolution. Moreover, this ITS shows patterns of biogeography between isolates of *Thermococcus* sampled in the Pacific Ocean (Huber et al., 2006). Other studies also showed these biogeographic patterns in *Thermococcus* (Lepage et al., 2004; Price et al., 2015).

From the collection of 273 sequenced isolates, 2 groups of closely related isolates were selected. Group I validated all our selection criteria. Indeed, the 21 isolates were from the Atlantic, the Pacific and the South Indian Oceans. Group II validated the two first criteria, which are multiple isolates from multiple origins. Only 12 isolates were monophyletic at the 16S rRNA gene-ITS level. But the 27 isolates composing group II were all close to *Thermococcus nautili*, the isolate cited above which harbors 3 plasmids (Gorlas et al., 2014). One of those plasmids code for an integrase, and this protein was shown to be responsible for shuffling the strain chromosome (Cossu et al., 2017). For authors, this can be a mean of rapid adaptation to environmental changes. Indeed, this shuffling could place genes in a different genomic context, what may cause repression or activation, which can lead to isolates with higher fitness or abilities to colonize a new ecological niche.

5) Conclusion

To conclude this work, we started from 305 isolates of *Thermococcales* present in the UBO Culture Collection. Now nearly all isolates have a taxonomic assignment, and a phylogenetic tree was built with the concatenation of 16S rRNA gene and ITS sequences of each isolates. We selected two groups of *Thermococcus*, composed of 21 and 27 isolates. Genomes of these 48 isolates will be fully sequenced to study their genomic diversity.

6) Materials and Methods

All strains were cultured anaerobically in penicillin vials using two media, TRM (Zeng et al., 2009) and Ravot (Gorlas et al., 2013), depending on the isolates.

Genomic DNA of each isolate was extracted using a classic phenol-chloroform technique, with 50mL of fresh culture. Briefly, centrifuge culture at 8 000 rpm for 5 min at 4°C and discard supernatant. Homogenize the cell pellet with 1mL of TNE 1X buffer

(50 mL Tris-HCl 1 M; 50 mL EDTA 0.5 M pH 8; 10 mL NaCl 5M; q.s.p 500 mL). Add the lysis solution (100 µL of *N*-Lauroylsarcosine sodium salt 10%; 100 µL of Sodium Dodecyl Sulfate 10%; 50 µL of Proteinase K at 20 $mg.mL^{-1}$ (Promega®)) and incubate at 55°C for 1 h in a water bath, mix slowly from time to time. Add 20 µL of RNase A at 10 $mg.mL^{-1}$ (Ref: 02101076, MP Biomedicals) and incubate 20 to 30 min at 37°C. Then add 1 mL of Phenol-Chloroform-Isoamyl alcohol (PCI) (25:24:1), shake by turning for 45 sec and centrifuge at 14 000 rpm for 15 min at 4°C. Recover the aqueous phase (upper phase). If the interface contains a lot of cell fragments, repeat the operation 1 time (1 mL PCI, mix 45 sec, centrifuge, recover supernatant). Then add 1 mL of chloroform and shake by turning for 45 sec and centrifuge at 14 000 rpm for 15 min at 4°C. Recover the aqueous phase (upper phase) and add 0.7 volume of frozen isopropanol (−20°C). Put the tube at −20°C for at least 1 h or overnight. Centrifuge at 14 000 rpm during 20 min at 4°C, discard the supernatant and resuspend the DNA pellet with 0.5 mL of frozen 75% (v/v) ethanol. Centrifuge at 14 000 rpm during 20 min at 4°C, discard the supernatant and dry the DNA pellet at room temperature. To finish, resuspend the DNA pellet by adding 50 to 100 µL of low EDTA buffer, for example buffer EB (Ref: 19086, Qiagen).

For the 16S-ITS amplification, we used primers A4F: TCC GGT TGA TCC TGC CRG, Tm=60°C (Reysenbach et al., 2000a), and A71R: TCG GYG CCC GAG CCG AGC CAT CC, Tm=62.6°C (Casamayor et al., 2002). The PCR template was as follow: Initial denaturation for 5 min at 95°C; then 30 cycles: 95°C 1 min, hybridization at 60°C during 1 min, elongation during 2 min at 72°C. Then, the PCR ends with a last elongation step during 10 min at 72°C.

Beckman Coulter Genomics (Takeley, UK) sequenced all 16S-ITS sequences. Three runs per isolates, with primers A4F, A71R and A1492R: GGC TAC CTT GTT ACG ACT T,

Tm=56°C (Teske et al., 2002). Reads were assembled with a homemade Perl script and quality of reads was assessed manually.

For the phylogenetic tree, were aligned 16S sequences with SINA v1.2.11, *profile* parameter set to *Archaea* (Pruesse et al., 2012). All ITS were aligned with MUSCLE v3.8.31, with default parameters (Edgar, 2004). Then both parts of each sequence were concatenated. The tree was built with MrBayes v3.2.6 (Ronquist and Huelsenbeck, 2003) with the following parameters: *nst=2*, *rates=invgamma*, *ngen=60000000*, *samplefreq=2000* and default burnin of 25%. Then, the tree was visualized with ARB (Ludwig et al., 2004) and FigTree v1.4.2 (http://tree.bio.ed.ac.uk/software/figtree).

II) Comparative genomics of closely related isolates to identify genetic and genomic markers of diversification

1) Abstract

The hyperthermophilic archaeon *Thermococcus* is always present in deep-sea hydrothermal vents. It is also a pioneering species during the colonization of new deep-sea chimneys. In the laboratory, it can be easily cultivated and new strains are isolated on a regular basis. In a previous study, we selected two groups of around 24 isolates from different geographical locations and sequenced their genomes. Isolates from group I originated from Pacific, Atlantic and South Indian Oceans, whereas group II isolates were sampled in two locations from Pacific Ocean (EPR). Here we are interested in the genomic diversity of closely related isolates, to identify markers of differentiation and pinpoint genes and pathways involved in this differentiation. To answer this, we used both pangenomics and phylogenomics with all genomes of *Thermococcales* available. From group I, isolates are well clustered by their geographic origin, and each cluster corresponds to a microbial species, while isolates from group II do not follow this trend. This second group contains 6 distinct species, among which 2 are present on two hydrothermal sites. Globally, many clade specific genes are involved in amino acids, energy and carbohydrates metabolisms, which must reflects selection pressure that these organisms encounter in hydrothermal environments. For geographically distant isolates, the distance parameter is a driver of diversification, which probably translates into a decrease or absence of gene flow, while niche partitioning should explain intra-site patterns observed in the group II. All specific genes highlighted here can serve as targets for future genetic studies.

2) Introduction

Deep-sea hydrothermal vents (DSHV) host a huge diversity of microorganisms. However, this environment is described as extreme, due to the high hydrostatique

pressure and hot hydrothermal fluid. In addition to the large geochemical gradients present, DSHV are dynamic geochemical and biological structures (Anderson et al., 2017b; Wang et al., 2009). Consequently, microorganisms have to adapt to quick environmental changes in their ecological niche.

Genomes of DSHV microorganisms are highly variable. A study reported that deep-sea metagenomes are enriched with mobile genetic elements such as genes encoding transposases (Brazelton and Baross, 2009), suggesting a need for adaptation over short periods of time. For example, the *Archaea Thermococcus nautili* harbors three plasmids, of which one code for an integrase. This particular protein is involved in the chromosome dynamics and shuffling of its host (Cossu et al., 2017). Authors proposed that this rapid reorganization of the chromosome can acts on the fitness of a particular genome within the population. In general, *Thermococcales* harbor a large number of mobile genetic elements, plasmids or viruses, potentially involved in genetic transfers between isolates (Erauso et al., 1996; Gorlas et al., 2012; Krupovic et al., 2013; Wagner et al., 2017). Moreover, *Pyrococcus* genomes show extensive rearrangements, and presence of horizontal gene transfer (White et al., 2008). While *Thermococcus* genomes are diverse and dynamic, core functions are conserved in the same genomic context (Cossu et al., 2015).

The genomic diversification depends on multiple mechanisms, notably the gene flow between isolates (Cordero and Polz, 2014; Shapiro et al., 2012). In case of sufficiently high rate of gene flow, the population should stay homogeneous, but the establishment of barriers will result in the emergence of at least two new species, populations or ecotypes (Cohan and Perry, 2007; Fraser et al., 2009; Shapiro et al., 2012). There are several barriers, like colonization of a new ecological niche or geographic isolation (Anderson et al., 2017a; Cadillo-Quiroz et al., 2012; Whitaker et al., 2003).

Here we aimed at identifying genomic markers of diversification in closely related isolates of *Thermococcus* subjected to different drivers of diversification. Thus, we pinpointed genes and pathways involved in this differentiation to learn more about the selection pressures that apply to these isolates. For sympatric isolates, specific loci accounting for differences between evolutionary adjacent clades (species or populations) are indeed likely to be those under selective pressure. Their function should therefore inform about the evolutionary processes leading to population differentiation and eventually, speciation. For allopatric isolates, those clade-specific loci will provide insights about the effects of gene flow limitations on early genomic differentiation, e.g. what are the loci most sensitive to genetic isolation.

In this work, we used a pangenomics approach to establish a phylogeny based on core genome of all *Thermococcales* genomes available at the time of this study. This approach allowed defining *de novo* a set of genes for building a phylogenomic tree. It thus represents an improvement over methods relying on universal SCGs gene sets like AMPHORA of PhyloSift (Darling et al., 2014; Wu and Eisen, 2008) as it strongly improved the robustness of the phylogenomic reconstruction.

In addition, this set of genes was selected without *a priori* about functions, because it was based on the core-genome, which are all genes shared by all genomes considered (Tettelin et al., 2005). In pangenomics, the core-genome is opposed to the accessory genome (also called specific genome), which include all genes not shared by all studied genomes. The union of the core-and accessory-genomes represents the pan-genome. The size of the core genome relies on the number of genomes included in the comparison. Indeed, the number of shared (core-) genes decreases with the addition of new genomes within the comparison, until reaching a plateau, whereas the size of the pan-genome can constantly increase (Tettelin et al., 2005; Touchon et al., 2009). In this

study, we reconstructed a phylogeny based on core-genes that are present in single copy within these *Thermococcales* genomes. We then applied this pan-genomics approach to perform comparative genomics analysis and identify protein cluster that are differentially represented in closely related clades.

3) Results

a) Genome sequencing and assembly

We constructed and sequenced Illumina libraries for each of the 48 genomes selected in a previous study. Among all genomes, 19 were circular, 15 were available in one linear contig, and 13 genomes were assembled in a number of contigs ranging from 2 to 57 (Table 5). We then looked for the presence of putative plasmids in the genomes sequenced here. A contig is considered as a potential plasmid if the coverage of the latter is at least 3 times greater than the average coverage of the chromosome (Figure 22). In total, 17 contigs are potential plasmids. Moreover, we expect the presence of such mobile genetic element integrated in two genomes, *T.* sp EXT10c and *T.* sp EXT11c, For *T.* sp EXT11c, we observed a sharp increase in coverage between positions 398 000 and 412 000, from 230X on average in the genome to 3000X between these positions (Figure 22). The same observation was made for *T*. sp EXT10c: the coverage increase from 200X to 650X between positions 1 896 100 to the end (1 909 014).

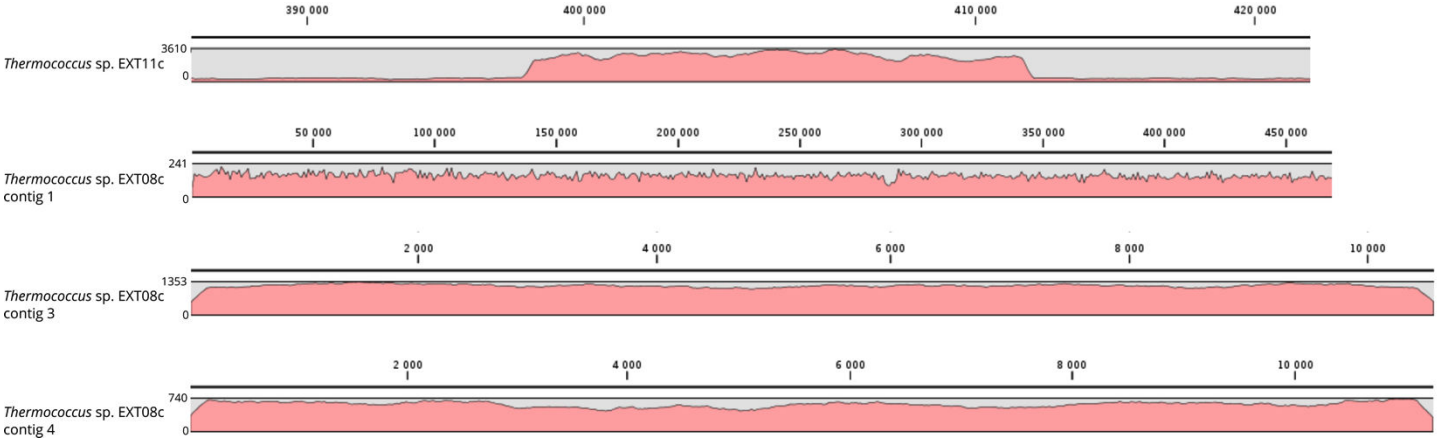


Figure 22: Examples of normal *vs*. potential mobile genetic elements coverages

For each line, the pink curve shows the genome coverage at a given position. The highest value is displayed on the left. For *Thermococcus* sp. EXT11c, the mean genome coverage is about 230X while there is a 13 kb region experiencing a 10-fold coverage increase, suggesting the presence of a mobile genetic element present in multiple copies within the cell and also inserted in the chromosome. *Thermococcus* sp. EXT08c contig 1: the mean genome coverage (about 148X); contigs 3 and 4: the coverage of two suspected mobile genetic elements, 1180X and 600X respectively.

Table 5: Summary of genomes sequenced in this study

Strain	Number of contigs	Number of potential plasmid	Assembly stage	Genome G+C content (in %)	Genome length (in bp)	Number of genes	Potential Plasmid length (in bp)	Group
AMTc09	9	-	In contigs	55.82	2 041 610	2 263	-	II
AMTc102	6	-	In contigs	55.79	2 040 569	2 231	-	II
AMTc19	3	1	In contigs	54.82	1 876 828	2 077	4 875	II
AMTc29	4	-	In contigs	55.79	2 042 352	2 238	-	II
AMTc30	1	-	Circular	54.89	1 941 830	2 126	-	II
AMTc51	1	-	Circular	53.93	1 941 015	2 096	-	II
AMTc52	1	-	Circular	53.93	1 941 015	2 096	-	II
AMTc67	1	-	Circular	55.05	1 887 775	2 077	-	II
AMTc79	1	-	Not circular	54.52	1 977 287	2 210	-	II
AMTc85	1	-	Circular	54.48	1 933 709	2 135	-	II
AMTc94	2	1	Not circular	55.00	1 968 052	2 173	7 354	II
AMTc95	3	1	Circular	54.83	1 876 541	2 076	4 857	II
E10P11	3	2	Not circular	54.95	1 963 793	2 163	11 631/3 839	II
E10P7	3	2	Circular	54.95	1 963 815	2 165	3 832/11 631	II
E10P8	3	2	Not circular	54.95	1 963 773	2 163	3 837/11 638	II
E14P19	3	2	Not circular	54.95	1 963 796	2 165	11 638/3 839	II
E15P25	29	-	In contigs	54.30	4 070 808	-	-	II
E15P29	2	1	Circular	54.77	2 048 193	2 278	13 976	II
E15P30	3	1	In contigs	54.77	2 048 296	2 279	14 055	II
E15P35	3	2	Not circular	54.77	2 048 179	2 279	10 153/3 880	II
E15P6	1	-	Not circular	54.76	2 049 600	2 279	-	II
EXT08c	4	2	In contigs	54.83	1 912 858	2 165	10 557/11 228	II
EXT10c	1	1 (Integrated)	Circular	54.95	1 909 014	2 128	~ 13 000	II
EXT11c	1	1 (Integrated)	Circular	54.95	1 909 018	2 132	~ 13 000	II
EXT12c	1	-	Circular	54.58	2 155 760	2 338	-	II
EXT13c	1	-	Circular	54.58	2 155 756	2 336	-	II
IRI06c	9	-	In contigs	55.02	2 135 086	2 302	-	II
AMTc70	1	-	Not circular	55.96	1 954 101	2 089	-	I
AMTc71	2	-	In contigs	55.96	1 954 215	2 086	-	I
AMTc72	2	-	In contigs	55.96	1 953 306	2 082	-	I
AMTc73	1	-	Not circular	55.96	1 954 120	2 090	-	I

IRI07c	1	-	Circular	54.24	2 124 406	2 303	-	I
IRI09c	1	-	Circular	54.24	2 124 338	2 304	-	I
IRI10c	1	-	Not circular	54.24	2 124 377	2 303	-	I
IRI14c	1	-	Not circular	54.59	2 064 797	2 235	-	I
IRI15c	1	-	Circular	54.60	2 061 457	2 233	-	I
IRI24c	1	-	Not circular	54.54	2 086 111	2 275	-	I
IRI25c	1	-	Not circular	54.58	2 081 805	2 256	-	I
IRI26c	1	-	Circular	54.54	2 086 247	2 275	-	I
IRI27c2	1	-	Not circular	54.54	2 086 145	2 275	-	I
IRI29c	1	-	Not circular	54.54	2 085 431	2 277	-	I
IRI35c	1	-	Circular	54.63	2 022 529	2 185	-	I
IRI35c2	1	-	Circular	54.63	2 022 531	2 185	-	I
IRI36c	1	-	Circular	54.63	2 022 486	2 184	-	I
MC4	55	-	In contigs	54.31	2 214 575	2 508	-	I
MC5	-	-	-		-		-	I
MC8	57	-	In contigs	54.25	2 230 429	2 536	-	I
MC9	55	-	In contigs	54.32	2 211 370	2 509	-	I

Table 6: Summary of other genomes used in pangenomics study

Strain	Genome length (in bp)	Genome G+C content (in %)	Number of genes	Group	Accession	Reference
Thermococcus celericrescens	2 337 139	54.29	2 626	I	NZ_LLYW00000000	Hong et al., 2015
BPK M *Thermococcus* EXT09c	1 922 260	54.82	2 169	II	-	Oberto J, Da Cunha V. and Forterre P
BPK P *Thermococcus* sp	1 924 085	54.73	2 145	II	-	Oberto J, Da Cunha V. and Forterre P
BPK E *Thermococcus* 29 3	1 969 639	54.80	2 164	II	-	Oberto J, Da Cunha V. and Forterre P
BKX A *Thermococcus nautili*	1 976 351	54.84	2 188	II	-	Oberto J, Da Cunha V. and Forterre P
BPK A *Thermococcus* 33 3	2 013 327	54.60	2 238	II	-	Oberto J, Da Cunha V. and Forterre P
BPK S *Thermococcus* 9 3	2 021 135	54.85	2 253	II	-	Oberto J, Da Cunha V. and Forterre P
BPK B *Thermococcus* 26 2	2 030 483	55.89	2 247	II	-	Oberto J, Da Cunha V. and Forterre P
BPK I *Thermococcus* E15p33	2 048 119	54.77	2 279	II	-	Oberto J, Da Cunha V. and Forterre P
BPI D *Pyrococcus* GE 23	1 720 484	44.72	1 874	-	-	Oberto J, Da Cunha V. and Forterre P

Thermococcus zilligii AN1	1 764 559	54.48	1 875	-	NZ_AJLF01000000	(Kim et al., 2012)
BPI E *Pyrococcus* IRI 42C	1 814 925	41.04	1 984	-	-	Oberto J, Da Cunha V. and Forterre P
BKT A *Pyrococcus abyssi* GE2	1 815 336	44.56	1 959	-	-	Oberto J, Da Cunha V. and Forterre P
BPI F *Pyrococcus* EXT 15c	1 851 203	45.01	2 080	-	-	Oberto J, Da Cunha V. and Forterre P
BKY A *Thermococcus* 23 2	1 854 909	55.69	1 989	-	-	Oberto J, Da Cunha V. and Forterre P
BPK R *Thermococcus* 5 4B	1 854 909	55.69	1 990	-	-	Oberto J, Da Cunha V. and Forterre P
BPK F *Thermococcus* 5 4A	1 854 910	55.69	1 990	-	-	Oberto J, Da Cunha V. and Forterre P
BPK C *Thermococcus* 15 2	1 881 790	54.44	2 059	-	-	Oberto J, Da Cunha V. and Forterre P
BPK G *Thermococcus prieuri* Col3	1 927 526	54.27	2 135	-	-	Oberto J, Da Cunha V. and Forterre P
BPK Q *Thermococcus prieuri* Brest	1 927 535	54.28	2 136	-	-	Oberto J, Da Cunha V. and Forterre P
BKW A *Pyrococcus* EXT16	1 952 677	45.05	2 171	-	-	Oberto J, Da Cunha V. and Forterre P
BPI A *Pyrococcus* 32 4	1 989 130	44.97	2 184	-	-	Oberto J, Da Cunha V. and Forterre P
BPI C *Pyrococcus* 32 3	1 989 130	44.97	2 180	-	-	Oberto J, Da Cunha V. and Forterre P
BPI B *Pyrococcus* 32 1	1 989 216	44.97	2 179	-	-	Oberto J, Da Cunha V. and Forterre P
Thermococcus sp MF15	1 998 809	51.97	2 169	-	-	Le Guellec et al., In prep
BPK H *Thermococcus* IRI 33c	2 115 565	52.80	2 379	-	-	Oberto J, Da Cunha V. and Forterre P
BPK N *Thermococcus* IRI 05c	2 119 293	53.14	2 340	-	-	Oberto J, Da Cunha V. and Forterre P
BPK L *Thermococcus* CIR 10a	2 201 239	41.38	2 459	-	-	Oberto J, Da Cunha V. and Forterre P
Palaeococcus ferrophilus	2 206 431	54.31	2 358	-	NZ_LANF00000000	Eisen et al., 2015
BPK K *Thermococcus* CIR 03a	2 207 808	53.55	2 373	-	-	Oberto J, Da Cunha V. and Forterre P
BKU A *Thermococcus* AMTc11	2 378 328	54.47	2 607	-	-	Oberto J, Da Cunha V. and Forterre P

Of these 48 sequenced genomes, only 45 were used for the pan-genomics analysis of group I and group II isolates. The IRI06c isolate genome was assembled into 9 contigs. It was expected to be closely related to *Thermococcus nautili* (16S-ITS phylogenetic tree), but on a core-genes based tree, IRI06c was no longer close to *T. nautili*. About this isolate, the 16S rRNA gene data was already available in the laboratory. We sequenced the ITS with DNA already extracted. However, we found 15 different nucleotides between the previously known 16S rRNA gene sequence and the one from the fresh cultured IRI06c genome, indicating that they are different isolates. Nevertheless, this genome will be integrated in the *Thermococcales* pan-genome. For the MC5 strain, the raw data contained only sequences for PhiX, a control used on Illumina sequencing systems. Several reasons may be advanced: a mistake may have occurred during the library preparation, nevertheless the library passed quality and quantity controls. Alternatively, the error may have occurred during the demultiplexing step, with a mistake in the barcode assignation. The last non-exploitable isolate was E15P25. In this case, the total length of assembled contigs was 4Mb, which is about twice the average length of a standard *Thermococcus* genome. Within the 29 contigs, a set of 162 archaeal genes that should be present in single copy (Rinke et al., 2013) were duplicated. Two ribosomal operons were also present. This strongly suggests the presence of two genomes within this sample, harboring the same 16S rRNA gene sequence and therefore an incomplete isolation of this strain or a contamination during growth experiment.

Concerning the other sequenced genomes, their average length is 2.02 Mbp, with a minimum of 1 876 541 bp (AMTc95) and a maximum of 2 230 429 bp (MC8). The gene number is on average 2 219 genes, with a minimum of 2 076 genes (AMTc95) and a maximum of 2 536 genes for MC8. The G+C content is on average 54.82%. Genomes

with the lowest G+C content are AMTc51-52, with 53.93%. And the genomes with higher G+C content are AMTc70-71-72-73, with 55.96% (Table 5).

b) Definition of closely related clades: what is the evolutionary history of our genomes?

The first step in our study was to place sequenced genomes within an evolutionary context. Here, we decided to use the largest set of genes common to all genomes, *i.e* the core-genome, instead of a predefined or universal set of genes to build a phylogenomic tree based on a set of "Single-copy Core Genes" (SCGs). To recover these genes, we decided to use a comparative genomics approach to obtain the pan-genome, and from this latter, identify SCGs without *a priori* knowledge of their annotation..

The figure 23 is a representation of this pan-genome. It is composed of 114 *Thermococcales* genomes, distributed as follow: 2 *Palaeococcus*, 17 *Pyrococcus* and 95 *Thermococcus*. All these genomes come from: (i) public databases (Table 2,6), (ii) this study (Table 5) or (iii) unpublished genomes sequenced by the laboratory MBGE from Pasteur Institute (Paris, France). The name of these latters starts with "*Bxx x*" (Table 6). This analysis grouped all 246 617 predicted proteins in 13 431 protein clusters (PCs). Among them, 778 PC belonged to the core genome (90 611 genes), of which 602 PCs represented SCGs, highlighted in light blue (Figure 23). Completion and redundancy for the 114 genomes was also evaluated based on a set of 162 archaeal genes that should be present in a single copy in any *Archaea as defined by REF*. Regarding completion, 90 genomes were complete. The 24 remaining genomes harbored completion between 99.38% and 97.53%, meaning that 1 to 4 SCG were missing (Figure 23). On the other hand, the redundancy in all genomes spanned from 2.47% to 6.17% respectively, meaning that 4 and 10 genes were found at least duplicated (Figure 23).

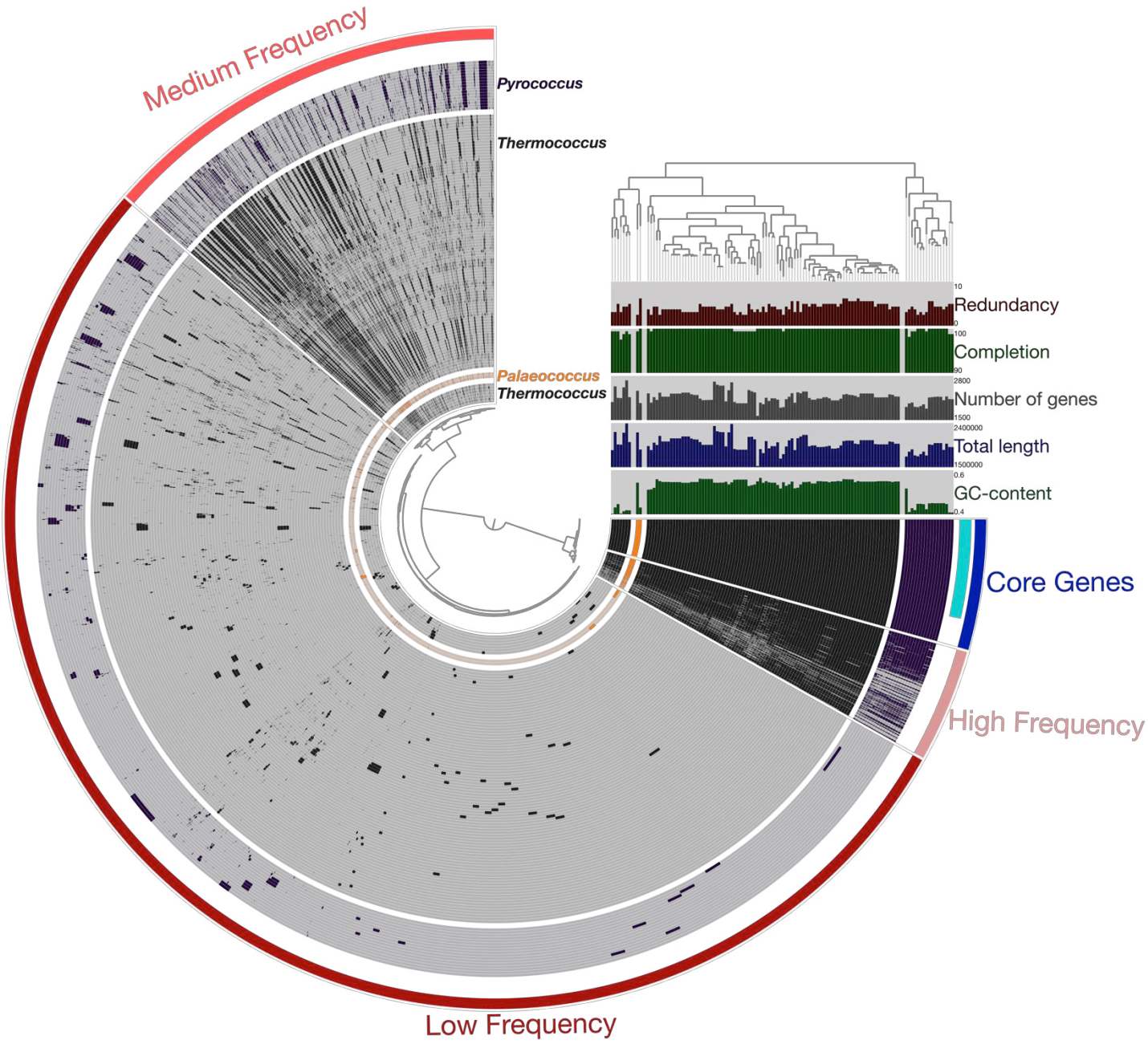


Figure 23: Pangenomics analysis of 114 *Thermococcales* genomes

On this figure, genomes are displayed as circle layers: 14 *Pyrococcus*, 95 *Thermococcus*, and 2 *Palaeococcus*. Each leaf of the centered dendrogram represents a PC, organized by hierarchical clustering based on their presence/absence across genomes. When a PC is present within a genome, it results in the presence of a dark colored bar on the genome's layer. On the upper right, are displayed metadata: G+C content, length and the number of genes for each genome. The Completion and Redundancy are two metrics to assess the quality of the genome assembly. They are based on a dataset of 162 archaeal single-copy genes (Rinke et al., 2013). Redundancy reflects the "*multiple occurrence of one or more single-copy genes*" in a genome (Eren et al., 2015). Then, layers or genomes are organized based on a phylogeny (top right tree). Besides, we highlighted groups of PC based on their frequency: In blue, all core genes, including in light blue the subpart of SCGs. The three shades of red represent the accessory genome, divided in tree by their relative frequency.

c) From pan-genome to phylogenomic tree

The 16S rRNA gene and the rest of the genome have different molecular clocks (Case et al., 2007). In addition, this marker gene lacks phylogenetic resolution for *Thermococcales*. We therefore used genomic information to build a phylogenomic tree. This tree (Figure 24A) was built by maximum likelihood from the concatenation of protein sequences from the 602 SCGs. The alignment matrix was composed of 157 675 AA positions, of which 92 510 were not redundant.

Concerning the tree topology, regarding the taxonomy of isolates used to build it, three major clades are present. A first grouped all *Pyrococcus*, a second mixed together the two *Palaeococcus* genomes plus *Thermococcus* with a low G+C content (see Figure 23), and the last was composed of all other *Thermococcus*.

On this tree, all isolates sequenced in this study are separated in two different groups, as expected with the 16S rRNA gene phylogenetic tree. All group I genomes (20) clustered together (Figure 24B), with also two genomes from the public databases: *T.* sp. 4557 and *T. celericrescens*. In this group, the diversification pattern is concomitant with the strains geographical origin. Concerning the group II, all 25 genomes clustered with a reference strain, *T. nautili*, and with 8 unpublished isolates ("*Bxx x"*), for a total of 34 genomes (Figure 24B, Tables 5,6). Interrestingly, the history (=phylogenetic relationship) of these isolates is not correlated with their geographic origin.

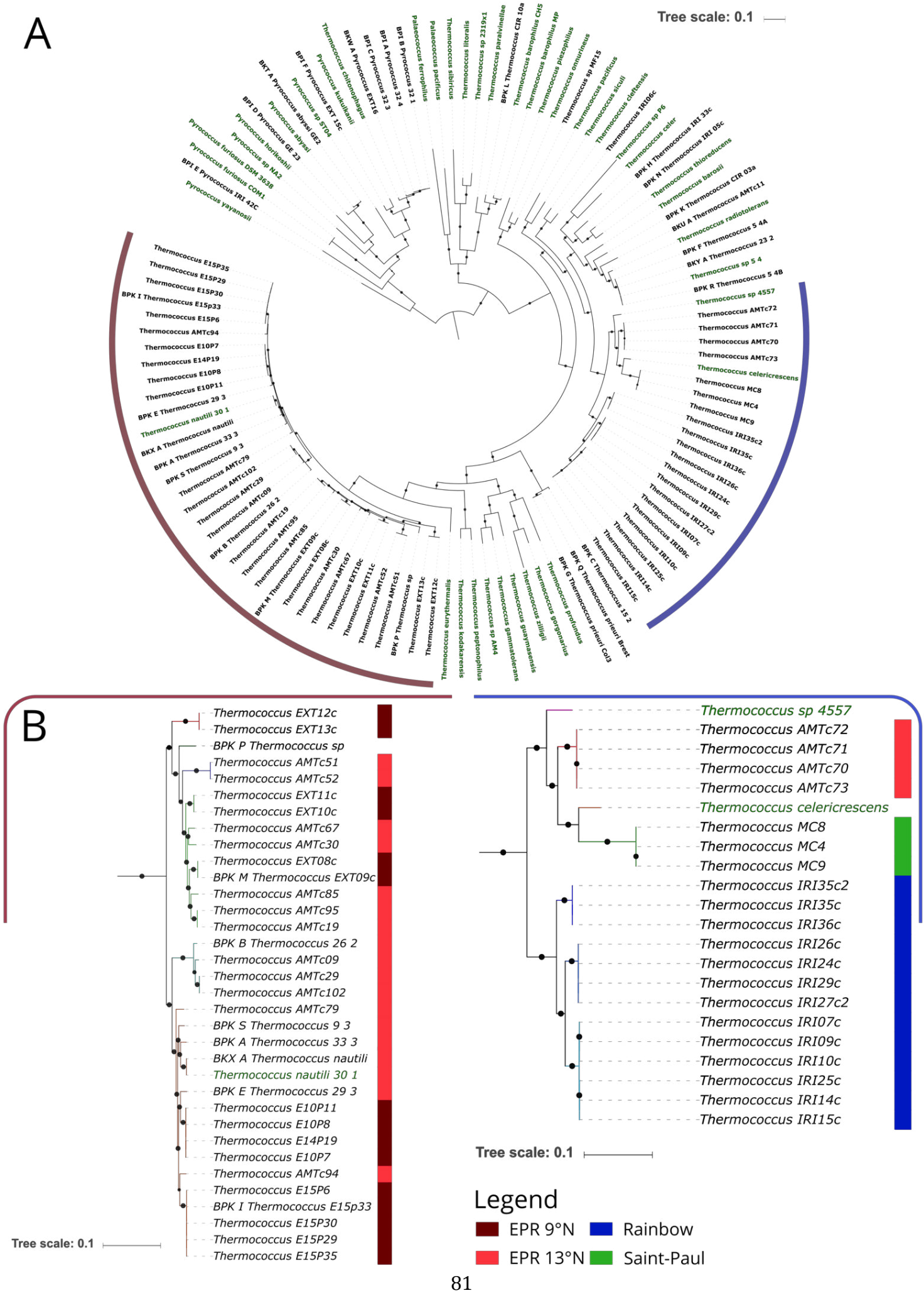

A
Tree scale: 0.1
Thermococcus E15P35
Thermococcus E15P29
Thermococcus E15P30
BPK I Thermococcus E15p33
Thermococcus E15P6
Thermococcus AMTc94
Thermococcus E10P7
Thermococcus E14P19
Thermococcus E10P8
Thermococcus E10P11
BPK E Thermococcus 29 3
Thermococcus nautili 30 1
BKX A Thermococcus nautili
BPK A Thermococcus 33 3
BPK S Thermococcus 9 3
Thermococcus AMTc79
Thermococcus AMTc102
Thermococcus AMTc29
Thermococcus AMTc09
BPK B Thermococcus 26 2
Thermococcus AMTc19
Thermococcus AMTc95
Thermococcus AMTc85
BPK M Thermococcus EXT09c
Thermococcus EXT08c
Thermococcus AMTc30
Thermococcus AMTc67
Thermococcus EXT10c
Thermococcus EXT11c
Thermococcus AMTc52
Thermococcus AMTc51
BPK P Thermococcus sp
Thermococcus EXT13c
Thermococcus EXT12c
Thermococcus sp 2319x1
Thermococcus siculi
Thermococcus kodakarensis
Thermococcus peptonophilus
Thermococcus sp AM4
Thermococcus gammatolerans
Thermococcus guaymasensis
Thermococcus zilligii
Thermococcus gorgonarius
Thermococcus profundus
BPK G Thermococcus prieurii Col3
BPK Q Thermococcus prieurii Brest
BPK C Thermococcus 15 2
Thermococcus IRI15c
Thermococcus IRI14c
Thermococcus IRI25c
Thermococcus IRI10c
Thermococcus IRI09c
Thermococcus IRI07c
Thermococcus IRI27c2
Thermococcus IRI29c
Thermococcus IRI24c
Thermococcus IRI26c
Thermococcus IRI36c
Thermococcus IRI135c
Thermococcus IRI135c2
Thermococcus MC9
Thermococcus MC4
Thermococcus MC8
Thermococcus celericrescens
Thermococcus AMTc73
Thermococcus AMTc70
Thermococcus AMTc71
Thermococcus AMTc72
Thermococcus sp 4557
BPK R Thermococcus 5 4B
Thermococcus sp 5 4
BKY A Thermococcus 23 2
BPK F Thermococcus 5 4A
BKU A Thermococcus radiotolerans
BPK K Thermococcus AMTc11
BPK N Thermococcus CIR 03a
Thermococcus barosii
Thermococcus thioreducens
BPK H Thermococcus IRI 05c
Thermococcus celer
Thermococcus sp P6
Thermococcus IRI06c
Thermococcus cleftensis
Thermococcus pacificus
Thermococcus sp MF15
Thermococcus onnurineus
Thermococcus piezophilus
Thermococcus barophilus MP
Thermococcus barophilus CH5
BPK L Thermococcus CIR 10a
Thermococcus paralvinellae
Thermococcus sp 2319x1
Thermococcus litoralis
Thermococcus sibiricus
Palaeococcus pacificus
Palaeococcus ferrophilus
BPI A Pyrococcus 32 4
BPI B Pyrococcus 32 1
BPI C Pyrococcus 32 3
BKW A Pyrococcus EXT16
Thermococcus chitonophagus
Pyrococcus kukulkanii
BPI F Pyrococcus EXT 15c
Pyrococcus sp ST04
BKT A Pyrococcus abyssi GE2
Pyrococcus abyssi GE 23
BPI D Pyrococcus GE 23
Pyrococcus horikoshii
Pyrococcus sp NA2
Pyrococcus furiosus DSM 3638
Pyrococcus furiosus COM1
BPI E Pyrococcus IRI 42C
Pyrococcus yayanosii
B
Thermococcus EXT12c
Thermococcus EXT13c
BPK P Thermococcus sp
Thermococcus AMTc51
Thermococcus AMTc52
Thermococcus EXT11c
Thermococcus EXT10c
Thermococcus AMTc67
Thermococcus AMTc30
Thermococcus EXT08c
BPK M Thermococcus EXT09c
Thermococcus AMTc85
Thermococcus AMTc95
Thermococcus AMTc19
BPK B Thermococcus 26 2
Thermococcus AMTc09
Thermococcus AMTc29
Thermococcus AMTc102
Thermococcus AMTc79
BPK S Thermococcus 9 3
BPK A Thermococcus 33 3
BKX A Thermococcus nautili
Thermococcus nautili 30 1
BPK E Thermococcus 29 3
Thermococcus E10P11
Thermococcus E10P8
Thermococcus E14P19
Thermococcus E10P7
Thermococcus AMTc94
Thermococcus E15P6
BPK I Thermococcus E15p33
Thermococcus E15P30
Thermococcus E15P29
Thermococcus E15P35
Tree scale: 0.1
Thermococcus sp 4557
Thermococcus AMTc72
Thermococcus AMTc71
Thermococcus AMTc70
Thermococcus AMTc73
Thermococcus celericrescens
Thermococcus MC8
Thermococcus MC4
Thermococcus MC9
Thermococcus IRI135c2
Thermococcus IRI135c
Thermococcus IRI36c
Thermococcus IRI26c
Thermococcus IRI24c
Thermococcus IRI29c
Thermococcus IRI27c2
Thermococcus IRI07c
Thermococcus IRI09c
Thermococcus IRI10c
Thermococcus IRI25c
Thermococcus IRI14c
Thermococcus IRI15c
Tree scale: 0.1
Legend
EPR 9°N
Rainbow
EPR 13°N
Saint-Paul

Figure 24: Phylogenomic tree of *Thermococcales* genomes

A, Global *Thermococcales* phylogeny built by maximum likelihood under LG model. Green leaves represent genomes from public databases and black leaves those not published. Leaves highlighted in blue indicate group I and group II genomes and are highlighted in red. The dots on branches represent bootstrap values > 50, scale is 1 substitution per 1000 sites. B, left side: group II genomes sub-tree; right side: group I genomes sub-tree. Isolates geographic origins are displayed as colored bars adjacent to trees.

d) Do groups represent species or subspecies ?

On the phylogenomic tree, both groups of selected isolates were monophyletic and well supported, and were thus suitable for the genomic diversity study of these closely related isolates. We then carried out two pangenomics studies. There, we only focused on within group genomes (group I & II as defined in this study). All plasmid-contigs have been removed, except for integrated mobile genetic elements.

Since both groups are well supported and monophyletic in the phylogenetic tree, we inspected if geographical isolation could explain their differentiation, and in a second time, wether each group correspond to new species or subspecies. To investigate this, we first considered the tree topology, and then used two metrics: the Average Nucleotide Identity (ANI) and *in silico* DNA DNA hybridization (DDH), with the thresholds 96% and 70% respectively.

On the phylogenomic tree, group I isolates formed 6 different clusters. Each of these monophyletic groups was matching a different geographical origin (Figure 24B), suggesting a differentiation pattern based on allopatry. In addition, strains originating from the Rainbow hydrothermal vent formed 2 distincts clades, suggesting sympatric differentiation event.

According to ANI and DDH, isolates “AMTc” from EPR 13°N formed a species; isolates “MC” from Saint-Paul also represented a new species, and isolates from Menez-Gwen too (Figure 25A, C). The species “*T.* sp IRI-1” refer to these three strains from Menez Gwen. Isolates from Rainbow were organized in 2 distinct clades (Figure 24B, light-blue branches). The 10 IRI isolates from Rainbow (07c, 09c, 10c, 14c, 15c, 24c, 25c, 26c, 27c2 and 29c) had ANI values higher than the threshold generally accepted for species delineation, namely 96% (Figure 25A). Within these 10 isolates, two groups emmerged. They were, on the one hand, *T.* sp. IRI24c, *T.* sp. IRI26c, *T.* sp. IRI27c2, *T.* sp. IRI29c,

named "*T.* sp IRI-2", and on the other hand, *T.* sp. IRI07c, *T.* sp. IRI09c, *T.* sp. IRI10c, *T.* sp. IRI14c, *T.* sp. IRI15c and *T.* sp. IRI25c, named "*T.* sp IRI-3" (Figure 25A). ANI values were close to 100% within these clades and closed to 96.5% between them. Based on DDH, *T.* sp IRI-2 and *T.* sp IRI-3 formed two species (Figure 25C).

The geographic origin of group II isolates could not explain the observed tree topology (Figure 24B). In this group, we could define 6 species (Figure 25B,D), 5 potentially new ones and 14 strains belonging to *T. nautili* species. These 6 species are highlighted by different branches colors on the phylogenetic tree (Figure 24B), and were named as follow (descending order in the tree, figure 24B) "Nautili-5", "Nautili-6", "Nautili-4", "Nautili-3", "Nautili-2", and "Nautili-1". Four of them are only present within one hydrothermal site, while "Nautili-1" and "Nautili-3" are present over the two EPR 9°N and EPR 13°N hydrothermal sites.

e) What are the consequences of the differentiations at the genomic level?

After establishing clades of closely related strains, we used comparative genomics on each group, to highlight genes and pathways involved in these differentiation processes. The presence of specific genes may be the result of (i) a reduction in genetic flow due to geographical isolation (possibly due to allopatric differentiation) or (ii) colonization of a new ecological niches (sympatric differentiation). We conducted a pan-genome analysis on group I and II strains separately. In this framework, we indentified PCs specific to previoulsly defined clades, and focused on the functions present in these specifics PCs (SPC). We only took in consideration functions that were uniq to a clade, implying that the function was absent from the rest of the pan-genome.

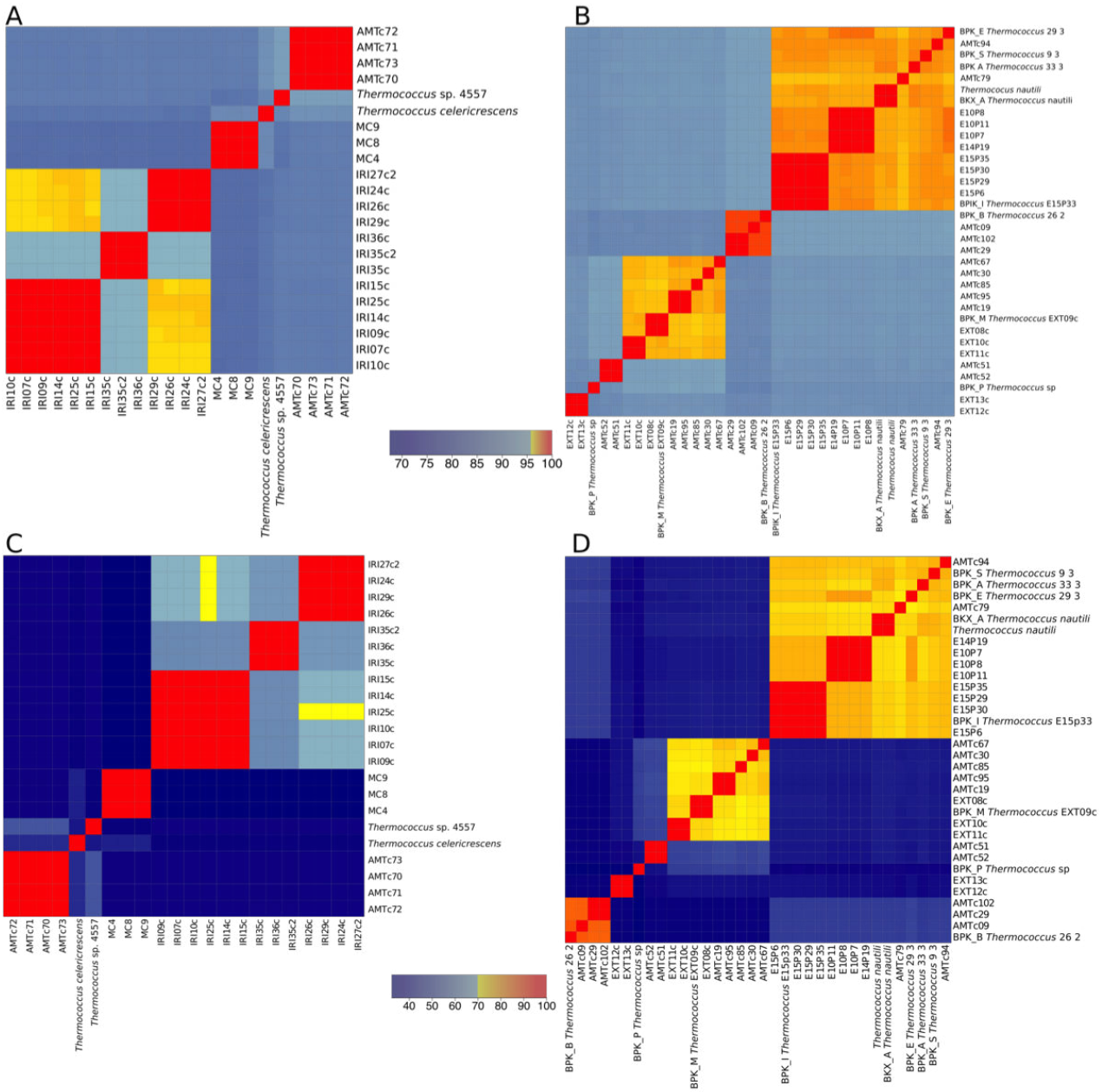


Figure 25: Genomic similarity of isolates from groups I and II

A,B: ANI 96% threshold. C,D: DDH 70% threshold. A,C: Group I. B,D: Group II. Each legend is displayed between heatmaps.

i) Group I pan-genome

This pan-genome grouped 49 945 genes within 22 genomes (Figure 26). All these genes were pooled within 3 948 PCs. The pan-genome was organized as follow: 1 459 PCs in the core-genome (32 568 genes), of which 1 352 PCs were only composed of SCGs (29 744 genes). Then the accessory genome was composed of 2 489 PCs, grouping 17 377 genes. Although the accessory genome grouped two-thirds of PCs, it only represented one-third in terms of genes number. All SPCs resulting from differentiation events were present in this accessory genome.

Here, we first investigated species SPCs. All studied bins of SPC are highlighted on the pan-genome figure (Figure 26). For *T.* AMTc isolates, 153 SPCs were found, of which 1 was involved in cell cycle, 3 in amino acid (AA) metabolism, 1 in transcription, 1 in cytoskeleton, 1 in secondary structure and 4 poorly characterized (Table 7). For the *T.* MC isolates, 283 SPCs were present, of which, 5 functions were involved in energetic metabolism, 29 in AA metabolism (Histidine and Tryptophan full biosynthesis pathways), 3 in nucleotide metabolism, 3 in carbohydrate metabolism, 2 in coenzyme metabolism, 2 in translation, 1 in transcription, 5 in inorganic ion transport, and 1 was of unknown function. *T*. sp. 4557 had 148 SPCs, and following functions were present: 1 was involved in chromatin structure and dynamics, 3 in energy metabolism, 1 in cell cycle, 2 in carbohydrate metabolism, 1 in membrane biogenesis, 2 in secondary structure and 2 were poorly characterized. *T. celericrescens* had 336 SPCs, and 2 had functions involved in energy metabolism, 2 in carbohydrate metabolism, 1 in replication and repair, 2 in membrane biogenesis, 2 in post-transcriptional modification, 1 in signal transduction, 1 in defense mechanism, 1 was related to mobile genetic element, and 7 were poorly characterized (Table 7).

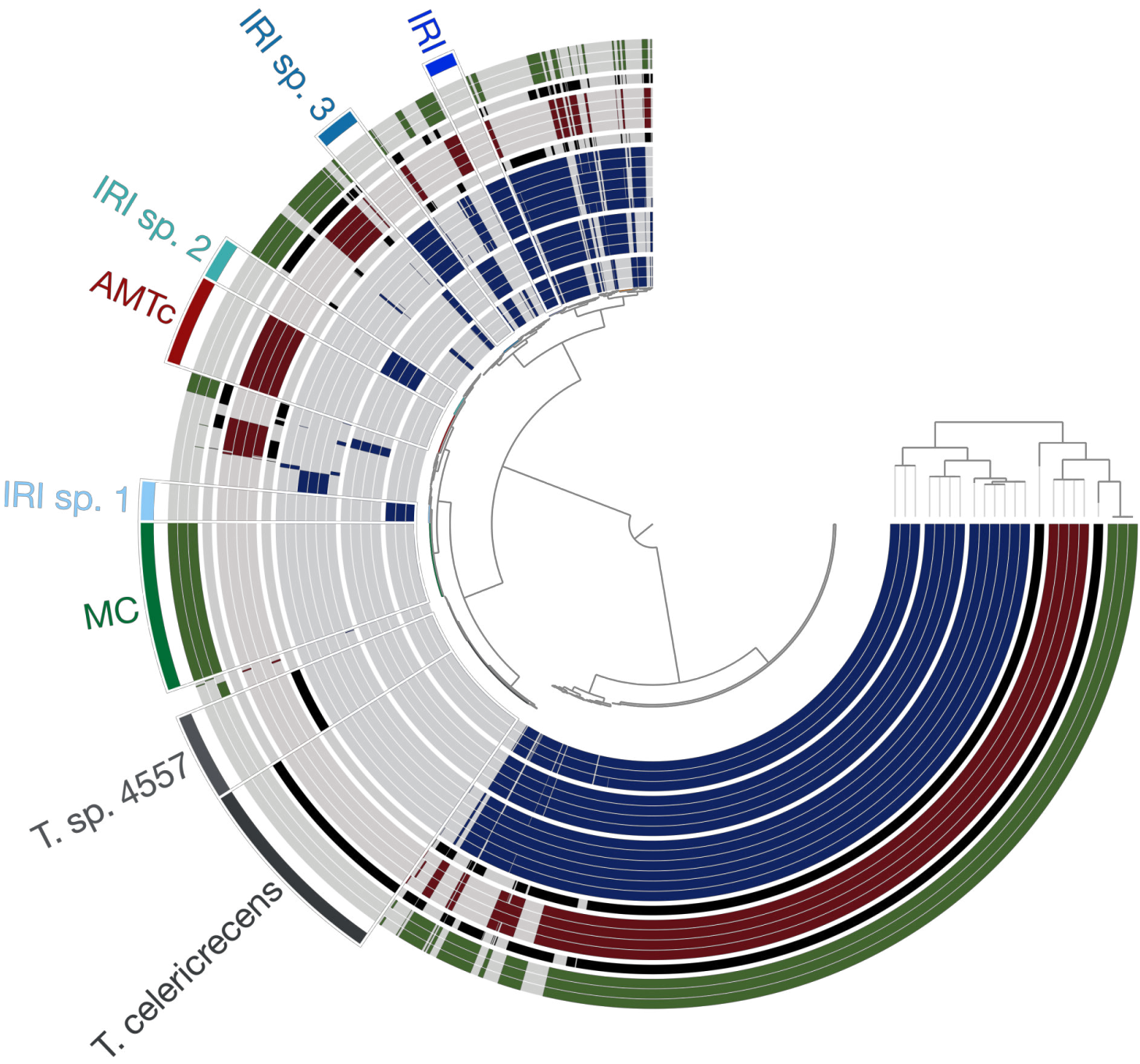


Figure 26: Group I pan-genome overview

Group I pan-genome comprised 3 948 PCs (49 945 genes). On this pan-genome, genomes (concentric circles) are grouped as species defined previously, and colored based on their geographic origin. They are organized based on the SCGs phylogeny. Bins, outermost layer, represent PCs specific to a geographical cluster and / or species.

In species *T.* sp. IRI-1, there were 68 SPCs, comprising only 1 function that was only present within this species, but belonging to a poorly characterized COG category (Table 7). Among the 73 SPCs found in *T.* sp. IRI-2, no SPC carried known function. Finally, *T.* sp. IRI-3 had 71 SPCs and only 1 had a specific function, related to the translation mechanism (Table 7).

Then, among all from Rainbow hydrothermal vent isolates (*T.* IRI), no specific function was found among the 48 SPCs.

ii) Group II pan-genome

The group II pan-genome included 34 *Thermococcus* genomes, of which 25 were sequenced for this study, 1 was the reference strain *T. nautili* plus a re-sequencing of this strain and 7 genomes came from unpublished data (Figure 27, Tables 2,5,6). In total, this pan-genome was composed of 4 255 PCs (74 782 genes) organized as follow: 1425 PCs (48 955 genes) present in the core-genome, whose 1 364 are SCGs (46 376 genes). The accessory genome was composed of 2 830 PCs, grouping 25 827 genes. For this pan-genome, we can make the same observation as for the group I pan-genome: nevertheless, even if the accessory genome represented two-thirds of PCs, it only grouped one-third of the total number of genes.

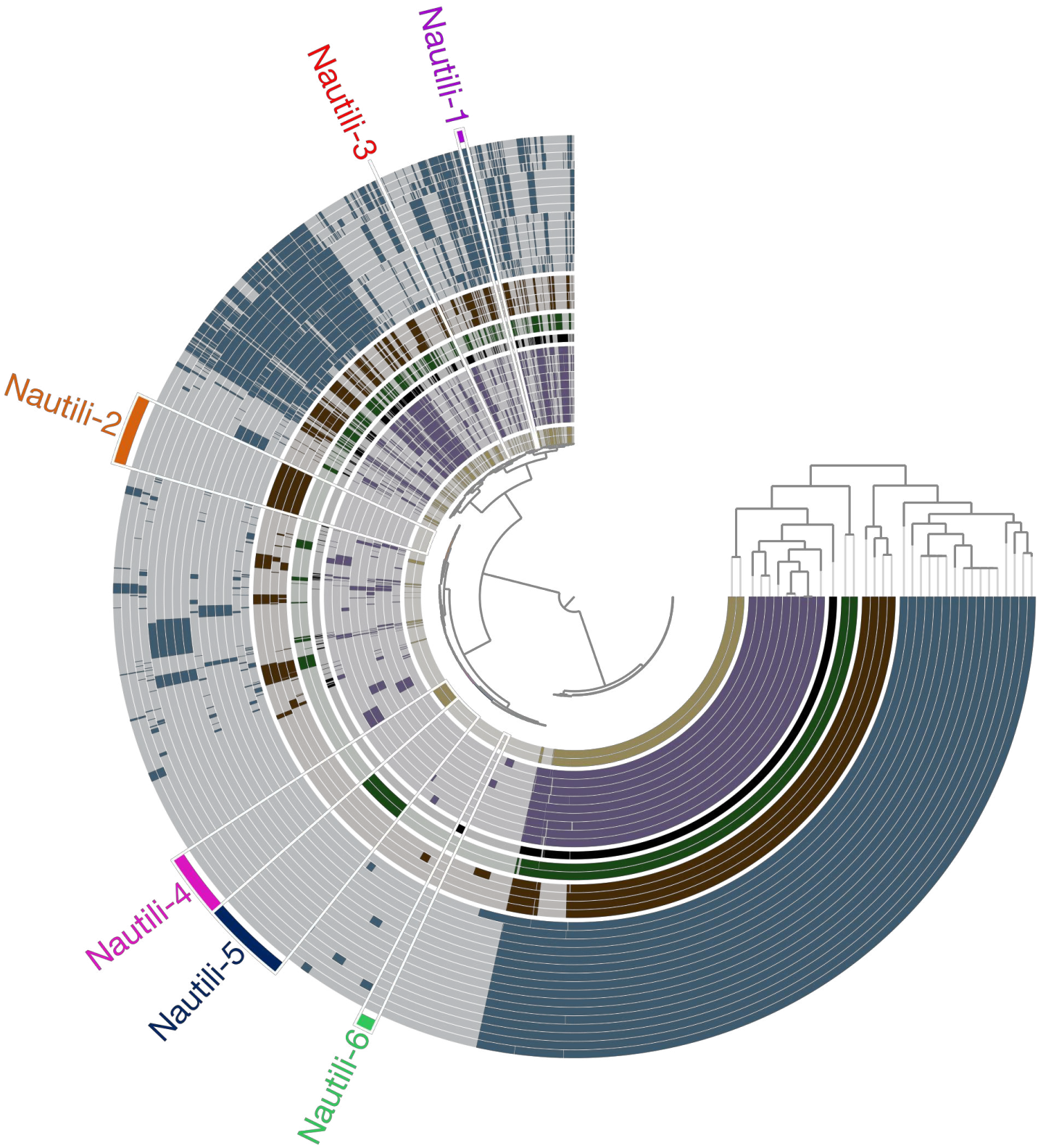


Figure 27: Group II pan-genome overview

Group II pan-genome comprises 4255 PCs (74 782 genes). Here, genomes (concentric circles) are colored as separated species, and organized based on the SCGs phylogeny. Bins (outermost layer) represent PCs specific to a species.

We then highlighted SPCs for each species and find species-specific functions (Figure 27). The first species grouped 15 isolates (*T.* sp. E15P35, *T.* sp. E15P30, *T.* sp. E15P29, *T.* sp. E15P6, *T.* sp. E15P33, *T.* sp. AMTc94, *T.* sp. AMTc79, *T.* sp. E10P7, *T.* sp. E10P8, *T.* sp. E10P11, *T.* sp. 29-3, *T.* sp. 33-3, *T.* sp. 9-3 and two *T. nautili*), and possessed 15 SPCs, with no assigned specific function (Table 7). The second species grouped 4 isolates (*T.* sp. AMTc102, *T.* sp. AMTc29, *T.* sp. AMTc09 and *T.* sp. 26-2) that shared 158 SPCs. Among them: 5 functions related to AA metabolism, 1 to carbohydrate metabolism, 1 to coenzyme metabolism, 2 to translation metabolism, 1 to post-transcriptional processes, 1 to inorganic ion metabolism and 7 corresponds to poorly characterized functions (Table 7). The third species grouped 9 isolates (*T.* sp. AMTc19, *T.* sp. AMTc95, *T.* sp. AMTc85, *T.* sp. AMTc30, *T.* sp. AMTc67, *T.* sp. EXT08c, *T.* sp. EXT09c, *T.* sp. EXT10c and *T.* sp. EXT11c), and only 1 SPC without specific function (Table 7). The fourth species grouped 2 isolates (*T.* sp. AMTc51 and *T.* sp. AMTc52). There were 124 SPCs, encoding for the following specific functions: 1 was assigned to RNA processing and modification, 3 to carbohydrate metabolism, 1 to inorganic ion metabolism, 1 to intracellular trafficking, secretion and vesicular transport, and 4 were poorly characterized functions. The fifth species grouped 2 isolates (T. sp. EXT12c and T. sp. EXT13c). 156 SPCs were found, among them, 14 corresponded to specific functions related to AA metabolism, 2 to carbohydrate metabolism, 2 to coenzyme metabolism, 2 to mobilome (= mobile genetic elements) and 1 was poorly characterized. To finish, the sixth species was composed of the isolate *Thermococcus sp.*, of which 36 SPCs displayed 1 specific function related to replication, recombination and repair, 1 to mobilome and 1 was annotated as poorly characterized (Table 7).

Table 7: Specific protein clusters and classification of specific functions

Clades		AMTc	MC	IRI	*T.* sp. 4557	*T. celericrecens*	IRI sp. 1	IRI sp. 2	IRI sp. 3	Nautili-1	Nautili-2	Nautili-3	Nautili-4	Nautili-5	Nautili-6	COG classification
Number of SPC		153	283	53	148	336	68	73	71	15	138	1	124	165	36	
COG functions	A												1			RNA processing and modification
	B				1											Chromatin structure and dynamics
	C		5		3	2										Energy production and conversion
	D	1			1											Cell cycle control, cell division, chromosome partitioning
	E	3	29								5			14		Amino acid transport and metabolism
	F		3													Nucleotide transport and metabolism
	G		3		2	2					1		3	2		Carbohydrate transport and metabolism
	H		2								1			2		Coenzyme transport and metabolism
	I															Lipid transport and metabolism
	J		2						1		2					Translation, ribosomal structure and biogenesis
	K	1	1													Transcription
	L					1									1	Replication, recombination and repair
	M			1	1	2										Cell wall/membrane/envelope biogenesis
	N															Cell motility
	O					2					1					Post translational modification, protein turnover, chaperones
	P		5								1		1			Inorganic ion transport and metabolism
	Q	1			2											Secondary metabolites biosynthesis, transport and catabolism
	R	4			2	3					4		1	1	1	General function prediction only
	S		1			4	1				3		3			Function unknown
	T					1										Signal transduction mechanisms
	U												1			Intracellular trafficking, secretion, and vesicular transport
	V					1										Defense mechanisms
	W												1			Extracellular structures
	X					1								2	1	Mobilome: prophages, transposons
	Y															Nuclear structure
	Z	1														Cytoskeleton

4) Discussion

In this work, we studied the consequences of differentiation between closely related isolates at the genomic level. We sequenced 45 new *Thermococcus* strains, present on two different branches on the *Thermococcales* phylogenetic tree (Figure 24A). We assembled 33 complete or near complete genomes, and 12 are fragmented in 2 two 57 contigs. The strains's genome size is around 2 Mbp and they average 2100 genes. These values are in agreement with the other *Thermococcales* genomes. The *Thermococcus* genus is known to host a wide diversity of mobile genetic elements (for example (Gorlas et al., 2013a; Kuprovic et al., 2013)). In genome sequenced for this study, we found 19 potential mobile genetic elements. Interestingly, they are only present in genomes from the group II, which are close to *T. nautili*. In a 2004 survey where authors investigated the *Thermococcales* molecular diversity in EPR 13°N, they detected extracellular elements in 36 strains among 70 new isolates (Lepage et al., 2004). This can suggest that these strains use mobile genetic elements to share genetic information.

We wanted to form groups of closely related *Thermococcus* with strains we selected, so we used a phylogenetic tree to organize them, as well as all *Thermococcales* genomes available. To build a robust phylogenetic tree, we decided to employ the largest gene set shared by all organisms, thus only using genes in single copy. We conducted a pan-genome analysis of all *Thermococcales* genomes available. This first pan-genome also allowed us to carry genus wide observations. In general for a pan-genomics study, the more genomes included, the smaller the core genome is. In contrast, the accessory genome increases in size (Tettelin et al., 2005; Touchon et al., 2009). Here the accessory genome seemed highly structured. We divided it in three. The first part was composed of 682 high frequency PCs. We supposed that this is likely due to the loss of these genes in a limited number of genomes. Then 2 437 PCs could be classified as "medium-

frequency" within this pan-genome. In this case, it could be genes that were in the process of being lost from those genomes (Cordero and Polz, 2014). The largest part of this pangenome was composed by the 9 354 PCs classified as "low-frequency" that looked like singletons. This could be attributed to a gene acquisition from horizontal transfer, or genes temporary present in genomes. In a previous *Thermococcales* comparative genomics study with 21 genomes, authors found 790 core genes, among them, 668 were present in single-copy (Cossu et al., 2015). Compared to our results, 778 core genes and 602 SCGs, the addition of about 100 genomes did not significantly decrease the size of the core-genome. We have no reference concerning the accessory genome size.

While methods exist to provide universal set of genes to reconstruct a phylogeny (Darling et al., 2014; Rinke et al., 2013; Wu and Eisen, 2008), we decided to use SCGs from the 114 *Thermococcales* genomes available. Our SCGs set provided a richer collection with 602 genes chose without *a priori*. We also selected SCGs because working with universal genes sets presented the risk of having duplicated or absent genes from certain genomes. A surprising result from this phylogenomic analysis was that *Thermococcus* genus is not monophyletic. Indeed, the two *Palaeococcus* genomes were branched within the *Thermococcus* genus. The main difference between the two *Thermococcus* clades is the G+C content. The clade with the higher strain number had an average G+C content of 55%, whereas the other clade had an average G+C content of 41%. At the species level, the difference in G+C should be less than 3% (Meier-Kolthoff et al., 2014), but no thresholds are available for genus delineation. This *Thermococcus* low G+C clade either might represent a new genus or may be reclassified as *Palaeococcus*. This should be the subject of a future study.

Then we investigated whether the difference of geographic origin, and therefore gene flow reduction or absence, can contribute to the formation of distinct monophyletic clades. Concerning group I, isolates were effectively clustered according to this criterion. *Thermococcus* isolates biogeographic clustering has already been investigated at small or larger scale. These studies showed a relationship between *Thermococcus* clones and the vent chemistry. It, also suggested the colonization of multiple ecological niches within the same hydrothermal vent (Huber et al., 2006; Lepage et al., 2004; Price et al., 2015). Conversely, group II isolates did not follow a clustering pattern according to the geographic origin. This can suggest the presence of gene flow between EPR 9°N and EPR 13°N hydrothermal sites.

Based on these results, we investigated the presence of microbial species within both groups. In other words, we tried to determine whether group I and group II represented species, populations, or multiple species? For each group, we used both ANI and DDH with thresholds 96% and 70% respectively. These metrics should delineate microbial species from genomic data (Auch et al., 2010a, 2010b; Konstantinidis and Tiedje, 2005; Rosselló-Mora, 2006; Stackebrandt et al., 2002). On group I isolates, 5 clusters were congruent with both methods, so we considered them as 5 species. The two remaining clusters, composed of 10 *T.* sp. IRI isolates, were not congruent by ANI and DDH. Indeed, by ANI, they represented only one species, even if isolates *T.* sp. IRI24c, *T.* sp. IRI26c, *T.* sp. IRI27c2 and *T.* sp. IRI29c were closer together than *T.* sp. IRI07c, *T.* sp. IRI09c, *T.* sp. IRI10c, *T.* sp. IRI14c, *T.* sp. IRI15c and *T.* sp. IRI25c, and *vice versa*. ANI values were close to 100% within those groups, and around 96.5% between them. This may evoke the final phase of a speciation event. This was confirmed when looking at DDH results. They showed that the two latter clades were separated. However, one exception remained: *T.* sp. IRI25c was above the threshold (DDH value equal to 70.1%), which supports the

hypothesis that these two species are at the end of speciation process. Nevertheless, these two clades were each monophyletic in the phylogenetic tree. These combined results lead us to think that each clade is a species that seemed to diverge in a sympatric way, *i.e.*, species that are present within the same environment. This suggests the presence of distinct ecological niches within DSHV chimneys colonized by closely related *Thermococcales* lineages harboring few distinct genes in relation to different ecological contrains.

Isolates of the group II represented 6 species based on both methods. To reinforce this result, all species corresponded to monophyletic clades. But here according to the SPC number found in each species that belong to group II, the species definition is too loose for clades we called “Nautili-1” and ‘Nautili-3”. Only 15 and 1 SPCs were found respectively, while 128, 124, 165 and 36 SPCs were found within “Nautili-2”, “Nautili-4”, “Nautili-5” and “Nautili-6” respectively. The high isolates number within the species could also explain this difference for “Nautili-1” and “Nautili-3” compared to the 4 other.

Lastly, we indentified SPCs for each microbial species within each group. This was analyzed in order to learn more about selection pressures applied on these isolates. W did not take into account redundant functions from SPCs (= functions that were also found on the rest of the pan-genome). What emerged from these unique functions is that the amino-acid metabolism seemed to be a major factor for differentiation between species. In group I, MC isolates remarkably harbored complete tryptophan and nearly complete histidine biosynthesis pathways, whereas AMTc isolates had symporters for alanine or glycine. In the group II, “Nautili-5” harbors the complete tryptophan biosynthesis pathway too. A recent transcriptomic study showed that within a piezophilic microorganism, the tryptophan biosynthesis was down regulated when the strain was cultivated at 25 MPa compared to atmospheric pressure (Amrani et al.,

2014). This would mean that MC isolates, from shallow hot springs, can synthesize tryptophan while EXT12c and EXT13c isolates (Nautili-5) from deep-sea hydrothermal vents, can no longer synthesize this amino acid. This could be related to the higher energy cost of this pathway, no longer sustainable under hydrostatic pressure stress. Moreover, in *Pyrococcus*, proteins from the piezophilic *P. abyssi* are composed of relatively fewer large amino acids, such as tryptophan or tyrosine, compared to proteins of the non-piezophilic *P. furiosus*. (Di Giulio, 2005). This characteristic of proteins from piezophilic organisms could be verified in studied genomes by comparing the amino acid composition of SCGs.

Within all conspicuous SPCs identified in this study, 50 to 65% of PCs did not have annotations, whereas these genes are likely to have key roles in differentiation and adaptation between *Thermococcus* studied here. To study these unknown genes in more details, sequence similarity networks should be the next step to explore the presence of these PCs in other organisms (Atkinson et al., 2009; Meng et al., 2017). In addition, phylogentic reconstruction of SPCs orthologous genes should also inform us about their origin, e.g. gene loss in other clades or aquisition through horizontal gene transfer.

5) Conclusion

From two sets of isolates belonging to *Thermococcus*, we built a phylogenomic tree based on core-gene present in single copy across all known genomes to date. From this tree, the genus *Thermococcus* was not monophyletic, and further studies are necessary to better delineate *Thermococcales* phylogeny.

Contrary to biogeographic studies on 16S rRNA gene and ITS sequences, isolates were not necessarily grouped according to their geographical origin, and surprisingly, one species may be present at several hydrothermal sites. This result was obtained with complete genome sequence of closely related organisms. From the SPCs highlighted as

unique to the defined clades (species or location), these isolates are mainly differentiated based on amino acids metabolism, energy and carbohydrates metabolism genetic potentials. Despite the high level of unannotated proteins, it would be interesting to take their annotation further, using another approach such as similarity network.

To continue this work, it will be interesting to identify genomic context of these SPCs, to determine for example wether they are structured in operons or genomic islands, or whether they originate from horizontal transfers, and if it is the case, from which organism they acquired their genes.

6) Materials and Methods

For all isolates, genomic DNA was extracted as in the chapter "Screening culture collection for *Thermococcus*". Illumina libraries were prepared using the kit *TruSeq*® *DNA PCR-free lib prep* according to recommendation of the manufacturer, and 550 bp as insert size. Sequencing was performed at the Marine Biological Laboratory (Woods Hole, MA, USA), on an Illumina MiSeq system and MiSeq v3 reagent kit to get 300 bp paired-end reads.

Read quality was assessed with *illumina-utils* with the command *iu-filter-quality-minoche* (Eren et al., 2013; Minoche et al., 2011), and genome assembly was carried out with CLC Genomics Workbench v8.5.1 (https://www.qiagenbioinformatics.com/products/clc-genomics-workbench/), using size of *k-mer* ranging from 21 (default) to 63 (system limitation).

Pangenomics pipeline was carried out using anvi'o v2.3.2 following the author's suggestion available here: http://merenlab.org/2016/11/08/pangenomics-v2/ (Eren et al., 2015). Briefly, we predict CDS with Prodigal v2.6.2, and annotate each CDS thanks to COG database through DIAMOND v0.9.8 (Buchfink et al., 2015; Galperin et al., 2014;

Hyatt et al., 2010). To assess the completion and redundancy of genomes, we used HMMER v3.1b2 and a set of 162 archaeal single copy genes (Mistry et al., 2013; Rinke et al., 2013). Then protein clusters (PC) are computed based on all-against-all BLASTP (v2.2.31+), based on these results a graph is build and resolved with MCL and inflation parameter (--*mcl-inflation* in anvi'o, -I in MCL) set to 6 for all *Thermîococcales* pangenomics and 8 for group specific pangenomics (Camacho et al., 2009; Van Dongen and Abreu-Goodger, 2012). All pangenomes are displayed with the command *anvi-display-pan* in anvi'o and InkScape v0.91 was used to for figures refinement (https://inkscape.org/en/).

Phylogenomic trees were was built by maximum likelihood with concatenation of single copy core-genes previously aligned with MAFFT v7.055b (parameters --*maxiterate 1000* --*localpair*) and trimmed with BMGE v1.12 default parameters (Criscuolo and Gribaldo, 2010; Katoh and Standley, 2013). PhyML v3 was used for the tree, as well as the SMS option to select the best evolution model, and aLRT as bootstrap method (Anisimova and Gascuel, 2006; Guindon et al., 2010; Lefort et al., 2017). All trees were visualized on iToL (Letunic and Bork, 2016). All gene were also annotated with the KEGG Automatic Annotation Server (KAAS) to recover metabolic pathways, using the Best BLAST Hit method and following list of organisms as reference: *hsa, dme, ath, sce, pfa, eco, sty, hin, pae, nme, hpy, rpr, mlo, bsu, sau, lla, spn, cac, mge, mtu, ctr, bbu, syn, aae, mja, afu, pho, ape, ton, tko, tga, tsi, tba, pab, pfu* (Moriya et al., 2007).

Species delineation was assessed with ANI, computed with OrthoANI v0.93 default parameters, and DDH, computed with GGDC v2.1 default parameters (Auch et al., 2010a, 2010b; Meier-Kolthoff et al., 2014). All heatmap were built with R v3.2.2 (R Core Team, 2015) and the package *gplots* v3.0.1 and the function *heatmap.2*.

III) *In situ* distribution of *Thermococcales* by a metapangenomics approach

1) Abstract

Thermococcales represent a hyperthermophilic archaeal order mainly found in hydrothermal vent environments. It may spread into the environment through marine currents and colonizes new hydrothermal systems. Its distribution is however of particular interest given its hyperthermophilic and strict anaerobic lifestyle limiting its dispersion by and survival in seawater. In this study, we wanted identify *Thermococcales* strains distribution in the environment by mapping published and available deep-sea hydrothermal vent and terrestrial hot spring metagenomes or metatranscriptomes on the *Thermococcales* genome collection established during thos work. From these mapping results, we detected more signals when metagenomes wrer obtained from a strain isolation site. We also identified two remarkable distribution patterns: either the genomes are found only in one single location, or they can be detected in several metagenomes of very diverse geographical origins. This is the case for example for *Thermococcus cleftensis*, which was detected in metagenomes from the Juan de Fuca Ridge and the Cayman Rise. In the future, with more metagenomes, this work will allows us to better indentify distribution of *Thermococcales* species, but also to give clues about the presence in the environment of less frequent strains such as ones of the *Palaeococcus* genus.

2) Introduction

Metagenomics is the study of DNA sequences found in the environment, without any *a priori* about the organism that owns this sequence. This tool revolutionized microbial ecology, with a tremendous amount of data available, but these data are harder to grasp owing to their complexity relative genomic data.

Deep-sea hydrothermal environments are also studied thanks to metagenomics. In this extreme environment, the archaeal order *Thermococcales* is widespread. These hyperthermophilic *Archaea* are among the first colonizers of newly formed hydrothermal chimney (McCliment et al., 2006; Nercessian et al., 2003; Pagé et al., 2008; Reysenbach et al., 2000b), they are detected within the first 4-5 days. They colonize walls of the chimney, where temperature is hot enough for their metabolism, but not too hot to kill them. In pure culture, the optimal growth temperature ranges from 75°C to 105°C (Callac et al., 2016; Dalmasso et al., 2016a). In general, they grow better and faster under high hydrostatic pressure in lab conditions.

Thermococcales are widespread in deep-sea hydrothermal vents, thus we investigated sites where they can be present using metagenomes collected at both deep-sea hydrothermal vents and terrestrial hot springs, and all available *Thermococcales* genomes. With this framework, we aimed to visualize the dispersion of strains/isolates in the environment, and see if there were any dispersal barriers.

3) Results

In this chapter, we mapped 354 sets of environmental data, metagenomes and metatranscriptomes, from deep-sea hydrothermal vents, shallow hydrothermal vents and terrestrial hot springs (Appendix 2). These metagenomes were collected at different places: 259 from Juan de Fuca (JdF) in North Pacific Ocean (AxialSeamount), 52 from the Cayman rise, 12 from Lau Basin in South Pacific Ocean, 3 from Menez Gwen at the Mid Atlantic Ridge (MAR), 2 from Lost City at the MAR, 3 from Loki's castle in North Atlantic, 3 from the Guaymas Basin, for the deep-sea context. Two other metagenomes were obtained from the bay of Prony in New Caledonia, from shallow and low temperature hydrothermal vents. Then, the remaining metagenomes were sampled from hot springs: 11 metagenomes sampled in the Yellowstone National Park, 3 from

Taïwan, 2 from South Africa, 1 from Malaysia and 1 from Oklahoma (USA) (Appendix 2). All these metagenomes were mapped onto 102 *Thermococcales* genomes distributed as follow: 1 *Palaeococcus*, 18 *Pyrococcus* and 83 *Thermococcus*. These 102 *Thermococcales* genomes corresponded to the 46 genomes sequenced during this project, the 26 unpublished genomes (7 *Pyrococcus* and 19 *Thermococcus*), and all 30 published genomes publicly available at the time of the study (1 *Palaeococcus*, 11 *Pyrococcus* and 18 *Thermococcus*).

For each genome, we looked for the percentage of detection of this genome across all metagenomes, or the proportion of a given genome that is covered at least 1X by mapped reads (Figure 28). From this analysis, sites have little or no detection for all genomes used. Like samples from JdF, hydrothermal fields Ashe and N3 (Ashe-2011, N3-2013) or samples from Lost City (LoscityMAR) and all samples collected in terrestrial hot spring (xxxHotSpring). In addition, the control sample from JdF in 2014 (CTD1200), *i.e.* seawater collected far from any hydrothermal vent, does not detect any *Thermococcales* genomes. Other sites exhibited much more detection and diversity of genome detected, like sites Anemone from JdF (2013_Anemone, 2014_Anemone), and Menez Gwen from Atlantic Ocean (MenewGwennMAR). The site Anemone was sampled 3 times, in 2012, 2013 and 2014. The number of genomes detected was higher in 2013 than in 2012 or 2014, but the genome of *Thermococcus cleftensis* has been detected the 3 times, with an average percentage of detection of 3% for 2012 and 2014, and 9% for 2013. In the site Menez Gwen, (MenezGwennMAR), all isolates "IRI" were well detected, from 1.5% for *Thermococcus* sp. IRI09c to 8.2% for BPK H *Thermococcus* sp IRI33c. To finish, both sites from the Cayman rise (CaymanVonDamm, CaymanPiccard) encompassed well two isolates, *Thermococcus cleftensis,* respectively 7% and 1.2%, and *Thermococcus piezophilus*, respectively 1.5% and 0.7%. In hydrothermal site Von Damm

from Cayman trench (CaymanVonDamm), two other *Thermococcus* were detected, *T. barophilus* and BPK L *Thermococcus* Cir10a, both with an average detection rate of 1.5%.

The genus *Palaeococcus* represented here by *P. pacificus* was only detected in samples from JdF, Dependable hydrothermal field (JdF_2013_Dependable). Here an average of 9.5% of the genome was detected, which was the highest detection value forall genomes. Genomes of *Pyrococcus* were not detected frequently. Two isolates were more detected than the others: *Pyrococcus* sp. ST04 at on average 5% and 3% in metagenomes from Anemone hydrothermal site in JdF (JdF_2013_Anemone, JdF_2014_Anemone). The second best detected *Pyrococcus* was BPI E *Pyrococcus* sp. IRI42c from Menez Gwen (MenezGwennMAR). An average of 1.5% of the genome was detected.

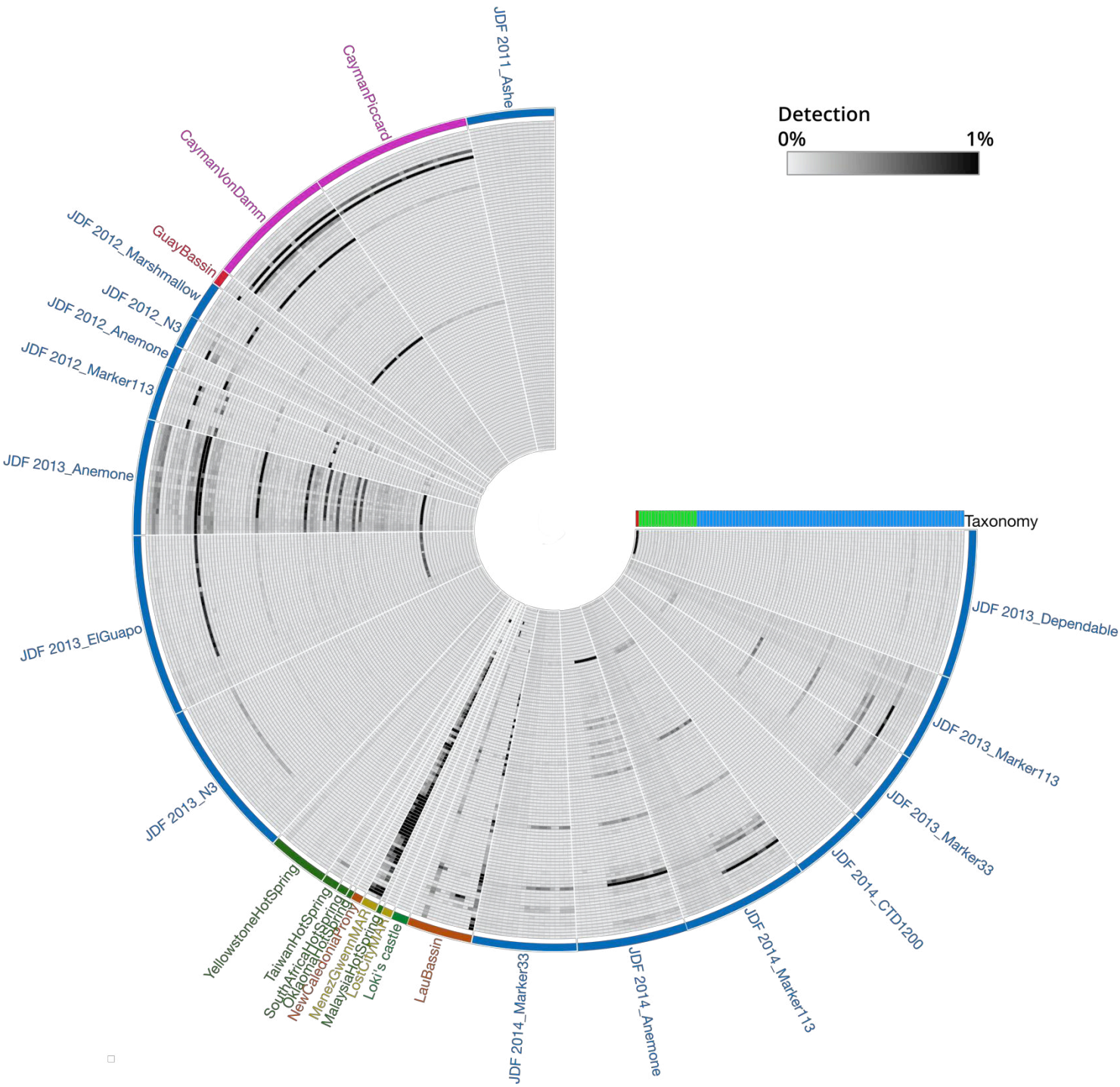


Figure 28: Mapping of metagenomes on *Thermococcales* genomes

This figure shows the detection of 102 *Thermococcales* genomes in 354 metagenomes (hydrothermal vent and terrestrial hot spring), that is the proportion of a given genome that is covered at least 1X. The colored scale ranges from 0% (light grey), to 1% of detection (black). Each layer of the disk represents a genome; they are ordered based on their taxonomy: Red = *Palaeococcus*, Green = *Pyrococcus*, Blue = *Thermococcus*. Each radius represents a metagenome or metatranscriptome, and replicates are grouped at the same place on the figure. JdF: Juan de Fuca.

4) Discussion

In this short study, the aim was to track the presence of *Thermococcales* in different hydrothermal places. Metagenomes and metatranscriptomes derived mainly from the North Pacific at Juan De Fuca ridge (Fortunato and Huber, 2016), and this might decrease the number of genomes detected, because many isolates with sequenced genomes come from other places like the Atlantic Ocean.

In metagenomes, no *Thermococcales* was detected, like at Loki's Castle, or at Lost City (LostCityMAR), whereas they were present according to studies based on the 16S rRNA genes (Brazelton et al., 2006; Jaeschke et al., 2012).

A parameter that we did not take into account here is the quantity of reads per metagenomes. We did not perform any normalization on the quantity of reads. This may add biases, because the more data there are in a metagenome, the greater is the chance of having detected genomes. This could explain why we detect traces of genomes (less than 0.001%) within terrestrial hot springs.

From this plot, it seems that some Thermococcales are widespread, while other only live in a unique place, like the *Palaeococcus* used here. This could explain why this genus is under-represented among *Thermococcales* isolates. But this *Palaeococcus pacificus* was characterized from the EPR 1°S hydrothermal vent (Zeng et al., 2013), suggesting the presence of *Palaeococcus* isolates in JdF. Moreover, one 16S rRNA gene sequence and ITS belonging to *Palaeococcus* was found in Axial Seamount from JdF (accession: AY559124), indicating their presence at this geographic location.

In samples from the Cayman trench, detection shed light on two genomes: *T. piezophilus* and *T. paralvinellae*. The first one was isolated from this place (site Piccard), while the second was isolated from a sample from JdF Endeavour (Dalmasso et al., 2016a;

Hensley et al., 2014). This tends to indicate that this strain might be able of moving over long distances, or that close isolates are present in these locations.

This archaeal order of *Thermococcales* acts as one of the first colonizers in newly formed hydrothermal vents (McCliment et al., 2006; Nercessian et al., 2003; Pagé et al., 2008; Reysenbach et al., 2000). A hypothesis on how they colonize is that they disseminate in the environment through oceanic currents and they reach a new site randomly (Wirth, 2017).

5) Conclusion

In this exploratory study, we aimed at tracking the presence of known *Thermococcales* in different geographical places. Despite the low detection of genomes, broad outlines emerged. These isolates are not present within terrestrial hot spring, certainly by the lack of a link between them and deep-sea environments. It is also likely that at temporal dynamics exists, as *Thermococcales* are not present at the same rate over the years. In some places, the diversity of *Thermococcales* is high, like in Menez Gwen (MenezGwenMAR), whereas other metagenomes did not captured sequences affiliated to *Thermococcales*. To finish, this work brings us clues on places where to search for new strains, like for *Palaeococcus* that seems to be present in the hydrothermal site Dependable at the Juan de Fuca ridge, whereas no isolates was characterized from this location.

6) Materials and Methods

Metagenomes and metatranscriptomes were recovered with SRAtoolkit v2.8.2: we downloaded raw data using the *prefetch* command with the SRA id or each metagenome as argument. Then we used *fastq-dump* with the parameter *--split-files* to convert raw file to two FastQ files (one file for each paired-end reads). Then each data set went through quality control with illumina-*utils* and command *iu-filter-quality-minoche* (Eren

et al., 2013; Minoche et al., 2011). Metagenomes and metatranscriptomes were mapped onto genomes with bowtie2 v2.2.9, default parameters (Langmead and Salzberg, 2012). Anvi'o was used to merge all mapping results, following profiling and merging parts of the tutorial available here: http://merenlab.org/2016/06/22/anvio-tutorial-v2/ (Eren et al., 2015). The data matrix about genomes detection was recovered from the command *anvi-summarize*. To obtain the figure presented in this chapter, we used the command *anvi-interactive*, with the transposed genomes detection matrix and a custom cladogram that grouped together all replicates for a metagenome.

V) Published data

During this project, 48 genomes were sequenced, some of which probably represent new species of *Thermococcus*. We published the genome of *Thermococcus* sp. EXT12c, isolated from a deep-sea hydrothermal vent rock sample at the EPR 9°N. Its closest relative is *T. nautili*, but the low DNA similarity values (ANI and DDH) with known strains strongly suggest that this isolate represents a new species. In comparative genomics study, it appeared in SPCs that this isolate had the complete biosynthesis pathway for tryptophan, an energy costly amino-acid synthesis pathway absent in several piezophilic microorganisms. Up to date, it was shown in a piezophilic bacterium that the synthesis of this amino acid is greatly reduced when the bacterium is cultivated under high hydrostatic pressure (Amrani et al., 2014). Concerning *Archaea*, a transcriptomics analysis carried out in *T. barophilus* and *T. kodakaraensis* revealed only a slight difference of expression for a few genes involved in this biosynthesis pathway. It would be interesting to explore this feature by transcriptomics analysis over a range of hydrostatic pressure culture conditions using a method with high sequencing depth.

jga-genome/jga99917/jga6112d17z | xppws | S=3 | 11/10/17 | 3:52 | ArtID: | DOI:10.1128/genomeA.01385-17 | CE: JJ

Editor: | Section: **Prokaryotes** | Designation:

AMERICAN SOCIETY FOR MICROBIOLOGY | genomeAnnouncements™

PROKARYOTES

Complete Genome Sequence of Hyperthermophilic Archeon *Thermococcus* sp. EXT12c, Isolated from East Pacific Rise 9°N

AQ:A

AQ: au Damien Courtine,[a,b,c] Karine Alain,[b,c,a] Myriam Georges,[a,b,c] Nadège Bienvenu,[a,b,c] Hilary G. Morrison,[d] A. Murat Eren,[d,e] Loïs Maignien[a,b,c,d]

AQ: aff Université de Bretagne Occidentale (UBO, UBL), Institut Universitaire Européen de la Mer (IUEM), UMR6197, Laboratoire de Microbiologie des Environnements Extrêmes (LM2E), Plouzané, France[a]; CNRS, IUEM–UMR 6197, Laboratoire de Microbiologie des Environnements Extrêmes (LM2E), Plouzané, France[b]; Ifremer, UMR6197, Laboratoire de Microbiologie des Environnements Extrêmes (LM2E), Plouzané, France[c]; Josephine Bay Paul Center for Comparative Molecular Biology and Evolution, Marine Biological Laboratory, Woods Hole, Massachusetts, USA[d]; Department of Medicine, University of Chicago, Chicago, Illinois, USA[e]

ABSTRACT We report the genome sequence of *Thermococcus* sp. EXT12c isolated from a deep-sea hydrothermal vent at the East Pacific Rise 9°N. Microbes in the genus *Thermococcus* are able to grow anaerobically at high temperature, around neutral pH, and some of them under high hydrostatic pressure.

Received 4 November 2017 **Accepted** 6 November 2017 **Published** XXX

Citation Courtine D, Alain K, Georges M, Bienvenu N, Morrison HG, Eren AM, Maignien L. 2017. Complete genome sequence of hyperthermophilic archeon *Thermococcus* sp. EXT12c, isolated from East Pacific Rise 9°N. Genome Announc 5:e01385-17. https://doi.org/10.1128/genomeA.01385-17.

Address correspondence to Loïs Maignien, lois.maignien@univ-brest.fr.

We isolated *Thermococcus* sp. EXT12c from a hydrothermal chimney rock sample collected from 2496-m depth near the East Pacific Rise 9°N (9°50′40.2″N; 104°17′37.798″W) during the oceanographic cruise EXTREME (Oct. 2001). *T.* sp. EXT12c is able to grow under anaerobic conditions, at 85°C and pH 6.8 in TRM medium (1). The strain is available in the UBOCC culture collection (Brest, France) under the reference UBOCC-M-2417.

We used a phenol-chloroform technique for DNA extraction and the TruSeq DNA PCR-free kit (Illumina, USA) to prepare paired-end sequencing libraries with an average insert size of 550 nt. Whole-genome sequencing at the Marine Biological Laboratory (Woods Hole, MA, USA) using an Illumina MiSeq machine (MiSeq reagent kit v3) produced 1,335,523 2 × 300 bp reads after quality filtering (2). Our *de novo* assembly with CLC Genomics Workbench v8.5.1 (https://www.qiagenbioinformatics.com/products/clc-genomics-workbench) resulted in a single chromosome with 2,155,760 nt and a GC content of 54.58%. This single chromosome recruited 98.8% of the short reads, with an average coverage of 350×.

The MaGe genome annotation platform (3–14) identified 2,365 coding sequences, a single 16S-23S operon, two 5S rRNA, 46 tRNA, and 16 miscellaneous RNA genes. InterProScan identified one integrase, 6 transposases, and 3 clustered regularly interspaced short palindromic repeat (CRISPR) loci associated with *cas* genes (*cas*, *cst*, and *cmr*), suggesting that the strain probably carries two types of CRISPR systems, class I type I and type III (15). These features suggest that the strain has a certain genomic plasticity.

Among the *Thermococcus* species with published genomes, *T.* sp. EXT12c is most closely related to *T. nautili* strain 30-1^T (16). These genomes have a DNA-DNA hybridization value of 43.6% and an average nucleotide identity of 91.18%, as predicted with GGDC v2.1 (17–19) and OrthoANI v1.20 (20), respectively.

T. sp. EXT12c possesses some complete metabolic pathways like the glycolysis and amino acid biosynthesis pathways for alanine, asparagine, glycine, glutamate, and tryptophan. To date, only ten *Thermococcales*, including *T. kodakaraensis*, *T. litoralis*,

jga-genome/jga99917/jga6112d17z	xppws	S=3	11/10/17	3:52	ArtID:	DOI:10.1128/genomeA.01385-17	CE: JJ

Pyrococcus furiosus, and *P. abyssi* (21–25), harbor a complete tryptophan biosynthesis pathway.

In this metabolic pathway, a single locus contains genes leading to the synthesis of chorismate, an intermediate for multiple metabolic pathways (TEXT12C_2159 to TEXT12C_2167), and tryptophan (TEXT12C_2174 to TEXT12C_2168, genes *trpCDEGFBA*). Within the *T. kodakaraensis* genome, genes that code for tryptophan biosynthesis are present in a single locus too (26). A transcriptomic study showed that the chorismate synthesis is downregulated when *T. kodakaraensis* is cultivated under high hydrostatic pressure, compared to atmospheric pressure (27). However, in the same study, the gene *trpC*, labeled TK0252 in *T. kodakaraensis*, is upregulated, indicating that the strain could continue to produce tryptophan and compensate the decrease of chorismate production under high pressure. Therefore, the regulation of tryptophan biosynthesis in *T.* sp. EXT12c under high-pressure conditions requires further investigations.

Accession number(s). This genome sequence has been deposited in DDBJ/ENA/GenBank under the accession no. LT900021. The version described in this paper is the first version.

ACKNOWLEDGMENTS

This work was supported by the "Laboratoire d'Excellence" LabexMER (ANR-10-LABX-19) and co-funded by a grant from the French Government to L.M. under the program "Investissements d'Avenir," a grant from the Regional Council of Brittany to L.M., and the Frank R. Lillie Research Innovation Award from the Marine Biological Laboratory to A.M.E.

REFERENCES

1. Zeng X, Birrien J-L, Fouquet Y, Cherkashov G, Jebbar M, Querellou J, Oger P, Cambon-Bonavita M-A, Xiao X, Prieur D. 2009. *Pyrococcus* CH1, an obligate piezophilic hyperthermophile: extending the upper pressure-temperature limits for life. ISME J 3:873–876. https://doi.org/10.1038/ismej.2009.21.
2. Eren AM, Maignien L, Sul WJ, Murphy LG, Grim SL, Morrison HG, Sogin ML. 2013. Oligotyping: differentiating between closely related microbial taxa using 16S rRNA gene data. Methods Ecol Evol 4:1111–1119.
3. Bocs S, Cruveiller S, Vallenet D, Nuel G, Médigue C. 2003. AMIGene: annotation of MIcrobial genes. Nucleic Acids Res 31:3723–3726. https://doi.org/10.1093/nar/gkg590.
4. Gardner PP, Daub J, Tate JG, Nawrocki EP, Kolbe DL, Lindgreen S, Wilkinson AC, Finn RD, Griffiths-Jones S, Eddy SR, Bateman A. 2009. Rfam: updates to the RNA families database. Nucleic Acids Res 37:D136–D140. https://doi.org/10.1093/nar/gkn766.
5. Lagesen K, Hallin P, Rødland EA, Staerfeldt H-H, Rognes T, Ussery DW. 2007. RNAmmer: consistent and rapid annotation of ribosomal RNA genes. Nucleic Acids Res 35:3100–3108. https://doi.org/10.1093/nar/gkm160.
6. Lowe TM, Eddy SR. 1997. TRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. Nucleic Acids Res 25:955–964. https://doi.org/10.1093/nar/25.5.0955.
7. Claudel-Renard C, Chevalet C, Faraut T, Kahn D. 2003. Enzyme-specific profiles for genome annotation: PRIAM. Nucleic Acids Res 31:6633–6639. https://doi.org/10.1093/nar/gkg847.
8. Bendtsen JD, Nielsen H, von Heijne G, Brunak S. 2004. Improved prediction of signal peptides: SignalP 3.0. J Mol Biol 340:783–795. https://doi.org/10.1016/j.jmb.2004.05.028.
9. Gardy JL, Laird MR, Chen F, Rey S, Walsh CJ, Ester M, Brinkman FSL. 2005. PSORTb v.2.0: expanded prediction of bacterial protein subcellular localization and insights gained from comparative proteome analysis. Bioinformatics 21:617–623. https://doi.org/10.1093/bioinformatics/bti057.
10. Hunter S, Apweiler R, Attwood TK, Bairoch A, Bateman A, Binns D, Bork P, Das U, Daugherty L, Duquenne L, Finn RD, Gough J, Haft D, Hulo N, Kahn D, Kelly E, Laugraud A, Letunic I, Lonsdale D, Lopez R, Madera M, Maslen J, McAnulla C, McDowall J, Mistry J, Mitchell A, Mulder N, Natale D, Orengo C, Quinn AF, Selengut JD, Sigrist CJA, Thimma M, Thomas PD, Valentin F, Wilson D, Wu CH, Yeats C. 2009. InterPro: the integrative protein signature database. Nucleic Acids Res 37:D211–D215. https://doi.org/10.1093/nar/gkn785.
11. Karp PD, Paley S, Romero P. 2002. The Pathway Tools software. Bioinformatics 18:S225–S232. https://doi.org/10.1093/bioinformatics/18.suppl_1.S225.
12. Sonnhammer EL, von Heijne G, Krogh A. 1998. A hidden Markov model for predicting transmembrane helices in protein sequences. Proc Int Conf Intell Syst Mol Biol 6:175–182.
13. Tatusov RL, Fedorova ND, Jackson JD, Jacobs AR, Kiryutin B, Koonin EV, Krylov DM, Mazumder R, Mekhedov SL, Nikolskaya AN, Rao BS, Smirnov S, Sverdlov AV, Vasudevan S, Wolf YI, Yin JJ, Natale DA. 2003. The COG database: an updated version includes eukaryotes. BMC Bioinformatics 4:41. https://doi.org/10.1186/1471-2105-4-41.
14. Bairoch A, Apweiler R, Wu CH, Barker WC, Boeckmann B, Ferro S, Gasteiger E, Huang H, Lopez R, Magrane M, Martin MJ, Natale DA, O'Donovan C, Redaschi N, Yeh L-SL. 2005. The universal protein resource (UniProt). Nucleic Acids Res 33:D154–D159. https://doi.org/10.1093/nar/gki070.
15. Makarova KS, Wolf YI, Alkhnbashi OS, Costa F, Shah SA, Saunders SJ, Barrangou R, Brouns SJJ, Charpentier E, Haft DH, Horvath P, Moineau S, Mojica FJ, Terns RM, Terns MP, White MF, Yakunin AF, Garrett RA, van der Oost J, Backofen R, Koonin EV. 2015. An updated evolutionary classification of CRISPR-Cas systems. Nat Rev Microbiol 13:722–736. https://doi.org/10.1038/nrmicro3569.
16. Oberto J, Gaudin M, Cossu M, Gorlas A, Slesarev A, Marguet E, Forterre P. 2014. Genome sequence of a hyperthermophilic archaeon, *Thermococcus nautili* 30–1, that produces viral vesicles. Genome Announc 2(2):e00243-14. https://doi.org/10.1128/genomeA.00243-14.
17. Auch AF, von Jan M, Klenk H-P, Göker M. 2010. Digital DNA-DNA hybridization for microbial species delineation by means of genome-to-genome sequence comparison. Stand Genomic Sci 2:117–134. https://doi.org/10.4056/sigs.531120.
18. Auch AF, Klenk H-P, Göker M. 2010. Standard operating procedure for calculating genome-to-genome distances based on high-scoring segment pairs. Stand Genomic Sci 2:142–148. https://doi.org/10.4056/sigs.541628.
19. Meier-Kolthoff JP, Klenk H-P, Göker M. 2014. Taxonomic use of DNA G+C content and DNA-DNA hybridization in the genomic age. Int J Syst Evol Microbiol 64:352–356. https://doi.org/10.1099/ijs.0.056994-0.
20. Lee I, Kim YO, Park S-C, Chun J. 2015. OrthoANI: an improved algorithm and software for calculating average nucleotide identity. Int J Syst Evol Microbiol 66:1100–1103. https://doi.org/10.1099/ijsem.0.000760.
21. Maeder DL, Weiss RB, Dunn DM, Cherry JL, González JM, DiRuggiero J,

jga-genome/jga99917/jga6112d17z	xppws	S=3	11/10/17	3:52	ArtID:	DOI:10.1128/genomeA.01385-17	CE: JJ

Robb FT. 1999. Divergence of the hyperthermophilic archaea *Pyrococcus furiosus* and *P. horikoshii* inferred from complete genomic sequences. Genetics 152:1299–1305.

22. Bridger SL, Lancaster WA, Poole FL, Schut GJ, Adams MWW. 2012. Genome sequencing of a genetically tractable *Pyrococcus furiosus* strain reveals a highly dynamic genome. J Bacteriol 194:4097–4106. https://doi.org/10.1128/JB.00439-12.
23. Fukui T, Atomi H, Kanai T, Matsumi R, Fujiwara S, Imanaka T. 2005. Complete genome sequence of the hyperthermophilic archaeon *Thermococcus kodakaraensis* KOD1 and comparison with *Pyrococcus* genomes. Genome Res 15:352–363. https://doi.org/10.1101/gr.3003105.
24. Gardner AF, Kumar S, Perler FB. 2012. Genome sequence of the model hyperthermophilic archaeon *Thermococcus litoralis* NS-C. J Bacteriol 194:2375–2376. https://doi.org/10.1128/JB.00123-12.
25. Cohen GN, Barbe V, Flament D, Galperin M, Heilig R, Lecompte O, Poch O, Prieur D, Quérellou J, Ripp R, Thierry J-C, Van der Oost J, Weissenbach J, Zivanovic Y, Forterre P. 2003. An integrated analysis of the genome of the hyperthermophilic archaeon *Pyrococcus abyssi*. Mol Microbiol 47:1495–1512. https://doi.org/10.1046/j.1365-2958.2003.03381.x.
26. Tang X, Ezaki S, Fujiwara S, Takagi M, Atomi H, Imanaka T. 1999. The tryptophan biosynthesis gene cluster trpCDEGFBA from *Pyrococcus kodakaraensis* KOD1 is regulated at the transcriptional level and expressed as a single mRNA. Mol Gen Genet 262:815–821. https://doi.org/10.1007/s004380051145.
27. Vannier P, Michoud G, Oger P, Marteinsson VÞ, Jebbar M. 2015. Genome expression of *Thermococcus barophilus* and *Thermococcus kodakarensis* in response to different hydrostatic pressure conditions. Res Microbiol 166:717–725. https://doi.org/10.1016/j.resmic.2015.07.006.

General synthesis

General synthesis

This research has been designed to study the ecology of microorganisms in deep environments using high throughput approaches such as complete genome sequencing, metagenomics, and metatranscriptomics. In this "ecogenomics" framework, I have been interested in the genomic diversity of closely related *Thermococcus* isolates. They are defined as closely related based on a phylogenetic point of view. We wanted to explore hypotheses about the mechanisms that influence the diversification of these genomes, and also learn more about the selection pressures that such organisms with restricted ecological niches can face in deep-sea hydrothermal environments.

The first phase of this work was to use the culture collection available in the laboratory, UBOCC. The marine part of the collection consists of approximately 1,300 isolates collected from various marine samples collected over the years during oceanographic cruises. Of these isolates, about 300 were annotated as belonging to the *Thermococcales* order. This order is composed of hyperthermophilic *Archaea* mainly found in deep marine hydrothermal vents. This assignment was based on morphological and cultural criteria, *i.e.* isolates that grow in anaerobic conditions at a temperature of about 85°C, on a medium rich in organic matter, mobile and having a coccoid morphology. In order to confirm and refine this first assignment, we used a ubiquitous marker: the gene coding for the small ribosome subunit RNA, *i.e.* 16S rRNA. This marker, once sequenced, was used to construct a phylogenetic tree to infer taxonomic affiliation at the genus level. All isolates were cultured and incubated at 80 or 85°C for 16 to 18 hours. Then the DNA of each isolate was extracted and we amplified by PCR the sequence of the gene coding for 16S rRNA, as well as the sequence between 16S-23S rRNA genes, called Internal Transcribed Spacer (ITS). The interest of sequencing the ITS is to have an additional marker to build a more robust phylogenetic tree. In total, for each isolate, 3

sequencings according to the Sanger method were required to obtain the complete sequence of each "16S-ITS". The sequences were assembled and their quality checked. A total of 273 sequences were completed and suitable to proceed to the next stage of phylogenetic tree construction. Sequences of 16S-ITS from the representatives of the three *Thermococcales* genera (*Pyrococcus*, *Thermococcus*, and *Palaeococcus*) were downloaded from the public databases. These and the 273 sequences obtained during this work were aligned and a tree was constructed. Figure 20 shows a simplified version of this tree. Of the 273 isolates, 14 were classified as belonging to the genus *Pyrococcus* and the remaining 259 were affiliated to *Thermococcus* (Appendix 1). Finally, from this tree, two groups of genomes were selected according to the following criteria: (i) several geographical origins, (ii) several genomes for the same geographical origin, (iii) monophyletic group if possible. Following these criteria, the first selected group included 21 isolates from the East Pacific Ridge 13°N (EPR 13°N), the Rainbow hydrothermal field located on the Mid-Atlantic Ridge, and Saint-Paul Island located in South Indian Ocean (Figure 21). The second group contained 27 genomes close to *Thermococcus nautili*, which originated from EPR 9°N and EPR 13°N in the Pacific Ocean (Figure 21).

The second part of my thesis began with the sequencing of the 48 genomes selected. All DNA samples were extracted at the LM2E. Half of the genomes were sent directly to the Marine Biological Laboratory (MBL) in Woods Hole, USA, for sequencing. For the second half of the genomes, the Illumina sequencing libraries were prepared at the LM2E and sent to the MBL for sequencing. We have chosen to sequence on Illumina MiSeq in pair end 2x300 bp. I then assembled all genomes using CLC Genomics Workbench, using different sizes of *k-mer*. A total of 46 genomes were successfully assembled, of which 19 were successfully assembled within a single circular contig, 15 within a non-

circularized contig and 13 genomes remained fragmented (2-57 contigs). Sequencing failed for a group I genome (MC5), and it appeared that the genome of *Thermococcus* sp. E15P25 (Group II) was contaminated, two genomes seeming to be present, making its use impossible.

In a second phase, we built a phylogenomic tree to put all these genomes within an evolutionary context. The phylogeny realized was based on the single-copy core-genome genes, because it is a rich data set defined without *a priori* on the genes function. To obtain this gene set, it was necessary to carry out a pangenomics analysis, which made it possible to define on the one hand, all the genes shared by all the genomes studied (= core-genome) and on the other hand, all the other genes, *i.e.* the accessory genome. The union of these two categories formed the pangenome. The latter was established with anvi'o. Briefly, we identified coding sequences (CDS) in all genomes and then compared all the sequences. This result was then provided to MCL, an algorithm that aggregates genes *via* a graph approach. A list of clusters of genes (PC, Protein Clusters) emerged. In a PC, if there was a gene of each genome, that PC belonged to the core-genome. When a PC in the core genome contained a unique gene from each of the genomes, this PC belonged to the "single copy core-gene (SCGs) genes". All non-core PCs were attributed to the accessory genome. Figure 23 shows a representation of this pangenome and the distribution of accessory PCs. After this step, the 602 SCGs were extracted, aligned, trimmed and concatenated. This alignment of approximately 92,000 non-redundant positions has been used to construct the phylogeny of these 114 genomes, using the maximum likelihood method (Figure 24A). In this tree, the 21 isolates of Group I formed a monophyletic clade. Two genomes from the public databases were added to this group: *Thermococcus celericrescens* and *T.* sp. 4557. Group II was composed of 34 genomes: 25 sequenced in this project, 1 reference genome (*T.*

nautili) and 8 unpublished genomes provided by the MBGE laboratory (Pasteur Institute). This group was supposed to contain 26 of the isolates sequenced during this work, but the strain *T.* sp. IRI06c was branched elsewhere in the tree. After verification, the 16S rRNA gene sequence in the genome and the sequence obtained in the first part of the thesis differed by 15 nucleotides indicating that it was not the same isolate. This genome was therefore simply retained for the global pangenome, but was not used for the rest of this work.

As this tree confirms the existence of these two distinct groups, we questioned the nature of the parameters that could explain the organization of isolates in each group. Initially, the impact of geographical origin was analyzed. For group I, this factor alone explained the organization of these genomes in clades (group with a common ancestor) in this tree (Figure 24B). The genomes of group II did not follow this trend, as the geographical origin did not explain their organization in this tree. The second parameter studied was the presence of microbial species *via* the two metrics ANI (average nucleotide identity between genomes) and DNA hybridization (DDH). Group I was composed of 7 species according to these two metrics: one species for each geographic origin, except for the Rainbow site, which had 3 sympatric species (Figure 25). Group II was composed of 6 species, always according to the ANI and the DDH (Figure 25).

The final step in this study was to identify genes and metabolic pathways involved in these differentiation processes. For each group, a new genome was established (Figures 26-27). From there, the focus was then on identifying genes specific to each species and identifying functions found only in those genes. In summary, there were a very variable number of specific genes, from 1 to 336 depending on the species. About half had no annotation, and of the remaining genes, many functions were redundant between

species. Nevertheless, the specific functions were still interesting. They were mainly associated with amino acid metabolism, energy metabolism, carbohydrate metabolism or the transport of inorganic ions. All of this has provided us information on the selection pressures that can be applied to these microorganisms in deep environments. This has also informed us about the metabolisms acquired or lost that lead to the formation of new microbial species in the deep marine hydrothermal environment.

The last part of this thesis was about the distribution of *Thermococcales* in the environment. To do this, metagenomes and metatranscriptomes from deep hydrothermal environments and terrestrial hot springs were mapped onto *Thermococcales* genomes (Figure 28). Overall, genomes were found in the seabed metagenomes but not in those from terrestrial hot springs, the barrier between these two environments being probably too complex to cross simply because of chance and marine streams. Unlike *Thermococcus*, *Pyrococcus* and *Palaeococcus* genera were much less detected in metagenomes. This suggests that these *Archaea* inhabit more restricted and low abundant ecological niches in the environment. Finally, some *Thermococcus* strains, such as *T. cleftensis*, appeared to be present in several locations (Northeast Pacific Ocean and Cayman Trough in the Atlantic Ocean), suggesting that they would be able to migrate over long distances, while other strains appear to remain at only one place. The novelty of the latter study was to provide clues to isolate interesting new taxa, such as *Palaeococcus*, which are poorly represented compared to the vast majority of isolated, characterized and sequenced *Thermococcus*.

Conclusion

Conclusion

In conclusion, during these three years of work, several aspects of the microbial ecology were covered. The first step of this work was to classify uncharacterized isolates from our culture collection to select closely related isolates of *Thermococcus* originating from different locations. These isolates were investigated in great details in the second part of the work. Here, comparative genomics allowed us to highlight mechanisms and genes resulting from early stages of genome differentiation. The objective of the last part of the thesis was to identify the presence of genomes in environmental metagenomes, and thereby to focus on their biogeography.

The first part of the work required a year and a half to select and sequence the 48 genomes of isolates from the culture of all UBOCC isolates. This involved re-culturing and DNA extraction for all isolates. Then, the tools available to make assembly of read Sanger reads can easily take into account two sequences (1F / 1R). But I could not find any tools able to assemble 3 reads (1F / 2R). Therefore, this step was done manually, as well as quality control. Finally, many tree-building methods were tested before reaching our final result. Although it took a long time, this initial work was necessary to know the phylogenetic diversity of the isolates in our possession.

Several results emerge from comparative genomics. First of all, it would seem that the phylogeny of *Thermococcales* needs to be revised. The genus *Thermococcus* does not constitute a monophyletic group in the phylogenomic tree. Further consideration should be given to this subject, using other approaches and taxonomic markers. Second, with regard to the genomic diversity of the two groups of isolates studied, geographical isolation may explain the differentiation of genomes for one group. Concerning the second group, isolates are not organized according to this parameter. The colonization of different ecological niches might explain their evolutionary history. Among these

groups, several species are probably present. The characterization of their phenotypes will be necessary to confirm it as the species definition of prokaryotic is based on phylo-phenetic features. The analysis of their specific genes revealed several metabolisms involved in the differentiation of genomes, including amino acid metabolism and energy production metabolism. Many functions are still unknown. In these gene pools, there are probably interesting functions. Further analysis using similarity networks can shed light on these functions.

The last section of the manuscript remains to be investigated in greater details, in particular by bringing new metagenomes. This new approach as the benefit to provide clues to target geographic areas where new close phylotypes of a reference strain may be present.

References

References

Adam, P.S., Borrel, G., Brochier-Armanet, C., and Gribaldo, S. (2017). The growing tree of Archaea: new perspectives on their diversity, evolution and ecology. ISME J.

Altschul, S.F., Gish, W., Miller, W., Myers, E.W., and Lipman, D.J. (1990). Basic local alignment search tool. J. Mol. Biol. *215*, 403–410.

Amend, J.P., Meyer-Dombard, D.R., Sheth, S.N., Zolotova, N., and Amend, A.C. (2003). Palaeococcus helgesonii sp. nov., a facultatively anaerobic, hyperthermophilic archaeon from a geothermal well on Vulcano Island, Italy. Arch. Microbiol. *179*, 394–401.

Amrani, A., Bergon, A., Holota, H., Tamburini, C., Garel, M., Ollivier, B., Imbert, J., Dolla, A., and Pradel, N. (2014). Transcriptomics Reveal Several Gene Expression Patterns in the Piezophile Desulfovibrio hydrothermalis in Response to Hydrostatic Pressure. PLOS ONE *9*, e106831.

Anantharaman, K., Breier, J.A., and Dick, G.J. (2016). Metagenomic resolution of microbial functions in deep-sea hydrothermal plumes across the Eastern Lau Spreading Center. ISME J. *10*, 225–239.

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

Anderson, R.E., Kouris, A., Seward, C.H., Campbell, K.M., and Whitaker, R.J. (2017a). Structured Populations of Sulfolobus acidocaldarius with Susceptibility to Mobile Genetic Elements. Genome Biol. Evol. *9*, 1699–1710.

Anderson, R.E., Reveillaud, J., Reddington, E., Delmont, T.O., Eren, A.M., McDermott, J.M., Seewald, J.S., and Huber, J.A. (2017b). Genomic variation in microbial populations inhabiting the marine subseafloor at deep-sea hydrothermal vents. Nat. Commun. *8*, 1114.

Anisimova, M., and Gascuel, O. (2006). Approximate Likelihood-Ratio Test for Branches: A Fast, Accurate, and Powerful Alternative. Syst. Biol. *55*, 539–552.

Arab, H., Völker, H., and Thomm, M. (2000). Thermococcus aegaeicus sp. nov. and Staphylothermus hellenicus sp. nov., two novel hyperthermophilic archaea isolated from geothermally heated vents off Palaeochori Bay, Milos, Greece. Int. J. Syst. Evol. Microbiol. *50*, 2101–2108.

Atkinson, H.J., Morris, J.H., Ferrin, T.E., and Babbitt, P.C. (2009). Using Sequence Similarity Networks for Visualization of Relationships Across Diverse Protein Superfamilies. PLOS ONE *4*, e4345.

Atomi, H., Fukui, T., Kanai, T., Morikawa, M., and Imanaka, T. (2004). Description of Thermococcus kodakaraensis sp. nov., a well studied hyperthermophilic archaeon previously

reported as Pyrococcus sp. KOD1. Archaea *1*, 263–267.

Auch, A.F., Jan, M. von, Klenk, H.-P., and Göker, M. (2010a). Digital DNA-DNA hybridization for microbial species delineation by means of genome-to-genome sequence comparison. Stand. Genomic Sci. *2*, 117.

Auch, A.F., Klenk, H.-P., and Göker, M. (2010b). Standard operating procedure for calculating genome-to-genome distances based on high-scoring segment pairs. Stand. Genomic Sci. *2*, 142.

Baker, B.J., Sheik, C.S., Taylor, C.A., Jain, S., Bhasi, A., Cavalcoli, J.D., and Dick, G.J. (2013). Community transcriptomic assembly reveals microbes that contribute to deep-sea carbon and nitrogen cycling. ISME J. *7*, 1962–1973.

Balch, W.E., Fox, G.E., Magrum, L.J., Woese, C.R., and Wolfe, R.S. (1979). Methanogens: reevaluation of a unique biological group. Microbiol. Rev. *43*, 260–296.

Barbier, G., Godfroy, A., Meunier, J.-R., Quérellou, J., Cambon, M.-A., Lesongeur, F., Grimont, P.A.D., and Raguénès, G. (1999). Pyrococcus glycovorans sp. nov., a hyperthermophilic archaeon isolated from the East Pacific Rise. Int. J. Syst. Evol. Microbiol. *49*, 1829–1837.

Birrien, J.-L., Zeng, X., Jebbar, M., Cambon-Bonavita, M.-A., Quérellou, J., Oger, P., Bienvenu, N., Xiao, X., and Prieur, D. (2011). Pyrococcus yayanosii sp. nov., an obligate piezophilic hyperthermophilic archaeon isolated from a deep-sea hydrothermal vent. Int. J. Syst. Evol. Microbiol. *61*, 2827–2881.

Brazelton, W.J., and Baross, J.A. (2009). Abundant transposases encoded by the metagenome of a hydrothermal chimney biofilm. ISME J. *3*, 1420–1424.

Brazelton, W.J., Schrenk, M.O., Kelley, D.S., and Baross, J.A. (2006). Methane- and Sulfur-Metabolizing Microbial Communities Dominate the Lost City Hydrothermal Field Ecosystem. Appl. Environ. Microbiol. *72*, 6257–6270.

Bridger, S.L., Lancaster, W.A., Poole, F.L., Schut, G.J., and Adams, M.W.W. (2012). Genome sequencing of a genetically tractable Pyrococcus furiosus strain reveals a highly dynamic genome. J. Bacteriol. *194*, 4097–4106.

Brock, T.D., and Freeze, H. (1969). Thermus aquaticus gen. n. and sp. n., a Nonsporulating Extreme Thermophile. J. Bacteriol. *98*, 289–297.

Buchfink, B., Xie, C., and Huson, D.H. (2015). Fast and sensitive protein alignment using DIAMOND. Nat. Methods *12*, 59–60.

Cadillo-Quiroz, H., Didelot, X., Held, N.L., Herrera, A., Darling, A., Reno, M.L., Krause, D.J., and Whitaker, R.J. (2012). Patterns of gene flow define species of thermophilic Archaea. PLoS Biol. *10*, e1001265.

Callac, N., Oger, P., Lesongeur, F., Rattray, J.E., Vannier, P., Michoud, G., Beauverger, M., Gayet, N., Rouxel, O., Jebbar, M., et al. (2016). Pyrococcus

kukulkanii sp. nov., a hyperthermophilic, piezophilic archaeon isolated from a deep-sea hydrothermal vent. Int. J. Syst. Evol. Microbiol. *66*, 3142–3149.

Camacho, C., Coulouris, G., Avagyan, V., Ma, N., Papadopoulos, J., Bealer, K., and Madden, T.L. (2009). BLAST+: architecture and applications. BMC Bioinformatics *10*, 421.

Cambon-Bonavita, M.-A., Lesongeur, F., Pignet, P., Wery, N., Lambert, C., Godfroy, A., Querellou, J., and Barbier, G. (2003). Extremophiles, Thermophily section, species description Thermococcus atlanticus sp. nov., a hyperthermophilic Archaeon isolated from a deep-sea hydrothermal vent in the Mid-Atlantic Ridge. Extremophiles *7*, 101–109.

Canganella, F., Jones, W.J., Gambacorta, A., and Antranikian, G. (1998). Thermococcus guaymasensis sp. nov. and Thermococcus aggregans sp. nov., two novel thermophilic archaea isolated from the Guaymas Basin hydrothermal vent site. Int. J. Syst. Bacteriol. *48*, 1181–1185.

Casamayor, E.O., Massana, R., Benlloch, S., Øvreås, L., Díez, B., Goddard, V.J., Gasol, J.M., Joint, I., Rodríguez-Valera, F., and Pedrós-Alió, C. (2002). Changes in archaeal, bacterial and eukaryal assemblages along a salinity gradient by comparison of genetic fingerprinting methods in a multipond solar saltern. Environ. Microbiol. *4*, 338–348.

Case, R.J., Boucher, Y., Dahllöf, I., Holmström, C., Doolittle, W.F., and Kjelleberg, S. (2007). Use of 16S rRNA and rpoB Genes as Molecular Markers for Microbial Ecology Studies. Appl. Environ. Microbiol. *73*, 278–288.

Castelle, C.J., Wrighton, K.C., Thomas, B.C., Hug, L.A., Brown, C.T., Wilkins, M.J., Frischkorn, K.R., Tringe, S.G., Singh, A., Markillie, L.M., et al. (2015). Genomic Expansion of Domain Archaea Highlights Roles for Organisms from New Phyla in Anaerobic Carbon Cycling. Current Biology *25*, 690–701.

Castelle, C.J., and Banfield, J.F. (2018). Major New Microbial Groups Expand Diversity and Alter our Understanding of the Tree of Life. Cell *172*, 1181–1197.

Chan, C.S., Chan, K.-G., Tay, Y.-L., Chua, Y.-H., and Goh, K.M. (2015). Diversity of thermophiles in a Malaysian hot spring determined using 16S rRNA and shotgun metagenome sequencing. Front. Microbiol. *6*.

Chien, A., Edgar, D.B., and Trela, J.M. (1976). Deoxyribonucleic acid polymerase from the extreme thermophile Thermus aquaticus. J. Bacteriol. *127*, 1550–1557.

Choi, D.H., Kwon, Y.M., Chiura, H.X., Yang, E.C., Bae, S.S., Kang, S.G., Lee, J.-H., Yoon, H.S., and Kim, S.-J. (2015). Extracellular Vesicles of the Hyperthermophilic Archaeon "Thermococcus onnurineus" NA1T. Appl. Environ. Microbiol. *81*, 4591–4599.

Cohan, F.M. (2002). What are Bacterial Species? Annu. Rev. Microbiol. *56*, 457–

487.

Cohan, F.M., and Perry, E.B. (2007). A Systematics for Discovering the Fundamental Units of Bacterial Diversity. Curr. Biol. *17*, R373–R386.

Cohen, G.N., Barbe, V., Flament, D., Galperin, M., Heilig, R., Lecompte, O., Poch, O., Prieur, D., Quérellou, J., Ripp, R., et al. (2003). An integrated analysis of the genome of the hyperthermophilic archaeon Pyrococcus abyssi. Mol. Microbiol. *47*, 1495–1512.

Compeau, P.E.C., Pevzner, P.A., and Tesler, G. (2011). How to apply de Bruijn graphs to genome assembly. Nat. Biotechnol. *29*, 987–991.

Connelly, D.P., Copley, J.T., Murton, B.J., Stansfield, K., Tyler, P.A., German, C.R., Van Dover, C.L., Amon, D., Furlong, M., Grindlay, N., et al. (2012). Hydrothermal vent fields and chemosynthetic biota on the world's deepest seafloor spreading centre. Nat. Commun. *3*, 620.

Cordero, O.X., Wildschutte, H., Kirkup, B., Proehl, S., Ngo, L., Hussain, F., Roux, F.L., Mincer, T., and Polz, M.F. (2012). Ecological Populations of Bacteria Act as Socially Cohesive Units of Antibiotic Production and Resistance. Science *337*, 1228–1231.

Cord-Ruwisch, R. (1985). A quick method for the determination of dissolved and precipitated sulfides in cultures of sulfate-reducing bacteria. J. Microbiol. Methods *4*, 33–36.

Corliss, J.B., and Ballard, R.D. (1977). Oases of Life in the Cold Abyss. Nationnal Geogr. *152*, 441–453.

Corliss, J.B., Dymond, J., Gordon, L.I., Edmond, J.M., Herzen, R.P. von, Ballard, R.D., Green, K., Williams, D., Bainbridge, A., Crane, K., et al. (1979). Submarine Thermal Sprirngs on the Galápagos Rift. Science *203*, 1073–1083.

Cossu, M., Da Cunha, V., Toffano-Nioche, C., Forterre, P., and Oberto, J. (2015). Comparative genomics reveals conserved positioning of essential genomic clusters in highly rearranged Thermococcales chromosomes. Biochimie *118*, 313–321.

Cossu, M., Badel, C., Catchpole, R., Gadelle, D., Marguet, E., Barbe, V., Forterre, P., and Oberto, J. (2017). Flipping chromosomes in deep-sea archaea. PLoS Genet. *13*.

Criscuolo, A., and Gribaldo, S. (2010). BMGE (Block Mapping and Gathering with Entropy): a new software for selection of phylogenetic informative regions from multiple sequence alignments. BMC Evol. Biol. *10*, 210.

Da Cunha, V., Gaia, M., Gadelle, D., Nasir, A., and Forterre, P. (2017). Lokiarchaea are close relatives of Euryarchaeota, not bridging the gap between prokaryotes and eukaryotes. PLoS Genet. *13*, e1006810.

Dalmasso, C., Oger, P., Selva, G., Courtine, D., L'Haridon, S., Garlaschelli, A., Roussel, E., Miyazaki, J., Reveillaud, J., Jebbar, M., et al. (2016a). Thermococcus piezophilus sp. nov., a novel hyperthermophilic and piezophilic archaeon with a broad

pressure range for growth, isolated from a deepest hydrothermal vent at the Mid-Cayman Rise. Syst. Appl. Microbiol. *39*, 440–444.

Dalmasso, C., Oger, P., Courtine, D., Georges, M., Takai, K., Maignien, L., and Alain, K. (2016b). Complete Genome Sequence of the Hyperthermophilic and Piezophilic Archeon Thermococcus piezophilus CDGST, Able To Grow under Extreme Hydrostatic Pressures. Genome Announc. *4*.

Darling, A.E., Jospin, G., Lowe, E., Matsen, F.A., Bik, H.M., and Eisen, J.A. (2014). PhyloSift: phylogenetic analysis of genomes and metagenomes. PeerJ *2*, e243.

Di Giulio, M. (2005). A comparison of proteins from Pyrococcus furiosus and Pyrococcus abyssi: barophily in the physicochemical properties of amino acids and in the genetic code. Gene *346*, 1–6.

Dirmeier, R., Keller, M., Hafenbradl, D., Braun, F.-J., Rachel, R., Burggraf, S., and Stetter, K.O. (1998). Thermococcus acidaminovorans sp. nov., a new hyperthermophilic alkalophilic archaeon growing on amino acids. Extremophiles *2*, 109–114.

Doolittle, W.F., and Papke, R.T. (2006). Genomics and the bacterial species problem. Genome Biol. *7*, 116.

Duffaud, G.D., d'Hennezel, O.B., Peek, A.S., Reysenbach, A.L., and Kelly, R.M. (1998). Isolation and characterization of Thermococcus barossii, sp. nov., a hyperthermophilic archaeon isolated from a hydrothermal vent flange formation. Syst. Appl. Microbiol. *21*, 40–49.

Edgar, R.C. (2004). MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res. *32*, 1792–1797.

Edmond, J.M., Von Damm, K.L., McDuff, R.E., and Measures, C.I. (1982). Chemistry of hot springs on the East Pacific Rise and their effluent dispersal. Nature *297*, 187–191.

Erauso, G., Reysenbach, A.-L., Godfroy, A., Meunier, J.-R., Crump, B., Partensky, F., Baross, J.A., Marteinsson, V., Barbier, G., Pace, N.R., et al. (1993). Pyrococcus abyssi sp. nov., a new hyperthermophilic archaeon isolated from a deep-sea hydrothermal vent. Arch. Microbiol. *160*, 338–349.

Erauso, G., Marsin, S., Benbouzid-Rollet, N., Baucher, M.F., Barbeyron, T., Zivanovic, Y., Prieur, D., and Forterre, P. (1996). Sequence of plasmid pGT5 from the archaeon Pyrococcus abyssi: evidence for rolling-circle replication in a hyperthermophile. J. Bacteriol. *178*, 3232–3237.

Eren, A.M., Vineis, J.H., Morrison, H.G., and Sogin, M.L. (2013). A Filtering Method to Generate High Quality Short Reads Using Illumina Paired-End Technology. PLOS ONE *8*, e66643.

Eren, A.M., Esen, Ö.C., Quince, C., Vineis, J.H., Morrison, H.G., Sogin, M.L., and Delmont, T.O. (2015). Anvi'o: an advanced analysis and visualization

platform for 'omics data. PeerJ *3*, e1319.

Ewing, B., and Green, P. (1998). Base-Calling of Automated Sequencer Traces Using Phred. II. Error Probabilities. Genome Res. *8*, 186–194.

Ewing, B., Hillier, L., Wendl, M.C., and Green, P. (1998). Base-Calling of Automated Sequencer Traces UsingPhred. I. Accuracy Assessment. Genome Res. *8*, 175–185.

Felsenstein, J. (1981). Evolutionary trees from DNA sequences: A maximum likelihood approach. J. Mol. Evol. *17*, 368–376.

Fiala, G., and Stetter, K.O. (1986). Pyrococcus furiosus sp. nov. represents a novel genus of marine heterotrophic archaebacteria growing optimally at 100°C. Arch. Microbiol. *145*, 56–61.

Flores, G.E., and Reysenbach, A.-L. (2011). Hydrothermal Environments, Marine. In Encyclopedia of Geobiology, J. Reitner, and V. Thiel, eds. (Springer Netherlands), pp. 456–467.

Flores, G.E., Campbell, J.H., Kirshtein, J.D., Meneghin, J., Podar, M., Steinberg, J.I., Seewald, J.S., Tivey, M.K., Voytek, M.A., Yang, Z.K., et al. (2011). Microbial community structure of hydrothermal deposits from geochemically different vent fields along the Mid-Atlantic Ridge. Environ. Microbiol. *13*, 2158–2171.

Flores, G.E., Wagner, I.D., Liu, Y., and Reysenbach, A.-L. (2012). Distribution, abundance, and diversity patterns of the thermoacidophilic "deep-sea hydrothermal vent euryarchaeota 2." Front. Microbiol. *3*, 47.

Forterre, P. (2015). The universal tree of life: an update. Microb. Physiol. Metab. 717.

Fortunato, C.S., and Huber, J.A. (2016). Coupled RNA-SIP and metatranscriptomics of active chemolithoautotrophic communities at a deep-sea hydrothermal vent. ISME J. *10*, 1925–1938.

Fox, G.E., Stackebrandt, E., Hespell, R.B., Gibson, J., Maniloff, J., Dyer, T.A., Wolfe, R.S., Balch, W.E., Tanner, R.S., Magrum, L.J., et al. (1980). The phylogeny of prokaryotes. Science *209*, 457–463.

Fox, G.E., Wisotzkey, J.D., and Jurtshuk, P. (1992). How Close Is Close: 16S rRNA Sequence Identity May Not Be Sufficient To Guarantee Species Identity. Int. J. Syst. Evol. Microbiol. *42*, 166–170.

Fraser, C., Alm, E.J., Polz, M.F., Spratt, B.G., and Hanage, W.P. (2009). The bacterial species challenge: making sense of genetic and ecological diversity. Science *323*, 741–746.

Fukui, T., Atomi, H., Kanai, T., Matsumi, R., Fujiwara, S., and Imanaka, T. (2005). Complete genome sequence of the hyperthermophilic archaeon Thermococcus kodakaraensis KOD1 and comparison with Pyrococcus genomes. Genome Res. *15*, 352–363.

Galperin, M.Y., Makarova, K.S., Wolf, Y.I., and Koonin, E.V. (2014). Expanded microbial genome coverage and improved protein family annotation in the COG database. Nucleic Acids Res.

gku1223.

Galtier, N., and Lobry, J.R. (1997). Relationships Between Genomic G+C Content, RNA Secondary Structures, and Optimal Growth Temperature in Prokaryotes. J. Mol. Evol. *44*, 632–636.

García-Martínez, J., and Rodríguez-Valera, F. (2000). Microdiversity of uncultured marine prokaryotes: the SAR11 cluster and the marine Archaea of Group I. Mol. Ecol. *9*, 935–948.

García-Martínez, J., Acinas, S.G., Massana, R., and Rodríguez-Valera, F. (2002). Prevalence and microdiversity of Alteromonas macleodii-like microorganisms in different oceanic regions. Environ. Microbiol. *4*, 42–50.

Gardner, A.F., Kumar, S., and Perler, F.B. (2012). Genome Sequence of the Model Hyperthermophilic Archaeon Thermococcus litoralis NS-C. J. Bacteriol. *194*, 2375–2376.

Gascuel, O. (1997). BIONJ: an improved version of the NJ algorithm based on a simple model of sequence data. Mol. Biol. Evol. *14*, 685–695.

Gaudin, M., Krupovic, M., Marguet, E., Gauliard, E., Cvirkaite-Krupovic, V., Le Cam, E., Oberto, J., and Forterre, P. (2014). Extracellular membrane vesicles harbouring viral genomes. Environ. Microbiol. *16*, 1167–1175.

Gavrilov, S.N., Stracke, C., Jensen, K., Menzel, P., Kallnik, V., Slesarev, A., Sokolova, T., Zayulina, K., Bräsen, C., Bonch-Osmolovskaya, E.A., et al. (2016). Isolation and Characterization of the First Xylanolytic Hyperthermophilic Euryarchaeon Thermococcus sp. Strain 2319x1 and Its Unusual Multidomain Glycosidase. Front. Microbiol. *7*, 552.

Geslin, C., Romancer, M.L., Erauso, G., Gaillard, M., Perrot, G., and Prieur, D. (2003). PAV1, the First Virus-Like Particle Isolated from a Hyperthermophilic Euryarchaeote, "Pyrococcus abyssi." J. Bacteriol. *185*, 3888–3894.

Glenn, T.C. (2011). Field guide to next-generation DNA sequencers. Mol. Ecol. Resour. *11*, 759–769.

Godfroy, A., Meunier, J.-R., Guezennec, J., Lesongeur, F., Raguénès, G., Rimbault, A., and Barbier, G. (1996). Thermococcus fumicolans sp. nov., a New Hyperthermophilic Archaeon Isolated from a Deep-Sea Hydrothermal Vent in the North Fiji Basin. Int. J. Syst. Evol. Microbiol. *46*, 1113–1119.

Godfroy, A., Lesongeur, F., Raguénès, G., Quérellou, J., Antoine, E., Meunier, J.-R., Guezennec, J., and Barbier, G. (1997). Thermococcus hydrothermalis sp. nov., a New Hyperthermophilic Archaeon Isolated from a Deep-Sea Hydrothermal Vent. Int. J. Syst. Evol. Microbiol. *47*, 622–626.

González, J.M., Kato, C., and Horikoshi, K. (1995). Thermococcus peptonophilus sp. nov., a fast-growing, extremely thermophilic archaebacterium isolated from deep-sea hydrothermal vents. Arch. Microbiol. *164*, 159–164.

González, J.M., Masuchi, Y., Robb, F.T.,

Ammerman, J.W., Maeder, D.L., Yanagibayashi, M., Tamaoka, J., and Kato, C. (1998). Pyrococcus horikoshii sp. nov., a hyperthermophilic archaeon isolated from a hydrothermal vent at the Okinawa Trough. Extremophiles *2*, 123–130.

González, J.M., Sheckells, D., Viebahn, M., Krupatkina, D., Borges, K.M., and Robb, F.T. (1999). Thermococcus waiotapuensis sp. nov., an extremely thermophilic archaeon isolated from a freshwater hot spring. Arch. Microbiol. *172*, 95–101.

Goris, J., Konstantinidis, K.T., Klappenbach, J.A., Coenye, T., Vandamme, P., and Tiedje, J.M. (2007). DNA–DNA hybridization values and their relationship to whole-genome sequence similarities. Int. J. Syst. Evol. Microbiol. *57*, 81–91.

Gorlas, A., Koonin, E.V., Bienvenu, N., Prieur, D., and Geslin, C. (2012). TPV1, the first virus isolated from the hyperthermophilic genus Thermococcus. Environ. Microbiol. *14*, 503–516.

Gorlas, A., Krupovic, M., Forterre, P., and Geslin, C. (2013a). Living Side by Side with a Virus: Characterization of Two Novel Plasmids from Thermococcus prieurii, a Host for the Spindle-Shaped Virus TPV1. Appl. Environ. Microbiol. *79*, 3822–3828.

Gorlas, A., Alain, K., Bienvenu, N., and Geslin, C. (2013b). Thermococcus prieurii sp. nov., a hyperthermophilic archaeon isolated from a deep-sea hydrothermal vent. International Journal of Systematic and Evolutionary Microbiology *63*, 2920–2926.

Gorlas, A., Croce, O., Oberto, J., Gauliard, E., Forterre, P., and Marguet, E. (2014). Thermococcus nautili sp. nov., a hyperthermophilic archaeon isolated from a hydrothermal deep-sea vent. Int. J. Syst. Evol. Microbiol. *64*, 1802–1810.

Gouy, M., Guindon, S., and Gascuel, O. (2010). SeaView version 4: A multiplatform graphical user interface for sequence alignment and phylogenetic tree building. Mol. Biol. Evol. *27*, 221–224.

Grote, R., Li, L., Tamaoka, J., Kato, C., Horikoshi, K., and Antranikian, G. (1999). Thermococcus siculi sp. nov., a novel hyperthermophilic archaeon isolated from a deep-sea hydrothermal vent at the Mid-Okinawa Trough. Extrem. Life Extreme Cond. *3*, 55–62.

Guindon, S., and Gascuel, O. (2003). A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. Syst. Biol. *52*, 696–704.

Guindon, S., Dufayard, J.-F., Lefort, V., Anisimova, M., Hordijk, W., and Gascuel, O. (2010). New Algorithms and Methods to Estimate Maximum-Likelihood Phylogenies: Assessing the Performance of PhyML 3.0. Syst. Biol. *59*, 307–321.

Guri, M., Durand, L., Cueff-Gauchard, V., Zbinden, M., Crassous, P., Shillito, B., and Cambon-Bonavita, M.-A. (2012). Acquisition of epibiotic bacteria along the life cycle of the hydrothermal

shrimp Rimicaris exoculata. ISME J. *6*, 597–609.

Guy, L., and Ettema, T.J.G. (2011). The archaeal 'TACK' superphylum and the origin of eukaryotes. Trends Microbiol. *19*, 580–587.

Guyer, R.L., and Koshland, D.E. (1989). The Molecule of the Year. Science *246*, 1543–1546.

Hasegawa, M., Kishino, H., and Yano, T. (1985). Dating of the human-ape splitting by a molecular clock of mitochondrial DNA. J. Mol. Evol. *22*, 160–174.

Hensley, S.A., Jung, J.-H., Park, C.-S., and Holden, J.F. (2014). Thermococcus paralvinellae sp. nov. and Thermococcus cleftensis sp. nov. of hyperthermophilic heterotrophs from deep-sea hydrothermal vents. Int. J. Syst. Evol. Microbiol. *64*, 3655–3659.

Horikoshi, P.K., and Bull, A.T. (2011). Prologue: Definition, Categories, Distribution, Origin and Evolution, Pioneering Studies, and Emerging Fields of Extremophiles. In Extremophiles Handbook, K. Horikoshi, ed. (Springer Japan), pp. 3–15.

Huber, J.A., Butterfield, D.A., and Baross, J.A. (2006). Diversity and distribution of subseafloor Thermococcales populations in diffuse hydrothermal vents at an active deep-sea volcano in the northeast Pacific Ocean. J. Geophys. Res. Biogeosciences *111*, G04016.

Huber, R., Stöhr, J., Hohenhaus, S., Rachel, R., Burggraf, S., Jannasch, H.W., and Stetter, K.O. (1995). Thermococcus chitonophagus sp. nov., a novel, chitin-degrading, hyperthermophilic archaeum from a deep-sea hydrothermal vent environment. Arch. Microbiol. *164*, 255–264.

Hug, L.A., Baker, B.J., Anantharaman, K., Brown, C.T., Probst, A.J., Castelle, C.J., Butterfield, C.N., Hernsdorf, A.W., Amano, Y., Ise, K., et al. (2016). A new view of the tree of life. Nature Microbiology *1*, nmicrobiol201648.

Hunt, D.E., David, L.A., Gevers, D., Preheim, S.P., Alm, E.J., and Polz, M.F. (2008). Resource Partitioning and Sympatric Differentiation Among Closely Related Bacterioplankton. Science *320*, 1081–1085.

Hyatt, D., Chen, G.-L., LoCascio, P.F., Land, M.L., Larimer, F.W., and Hauser, L.J. (2010). Prodigal: prokaryotic gene recognition and translation initiation site identification. BMC Bioinformatics *11*, 119.

Jaenicke, R., and Sterner, R. (2006). Life at High Temperatures. In The Prokaryotes, M. Dworkin, S. Falkow, E. Rosenberg, K.-H. Schleifer, and E. Stackebrandt, eds. (New York, NY: Springer New York), pp. 167–209.

Jaeschke, A., Jørgensen, S.L., Bernasconi, S.M., Pedersen, R.B., Thorseth, I.H., and Früh-Green, G.L. (2012). Microbial diversity of Loki's Castle black smokers at the Arctic Mid-Ocean Ridge. Geobiology *10*, 548–561.

Jannasch, H.W., and Taylor, C.D.

(1984). Deep-Sea Microbiology. Annu. Rev. Microbiol. *38*, 487–487.

Jannasch, H.W., and Wirsen, C.O. (1979). Chemosynthetic Primary Production at East Pacific Sea Floor Spreading Centers. BioScience *29*, 592–598.

Jolivet, E., L'Haridon, S., Corre, E., Forterre, P., and Prieur, D. (2003). Thermococcus gammatolerans sp. nov., a hyperthermophilic archaeon from a deep-sea hydrothermal vent that resists ionizing radiation. Int. J. Syst. Evol. Microbiol. *53*, 847–851.

Jolivet, E., Corre, E., L'Haridon, S., Forterre, P., and Prieur, D. (2004). Thermococcus marinus sp. nov. and Thermococcus radiotolerans sp. nov., two hyperthermophilic archaea from deep-sea hydrothermal vents that resist ionizing radiation. Extrem. Life Extreme Cond. *8*, 219–227.

Jun, X., Lupeng, L., Minjuan, X., Oger, P., Fengping, W., Jebbar, M., and Xiang, X. (2011). Complete genome sequence of the obligate piezophilic hyperthermophilic archaeon Pyrococcus yayanosii CH1. J. Bacteriol. *193*, 4297–4298.

Jung, J.-H., Lee, J.-H., Holden, J.F., Seo, D.-H., Shin, H., Kim, H.-Y., Kim, W., Ryu, S., and Park, C.-S. (2012a). Complete genome sequence of the hyperthermophilic archaeon Pyrococcus sp. strain ST04, isolated from a deep-sea hydrothermal sulfide chimney on the Juan de Fuca Ridge. J. Bacteriol. *194*, 4434–4435.

Jung, J.-H., Holden, J.F., Seo, D.-H., Park, K.-H., Shin, H., Ryu, S., Lee, J.-H., and Park, C.-S. (2012b). Complete genome sequence of the hyperthermophilic archaeon Thermococcus sp. strain CL1, isolated from a Paralvinella sp. polychaete worm collected from a hydrothermal vent. J. Bacteriol. *194*, 4769–4770.

Katoh, K., and Standley, D.M. (2013). MAFFT Multiple Sequence Alignment Software Version 7: Improvements in Performance and Usability. Mol. Biol. Evol. *30*, 772–780.

Kawarabayasi, Y., Sawada, M., Horikawa, H., Haikawa, Y., Hino, Y., Yamamoto, S., Sekine, M., Baba, S., Kosugi, H., Hosoyama, A., et al. (1998). Complete sequence and gene organization of the genome of a hyper-thermophilic archaebacterium, Pyrococcus horikoshii OT3. DNA Res. Int. J. Rapid Publ. Rep. Genes Genomes *5*, 55–76.

Keller, M., Braun, F.-J., Dirmeier, R., Hafenbradl, D., Burggraf, S., Rachel, R., and Stetter, K.O. (1995). Thermococcus alcaliphilus sp. nov., a new hyperthermophilic archaeum growing on polysulfide at alkaline pH. Arch. Microbiol. *164*, 390–395.

Kelley, D.S., Baross, J.A., and Delaney, J.R. (2002). Volcanoes, Fluids, and Life at Mid-Ocean Ridge Spreading Centers. Annu. Rev. Earth Planet. Sci. *30*, 385.

Kim, B.K., Lee, S.H., Kim, S.-Y., Jeong, H., Kwon, S.-K., Lee, C.H., Song, J.Y., Yu, D.S., Kang, S.G., and Kim, J.F. (2012). Genome sequence of an oligohaline

hyperthermophilic archaeon, Thermococcus zilligii AN1, isolated from a terrestrial geothermal freshwater spring. J. Bacteriol. *194*, 3765–3766.

Kobayashi, T. (2015). Thermococcus. In Bergey's Manual of Systematics of Archaea and Bacteria, (John Wiley & Sons, Ltd), p.

Kobayashi, T., Kwak, Y.S., Akiba, T., Kudo, T., and Horikoshi, K. (1994). Thermococcus profundus sp. nov., a new hyperthermophilic archaeon isolated from a deep-sea hydrothermal vent. Syst. Appl. Microbiol. *17*, 232–236.

Konstantinidis, K.T., and Tiedje, J.M. (2005). Genomic insights that advance the species definition for prokaryotes. Proc. Natl. Acad. Sci. U. S. A. *102*, 2567–2572.

Kozubal, M.A., Romine, M., Jennings, R. deM, Jay, Z.J., Tringe, S.G., Rusch, D.B., Beam, J.P., McCue, L.A., and Inskeep, W.P. (2013). Geoarchaeota: a new candidate phylum in the Archaea from high-temperature acidic iron mats in Yellowstone National Park. The ISME Journal *7*, 622–634.

Krause, D.J., Didelot, X., Cadillo-Quiroz, H., and Whitaker, R.J. (2014). Recombination Shapes Genome Architecture in an Organism from the Archaeal Domain. Genome Biol. Evol. *6*, 170–178.

Krupovic, M., Gonnet, M., Hania, W.B., Forterre, P., and Erauso, G. (2013). Insights into dynamics of mobile genetic elements in hyperthermophilic environments from five new Thermococcus plasmids. PloS One *8*, e49044.

Kuwabara, T., Minaba, M., Iwayama, Y., Inouye, I., Nakashima, M., Marumo, K., Maruyama, A., Sugai, A., Itoh, T., Ishibashi, J., et al. (2005). Thermococcus coalescens sp. nov., a cell-fusing hyperthermophilic archaeon from Suiyo Seamount. Int. J. Syst. Evol. Microbiol. *55*, 2507–2514.

Kuwabara, T., Minaba, M., Ogi, N., and Kamekura, M. (2007). Thermococcus celericrescens sp. nov., a fast-growing and cell-fusing hyperthermophilic archaeon from a deep-sea hydrothermal vent. Int. J. Syst. Evol. Microbiol. *57*, 437–443.

Langmead, B., and Salzberg, S.L. (2012). Fast gapped-read alignment with Bowtie 2. Nat. Methods *9*, 357–359.

Lee, H.S., Kang, S.G., Bae, S.S., Lim, J.K., Cho, Y., Kim, Y.J., Jeon, J.H., Cha, S.-S., Kwon, K.K., Kim, H.-T., et al. (2008). The Complete Genome Sequence of Thermococcus onnurineus NA1 Reveals a Mixed Heterotrophic and Carboxydotrophic Metabolism. J. Bacteriol. *190*, 7491–7499.

Lee, H.S., Bae, S.S., Kim, M.-S., Kwon, K.K., Kang, S.G., and Lee, J.-H. (2011). Complete genome sequence of hyperthermophilic Pyrococcus sp. strain NA2, isolated from a deep-sea hydrothermal vent area. J. Bacteriol. *193*, 3666–3667.

Lee, I., Kim, Y.O., Park, S.-C., and Chun, J. (2015). OrthoANI: An improved algorithm and software for calculating

average nucleotide identity. Int. J. Syst. Evol. Microbiol.

Lefort, V., Longueville, J.-E., and Gascuel, O. (2017). SMS: Smart Model Selection in PhyML. Mol. Biol. Evol. *34*, 2422–2424.

Lepage, E., Marguet, E., Geslin, C., Matte-Tailliez, O., Zillig, W., Forterre, P., and Tailliez, P. (2004). Molecular diversity of new Thermococcales isolates from a single area of hydrothermal deep-sea vents as revealed by randomly amplified polymorphic DNA fingerprinting and 16S rRNA gene sequence analysis. Appl. Environ. Microbiol. *70*, 1277–1286.

Letunic, I., and Bork, P. (2016). Interactive tree of life (iTOL) v3: an online tool for the display and annotation of phylogenetic and other trees. Nucleic Acids Res. *44*, W242-245.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., Durbin, R., and 1000 Genome Project Data Processing Subgroup (2009). The Sequence Alignment/Map format and SAMtools. Bioinforma. Oxf. Engl. *25*, 2078–2079.

Ludwig, W., Strunk, O., Westram, R., Richter, L., Meier, H., Yadhukumar, null, Buchner, A., Lai, T., Steppi, S., Jobb, G., et al. (2004). ARB: a software environment for sequence data. Nucleic Acids Res. *32*, 1363–1371.

Lundberg, K.S., Shoemaker, D.D., Adams, M.W., Short, J.M., Sorge, J.A., and Mathur, E.J. (1991). High-fidelity amplification using a thermostable DNA polymerase isolated from Pyrococcus furiosus. Gene *108*, 1–6.

Lutz, R.A., and Kennish, M.J. (1993). Ecology of deep-sea hydrothermal vent communities: A review. Rev. Geophys. *31*, 211–242.

Macelroy, R.D. (1974). Some comments on the evolution of extremophiles. Biosystems *6*, 74–75.

Madigan, M.T., Martinko, J.M., Stahl, D.A., and Clark, D.P. (2012). Brock biology of microorganisms (San Francisco: Benjamin Cummings).

Maeder, D.L., Weiss, R.B., Dunn, D.M., Cherry, J.L., González, J.M., DiRuggiero, J., and Robb, F.T. (1999). Divergence of the hyperthermophilic archaea Pyrococcus furiosus and P. horikoshii inferred from complete genomic sequences. Genetics *152*, 1299–1305.

Maheshwari, R., Bharadwaj, G., and Bhat, M.K. (2000). Thermophilic Fungi: Their Physiology and Enzymes. Microbiol. Mol. Biol. Rev. *64*, 461–488.

Mardanov, A.V., Ravin, N.V., Svetlitchnyi, V.A., Beletsky, A.V., Miroshnichenko, M.L., Bonch-Osmolovskaya, E.A., and Skryabin, K.G. (2009). Metabolic versatility and indigenous origin of the archaeon Thermococcus sibiricus, isolated from a siberian oil reservoir, as revealed by genome analysis. Appl. Environ. Microbiol. *75*, 4580–4588.

Marteinsson, V.T., Birrien, J.L., Reysenbach, A.L., Vernet, M., Marie, D., Gambacorta, A., Messner, P.,

Sleytr, U.B., and Prieur, D. (1999). Thermococcus barophilus sp. nov., a new barophilic and hyperthermophilic archaeon isolated under high hydrostatic pressure from a deep-sea hydrothermal vent. Int. J. Syst. Bacteriol. *49 Pt 2*, 351–359.

Martin, W., Baross, J., Kelley, D., and Russell, M.J. (2008). Hydrothermal vents and the origin of life. Nat. Rev. Microbiol. *6*, 805–814.

Mayr, E. (1942). Systematics and the Origin of Species, from the Viewpoint of a Zoologist (Harvard University Press).

McCliment, E.A., Voglesonger, K.M., O'Day, P.A., Dunn, E.E., Holloway, J.R., and Cary, S.C. (2006). Colonization of nascent, deep-sea hydrothermal vents by a novel Archaeal and Nanoarchaeal assemblage. Environ. Microbiol. *8*, 114–125.

McCollom, T.M. (2007). Geochemical Constraints on Sources of Metabolic Energy for Chemolithoautotrophy in Ultramafic-Hosted Deep-Sea Hydrothermal Systems. Astrobiology *7*, 933–950.

Mei, N., Postec, A., Monnin, C., Pelletier, B., Payri, C.E., Ménez, B., Frouin, E., Ollivier, B., Erauso, G., and Quéméneur, M. (2016). Metagenomic and PCR-Based Diversity Surveys of [FeFe]-Hydrogenases Combined with Isolation of Alkaliphilic Hydrogen-Producing Bacteria from the Serpentinite-Hosted Prony Hydrothermal Field, New Caledonia. Front. Microbiol. *7*.

Meier, D.V., Bach, W., Girguis, P.R., Gruber-Vodicka, H.R., Reeves, E.P., Richter, M., Vidoudez, C., Amann, R., and Meyerdierks, A. (2016). Heterotrophic Proteobacteria in the vicinity of diffuse hydrothermal venting. Environ. Microbiol. *18*, 4348–4368.

Meier-Kolthoff, J.P., Auch, A.F., Klenk, H.-P., and Göker, M. (2013). Genome sequence-based species delimitation with confidence intervals and improved distance functions. BMC Bioinformatics *14*, 60.

Meier-Kolthoff, J.P., Klenk, H.-P., and Göker, M. (2014). Taxonomic use of DNA G+C content and DNA–DNA hybridization in the genomic age. Int. J. Syst. Evol. Microbiol. *64*, 352–356.

Meng, J., Xu, J., Qin, D., He, Y., Xiao, X., and Wang, F. (2014). Genetic and functional properties of uncultivated MCG archaea assessed by metagenome and gene expression analyses. The ISME Journal *8*, 650–659.

Meng, A., Corre, E., Probert, I., Gutierrez-Rodriguez, A., Siano, R., Annamale, A., Alberti, A., Silva, C.D., Wincker, P., Crom, S.L., et al. (2017). Analysis Of The Genomic Basis Of Functional Diversity In Dinoflagellates Using A Transcriptome-Based Sequence Similarity Network. BioRxiv 211243.

Minoche, A.E., Dohm, J.C., and Himmelbauer, H. (2011). Evaluation of genomic high-throughput sequencing data generated on Illumina HiSeq and Genome Analyzer systems. Genome Biol. *12*, R112.

Miroshnichenko, M.L., Bonch-Osmolovskaya, E.A., Neuner, A., Kostrikina, N.A., Chernych, N.A., and Alekseev, V.A. (1989). Thermococcus stetteri sp. nov., a New Extremely Thermophilic Marine Sulfur-Metabolizing Archaebacterium. Syst. Appl. Microbiol. *12*, 257–262.

Miroshnichenko, M.L., Gongadze, G.M., Rainey, F.A., Kostyukova, A.S., Lysenko, A.M., Chernyh, N.A., and Bonch-Osmolovskaya, E.A. (1998). Thermococcus gorgonarius sp. nov. and Thermococcus pacificus sp. nov.: heterotrophic extremely thermophilic archaea from New Zealand submarine hot vents. Int. J. Syst. Evol. Microbiol. *48*, 23–29.

Miroshnichenko, M.L., Hippe, H., Stackebrandt, E., Kostrikina, N.A., Chernyh, N.A., Jeanthon, C., Nazina, T.N., Belyaev, S.S., and Bonch-Osmolovskaya, E.A. (2001). Isolation and characterization of Thermococcus sibiricus sp. nov. from a Western Siberia high-temperature oil reservoir. Extrem. Life Extreme Cond. *5*, 85–91.

Mistry, J., Finn, R.D., Eddy, S.R., Bateman, A., and Punta, M. (2013). Challenges in homology search: HMMER3 and convergent evolution of coiled-coil regions. Nucleic Acids Res. *41*, e121–e121.

Moriya, Y., Itoh, M., Okuda, S., Yoshizawa, A.C., and Kanehisa, M. (2007). KAAS: an automatic genome annotation and pathway reconstruction server. Nucleic Acids Res. *35*, W182-185.

Nakagawa, S., and Takai, K. (2008). Deep-sea vent chemoautotrophs: diversity, biochemistry and ecological significance. FEMS Microbiol. Ecol. *65*, 1–14.

Nercessian, O., Reysenbach, A.-L., Prieur, D., and Jeanthon, C. (2003). Archaeal diversity associated with in situ samplers deployed on hydrothermal vents on the East Pacific Rise (13°N). Environ. Microbiol. *5*, 492–502.

Neuner, A., Jannasch, H.W., Belkin, S., and Stetter, K.O. (1990). Thermococcus litoralis sp. nov.: A new species of extremely thermophilic marine archaebacteria. Arch. Microbiol. *153*, 205–207.

Oberto, J., Gaudin, M., Cossu, M., Gorlas, A., Slesarev, A., Marguet, E., and Forterre, P. (2014). Genome Sequence of a Hyperthermophilic Archaeon, Thermococcus nautili 30-1, That Produces Viral Vesicles. Genome Announc. *2*, e00243-14.

Oger, P., Sokolova, T.G., Kozhevnikova, D.A., Chernyh, N.A., Bartlett, D.H., Bonch-Osmolovskaya, E.A., and Lebedinsky, A.V. (2011). Complete genome sequence of the hyperthermophilic archaeon Thermococcus sp. strain AM4, capable of organotrophic growth and growth at the expense of hydrogenogenic or sulfidogenic oxidation of carbon monoxide. J. Bacteriol. *193*, 7019–7020.

Oger, P., Sokolova, T.G., Kozhevnikova, D.A., Taranov, E.A., Vannier, P., Lee, H.S., Kwon, K.K., Kang, S.G., Lee, J.-H., Bonch-

Osmolovskaya, E.A., et al. (2016). Complete Genome Sequence of the Hyperthermophilic and Piezophilic Archaeon Thermococcus barophilus Ch5, Capable of Growth at the Expense of Hydrogenogenesis from Carbon Monoxide and Formate. Genome Announc. *4*.

Oger, P.M., Callac, N., Oger-Desfeux, C., Hughes, S., Gillet, B., Jebbar, M., and Godfroy, A. (2017). Complete Genome Sequence of the Hyperthermophilic Piezophilic Archaeon Pyrococcus kukulkanii NCB100 Isolated from the Rebecca's Roost Hydrothermal Vent in the Guaymas Basin. Genome Announc. *5*.

Oger, P.M. (2018). Complete Genome Sequences of 11 Type Species from the *Thermococcus* Genus of Hyperthermophilic and Piezophilic Archaea. Genome Announc. *6*, e00037-18.

Orcutt, B.N., Sylvan, J.B., Knab, N.J., and Edwards, K.J. (2011). Microbial Ecology of the Dark Ocean above, at, and below the Seafloor. Microbiol. Mol. Biol. Rev. *75*, 361–422.

Pagé, A., Tivey, M.K., Stakes, D.S., and Reysenbach, A.-L. (2008). Temporal and spatial archaeal colonization of hydrothermal vent deposits. Environ. Microbiol. *10*, 874–884.

Papadimitriou, K., Baharidis, P.K., Georgoulis, A., Engel, M., Louka, M., Karamolegkou, G., Tsoka, A., Blom, J., Pot, B., Malecki, P., et al. (2016). Analysis of the complete genome sequence of the archaeon Pyrococcus chitonophagus DSM 10152 (formerly Thermococcus chitonophagus). Extremophiles *20*, 351–361.

Petitjean, C., Deschamps, P., López-García, P., Moreira, D., and Brochier-Armanet, C. (2015). Extending the Conserved Phylogenetic Core of Archaea Disentangles the Evolution of the Third Domain of Life. Mol. Biol. Evol. *32*, 1242–1254.

Pikuta, E.V., Marsic, D., Itoh, T., Bej, A.K., Tang, J., Whitman, W.B., Ng, J.D., Garriott, O.K., and Hoover, R.B. (2007). Thermococcus thioreducens sp. nov., a novel hyperthermophilic, obligately sulfur-reducing archaeon from a deep-sea hydrothermal vent. Int. J. Syst. Evol. Microbiol. *57*, 1612–1618.

Price, M.T., Fullerton, H., and Moyer, C.L. (2015). Biogeography and evolution of Thermococcus isolates from hydrothermal vent systems of the Pacific. Front. Microbiol. *6*.

Pruesse, E., Peplies, J., and Glöckner, F.O. (2012). SINA: Accurate high-throughput multiple sequence alignment of ribosomal RNA genes. Bioinformatics *28*, 1823–1829.

R Core Team (2015). R: The R Project for Statistical Computing (Vienna, Austria).

Raymann, K., Brochier-Armanet, C., and Gribaldo, S. (2015). The two-domain tree of life is linked to a new root for the Archaea. Proc. Natl. Acad. Sci. U. S. A. *112*, 6670–6675.

Reveillaud, J., Reddington, E., McDermott, J., Algar, C., Meyer, J.L., Sylva, S., Seewald, J., German, C.R.,

and Huber, J.A. (2016). Subseafloor microbial communities in hydrogen-rich vent fluids from hydrothermal systems along the Mid-Cayman Rise. Environ. Microbiol. n/a-n/a.

Reysenbach, A.-L., Ehringer, M., and Hershberger, K. (2000a). Microbial diversity at 83°C in Calcite Springs, Yellowstone National Park: another environment where the Aquificales and "Korarchaeota" coexist. Extremophiles *4*, 61–67.

Reysenbach, A.-L., Longnecker, K., and Kirshtein, J. (2000b). Novel Bacterial and Archaeal Lineages from an In Situ Growth Chamber Deployed at a Mid-Atlantic Ridge Hydrothermal Vent. Appl. Environ. Microbiol. *66*, 3798–3806.

Richter, M., and Rosselló-Móra, R. (2009). Shifting the genomic gold standard for the prokaryotic species definition. Proc. Natl. Acad. Sci. *106*, 19126–19131.

Rinke, C., Schwientek, P., Sczyrba, A., Ivanova, N.N., Anderson, I.J., Cheng, J.-F., Darling, A., Malfatti, S., Swan, B.K., Gies, E.A., et al. (2013). Insights into the phylogeny and coding potential of microbial dark matter. Nature *499*, 431–437.

Ronimus, R.S., Reysenbach, A.-L., Musgrave, D.R., and Morgan, H.W. (1997). The phylogenetic position of the Thermococcus isolate AN1 based on 16S rRNA gene sequence analysis: a proposal that AN1 represents a new species, Thermococcus zilligii sp. nov. Arch. Microbiol. *168*, 245–248.

Ronquist, F., and Huelsenbeck, J.P. (2003). MrBayes 3: Bayesian phylogenetic inference under mixed models. Bioinformatics *19*, 1572–1574.

Rosselló-Mora, R., and Amann, R. (2001). The species concept for prokaryotes. FEMS Microbiol. Rev. *25*, 39–67.

Rothschild, L.J., and Mancinelli, R.L. (2001). Life in extreme environments. Nature *409*, 1092–1101.

Roussel, E.G., Konn, C., Charlou, J.-L., Donval, J.-P., Fouquet, Y., Querellou, J., Prieur, D., and Cambon Bonavita, M.-A. (2011). Comparison of microbial communities associated with three Atlantic ultramafic hydrothermal systems. FEMS Microbiol. Ecol. *77*, 647–665.

Saiki, R.K., Scharf, S., Faloona, F., Mullis, K.B., Horn, G.T., Erlich, H.A., and Arnheim, N. (1985). Enzymatic amplification of beta-globin genomic sequences and restriction site analysis for diagnosis of sickle cell anemia. Science *230*, 1350–1354.

Seitz, K.W., Lazar, C.S., Hinrichs, K.-U., Teske, A.P., and Baker, B.J. (2016). Genomic reconstruction of a novel, deeply branched sediment archaeal phylum with pathways for acetogenesis and sulfur reduction. The ISME Journal *10*, 1696–1705.

Seki, K., and Toyoshima, M. (1998). Preserving tardigrades under pressure. Nature *395*, 853–854.

Shapiro, B.J., and Polz, M.F. (2014). Ordering microbial diversity into

ecologically and genetically cohesive units. Trends Microbiol. *22*, 235–247.

Shapiro, B.J., Friedman, J., Cordero, O.X., Preheim, S.P., Timberlake, S.C., Szabó, G., Polz, M.F., and Alm, E.J. (2012). Population genomics of early events in the ecological differentiation of bacteria. Science *336*, 48–51.

Skennerton, C.T., Ward, L.M., Michel, A., Metcalfe, K., Valiente, C., Mullin, S., Chan, K.Y., Gradinaru, V., and Orphan, V.J. (2015). Genomic Reconstruction of an Uncultured Hydrothermal Vent Gammaproteobacterial Methanotroph (Family Methylothermaceae) Indicates Multiple Adaptations to Oxygen Limitation. Front. Microbiol. *6*.

Sokolova, T.G., Jeanthon, C., Kostrikina, N.A., Chernyh, N.A., Lebedinsky, A.V., Stackebrandt, E., and Bonch-Osmolovskaya, E.A. (2004). The first evidence of anaerobic CO oxidation coupled with H2 production by a hyperthermophilic archaeon isolated from a deep-sea hydrothermal vent. Extremophiles *8*, 317–323.

Soler, N., Marguet, E., Verbavatz, J.-M., and Forterre, P. (2008). Virus-like vesicles and extracellular DNA produced by hyperthermophilic archaea of the order Thermococcales. Res. Microbiol. *159*, 390–399.

Spang, A., Saw, J.H., Jørgensen, S.L., Zaremba-Niedzwiedzka, K., Martijn, J., Lind, A.E., van Eijk, R., Schleper, C., Guy, L., and Ettema, T.J.G. (2015). Complex archaea that bridge the gap between prokaryotes and eukaryotes. Nature *521*, 173–179.

Stackebrandt, E., and Ebers, J. (2006). Taxonomic parameters revisited: tarnished gold standards. Microbiol. Today *33*, 152.

Stackebrandt, E., and Goebel, B.M. (1994). Taxonomic Note: A Place for DNA-DNA Reassociation and 16S rRNA Sequence Analysis in the Present Species Definition in Bacteriology. Int. J. Syst. Evol. Microbiol. *44*, 846–849.

Stackebrandt, E., Frederiksen, W., Garrity, G.M., Grimont, P.A.D., Kämpfer, P., Maiden, M.C.J., Nesme, X., Rosselló-Mora, R., Swings, J., Trüper, H.G., et al. (2002). Report of the ad hoc committee for the re-evaluation of the species definition in bacteriology. Int. J. Syst. Evol. Microbiol. *52*, 1043–1047.

Stetter, K.O. (2006). History of discovery of the first hyperthermophiles. Extrem. Life Extreme Cond. *10*, 357–362.

Stetter, K.O., and Huber, H. (2015). Pyrococcus. In Bergey's Manual of Systematics of Archaea and Bacteria, (John Wiley & Sons, Ltd), p.

Takai, K. (2015). Palaeococcus. In Bergey's Manual of Systematics of Archaea and Bacteria, (John Wiley & Sons, Ltd), p.

Takai, K., and Nakamura, K. (2011). Archaeal diversity and community development in deep-sea hydrothermal vents. Curr. Opin. Microbiol. *14*, 282–291.

Takai, K., Sugai, A., Itoh, T., and

Horikoshi, K. (2000). Palaeococcus ferrophilus gen. nov., sp. nov., a barophilic, hyperthermophilic archaeon from a deep-sea hydrothermal vent chimney. Int. J. Syst. Evol. Microbiol. *50*, 489–500.

Tansey, M.R., and Brock, T.D. (1972). The Upper Temperature Limit for Eukaryotic Organisms. Proc. Natl. Acad. Sci. *69*, 2426–2428.

Teske, A., Hinrichs, K.-U., Edgcomb, V., de Vera Gomez, A., Kysela, D., Sylva, S.P., Sogin, M.L., and Jannasch, H.W. (2002). Microbial Diversity of Hydrothermal Sediments in the Guaymas Basin: Evidence for Anaerobic Methanotrophic Communities. Appl. Environ. Microbiol. *68*, 1994–2007.

Tettelin, H., Masignani, V., Cieslewicz, M.J., Donati, C., Medini, D., Ward, N.L., Angiuoli, S.V., Crabtree, J., Jones, A.L., Durkin, A.S., et al. (2005). Genome analysis of multiple pathogenic isolates of Streptococcus agalactiae: implications for the microbial "pan-genome." Proc. Natl. Acad. Sci. U. S. A. *102*, 13950–13955.

Tettelin, H., Riley, D., Cattuto, C., and Medini, D. (2008). Comparative genomics: the bacterial pan-genome. Curr. Opin. Microbiol. *11*, 472–477.

Thiel, A., Michoud, G., Moalic, Y., Flament, D., and Jebbar, M. (2014). Genetic Manipulations of the Hyperthermophilic Piezophilic Archaeon Thermococcus barophilus. Appl. Environ. Microbiol. *80*, 2299–2306.

Thompson, C.C., Amaral, G.R., Campeão, M., Edwards, R.A., Polz, M.F., Dutilh, B.E., Ussery, D.W., Sawabe, T., Swings, J., and Thompson, F.L. (2015). Microbial taxonomy in the post-genomic era: rebuilding from scratch? Arch. Microbiol. *197*, 359–370.

Thorgersen, M.P., Stirrett, K., Scott, R.A., and Adams, M.W.W. (2012). Mechanism of oxygen detoxification by the surprisingly oxygen-tolerant hyperthermophilic archaeon, Pyrococcus furiosus. Proc. Natl. Acad. Sci. *109*, 18547–18552.

Tindall, B.J., Rosselló-Móra, R., Busse, H.-J., Ludwig, W., and Kämpfer, P. (2010). Notes on the characterization of prokaryote strains for taxonomic purposes. Int. J. Syst. Evol. Microbiol. *60*, 249–266.

Touchon, M., Hoede, C., Tenaillon, O., Barbe, V., Baeriswyl, S., Bidet, P., Bingen, E., Bonacorsi, S., Bouchier, C., Bouvet, O., et al. (2009). Organised Genome Dynamics in the Escherichia coli Species Results in Highly Diverse Adaptive Paths. PLOS Genet *5*, e1000344.

Van Dongen, S. (2000). A Cluster algorithm for graphs. Rep. - Inf. Syst. 1–40.

Van Dongen, S., and Abreu-Goodger, C. (2012). Using MCL to Extract Clusters from Networks. In Bacterial Molecular Networks, J. van Helden, A. Toussaint, and D. Thieffry, eds. (New York, NY: Springer New York), pp. 281–295.

Vandamme, P., and Peeters, C. (2014).

Time to revisit polyphasic taxonomy. Antonie Van Leeuwenhoek *106*, 57–65.

Vannier, P., Marteinsson, V.T., Fridjonsson, O.H., Oger, P., and Jebbar, M. (2011). Complete genome sequence of the hyperthermophilic, piezophilic, heterotrophic, and carboxydotrophic archaeon Thermococcus barophilus MP. J. Bacteriol. *193*, 1481–1482.

Vanwonterghem, I., Evans, P.N., Parks, D.H., Jensen, P.D., Woodcroft, B.J., Hugenholtz, P., and Tyson, G.W. (2016). Methylotrophic methanogenesis discovered in the archaeal phylum Verstraetearchaeota. Nature Microbiology *1*, 16170.

Von Damm, K.L. (1995). Controls on the Chemistry and Temporal Variability of Seafloor Hydrothermal Fluids. In Seafloor Hydrothermal Systems: Physical, Chemical, Biological, and Geological Interactions, S.E. Humphris, R.A. Zierenberg, L.S. Mullineaux, and R.E. Thomson, eds. (American Geophysical Union), pp. 222–247.

Wagner, A., Whitaker, R.J., Krause, D.J., Heilers, J.-H., van Wolferen, M., van der Does, C., and Albers, S.-V. (2017). Mechanisms of gene flow in archaea. Nat. Rev. Microbiol. *advance online publication*.

Wang, F., Zhou, H., Meng, J., Peng, X., Jiang, L., Sun, P., Zhang, C., Nostrand, J.D.V., Deng, Y., He, Z., et al. (2009). GeoChip-based analysis of metabolic diversity of microbial communities at the Juan de Fuca Ridge hydrothermal vent. Proc. Natl. Acad. Sci. *106*, 4840–4845.

Wang, Q., Garrity, G.M., Tiedje, J.M., and Cole, J.R. (2007). Naive Bayesian classifier for rapid assignment of rRNA sequences into the new bacterial taxonomy. Appl. Environ. Microbiol. *73*, 5261–5267.

Wang, X., Gao, Z., Xu, X., and Ruan, L. (2011). Complete genome sequence of Thermococcus sp. strain 4557, a hyperthermophilic archaeon isolated from a deep-sea hydrothermal vent area. J. Bacteriol. *193*, 5544–5545.

Wayne, L.G., Brenner, D.J., Colwell, R.R., Grimont, P.A.D., Kandler, O., Krichevsky, M.I., Moore, L.H., Moore, W.E.C., Murray, R.G.E., Stackebrandt, E., et al. (1987). Report of the Ad Hoc Committee on Reconciliation of Approaches to Bacterial Systematics. Int. J. Syst. Evol. Microbiol. *37*, 463–464.

Welch, R.A., Burland, V., Plunkett, G., Redford, P., Roesch, P., Rasko, D., Buckles, E.L., Liou, S.-R., Boutin, A., Hackett, J., et al. (2002). Extensive mosaic structure revealed by the complete genome sequence of uropathogenic Escherichia coli. Proc. Natl. Acad. Sci. *99*, 17020–17024.

Wetzel, L.R., and Shock, E.L. (2000). Distinguishing ultramafic-from basalt-hosted submarine hydrothermal systems by comparing calculated vent fluid compositions. J. Geophys. Res. Solid Earth *105*, 8319–8340.

Whitaker, R.J., Grogan, D.W., and Taylor, J.W. (2003). Geographic Barriers Isolate Endemic Populations of

Hyperthermophilic Archaea. Science *301*, 976–978.

Whitaker, R.J., Grogan, D.W., and Taylor, J.W. (2005). Recombination Shapes the Natural Population Structure of the Hyperthermophilic Archaeon Sulfolobus islandicus. Mol. Biol. Evol. *22*, 2354–2361.

White, J.R., Escobar-Paramo, P., Mongodin, E.F., Nelson, K.E., and DiRuggiero, J. (2008). Extensive Genome Rearrangements and Multiple Horizontal Gene Transfers in a Population of Pyrococcus Isolates from Vulcano Island, Italy. Appl. Environ. Microbiol. *74*, 6447–6451.

Whittaker, R.H., and Margulis, L. (1978). Protist classification and the kingdoms of organisms. Biosystems *10*, 3–18.

Williams, T.A., Szöllősi, G.J., Spang, A., Foster, P.G., Heaps, S.E., Boussau, B., Ettema, T.J.G., and Embley, T.M. (2017). Integrative modeling of gene and genome evolution roots the archaeal tree of life. Proc. Natl. Acad. Sci. *114*, E4602–E4611.

Wirth, R. (2017). Colonization of Black Smokers by Hyperthermophilic Microorganisms. Trends Microbiol. *25*, 92–99.

Woese, C.R., and Fox, G.E. (1977). Phylogenetic structure of the prokaryotic domain: The primary kingdoms. Proc. Natl. Acad. Sci. *74*, 5088–5090.

Woese, C.R., Kandler, O., and Wheelis, M.L. (1990). Towards a natural system of organisms: proposal for the domains Archaea, Bacteria, and Eucarya. Proc. Natl. Acad. Sci. *87*, 4576–4579.

Wu, M., and Eisen, J.A. (2008). A simple, fast, and accurate method of phylogenomic inference. Genome Biol. *9*, R151.

Zaremba-Niedzwiedzka, K., Caceres, E.F., Saw, J.H., Bäckström, D., Juzokaite, L., Vancaester, E., Seitz, K.W., Anantharaman, K., Starnawski, P., Kjeldsen, K.U., et al. (2017). Asgard archaea illuminate the origin of eukaryotic cellular complexity. Nature *541*, 353–358.

Zeng, X., Birrien, J.-L., Fouquet, Y., Cherkashov, G., Jebbar, M., Querellou, J., Oger, P., Cambon-Bonavita, M.-A., Xiao, X., and Prieur, D. (2009). Pyrococcus CH1, an obligate piezophilic hyperthermophile: extending the upper pressure-temperature limits for life. ISME J. *3*, 873–876.

Zeng, X., Zhang, X., Jiang, L., Alain, K., Jebbar, M., and Shao, Z. (2013). Palaeococcus pacificus sp. nov., an archaeon from deep-sea hydrothermal sediment. Int. J. Syst. Evol. Microbiol. *63*, 2155–2159.

Zeng, X., Jebbar, M., and Shao, Z. (2015). Complete Genome Sequence of Hyperthermophilic Piezophilic Archaeon Palaeococcus pacificus DY20341T, Isolated from Deep-Sea Hydrothermal Sediments. Genome Announc. *3*.

Zhao, W., Zeng, X., and Xiao, X. (2015). Thermococcus eurythermalis sp. nov., a

conditional piezophilic, hyperthermophilic archaeon with a wide temperature range for growth, isolated from an oil-immersed chimney in the Guaymas Basin. Int. J. Syst. Evol. Microbiol. *65*, 30–35.

Zillig, W., and Reysenbach, A.-L. (2015). Thermococcales. In Bergey's Manual of Systematics of Archaea and Bacteria, (John Wiley & Sons, Ltd), p.

Zillig, W., Holz, I., Janekovic, D., Schäfer, W., and Reiter, W.D. (1983). The Archaebacterium Thermococcus celer Represents, a Novel Genus within the Thermophilic Branch of the Archaebacteria. Syst. Appl. Microbiol. *4*, 88–94.

Zillig, W., Holz, I., Klenk, H.-P., Trent, J., Wunderl, S., Janekovic, D., Imsel, E., and Haas, B. (1987). Pyrococcus woesei, sp. nov., an ultra-thermophilic marine archaebacterium, representing a novel order, Thermococcales. Syst. Appl. Microbiol. *9*, 62–70.

Zivanovic, Y., Armengaud, J., Lagorce, A., Leplat, C., Guérin, P., Dutertre, M., Anthouard, V., Forterre, P., Wincker, P., and Confalonieri, F. (2009). Genome analysis and genome-wide proteomics of Thermococcus gammatolerans, the most radioresistant organism known amongst the Archaea. Genome Biol. *10*, R70.

Appendix

Appendix

Appendix 1: *Thermococcales* isolates present in the UBOCC culture collection

strain name	16S-ITS sequenced?	RDP v2.11 classification	RDP support value	Genome sequenced?	Geographic Origin	precision
546	yes	*Pyrococcus*	100%	-	-	-
549	yes	*Pyrococcus*	100%	-	-	-
AMTc 01	yes	*Thermococcus*	100%	-	EPR 13°N	13N Grandbonum PP-52
AMTc 02	yes	*Thermococcus*	100%	-	EPR 13°N	13N Grandbonum PP-52
AMTc 03	yes	*Thermococcus*	100%	-	EPR 13°N	13N Grandbonum PP-52
AMTc 04	yes	*Thermococcus*	100%	-	EPR 13°N	13N Genesis PP-12
AMTc 05	yes	*Thermococcus*	100%	-	EPR 13°N	13N Genesis PP-12
AMTc 07	yes	*Thermococcus*	100%	-	EPR 13°N	13N Genesis PP-12
AMTc 09	yes	*Thermococcus*	100%	yes	EPR 13°N	13N Grandbonum PP-52
AMTc 10	yes	*Thermococcus*	100%	-	EPR 13°N	13N Pulsar PP-55
AMTc 101	yes	*Thermococcus*	100%	-	EPR 13°N	13N
AMTc 102	yes	*Thermococcus*	100%	yes	EPR 13°N	13N
AMTc 11	yes	*Thermococcus*	100%	-	EPR 13°N	13N Pulsar PP-55
AMTc 12	yes	*Thermococcus*	100%	-	EPR 13°N	13N Pulsar PP-55
AMTc 13	yes	*Thermococcus*	100%	-	EPR 13°N	13N Grandbonum PP-52
AMTc 14	yes	*Thermococcus*	100%	-	EPR 13°N	13N Grandbonum PP-52
AMTc 15	yes	*Thermococcus*	100%	-	EPR 13°N	13N Grandbonum PP-52
AMTc 16	yes	*Thermococcus*	100%	-	EPR 13°N	13N PP-Hot14
AMTc 17	yes	*Thermococcus*	100%	-	EPR 13°N	13N PP-Hot14
AMTc 18	yes	*Thermococcus*	100%	-	EPR 13°N	13N La chainette PP-57
AMTc 19	yes	*Thermococcus*	100%	yes	EPR 13°N	13N La chainette PP-57
AMTc 20	yes	*Thermococcus*	100%	-	EPR 13°N	13N PP-Hot14
AMTc 21	yes	*Thermococcus*	100%	-	EPR 13°N	13N PP-Hot14
AMTc 22	yes	*Thermococcus*	100%	-	EPR 13°N	13N Elsa
AMTc 23	yes	*Thermococcus*	100%	-	EPR 13°N	13N Elsa
AMTc 24	yes	*Thermococcus*	100%	-	EPR 13°N	13N Elsa
AMTc 26	yes	*Thermococcus*	100%	-	EPR 13°N	13N Elsa
AMTc 27	yes	*Thermococcus*	100%	-	EPR 13°N	13N Elsa
AMTc 29	yes	*Thermococcus*	100%	yes	EPR 13°N	13N Elsa
AMTc 30	yes	*Thermococcus*	100%	yes	EPR 13°N	13N Genesis PP-12
AMTc 31	yes	*Thermococcus*	100%	-	EPR 13°N	13N Genesis PP-12
AMTc 32	yes	*Thermococcus*	100%	-	EPR 13°N	13N Genesis PP-12
AMTc 33	yes	*Thermococcus*	100%	-	EPR 13°N	13N PP-Hot14
AMTc 34	yes	*Thermococcus*	100%	-	EPR 13°N	13N PP-Hot14
AMTc 35	yes	*Thermococcus*	100%	-	EPR 13°N	13N PP-Hot14
AMTc 36	yes	*Thermococcus*	100%	-	EPR 13°N	13N PP-Hot14
AMTc 38	yes	*Thermococcus*	100%	-	EPR 13°N	13N PP-Hot14
AMTc 40	yes	*Thermococcus*	100%	-	EPR 13°N	13N PP-Hot14
AMTc 41	yes	*Thermococcus*	100%	-	EPR 13°N	13N PP-Hot14
AMTc 42	yes	*Thermococcus*	100%	-	EPR 13°N	13N Grandbonum PP-52
AMTc 43	yes	*Thermococcus*	100%	-	EPR 13°N	13N Grandbonum PP-52
AMTc 44	yes	*Thermococcus*	100%	-	EPR 13°N	13N Grandbonum PP-52
AMTc 45	yes	*Thermococcus*	100%	-	EPR 13°N	13N Grandbonum PP-52
AMTc 46	yes	*Thermococcus*	100%	-	EPR 13°N	13N Grandbonum PP-52
AMTc 47	yes	*Thermococcus*	100%	-	EPR 13°N	13N Grandbonum PP-52
AMTc 48	yes	*Thermococcus*	100%	-	EPR 13°N	13N Grandbonum PP-52
AMTc 49	yes	*Thermococcus*	100%	-	EPR 13°N	13N Grandbonum PP-52

AMTc 50	yes	*Thermococcus*	100%	-	EPR 13°N	13N Grandbonum PP-52
AMTc 51	yes	*Thermococcus*	100%	yes	EPR 13°N	13N
AMTc 52	yes	*Thermococcus*	100%	yes	EPR 13°N	13N
AMTc 53	yes	*Thermococcus*	100%	-	EPR 13°N	13N
AMTc 54	yes	*Thermococcus*	100%	-	EPR 13°N	13N
AMTc 55	yes	*Thermococcus*	100%	-	EPR 13°N	13N
AMTc 57	yes	*Thermococcus*	100%	-	EPR 13°N	13N Totem
AMTc 59	yes	*Thermococcus*	100%	-	EPR 13°N	13N Totem
AMTc 62	yes	*Thermococcus*	100%	-	EPR 13°N	13N
AMTc 66	no	-	-	-	-	-
AMTc 67	yes	*Thermococcus*	100%	yes	EPR 13°N	13N
AMTc 69	yes	*Thermococcus*	100%	-	EPR 13°N	13N
AMTc 70	yes	*Thermococcus*	100%	yes	EPR 13°N	13N Genesis PP-12
AMTc 71	yes	*Thermococcus*	100%	yes	EPR 13°N	13N Genesis PP-12
AMTc 72	yes	*Thermococcus*	100%	yes	EPR 13°N	13N Genesis PP-12
AMTc 73	yes	*Thermococcus*	100%	yes	EPR 13°N	13N Genesis PP-12
AMTc 76	yes	*Thermococcus*	100%	-	EPR 13°N	13N Genesis PP-12
AMTc 77	yes	*Thermococcus*	100%	-	EPR 13°N	13N Genesis PP-12
AMTc 78	yes	*Thermococcus*	100%	-	EPR 13°N	13N Pulsar PP-55
AMTc 79	yes	*Thermococcus*	100%	yes	EPR 13°N	13N Pulsar PP-55
AMTc 83	yes	*Thermococcus*	100%	-	EPR 13°N	13N La chainette PP-57
AMTc 84	yes	*Thermococcus*	100%	-	EPR 13°N	13N La chainette PP-57
AMTc 85	yes	*Thermococcus*	100%	yes	EPR 13°N	13N La chainette PP-57
AMTc 87	yes	*Thermococcus*	100%	-	EPR 13°N	13N Pulsar PP-55
AMTc 90	yes	*Thermococcus*	100%	-	EPR 13°N	13N Pulsar PP-55
AMTc 92	yes	*Thermococcus*	100%	-	EPR 13°N	13N Grandbonum PP-52
AMTc 93	yes	*Thermococcus*	100%	-	EPR 13°N	13N Grandbonum PP-52
AMTc 94	yes	*Thermococcus*	100%	yes	EPR 13°N	13N La chainette PP-57
AMTc 95	yes	*Thermococcus*	100%	yes	EPR 13°N	13N La chainette PP-57
AMTc 96	yes	*Thermococcus*	100%	-	EPR 13°N	13N Elsa
AMTc 97	yes	*Thermococcus*	100%	-	EPR 13°N	13N PP-Hot14
AMTc 98	yes	*Thermococcus*	100%	-	EPR 13°N	13N Grandbonum PP-52
AMTc 99	yes	*Thermococcus*	100%	-	EPR 13°N	13N
CIR 02c	yes	*Thermococcus*	100%	-	Indian Ocean	-25; 75
CIR 03a	yes	*Thermococcus*	100%	-	Indian Ocean	-25; 75
CIR 04a	yes	*Thermococcus*	100%	-	Indian Ocean	-25; 75
CIR 05a	yes	*Thermococcus*	100%	-	Indian Ocean	-25; 75
CIR 06a	yes	*Thermococcus*	100%	-	Indian Ocean	-25; 75
CIR 07a	yes	*Thermococcus*	100%	-	Indian Ocean	-25; 75
CIR 08a	yes	*Thermococcus*	100%	-	Indian Ocean	-25; 75
CIR 09a	yes	*Thermococcus*	100%	-	Indian Ocean	-25; 75
CIR 10a	yes	*Thermococcus*	100%	-	Indian Ocean	-25; 75
CIR 12a	no	-	-	-	-	-
CIR 14a	yes	*Thermococcus*	100%	-	Indian Ocean	-25; 75
CIR 15a	yes	*Thermococcus*	100%	-	Indian Ocean	-25; 75
CIR 16a	yes	*Thermococcus*	100%	-	Indian Ocean	-25; 75
E10p10	no	-	-	-	-	-
E10p11	yes	*Thermococcus*	100%	yes	EPR 9°N	-
E10p12	yes	*Thermococcus*	100%	-	EPR 9°N	-
E10p13	no	-	-	-	-	-
E10p14	yes	*Thermococcus*	100%	-	EPR 9°N	-
E10p15	yes	*Thermococcus*	100%	-	EPR 9°N	-
E10p2	yes	*Thermococcus*	100%	-	EPR 9°N	-
E10p3	yes	*Thermococcus*	100%	-	EPR 9°N	-
E10p4	yes	*Thermococcus*	100%	-	EPR 9°N	-
E10p5	yes	*Thermococcus*	100%	-	EPR 9°N	-
E10p6	yes	*Thermococcus*	100%	-	EPR 9°N	-
E10p7	yes	*Thermococcus*	100%	yes	EPR 9°N	-

E10p8	yes	*Thermococcus*	100%	yes	EPR 9°N	-
E10p9	yes	*Thermococcus*	100%	-	EPR 9°N	-
E11d1	yes	*Thermococcus*	100%	-	EPR 9°N	-
E12d10	yes	*Thermococcus*	100%	-	EPR 9°N	-
E12d13	yes	*Thermococcus*	100%	-	EPR 9°N	-
E12d5	yes	*Thermococcus*	100%	-	EPR 9°N	-
E12d9	yes	*Thermococcus*	100%	-	EPR 9°N	-
E13p1	yes	*Thermococcus*	100%	-	EPR 9°N	-
E13p11	yes	*Thermococcus*	100%	-	EPR 9°N	-
E13p2	yes	*Thermococcus*	100%	-	EPR 9°N	-
E13p3	yes	*Thermococcus*	100%	-	EPR 9°N	-
E13p4	yes	*Thermococcus*	100%	-	EPR 9°N	-
E13p5	yes	*Thermococcus*	100%	-	EPR 9°N	-
E13p6	yes	*Thermococcus*	100%	-	EPR 9°N	-
E13p8	yes	*Thermococcus*	100%	-	EPR 9°N	-
E14d1	yes	*Thermococcus*	100%	-	EPR 9°N	-
E14d10	yes	*Thermococcus*	100%	-	EPR 9°N	-
E14d11	yes	*Thermococcus*	100%	-	EPR 9°N	-
E14d12	yes	*Thermococcus*	100%	-	EPR 9°N	-
E14d2	yes	*Thermococcus*	100%	-	EPR 9°N	-
E14d3	yes	*Thermococcus*	100%	-	EPR 9°N	-
E14d4	yes	*Thermococcus*	100%	-	EPR 9°N	-
E14d5	yes	*Thermococcus*	100%	-	EPR 9°N	-
E14d6	yes	*Thermococcus*	100%	-	EPR 9°N	-
E14d7	yes	*Thermococcus*	100%	-	EPR 9°N	-
E14d8	yes	*Thermococcus*	100%	-	EPR 9°N	-
E14d9	yes	*Thermococcus*	100%	-	EPR 9°N	-
E14p13	yes	*Thermococcus*	100%	-	EPR 9°N	-
E14p14	yes	*Thermococcus*	100%	-	EPR 9°N	-
E14p15	yes	*Thermococcus*	100%	-	EPR 9°N	-
E14p16	yes	*Thermococcus*	100%	-	EPR 9°N	-
E14p17	yes	*Thermococcus*	100%	-	EPR 9°N	-
E14p18	yes	*Thermococcus*	100%	-	EPR 9°N	-
E14p19	yes	*Thermococcus*	100%	yes	EPR 9°N	-
E14p20	no	-	-	-	-	-
E14p21	yes	*Thermococcus*	100%	-	EPR 9°N	-
E14p22	yes	*Thermococcus*	100%	-	EPR 9°N	-
E14p23	yes	*Thermococcus*	100%	-	EPR 9°N	-
E15d13	yes	*Thermococcus*	100%	-	EPR 9°N	-
E15d14	yes	*Thermococcus*	100%	-	EPR 9°N	-
E15d15	yes	*Thermococcus*	100%	-	EPR 9°N	-
E15d16	yes	*Thermococcus*	100%	-	EPR 9°N	-
E15d17	no	-	-	-	-	-
E15d18	no	-	-	-	-	-
E15d19	yes	*Thermococcus*	100%	-	EPR 9°N	-
E15d20	yes	*Thermococcus*	100%	-	EPR 9°N	-
E15d21	no	-	-	-	-	-
E15d22	yes	*Thermococcus*	100%	-	EPR 9°N	-
E15d23	yes	*Thermococcus*	100%	-	EPR 9°N	-
E15d24	no	-	-	-	-	-
E15p1	yes	*Thermococcus*	100%	-	EPR 9°N	-
E15p10	yes	*Thermococcus*	100%	-	EPR 9°N	-
E15p11	yes	*Thermococcus*	100%	-	EPR 9°N	-
E15p12	no	-	-	-	-	-
E15p2	yes	*Thermococcus*	100%	-	EPR 9°N	-
E15p25	yes	*Thermococcus*	100%	yes	EPR 9°N	-
E15p26	yes	*Thermococcus*	100%	-	EPR 9°N	-
E15p27	yes	*Thermococcus*	100%	-	EPR 9°N	-

E15p28	yes	*Thermococcus*	100%	-	EPR 9°N	-
E15p29	yes	*Thermococcus*	100%	yes	EPR 9°N	-
E15p3	yes	*Thermococcus*	100%	-	EPR 9°N	-
E15p30	yes	*Thermococcus*	100%	yes	EPR 9°N	-
E15p31	yes	*Thermococcus*	100%	-	EPR 9°N	-
E15p32	yes	*Thermococcus*	100%	-	EPR 9°N	-
E15p33	yes	*Thermococcus*	100%	-	EPR 9°N	-
E15p34	no	-	-	-	-	-
E15p35	yes	*Thermococcus*	100%	yes	EPR 9°N	-
E15p4	yes	*Thermococcus*	100%	-	EPR 9°N	-
E15p5	no	-	-	-	-	-
E15p6	yes	*Thermococcus*	100%	yes	EPR 9°N	-
E15p7	yes	*Thermococcus*	100%	-	EPR 9°N	-
E15p8	yes	*Thermococcus*	100%	-	EPR 9°N	-
E15p9	yes	*Thermococcus*	100%	-	EPR 9°N	-
E1d1	yes	*Thermococcus*	100%	-	EPR 9°N	-
E1d10	yes	*Thermococcus*	100%	-	EPR 9°N	-
E1d2	yes	*Thermococcus*	100%	-	EPR 9°N	-
E1d3	yes	*Thermococcus*	100%	-	EPR 9°N	-
E1d5	yes	*Thermococcus*	100%	-	EPR 9°N	-
E1d6	yes	*Thermococcus*	100%	-	EPR 9°N	-
E1d7	yes	*Thermococcus*	100%	-	EPR 9°N	-
E1d8	yes	*Thermococcus*	100%	-	EPR 9°N	-
E2p1	yes	*Thermococcus*	100%	-	EPR 9°N	-
E2p10	yes	*Thermococcus*	100%	-	EPR 9°N	-
E2p11	yes	*Thermococcus*	100%	-	EPR 9°N	-
E2p12	yes	*Thermococcus*	100%	-	EPR 9°N	-
E2p13	yes	*Thermococcus*	100%	-	EPR 9°N	-
E2p14	yes	*Thermococcus*	100%	-	EPR 9°N	-
E2p15	yes	*Thermococcus*	100%	-	EPR 9°N	-
E2p16	yes	*Thermococcus*	100%	-	EPR 9°N	-
E2p2	yes	*Thermococcus*	100%	-	EPR 9°N	-
E2p3	yes	*Thermococcus*	100%	-	EPR 9°N	-
E2p4	yes	*Thermococcus*	100%	-	EPR 9°N	-
E2p5	yes	*Thermococcus*	100%	-	EPR 9°N	-
E2p6	yes	*Thermococcus*	100%	-	EPR 9°N	-
E2p7	yes	*Thermococcus*	100%	-	EPR 9°N	-
E2p8	yes	*Thermococcus*	100%	-	EPR 9°N	-
E2p9	yes	*Thermococcus*	100%	-	EPR 9°N	-
E4d12	yes	*Thermococcus*	100%	-	EPR 9°N	-
E4d2	yes	*Thermococcus*	100%	-	EPR 9°N	-
E4d3	yes	*Thermococcus*	100%	-	EPR 9°N	-
E4d5	yes	*Thermococcus*	100%	-	EPR 9°N	-
E4d7	yes	*Thermococcus*	100%	-	EPR 9°N	-
E4d8	yes	*Thermococcus*	100%	-	EPR 9°N	-
E4p13	yes	*Thermococcus*	100%	-	EPR 9°N	-
E4p18	yes	*Thermococcus*	100%	-	EPR 9°N	-
E4p19	yes	*Thermococcus*	100%	-	EPR 9°N	-
E5p1	yes	*Thermococcus*	100%	-	EPR 9°N	-
E7p13	yes	*Thermococcus*	100%	-	EPR 9°N	-
EXT 01c	yes	*Thermococcus*	100%	-	EPR 9°N	-
EXT 02c	no	-	-	-	-	-
EXT 03c	no	-	-	-	-	-
EXT 04c	no	-	-	-	-	-
EXT 05c	yes	*Thermococcus*	100%	-	EPR 9°N	-
EXT 06c	yes	*Thermococcus*	100%	-	EPR 9°N	-
EXT 07c	no	-	-	-	-	-
EXT 08c	yes	*Thermococcus*	100%	yes	EPR 9°N	-

EXT 09c	yes	*Thermococcus*	100%	-	EPR 9°N	-
EXT 10c	yes	*Thermococcus*	100%	yes	EPR 9°N	-
EXT 11c	yes	*Thermococcus*	100%	yes	EPR 9°N	-
EXT 12c	yes	*Thermococcus*	100%	yes	EPR 9°N	-
EXT 13c	yes	*Thermococcus*	100%	yes	EPR 9°N	-
EXT 15c	yes	*Pyrococcus*	100%	-	EPR 9°N	-
EXT 16c	yes	*Pyrococcus*	100%	-	EPR 9°N	-
GE 01	yes	*Pyrococcus*	100%	-	-	-
GE 02	no	-	-	-	-	-
GE 03	no	-	-	-	-	-
GE 05	no	-	-	-	-	-
GE 05a	no	-	-	-	-	-
GE 06	yes	*Thermococcus*	100%	-	-	cruise Starmer; site White lady
GE 07	yes	*Pyrococcus*	100%	-	-	-
GE 08	yes	*Thermococcus*	100%	-	-	cruise Starmer; site White lady
GE 18	no	-	-	-	-	-
GE 19	yes	*Thermococcus*	100%	-	-	-
GE 20	no	-	-	-	-	-
GE 21	no	-	-	-	-	-
GE 22	no	-	-	-	-	-
GE 23	yes	*Pyrococcus*	100%	-	-	-
GE 25	no	-	-	-	-	-
GE 26	yes	*Pyrococcus*	100%	-	-	cruise Starmer; site White lady
GE 27	yes	*Pyrococcus*	100%	-	-	cruise Starmer; site White lady
GE 30	no	-	-	-	-	-
GE 31	yes	*Thermococcus*	100%	-	-	-
GE 32	no	-	-	-	-	-
IRI 01c	yes	*Thermococcus*	100%	-	Mid Atlantic Ridge	Rainbow
IRI 02c	yes	*Thermococcus*	100%	-	Mid Atlantic Ridge	Rainbow
IRI 03c	yes	*Thermococcus*	100%	-	Mid Atlantic Ridge	Rainbow
IRI 05c	yes	*Thermococcus*	100%	-	Mid Atlantic Ridge	Rainbow
IRI 06c	yes	*Thermococcus*	100%	yes	Mid Atlantic Ridge	Rainbow
IRI 07c	yes	*Thermococcus*	100%	yes	Mid Atlantic Ridge	Rainbow
IRI 09c	yes	*Thermococcus*	100%	yes	Mid Atlantic Ridge	Rainbow
IRI 10c	yes	*Thermococcus*	100%	yes	Mid Atlantic Ridge	Rainbow
IRI 14c	yes	*Thermococcus*	100%	yes	Mid Atlantic Ridge	Rainbow
IRI 15c	yes	*Thermococcus*	100%	yes	Mid Atlantic Ridge	Rainbow
IRI 17c	yes	*Thermococcus*	100%	-	Mid Atlantic Ridge	Rainbow
IRI 19c	yes	*Thermococcus*	100%	-	Mid Atlantic Ridge	Rainbow
IRI 21c	yes	*Thermococcus*	100%	-	Mid Atlantic Ridge	Rainbow
IRI 23c	no	-	-	-	-	-
IRI 24c	yes	*Thermococcus*	100%	yes	Mid Atlantic Ridge	Rainbow
IRI 25c	yes	*Thermococcus*	100%	yes	Mid Atlantic Ridge	Rainbow
IRI 26c	yes	*Thermococcus*	100%	yes	Mid Atlantic Ridge	Rainbow
IRI 27 c2	yes	*Thermococcus*	100%	yes	Mid Atlantic Ridge	Rainbow
IRI 29c	yes	*Thermococcus*	100%	yes	Mid Atlantic Ridge	Rainbow
IRI 30c	yes	*Thermococcus*	100%	-	Mid Atlantic Ridge	Rainbow
IRI 31b	yes	*Thermococcus*	100%	-	Mid Atlantic Ridge	Rainbow (IRIS 07)
IRI 32b	yes	*Thermococcus*	100%	-	Mid Atlantic Ridge	Rainbow (IRIS 09)
IRI 33c	yes	*Thermococcus*	100%	-	Mid Atlantic Ridge	Menez Gwen
IRI 34c	yes	*Thermococcus*	100%	-	Mid Atlantic Ridge	Menez Gwen
IRI 35c	yes	*Thermococcus*	100%	yes	Mid Atlantic Ridge	Menez Gwen
IRI 35c2	yes	*Thermococcus*	100%	yes	Mid Atlantic Ridge	Menez Gwen
IRI 36c	yes	*Thermococcus*	100%	yes	Mid Atlantic Ridge	Menez Gwen

IRI 37c	yes	*Pyrococcus*	100%	-	Mid Atlantic Ridge	Menez Gwen
IRI 38c	yes	*Pyrococcus*	100%	-	Mid Atlantic Ridge	Menez Gwen
IRI 39b	yes	*Pyrococcus*	100%	-	Mid Atlantic Ridge	Menez Gwen
IRI 40b	no	-	-	-	-	-
IRI 42c	yes	*Pyrococcus*	100%	-	Mid Atlantic Ridge	Rainbow
IRI 43a	yes	*Thermococcus*	100%	-	Mid Atlantic Ridge	Rainbow (IRIS 07)
IRI 44a	yes	*Thermococcus*	100%	-	Mid Atlantic Ridge	Rainbow (IRIS 07)
IRI 45a	yes	*Thermococcus*	100%	-	Mid Atlantic Ridge	Rainbow (IRIS 07)
IRI 47a	yes	*Thermococcus*	100%	-	Mid Atlantic Ridge	Rainbow (IRIS 07)
IRI 48a	yes	*Thermococcus*	100%	-	Mid Atlantic Ridge	Rainbow (IRIS 07)
IRI 49a	yes	*Thermococcus*	100%	-	Mid Atlantic Ridge	Rainbow (IRIS 07)
IRI 50a	yes	*Thermococcus*	100%	-	Mid Atlantic Ridge	Rainbow (IRIS 07)
IRI 51a	no	-	-	-	-	-
IRI 52a	yes	*Thermococcus*	100%	-	Mid Atlantic Ridge	Rainbow (IRIS 07)
IRI 53a	yes	*Thermococcus*	100%	-	Mid Atlantic Ridge	Rainbow (IRIS 07)
IRI 54a	yes	*Thermococcus*	100%	-	Mid Atlantic Ridge	Rainbow (IRIS 07)
IRI 55a	yes	*Thermococcus*	100%	-	Mid Atlantic Ridge	Rainbow (IRIS 07)
IRI 56a	yes	*Thermococcus*	100%	-	Mid Atlantic Ridge	Rainbow (IRIS 07)
IRI 57a	yes	*Pyrococcus*	100%	-	Mid Atlantic Ridge	Rainbow (IRIS 07)
IRI 58a	yes	*Thermococcus*	100%	-	Mid Atlantic Ridge	Rainbow (IRIS 07)
IRI 59a	yes	*Thermococcus*	100%	-	Mid Atlantic Ridge	Rainbow (IRIS 07)
IRI 60a	yes	*Thermococcus*	100%	-	Mid Atlantic Ridge	Rainbow (IRIS 07)
IRI 61a	yes	*Thermococcus*	100%	-	Mid Atlantic Ridge	Rainbow (IRIS 07)
KAZA	yes	*Thermococcus*	100%	-	Mid Atlantic Ridge	Rainbow
m 13	no	-	-	-	-	-
m 2	no	-	-	-	-	-
m 20	no	-	-	-	-	-
MC4	yes	*Thermococcus*	100%	yes	South Indian Ocean	Saint-Paul
MC5	yes	*Thermococcus*	100%	yes	South Indian Ocean	Saint-Paul
MC8	yes	*Thermococcus*	100%	yes	South Indian Ocean	Saint-Paul
MC9	yes	*Thermococcus*	100%	yes	South Indian Ocean	Saint-Paul

Appendix 2: List of all metagenomes mapped on *Thermococcales* genomes

Sample type	Location	Depth	SRA ID	Library type	Reference
deep_sea_hydrothermal_vent	Axial Seamount, Juan de Fuca	1500m	SRR1688294	Metatranscriptomic	-
deep_sea_hydrothermal_vent	Axial Seamount, Juan de Fuca	1500m	SRR1688295	Metatranscriptomic	-
deep_sea_hydrothermal_vent	Axial Seamount, Juan de Fuca	1500m	SRR1688296	Metatranscriptomic	-
deep_sea_hydrothermal_vent	Axial Seamount, Juan de Fuca	1500m	SRR1688298	Metatranscriptomic	-
deep_sea_hydrothermal_vent	Axial Seamount, Juan de Fuca	1500m	SRR1688300	Metatranscriptomic	-
deep_sea_hydrothermal_vent	Axial Seamount, Juan de Fuca	1500m	SRR1688303	Metatranscriptomic	-
deep_sea_hydrothermal_vent	Axial Seamount, Juan de Fuca	1500m	SRR1688306	Metatranscriptomic	-
deep_sea_hydrothermal_vent	Axial Seamount, Juan de Fuca	1500m	SRR1688313	Metatranscriptomic	-
deep_sea_hydrothermal_vent	Axial Seamount, Juan de Fuca	1500m	SRR1688314	Metatranscriptomic	-
deep_sea_hydrothermal_vent	Axial Seamount, Juan de Fuca	1500m	SRR1688315	Metatranscriptomic	-
deep_sea_hydrothermal_vent	Axial Seamount, Juan de Fuca	1500m	SRR1688316	Metatranscriptomic	-
deep_sea_hydrothermal_vent	Axial Seamount, Juan de Fuca	1500m	SRR1688317	Metatranscriptomic	-
deep_sea_hydrothermal_vent	Axial Seamount, Juan de Fuca	1500m	SRR1688318	Metatranscriptomic	-
deep_sea_hydrothermal_vent	Axial Seamount, Juan de Fuca	1500m	SRR1688319	Metatranscriptomic	-
deep_sea_hydrothermal_vent	Axial Seamount, Juan de Fuca	1500m	SRR1688320	Metatranscriptomic	-
deep_sea_hydrothermal_vent	Axial Seamount, Juan de Fuca	1500m	SRR1688321	Metatranscriptomic	-
deep_sea_hydrothermal_vent	Shrimp Gulley #2, Piccard, Mid Cayman rise	4940m	ERR868087	Metagenomic	Reveillaud et al., 2016
deep_sea_hydrothermal_vent	Shrimp Gulley #2, Piccard, Mid Cayman rise	4940m	ERR868088	Metagenomic	Reveillaud et al., 2016
deep_sea_hydrothermal_vent	Shrimp Gulley #2, Piccard, Mid Cayman rise	4940m	ERR868089	Metagenomic	Reveillaud et al., 2016
deep_sea_hydrothermal_vent	Shrimp Gulley #2, Piccard, Mid Cayman rise	4940m	ERR868090	Metagenomic	Reveillaud et al., 2016
deep_sea_hydrothermal_vent	Shrimp Gulley #2, Piccard, Mid Cayman rise	4940m	ERR868091	Metagenomic	Reveillaud et al., 2016
deep_sea_hydrothermal_vent	Shrimp Gulley #2, Piccard, Mid Cayman rise	4940m	ERR868092	Metagenomic	Reveillaud et al., 2016
deep_sea_hydrothermal_vent	Shrimp Gulley #2, Piccard, Mid Cayman rise	4940m	ERR868093	Metagenomic	Reveillaud et al., 2016
deep_sea_hydrothermal_vent	Shrimp Gulley #2, Piccard, Mid Cayman rise	4940m	ERR868094	Metagenomic	Reveillaud et al., 2016
deep_sea_hydrothermal_vent	Shrimp Gulley #2, Piccard, Mid Cayman rise	4940m	ERR868095	Metagenomic	Reveillaud et al., 2016
deep_sea_hydrothermal_vent	Shrimp Gulley #2, Piccard, Mid Cayman rise	4940m	ERR868096	Metagenomic	Reveillaud et al., 2016
deep_sea_hydrothermal_vent	Shrimp Gulley #2, Piccard, Mid Cayman rise	4940m	ERR868097	Metagenomic	Reveillaud et al., 2016
deep_sea_hydrothermal_vent	Shrimp Gulley #2, Piccard, Mid Cayman rise	4940m	ERR868098	Metagenomic	Reveillaud et al., 2016
deep_sea_hydrothermal_vent	Shrimp Gulley #2, Piccard, Mid Cayman rise	4940m	ERR868099	Metagenomic	Reveillaud et al., 2016
deep_sea_hydrothermal_vent	Shrimp Gulley #2, Piccard, Mid Cayman rise	4940m	ERR868100	Metagenomic	Reveillaud et al., 2016
deep_sea_hydrothermal_vent	Shrimp Gulley #2, Piccard, Mid Cayman rise	4940m	ERR868101	Metagenomic	Reveillaud et al., 2016
deep_sea_hydrothermal_vent	Shrimp Gulley #2, Piccard, Mid Cayman rise	4940m	ERR868102	Metagenomic	Reveillaud et al., 2016
deep_sea_hydrothermal_vent	Shrimp Gulley #2, Piccard, Mid Cayman rise	4940m	ERR868103	Metagenomic	Reveillaud et al., 2016
deep_sea_hydrothermal_vent	Shrimp Gulley #2, Piccard, Mid Cayman rise	4940m	ERR868104	Metagenomic	Reveillaud et al., 2016
deep_sea_hydrothermal_vent	Shrimp Gulley #2, Piccard, Mid Cayman rise	4940m	ERR868105	Metagenomic	Reveillaud et al., 2016
deep_sea_hydrothermal_vent	Shrimp Gulley #2, Piccard, Mid Cayman rise	4940m	ERR868106	Metagenomic	Reveillaud et al., 2016
deep_sea_hydrothermal_vent	Shrimp Gulley #2, Piccard, Mid Cayman rise	4940m	ERR868107	Metagenomic	Reveillaud et al., 2016

deep_sea_hydrothermal_vent	Shrimp Gulley #2, Piccard, Mid Cayman rise	4940m	ERR868108	Metagenomic	Reveillaud et al., 2016
deep_sea_hydrothermal_vent	Shrimp Gulley #2, Piccard, Mid Cayman rise	4940m	ERR868109	Metagenomic	Reveillaud et al., 2016
deep_sea_hydrothermal_vent	Shrimp Gulley #2, Piccard, Mid Cayman rise	4940m	ERR868110	Metagenomic	Reveillaud et al., 2016
deep_sea_hydrothermal_vent	Shrimp Gulley #2, Piccard, Mid Cayman rise	4940m	ERR868111	Metagenomic	Reveillaud et al., 2016
deep_sea_hydrothermal_vent	Shrimp Gulley #2, Piccard, Mid Cayman rise	4940m	ERR868112	Metagenomic	Reveillaud et al., 2016
deep_sea_hydrothermal_vent	Shrimp Gulley #2, Piccard, Mid Cayman rise	4940m	ERR868113	Metagenomic	Reveillaud et al., 2016
deep_sea_hydrothermal_vent	Shrimp Gulley #2, Piccard, Mid Cayman rise	4940m	ERR868114	Metagenomic	Reveillaud et al., 2016
deep_sea_hydrothermal_vent	Shrimp Gulley #2, Piccard, Mid Cayman rise	4940m	ERR868115	Metagenomic	Reveillaud et al., 2016
deep_sea_hydrothermal_vent	Ginger Castle, Von Damm, Mid Cayman rise	2308m	ERR868031	Metagenomic	Reveillaud et al., 2016
deep_sea_hydrothermal_vent	Ginger Castle, Von Damm, Mid Cayman rise	2308m	ERR868032	Metagenomic	Reveillaud et al., 2016
deep_sea_hydrothermal_vent	Ginger Castle, Von Damm, Mid Cayman rise	2308m	ERR868033	Metagenomic	Reveillaud et al., 2016
deep_sea_hydrothermal_vent	Ginger Castle, Von Damm, Mid Cayman rise	2308m	ERR868034	Metagenomic	Reveillaud et al., 2016
deep_sea_hydrothermal_vent	Ginger Castle, Von Damm, Mid Cayman rise	2308m	ERR868035	Metagenomic	Reveillaud et al., 2016
deep_sea_hydrothermal_vent	Ginger Castle, Von Damm, Mid Cayman rise	2308m	ERR868036	Metagenomic	Reveillaud et al., 2016
deep_sea_hydrothermal_vent	Ginger Castle, Von Damm, Mid Cayman rise	2308m	ERR868037	Metagenomic	Reveillaud et al., 2016
deep_sea_hydrothermal_vent	Ginger Castle, Von Damm, Mid Cayman rise	2308m	ERR868038	Metagenomic	Reveillaud et al., 2016
deep_sea_hydrothermal_vent	Ginger Castle, Von Damm, Mid Cayman rise	2308m	ERR868039	Metagenomic	Reveillaud et al., 2016
deep_sea_hydrothermal_vent	Ginger Castle, Von Damm, Mid Cayman rise	2308m	ERR868040	Metagenomic	Reveillaud et al., 2016
deep_sea_hydrothermal_vent	Ginger Castle, Von Damm, Mid Cayman rise	2308m	ERR868041	Metagenomic	Reveillaud et al., 2016
deep_sea_hydrothermal_vent	Ginger Castle, Von Damm, Mid Cayman rise	2308m	ERR868042	Metagenomic	Reveillaud et al., 2016
deep_sea_hydrothermal_vent	Ginger Castle, Von Damm, Mid Cayman rise	2308m	ERR868043	Metagenomic	Reveillaud et al., 2016
deep_sea_hydrothermal_vent	Ginger Castle, Von Damm, Mid Cayman rise	2308m	ERR868044	Metagenomic	Reveillaud et al., 2016
deep_sea_hydrothermal_vent	Ginger Castle, Von Damm, Mid Cayman rise	2308m	ERR868045	Metagenomic	Reveillaud et al., 2016
deep_sea_hydrothermal_vent	Ginger Castle, Von Damm, Mid Cayman rise	2308m	ERR868046	Metagenomic	Reveillaud et al., 2016
deep_sea_hydrothermal_vent	Ginger Castle, Von Damm, Mid Cayman rise	2308m	ERR868047	Metagenomic	Reveillaud et al., 2016
deep_sea_hydrothermal_vent	Ginger Castle, Von Damm, Mid Cayman rise	2308m	ERR868048	Metagenomic	Reveillaud et al., 2016
deep_sea_hydrothermal_vent	Ginger Castle, Von Damm, Mid Cayman rise	2308m	ERR868049	Metagenomic	Reveillaud et al., 2016
deep_sea_hydrothermal_vent	Ginger Castle, Von Damm, Mid Cayman rise	2308m	ERR868050	Metagenomic	Reveillaud et al., 2016
deep_sea_hydrothermal_vent	Ginger Castle, Von Damm, Mid Cayman rise	2308m	ERR868051	Metagenomic	Reveillaud et al., 2016
deep_sea_hydrothermal_vent	Ginger Castle, Von Damm, Mid Cayman rise	2308m	ERR868052	Metagenomic	Reveillaud et al., 2016
deep_sea_hydrothermal_vent	Ginger Castle, Von Damm, Mid Cayman rise	2308m	ERR868053	Metagenomic	Reveillaud et al., 2016
deep_sea_hydrothermal_vent	Guaymas Bassin	1600m	SRR453184	Metatranscriptomic	Baker et al., 2013
deep_sea_hydrothermal_vent	Guaymas Bassin	1950m	SRR452448	Metatranscriptomic	Baker et al., 2013
deep_sea_hydrothermal_sediment	Guaymas Bassin	NA	SRR5214303	Metagenomic	-
deep_sea_hydrothermal_vent	Axial Seamount, Juan de Fuca	NA	ERR420342	Metatranscriptomic	-
deep_sea_hydrothermal_vent	Axial Seamount, Juan de Fuca	NA	ERR420343	Metatranscriptomic	-
deep_sea_hydrothermal_vent	Axial Seamount, Juan de Fuca	NA	ERR420344	Metagenomic	-
deep_sea_hydrothermal_vent	Axial Seamount, Juan de Fuca	NA	ERR420345	Metagenomic	-
deep_sea_hydrothermal_vent	Axial Seamount, Juan de Fuca	NA	ERR420346	Metagenomic	-
deep_sea_hydrothermal_vent	Axial Seamount, Juan de Fuca	NA	ERR420347	Metagenomic	-
deep_sea_hydrothermal_vent	Axial Seamount, Juan de Fuca	NA	ERR420348	Metagenomic	-
deep_sea_hydrothermal_vent	Axial Seamount, Juan de Fuca	NA	ERR420349	Metatranscriptomic	-
deep_sea_hydrothermal_vent	Axial Seamount, Juan de Fuca	NA	ERR420350	Metatranscriptomic	-
deep_sea_hydrothermal_vent	Axial Seamount, Juan de Fuca	NA	ERR420351	Metatranscriptomic	-

deep_sea_hydrothermal_vent	Axial Seamount, Juan de Fuca	NA	ERR420352	Metatranscriptomic	-
deep_sea_hydrothermal_vent	Axial Seamount, Juan de Fuca	NA	ERR420353	Metagenomic	-
deep_sea_hydrothermal_vent	Axial Seamount, Juan de Fuca	NA	ERR420354	Metagenomic	-
deep_sea_hydrothermal_vent	Axial Seamount, Juan de Fuca	NA	ERR420355	Metatranscriptomic	-
deep_sea_hydrothermal_vent	Axial Seamount, Juan de Fuca	NA	ERR420356	Metatranscriptomic	-
deep_sea_hydrothermal_vent	Axial Seamount, Juan de Fuca	NA	ERR420357	Metagenomic	-
deep_sea_hydrothermal_vent	Axial Seamount, Juan de Fuca	NA	ERR420362	Metatranscriptomic	-
deep_sea_hydrothermal_vent	Axial Seamount, Juan de Fuca	NA	ERR420363	Metagenomic	-
deep_sea_hydrothermal_vent	Axial Seamount, Juan de Fuca	NA	ERR420365	Metagenomic	-
deep_sea_hydrothermal_vent	Axial Seamount, Juan de Fuca	NA	ERR420368	Metagenomic	-
deep_sea_hydrothermal_vent	Axial Seamount, Juan de Fuca	NA	ERR420366	Metagenomic	-
deep_sea_hydrothermal_vent	Axial Seamount, Juan de Fuca	NA	ERR420358	Metagenomic	-
deep_sea_hydrothermal_vent	Axial Seamount, Juan de Fuca	NA	ERR420359	Metatranscriptomic	-
deep_sea_hydrothermal_vent	Axial Seamount, Juan de Fuca	NA	ERR420360	Metatranscriptomic	-
deep_sea_hydrothermal_vent	Axial Seamount, Juan de Fuca	NA	ERR420361	Metatranscriptomic	-
deep_sea_hydrothermal_vent	Axial Seamount, Juan de Fuca	NA	ERR420364	Metagenomic	-
deep_sea_hydrothermal_vent	Axial Seamount, Juan de Fuca	NA	ERR420367	Metagenomic	-
deep_sea_hydrothermal_vent	Axial Seamount, Juan de Fuca	1543m	ERR694114	Metagenomic	-
deep_sea_hydrothermal_vent	Axial Seamount, Juan de Fuca	1543m	ERR694115	Metagenomic	-
deep_sea_hydrothermal_vent	Axial Seamount, Juan de Fuca	1543m	ERR694116	Metagenomic	-
deep_sea_hydrothermal_vent	Axial Seamount, Juan de Fuca	1543m	ERR694123	Metagenomic	-
deep_sea_hydrothermal_vent	Axial Seamount, Juan de Fuca	1543m	ERR694110	Metagenomic	-
deep_sea_hydrothermal_vent	Axial Seamount, Juan de Fuca	1543m	ERR694112	Metagenomic	-
deep_sea_hydrothermal_vent	Axial Seamount, Juan de Fuca	1543m	ERR694113	Metagenomic	-
deep_sea_hydrothermal_vent	Axial Seamount, Juan de Fuca	1543m	ERR694117	Metagenomic	-
deep_sea_hydrothermal_vent	Axial Seamount, Juan de Fuca	1543m	ERR694118	Metagenomic	-
deep_sea_hydrothermal_vent	Axial Seamount, Juan de Fuca	1543m	ERR694119	Metagenomic	-
deep_sea_hydrothermal_vent	Axial Seamount, Juan de Fuca	1917m	ERR694177	Metagenomic	-
deep_sea_hydrothermal_vent	Axial Seamount, Juan de Fuca	1917m	ERR694179	Metagenomic	-
deep_sea_hydrothermal_vent	Axial Seamount, Juan de Fuca	1917m	ERR694189	Metagenomic	-
deep_sea_hydrothermal_vent	Axial Seamount, Juan de Fuca	1917m	ERR694190	Metagenomic	-
deep_sea_hydrothermal_vent	Axial Seamount, Juan de Fuca	1917m	ERR694192	Metagenomic	-
deep_sea_hydrothermal_vent	Axial Seamount, Juan de Fuca	1917m	ERR694193	Metagenomic	-
deep_sea_hydrothermal_vent	Axial Seamount, Juan de Fuca	1917m	ERR694198	Metagenomic	-
deep_sea_hydrothermal_vent	Axial Seamount, Juan de Fuca	1917m	ERR694376	Metatranscriptomic	-
deep_sea_hydrothermal_vent	Axial Seamount, Juan de Fuca	1917m	ERR694373	Metatranscriptomic	-
deep_sea_hydrothermal_vent	Axial Seamount, Juan de Fuca	1917m	ERR694374	Metatranscriptomic	-
deep_sea_hydrothermal_vent	Axial Seamount, Juan de Fuca	1543m	ERR694120	Metagenomic	-
deep_sea_hydrothermal_vent	Axial Seamount, Juan de Fuca	1917m	ERR694375	Metatranscriptomic	-
deep_sea_hydrothermal_vent	Axial Seamount, Juan de Fuca	1522m	ERR694387	Metatranscriptomic	Fortunato and Huber, 2016
deep_sea_hydrothermal_vent	Axial Seamount, Juan de Fuca	1522m	ERR694383	Metatranscriptomic	Fortunato and Huber, 2016
deep_sea_hydrothermal_vent	Axial Seamount, Juan de Fuca	1522m	ERR694385	Metatranscriptomic	Fortunato and Huber, 2016
deep_sea_hydrothermal_vent	Axial Seamount, Juan de Fuca	1522m	ERR694386	Metatranscriptomic	Fortunato and Huber, 2016
deep_sea_hydrothermal_vent	Axial Seamount, Juan de Fuca	1522m	ERR694391	Metatranscriptomic	Fortunato and Huber, 2016
deep_sea_hydrothermal_vent	Axial Seamount, Juan de Fuca	1522m	ERR694388	Metatranscriptomic	Fortunato and Huber, 2016
deep_sea_hydrothermal_vent	Axial Seamount, Juan de Fuca	1522m	ERR694389	Metatranscriptomic	Fortunato and Huber, 2016

deep_sea_hydrothermal_vent	Axial Seamount, Juan de Fuca	1522m	ERR694384	Metatranscriptomic	Fortunato and Huber, 2016
deep_sea_hydrothermal_vent	Axial Seamount, Juan de Fuca	1522m	ERR694390	Metatranscriptomic	Fortunato and Huber, 2016
deep_sea_hydrothermal_vent	Axial Seamount, Juan de Fuca	1543m	ERR694126	Metagenomic	-
deep_sea_hydrothermal_vent	Axial Seamount, Juan de Fuca	1522m	ERR694207	Metagenomic	Fortunato and Huber, 2016
deep_sea_hydrothermal_vent	Axial Seamount, Juan de Fuca	1522m	ERR694208	Metagenomic	Fortunato and Huber, 2016
deep_sea_hydrothermal_vent	Axial Seamount, Juan de Fuca	1522m	ERR694209	Metagenomic	Fortunato and Huber, 2016
deep_sea_hydrothermal_vent	Axial Seamount, Juan de Fuca	1522m	ERR694210	Metagenomic	Fortunato and Huber, 2016
deep_sea_hydrothermal_vent	Axial Seamount, Juan de Fuca	1522m	ERR694211	Metagenomic	Fortunato and Huber, 2016
deep_sea_hydrothermal_vent	Axial Seamount, Juan de Fuca	1522m	ERR694212	Metagenomic	Fortunato and Huber, 2016
deep_sea_hydrothermal_vent	Axial Seamount, Juan de Fuca	1522m	ERR694213	Metagenomic	Fortunato and Huber, 2016
deep_sea_hydrothermal_vent	Axial Seamount, Juan de Fuca	1516m	ERR694200	Metagenomic	-
deep_sea_hydrothermal_vent	Axial Seamount, Juan de Fuca	1516m	ERR694199	Metagenomic	-
deep_sea_hydrothermal_vent	Axial Seamount, Juan de Fuca	1516m	ERR694203	Metagenomic	-
deep_sea_hydrothermal_vent	Axial Seamount, Juan de Fuca	1543m	ERR694111	Metagenomic	-
deep_sea_hydrothermal_vent	Axial Seamount, Juan de Fuca	1516m	ERR694201	Metagenomic	-
deep_sea_hydrothermal_vent	Axial Seamount, Juan de Fuca	1516m	ERR694202	Metagenomic	-
deep_sea_hydrothermal_vent	Axial Seamount, Juan de Fuca	1516m	ERR694204	Metagenomic	-
deep_sea_hydrothermal_vent	Axial Seamount, Juan de Fuca	1516m	ERR694205	Metagenomic	-
deep_sea_hydrothermal_vent	Axial Seamount, Juan de Fuca	1516m	ERR694206	Metagenomic	-
deep_sea_hydrothermal_vent	Axial Seamount, Juan de Fuca	1516m	ERR694377	Metatranscriptomic	-
deep_sea_hydrothermal_vent	Axial Seamount, Juan de Fuca	1516m	ERR694378	Metatranscriptomic	-
deep_sea_hydrothermal_vent	Axial Seamount, Juan de Fuca	1516m	ERR694379	Metatranscriptomic	-
deep_sea_hydrothermal_vent	Axial Seamount, Juan de Fuca	1516m	ERR694380	Metatranscriptomic	-
deep_sea_hydrothermal_vent	Axial Seamount, Juan de Fuca	1516m	ERR694381	Metatranscriptomic	-
deep_sea_hydrothermal_vent	Axial Seamount, Juan de Fuca	1543m	ERR694121	Metagenomic	-
deep_sea_hydrothermal_vent	Axial Seamount, Juan de Fuca	1516m	ERR694382	Metatranscriptomic	-
deep_sea_hydrothermal_vent	Axial Seamount, Juan de Fuca	1543m	ERR694122	Metagenomic	-
deep_sea_hydrothermal_vent	Axial Seamount, Juan de Fuca	1543m	ERR694124	Metagenomic	-
deep_sea_hydrothermal_vent	Axial Seamount, Juan de Fuca	1543m	ERR694125	Metagenomic	-
deep_sea_hydrothermal_vent	Axial Seamount, Juan de Fuca	1543m	ERR694356	Metatranscriptomic	-
deep_sea_hydrothermal_vent	Axial Seamount, Juan de Fuca	1543m	ERR694357	Metatranscriptomic	-
deep_sea_hydrothermal_vent	Axial Seamount, Juan de Fuca	1543m	ERR694358	Metatranscriptomic	-
deep_sea_hydrothermal_vent	Axial Seamount, Juan de Fuca	1543m	ERR694359	Metatranscriptomic	-
deep_sea_hydrothermal_vent	Axial Seamount, Juan de Fuca	1507m	ERR694129	Metagenomic	-
deep_sea_hydrothermal_vent	Axial Seamount, Juan de Fuca	1507m	ERR694135	Metagenomic	-
deep_sea_hydrothermal_vent	Axial Seamount, Juan de Fuca	1507m	ERR694137	Metagenomic	-
deep_sea_hydrothermal_vent	Axial Seamount, Juan de Fuca	1507m	ERR694142	Metagenomic	-
deep_sea_hydrothermal_vent	Axial Seamount, Juan de Fuca	1507m	ERR694145	Metagenomic	-
deep_sea_hydrothermal_vent	Axial Seamount, Juan de Fuca	1507m	ERR694151	Metagenomic	-
deep_sea_hydrothermal_vent	Axial Seamount, Juan de Fuca	1507m	ERR694152	Metagenomic	-
deep_sea_hydrothermal_vent	Axial Seamount, Juan de Fuca	1507m	ERR694132	Metagenomic	-
deep_sea_hydrothermal_vent	Axial Seamount, Juan de Fuca	1507m	ERR694133	Metagenomic	-
deep_sea_hydrothermal_vent	Axial Seamount, Juan de Fuca	1507m	ERR694134	Metagenomic	-
deep_sea_hydrothermal_vent	Axial Seamount, Juan de Fuca	1507m	ERR694136	Metagenomic	-
deep_sea_hydrothermal_vent	Axial Seamount, Juan de Fuca	1507m	ERR694140	Metagenomic	-
deep_sea_hydrothermal_vent	Axial Seamount, Juan de Fuca	1507m	ERR694144	Metagenomic	-

deep_sea_hydrothermal_vent	Axial Seamount, Juan de Fuca	1507m	ERR694146	Metagenomic	-
deep_sea_hydrothermal_vent	Axial Seamount, Juan de Fuca	1507m	ERR694150	Metagenomic	-
deep_sea_hydrothermal_vent	Axial Seamount, Juan de Fuca	1507m	ERR694127	Metagenomic	-
deep_sea_hydrothermal_vent	Axial Seamount, Juan de Fuca	1507m	ERR694128	Metagenomic	-
deep_sea_hydrothermal_vent	Axial Seamount, Juan de Fuca	1507m	ERR694130	Metagenomic	-
deep_sea_hydrothermal_vent	Axial Seamount, Juan de Fuca	1507m	ERR694131	Metagenomic	-
deep_sea_hydrothermal_vent	Axial Seamount, Juan de Fuca	1507m	ERR694138	Metagenomic	-
deep_sea_hydrothermal_vent	Axial Seamount, Juan de Fuca	1507m	ERR694139	Metagenomic	-
deep_sea_hydrothermal_vent	Axial Seamount, Juan de Fuca	1507m	ERR694141	Metagenomic	-
deep_sea_hydrothermal_vent	Axial Seamount, Juan de Fuca	1507m	ERR694143	Metagenomic	-
deep_sea_hydrothermal_vent	Axial Seamount, Juan de Fuca	1507m	ERR694147	Metagenomic	-
deep_sea_hydrothermal_vent	Axial Seamount, Juan de Fuca	1507m	ERR694148	Metagenomic	-
deep_sea_hydrothermal_vent	Axial Seamount, Juan de Fuca	1507m	ERR694149	Metagenomic	-
deep_sea_hydrothermal_vent	Axial Seamount, Juan de Fuca	1507m	ERR694360	Metatranscriptomic	-
deep_sea_hydrothermal_vent	Axial Seamount, Juan de Fuca	1507m	ERR694363	Metatranscriptomic	-
deep_sea_hydrothermal_vent	Axial Seamount, Juan de Fuca	1507m	ERR694365	Metatranscriptomic	-
deep_sea_hydrothermal_vent	Axial Seamount, Juan de Fuca	1507m	ERR694366	Metatranscriptomic	-
deep_sea_hydrothermal_vent	Axial Seamount, Juan de Fuca	1507m	ERR694362	Metatranscriptomic	-
deep_sea_hydrothermal_vent	Axial Seamount, Juan de Fuca	1507m	ERR694361	Metatranscriptomic	-
deep_sea_hydrothermal_vent	Axial Seamount, Juan de Fuca	1507m	ERR694364	Metatranscriptomic	-
deep_sea_hydrothermal_vent	Axial Seamount, Juan de Fuca	1530m	ERR694170	Metagenomic	-
deep_sea_hydrothermal_vent	Axial Seamount, Juan de Fuca	1530m	ERR694157	Metagenomic	-
deep_sea_hydrothermal_vent	Axial Seamount, Juan de Fuca	1530m	ERR694154	Metagenomic	-
deep_sea_hydrothermal_vent	Axial Seamount, Juan de Fuca	1530m	ERR694156	Metagenomic	-
deep_sea_hydrothermal_vent	Axial Seamount, Juan de Fuca	1530m	ERR694160	Metagenomic	-
deep_sea_hydrothermal_vent	Axial Seamount, Juan de Fuca	1530m	ERR694161	Metagenomic	-
deep_sea_hydrothermal_vent	Axial Seamount, Juan de Fuca	1530m	ERR694162	Metagenomic	-
deep_sea_hydrothermal_vent	Axial Seamount, Juan de Fuca	1530m	ERR694163	Metagenomic	-
deep_sea_hydrothermal_vent	Axial Seamount, Juan de Fuca	1530m	ERR694164	Metagenomic	-
deep_sea_hydrothermal_vent	Axial Seamount, Juan de Fuca	1530m	ERR694165	Metagenomic	-
deep_sea_hydrothermal_vent	Axial Seamount, Juan de Fuca	1530m	ERR694166	Metagenomic	-
deep_sea_hydrothermal_vent	Axial Seamount, Juan de Fuca	1530m	ERR694153	Metagenomic	-
deep_sea_hydrothermal_vent	Axial Seamount, Juan de Fuca	1530m	ERR694155	Metagenomic	-
deep_sea_hydrothermal_vent	Axial Seamount, Juan de Fuca	1530m	ERR694158	Metagenomic	-
deep_sea_hydrothermal_vent	Axial Seamount, Juan de Fuca	1530m	ERR694167	Metagenomic	-
deep_sea_hydrothermal_vent	Axial Seamount, Juan de Fuca	1530m	ERR694168	Metagenomic	-
deep_sea_hydrothermal_vent	Axial Seamount, Juan de Fuca	1530m	ERR694169	Metagenomic	-
deep_sea_hydrothermal_vent	Axial Seamount, Juan de Fuca	1530m	ERR694159	Metagenomic	-
deep_sea_hydrothermal_vent	Axial Seamount, Juan de Fuca	1530m	ERR694171	Metagenomic	-
deep_sea_hydrothermal_vent	Axial Seamount, Juan de Fuca	1530m	ERR694174	Metagenomic	-
deep_sea_hydrothermal_vent	Axial Seamount, Juan de Fuca	1530m	ERR694172	Metagenomic	-
deep_sea_hydrothermal_vent	Axial Seamount, Juan de Fuca	1530m	ERR694173	Metagenomic	-
deep_sea_hydrothermal_vent	Axial Seamount, Juan de Fuca	1530m	ERR694175	Metagenomic	-
deep_sea_hydrothermal_vent	Axial Seamount, Juan de Fuca	1530m	ERR694367	Metatranscriptomic	-
deep_sea_hydrothermal_vent	Axial Seamount, Juan de Fuca	1530m	ERR694371	Metatranscriptomic	-
deep_sea_hydrothermal_vent	Axial Seamount, Juan de Fuca	1530m	ERR694372	Metatranscriptomic	-

deep_sea_hydrothermal_vent	Axial Seamount, Juan de Fuca	1530m	ERR694368	Metatranscriptomic	-
deep_sea_hydrothermal_vent	Axial Seamount, Juan de Fuca	1530m	ERR694369	Metatranscriptomic	-
deep_sea_hydrothermal_vent	Axial Seamount, Juan de Fuca	1530m	ERR694370	Metatranscriptomic	-
deep_sea_hydrothermal_vent	Axial Seamount, Juan de Fuca	1917m	ERR694184	Metagenomic	-
deep_sea_hydrothermal_vent	Axial Seamount, Juan de Fuca	1917m	ERR694185	Metagenomic	-
deep_sea_hydrothermal_vent	Axial Seamount, Juan de Fuca	1917m	ERR694191	Metagenomic	-
deep_sea_hydrothermal_vent	Axial Seamount, Juan de Fuca	1917m	ERR694196	Metagenomic	-
deep_sea_hydrothermal_vent	Axial Seamount, Juan de Fuca	1917m	ERR694197	Metagenomic	-
deep_sea_hydrothermal_vent	Axial Seamount, Juan de Fuca	1917m	ERR694186	Metagenomic	-
deep_sea_hydrothermal_vent	Axial Seamount, Juan de Fuca	1917m	ERR694178	Metagenomic	-
deep_sea_hydrothermal_vent	Axial Seamount, Juan de Fuca	1917m	ERR694176	Metagenomic	-
deep_sea_hydrothermal_vent	Axial Seamount, Juan de Fuca	1917m	ERR694180	Metagenomic	-
deep_sea_hydrothermal_vent	Axial Seamount, Juan de Fuca	1917m	ERR694181	Metagenomic	-
deep_sea_hydrothermal_vent	Axial Seamount, Juan de Fuca	1917m	ERR694182	Metagenomic	-
deep_sea_hydrothermal_vent	Axial Seamount, Juan de Fuca	1917m	ERR694183	Metagenomic	-
deep_sea_hydrothermal_vent	Axial Seamount, Juan de Fuca	1917m	ERR694187	Metagenomic	-
deep_sea_hydrothermal_vent	Axial Seamount, Juan de Fuca	1917m	ERR694188	Metagenomic	-
deep_sea_hydrothermal_vent	Axial Seamount, Juan de Fuca	1917m	ERR694194	Metagenomic	-
deep_sea_hydrothermal_vent	Axial Seamount, Juan de Fuca	1917m	ERR694195	Metagenomic	-
deep_sea_hydrothermal_vent	Axial Seamount, Juan de Fuca	1200m	ERR1163068	Metagenomic	-
deep_sea_hydrothermal_vent	Axial Seamount, Juan de Fuca	1200m	ERR1163069	Metagenomic	-
deep_sea_hydrothermal_vent	Axial Seamount, Juan de Fuca	1200m	ERR1163070	Metagenomic	-
deep_sea_hydrothermal_vent	Axial Seamount, Juan de Fuca	1200m	ERR1163071	Metagenomic	-
deep_sea_hydrothermal_vent	Axial Seamount, Juan de Fuca	1200m	ERR1163072	Metagenomic	-
deep_sea_hydrothermal_vent	Axial Seamount, Juan de Fuca	1200m	ERR1163073	Metagenomic	-
deep_sea_hydrothermal_vent	Axial Seamount, Juan de Fuca	1200m	ERR1163074	Metagenomic	-
deep_sea_hydrothermal_vent	Axial Seamount, Juan de Fuca	1200m	ERR1163075	Metagenomic	-
deep_sea_hydrothermal_vent	Axial Seamount, Juan de Fuca	1200m	ERR1163076	Metagenomic	-
deep_sea_hydrothermal_vent	Axial Seamount, Juan de Fuca	1200m	ERR1163077	Metatranscriptomic	-
deep_sea_hydrothermal_vent	Axial Seamount, Juan de Fuca	1200m	ERR1163078	Metatranscriptomic	-
deep_sea_hydrothermal_vent	Axial Seamount, Juan de Fuca	1200m	ERR1163079	Metatranscriptomic	-
deep_sea_hydrothermal_vent	Axial Seamount, Juan de Fuca	1200m	ERR1163080	Metatranscriptomic	-
deep_sea_hydrothermal_vent	Axial Seamount, Juan de Fuca	1200m	ERR1163081	Metatranscriptomic	-
deep_sea_hydrothermal_vent	Axial Seamount, Juan de Fuca	1521m	ERR1163082	Metagenomic	-
deep_sea_hydrothermal_vent	Axial Seamount, Juan de Fuca	1521m	ERR1163083	Metagenomic	-
deep_sea_hydrothermal_vent	Axial Seamount, Juan de Fuca	1521m	ERR1163084	Metagenomic	-
deep_sea_hydrothermal_vent	Axial Seamount, Juan de Fuca	1521m	ERR1163085	Metagenomic	-
deep_sea_hydrothermal_vent	Axial Seamount, Juan de Fuca	1521m	ERR1163086	Metagenomic	-
deep_sea_hydrothermal_vent	Axial Seamount, Juan de Fuca	1521m	ERR1163087	Metagenomic	-
deep_sea_hydrothermal_vent	Axial Seamount, Juan de Fuca	1521m	ERR1163088	Metagenomic	-
deep_sea_hydrothermal_vent	Axial Seamount, Juan de Fuca	1521m	ERR1163089	Metagenomic	-
deep_sea_hydrothermal_vent	Axial Seamount, Juan de Fuca	1521m	ERR1163090	Metagenomic	-
deep_sea_hydrothermal_vent	Axial Seamount, Juan de Fuca	1521m	ERR1163091	Metagenomic	-
deep_sea_hydrothermal_vent	Axial Seamount, Juan de Fuca	1521m	ERR1163092	Metagenomic	-
deep_sea_hydrothermal_vent	Axial Seamount, Juan de Fuca	1521m	ERR1163093	Metagenomic	-
deep_sea_hydrothermal_vent	Axial Seamount, Juan de Fuca	1521m	ERR1163094	Metagenomic	-

deep_sea_hydrothermal_vent	Axial Seamount, Juan de Fuca	1521m	ERR1163095	Metatranscriptomic	-
deep_sea_hydrothermal_vent	Axial Seamount, Juan de Fuca	1521m	ERR1163096	Metatranscriptomic	-
deep_sea_hydrothermal_vent	Axial Seamount, Juan de Fuca	1521m	ERR1163097	Metatranscriptomic	-
deep_sea_hydrothermal_vent	Axial Seamount, Juan de Fuca	1521m	ERR1163098	Metatranscriptomic	-
deep_sea_hydrothermal_vent	Axial Seamount, Juan de Fuca	1521m	ERR1163099	Metatranscriptomic	-
deep_sea_hydrothermal_vent	Axial Seamount, Juan de Fuca	1521m	ERR1163100	Metatranscriptomic	-
deep_sea_hydrothermal_vent	Axial Seamount, Juan de Fuca	1521m	ERR1163101	Metatranscriptomic	-
deep_sea_hydrothermal_vent	Axial Seamount, Juan de Fuca	1521m	ERR1163102	Metatranscriptomic	-
deep_sea_hydrothermal_vent	Axial Seamount, Juan de Fuca	1521m	ERR1163103	Metatranscriptomic	-
deep_sea_hydrothermal_vent	Axial Seamount, Juan de Fuca	1521m	ERR1163104	Metatranscriptomic	-
deep_sea_hydrothermal_vent	Axial Seamount, Juan de Fuca	1539m	ERR1163105	Metagenomic	-
deep_sea_hydrothermal_vent	Axial Seamount, Juan de Fuca	1539m	ERR1163106	Metagenomic	-
deep_sea_hydrothermal_vent	Axial Seamount, Juan de Fuca	1539m	ERR1163107	Metagenomic	-
deep_sea_hydrothermal_vent	Axial Seamount, Juan de Fuca	1539m	ERR1163108	Metagenomic	-
deep_sea_hydrothermal_vent	Axial Seamount, Juan de Fuca	1539m	ERR1163109	Metagenomic	-
deep_sea_hydrothermal_vent	Axial Seamount, Juan de Fuca	1539m	ERR1163110	Metagenomic	-
deep_sea_hydrothermal_vent	Axial Seamount, Juan de Fuca	1539m	ERR1163111	Metagenomic	-
deep_sea_hydrothermal_vent	Axial Seamount, Juan de Fuca	1539m	ERR1163112	Metagenomic	-
deep_sea_hydrothermal_vent	Axial Seamount, Juan de Fuca	1539m	ERR1163113	Metagenomic	-
deep_sea_hydrothermal_vent	Axial Seamount, Juan de Fuca	1539m	ERR1163114	Metagenomic	-
deep_sea_hydrothermal_vent	Axial Seamount, Juan de Fuca	1539m	ERR1163115	Metagenomic	-
deep_sea_hydrothermal_vent	Axial Seamount, Juan de Fuca	1539m	ERR1163116	Metagenomic	-
deep_sea_hydrothermal_vent	Axial Seamount, Juan de Fuca	1539m	ERR1163117	Metagenomic	-
deep_sea_hydrothermal_vent	Axial Seamount, Juan de Fuca	1539m	ERR1163118	Metatranscriptomic	-
deep_sea_hydrothermal_vent	Axial Seamount, Juan de Fuca	1539m	ERR1163119	Metatranscriptomic	-
deep_sea_hydrothermal_vent	Axial Seamount, Juan de Fuca	1539m	ERR1163120	Metatranscriptomic	-
deep_sea_hydrothermal_vent	Axial Seamount, Juan de Fuca	1539m	ERR1163121	Metatranscriptomic	-
deep_sea_hydrothermal_vent	Axial Seamount, Juan de Fuca	1539m	ERR1163122	Metatranscriptomic	-
deep_sea_hydrothermal_vent	Axial Seamount, Juan de Fuca	1539m	ERR1163123	Metatranscriptomic	-
deep_sea_hydrothermal_vent	Axial Seamount, Juan de Fuca	1539m	ERR1163124	Metatranscriptomic	-
deep_sea_hydrothermal_vent	Axial Seamount, Juan de Fuca	1514m	ERR1163125	Metagenomic	-
deep_sea_hydrothermal_vent	Axial Seamount, Juan de Fuca	1514m	ERR1163126	Metagenomic	-
deep_sea_hydrothermal_vent	Axial Seamount, Juan de Fuca	1514m	ERR1163127	Metagenomic	-
deep_sea_hydrothermal_vent	Axial Seamount, Juan de Fuca	1514m	ERR1163128	Metagenomic	-
deep_sea_hydrothermal_vent	Axial Seamount, Juan de Fuca	1514m	ERR1163129	Metagenomic	-
deep_sea_hydrothermal_vent	Axial Seamount, Juan de Fuca	1514m	ERR1163130	Metagenomic	-
deep_sea_hydrothermal_vent	Axial Seamount, Juan de Fuca	1514m	ERR1163131	Metagenomic	-
deep_sea_hydrothermal_vent	Axial Seamount, Juan de Fuca	1514m	ERR1163132	Metagenomic	-
deep_sea_hydrothermal_vent	Axial Seamount, Juan de Fuca	1514m	ERR1163133	Metagenomic	-
deep_sea_hydrothermal_vent	Axial Seamount, Juan de Fuca	1514m	ERR1163134	Metagenomic	-
deep_sea_hydrothermal_vent	Axial Seamount, Juan de Fuca	1514m	ERR1163135	Metagenomic	-
deep_sea_hydrothermal_vent	Axial Seamount, Juan de Fuca	1514m	ERR1163136	Metatranscriptomic	-
deep_sea_hydrothermal_vent	Axial Seamount, Juan de Fuca	1514m	ERR1163137	Metatranscriptomic	-
deep_sea_hydrothermal_vent	Axial Seamount, Juan de Fuca	1514m	ERR1163138	Metatranscriptomic	-
deep_sea_hydrothermal_vent	Axial Seamount, Juan de Fuca	1514m	ERR1163139	Metatranscriptomic	-
deep_sea_hydrothermal_vent	Axial Seamount, Juan de Fuca	1514m	ERR1163140	Metatranscriptomic	-

deep_sea_hydrothermal_vent	Axial Seamount, Juan de Fuca	1514m	ERR1163141	Metatranscriptomic	-
deep_sea_hydrothermal_vent	Axial Seamount, Juan de Fuca	1514m	ERR1163142	Metatranscriptomic	-
deep_sea_hydrothermal_vent	Axial Seamount, Juan de Fuca	1514m	ERR1163143	Metatranscriptomic	-
deep_sea_hydrothermal_vent	Kilo Moana, Lau Bassin	2605m	SRR1217367	Metagenomic	Anantharaman et al., 2016
deep_sea_hydrothermal_vent	Mariner, Lau Bassin	1890m	SRR1217452	Metagenomic	Anantharaman et al., 2016
deep_sea_hydrothermal_vent	Kilo Moana, Lau Bassin	2440m	SRR1217459	Metagenomic	Anantharaman et al., 2016
deep_sea_hydrothermal_vent	Abe, Lau Bassin	1960m	SRR1217460	Metagenomic	Anantharaman et al., 2016
deep_sea_hydrothermal_vent	Tui Malila, Lau Bassin	1919m	SRR1217462	Metagenomic	Anantharaman et al., 2016
deep_sea_hydrothermal_vent	Abe, Lau Bassin	2155m	SRR1217463	Metagenomic	Anantharaman et al., 2016
deep_sea_hydrothermal_vent	Tahi Moana, Lau Bassin	2229m	SRR1217465	Metagenomic	Anantharaman et al., 2016
deep_sea_hydrothermal_vent	Kilo Moana, Lau Bassin	2639m	SRR1217564	Metagenomic	Anantharaman et al., 2016
deep_sea_hydrothermal_vent	Abe, Lau Bassin	2159m	SRR1217565	Metagenomic	Anantharaman et al., 2016
deep_sea_hydrothermal_vent	Mariner, Lau Bassin	1785m	SRR1217567	Metagenomic	Anantharaman et al., 2016
deep_sea_hydrothermal_vent	Kilo Moana, Lau Bassin	2315m	SRR1217566	Metagenomic	Anantharaman et al., 2016
deep_sea_hydrothermal_vent	Tu'i Malila vent field, Lau bassin	1876m	SRR4453772	Metagenomic	Skennerton et al., 2015
deep_sea_marine_sediment	Loki's castle, GC14, Norwegian Sea	3283m	SRR1555744	Metagenomic	Spang et al., 2015
deep_sea_marine_sediment	Loki's castle, GC14, Norwegian Sea	3283m	SRR1555743	Metagenomic	Spang et al., 2015
deep_sea_marine_sediment	Loki's castle, GC14, Norwegian Sea	3283m	SRR1555748	Metagenomic	Spang et al., 2015
deep_sea_hydrothermal_vent	Lost city, Mid Atlantic Ridge	733m	SRR1636508	Metagenomic	-
deep_sea_hydrothermal_vent	Lost city, Mid Atlantic Ridge	766m	SRR1636509	Metagenomic	-
terrestrial_hot_spring	Sungai Klah, Malaysia	0m	ERR372908	Metagenomic	Chan et al., 2015
deep_sea_hydrothermal_vent	Menez Gwen, Mid Atlantic Ridge	828m	ERR1078300	Metagenomic	Meier et al., 2016
deep_sea_hydrothermal_vent	Menez Gwen, Mid Atlantic Ridge	828m	ERR1078301	Metagenomic	Meier et al., 2016
deep_sea_hydrothermal_vent	Menez Gwen, Mid Atlantic Ridge	828m	ERR1078302	Metagenomic	Meier et al., 2016
hydrothermal_spring	Bay of Prony, ST09, New Caledonia	43m	SRR1636517	Metagenomic	Mei et al., 2016
hydrothermal_spring	Bay of Prony, ST09, New Caledonia	43m	SRR1636516	Metagenomic	Mei et al., 2016
terrestrial_hot_spring	Zodletone Spring, Oklahoma, USA	0m	SRR5214155	Metagenomic	-
terrestrial_hot_spring	Limpopo, South Africa	0m	SRR5214705	Metagenomic	-
terrestrial_hot_spring	Tshipise, South Africa	0m	SRR5214706	Metagenomic	-
terrestrial_hot_spring	Shi-Huang-Ping, Taïwan	0m	SRR1297185	Metagenomic	-
terrestrial_hot_spring	Shi-Huang-Ping, Taïwan	0m	SRR1297186	Metagenomic	-
terrestrial_hot_spring	Shi-Huang-Ping, Taïwan	0m	SRR1297203	Metagenomic	-
terrestrial_hot_spring	Octopus Spring, Yellowstone National Park	0m	SRR4030098	Metagenomic	-
terrestrial_hot_spring	Beowulf Spring, Yellowstone National Park	0m	SRR4030100	Metagenomic	-
terrestrial_hot_spring	Mammoth Spring, Yellowstone National Park	0m	SRR4030101	Metagenomic	-
terrestrial_hot_spring	Conch Spring, Yellowstone National Park	0m	SRR4030102	Metagenomic	-
terrestrial_hot_spring	Grendel Spring, Yellowstone National Park	0m	SRR4030106	Metagenomic	-
terrestrial_hot_spring	Yellowstone National Park	0m	SRR5207630	Metatranscriptomic	-
terrestrial_hot_spring	Yellowstone National Park	0m	SRR5207631	Metatranscriptomic	-
terrestrial_hot_spring	Yellowstone National Park	0m	SRR5207688	Metatranscriptomic	-
terrestrial_hot_spring	Yellowstone National Park	0m	SRR5207689	Metatranscriptomic	-
terrestrial_hot_spring	Yellowstone National Park	0m	SRR5208581	Metatranscriptomic	-
terrestrial_hot_spring	Yellowstone National Park	0m	SRR5248167	Metagenomic	-

Appendix 3: Protocole bouteille de Widdel

Laboratoire de Microbiologie des Environnements Extrêmes
Institut Universitaire Européen de la Mer

Répartition de milieu de culture à l'aide de la bouteille de Widdel

Protocole Manipulation	
Rédacteur du protocole	*Courtine Damien (Doctorant)*
Vérificateur du protocole	*Maignien Loïs (Chercheur)*
Date	*4 Novembre 2015*

Risques et dangers

Ces expérimentations doivent être effectuées, après accord avec le(s) responsable(s) de projets, en respectant les bonnes pratiques de laboratoire (cf. Guide des Bonnes Pratiques de Laboratoire) et après avoir été formé au poste de travail concerné. **Certaines manipulations et certains produits marqués d'un astérisque** sont particulièrement dangereux (cf. fiches de sécurité MSDS disponibles sur le site Sigma http://www.sigmaaldrich.com/france.html) et doivent donc être réalisées/manipulés avec précaution et dans des conditions optimales de sécurité. Ils sont listés ci-après :

Section ci-dessus à garder dans tous les protocoles

- **Utilisation de l'autoclave : uniquement le personnel possédant une habilitation**
- **Utilisation de la rampe à gaz : *cf* « Protocole rampe à gaz »**
- Risque de brûlure
- Port de la blouse et de gants de protections

Nettoyage – Gestion des déchets

Spécifications pour le nettoyage

Nettoyer la verrerie en accord avec les pratiques établies dans le laboratoire : 3 rinçages à l'eau déionisée puis 3 rinçages à l'eau MiliQ.

Nettoyer la zone de travail après utilisation

Veiller à bien couper l'alimentation en gaz et fermer les bouteilles (cf protocoles sur les bouteilles de gaz)

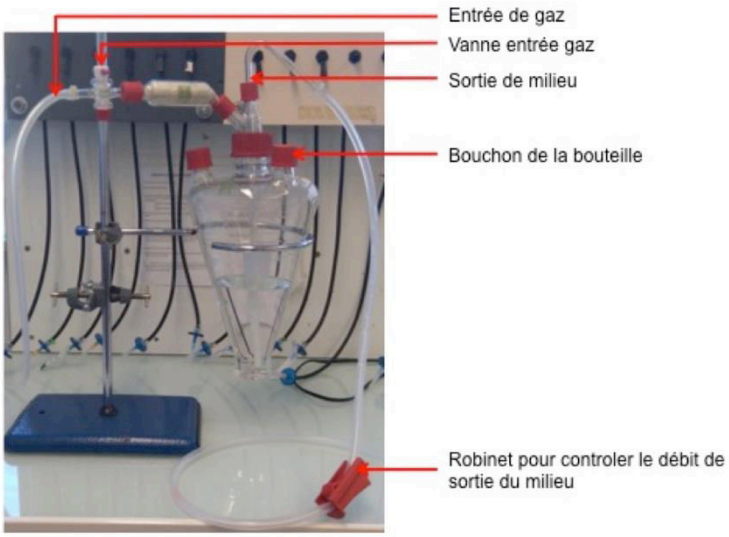
Gestion des déchets

Opercules en aluminium à jeter dans la poubelle prévue à cet effet.

Informations techniques

Principe - remarques

La bouteille de Widdel permet de répartir du milieu de culture stérile et réduit, environ 1,5 à 2L, dans des fioles pénicilline, en condition stérile et sous un flux de gaz constant (azote).



Matériel & réactifs – appareils associés

- Bouteille de Widdel, (Widdel and Bak, 1992) stockée derrière la rampe à gaz et son socle **(ref. 110012, Glasgerätebau Ochs)**
- Agitateur magnétique **(Jeiotech, MS-12B)**
- Milieu de culture et solutions de complémentation (si il y en a)
- Bécher 5L
- Aiguille rose **(BD spinal needle, 18G x 6.00 in (1,2x152mm))** ou aiguille en Inox sterilisable au four Pasteur.
- Fioles stériles
- Bouchons bleus ou gris
- Papier aluminium
- Autoclave, **salle A220**
- Poste 3, **salle A235**
- Rampe à gaz

Mode opératoire

Préparation de l'autoclavage :

- Vider l'eau résiduelle présente dans la bouteille
- Vérifier la présence de l'agitateur magnétique au fond de la bouteille
- Remplir avec du milieu neuf « fraîchement préparé », non autoclavé, non complémenté
- Sceller l'extrémité du tuyau d'entrée de gaz et celle du tuyau de sortie de milieu (le tuyau avec une cloche) à l'aide d'une feuille d'aluminium
- Visser les deux bouchons situés sur la bouteille et **les desserrer d'¼ de tour (très important !!)**
- Transférer la bouteille dans un bécher de 5L, avec les tuyaux
- Autoclaver dans le grand autoclave (salle A220), cycle humide.

Préparation de la bouteille et du milieu :

À la sortie de l'autoclave, replacer la bouteille sur son socle et disposer l'agitateur magnétique en dessous. Il est possible de mettre la bouteille dans bécher contenant de la glace afin de refroidir plus vite le milieu.

Ensuite

- Brancher l'entrée de gaz sur la rampe à gaz (le plus à droite en regardant de face), ne surtout pas toucher au filtre 0,22µm (disque bleu)
- Régler la pression à l'aide du manomètre de la bouteille à environ 0.2/0.3 Bar.
- **Afin d'éviter une sortie de milieu non voulue**, desserrer un bouchon de la bouteille
- Ouvrir les vannes (rampe à gaz et vanne gaz flacon Widdel)
- Mettre en marche l'agitation.

Ajouter les compléments nécessaire au milieu de culture si besoin, ainsi que l'agent réducteur, via un des bouchons de la bouteille. Quand tout les éléments ont été ajoutés, resserrer le bouchon en veillant à laisser légèrement ouvert afin d'éviter une surpression trop importante dans la bouteille.

Répartition du milieu dans les fioles :
Figure synthétique à la fin du paragraphe

- Mettre à disposition des fioles stériles et des bouchons (ouvrir le bocal à côté d'un bec Bunsen)
- Brancher une grande aiguille stérile sur une sortie de gaz, ne surtout pas toucher au filtre 0,22µm (disque bleu), et ouvrir la vanne de gaz
- Avec la grande aiguille, percer la feuille d'aluminium pour flusher une première fiole

- Disposer la fiole sous le tuyau de distribution du milieu et maintenir le flux de gaz dans la fiole
- Ouvrir le robinet qui contrôle la sortie du milieu réduit, et mettre la quantité de milieu désirée
- Fermer le robinet
- Prendre un bouchon, toujours maintenir le flux d'azote dans la fiole
- Apposer le bouchon en enlevant progressivement l'aiguille de la fiole
- Transférer le flux d'azote dans la fiole suivante
- Vérifier que le bouchon est bien positionné sur la fiole
- Mettre la fiole fermée de côté

Renouveler ces étapes jusqu'à épuisement du milieu dans la bouteille

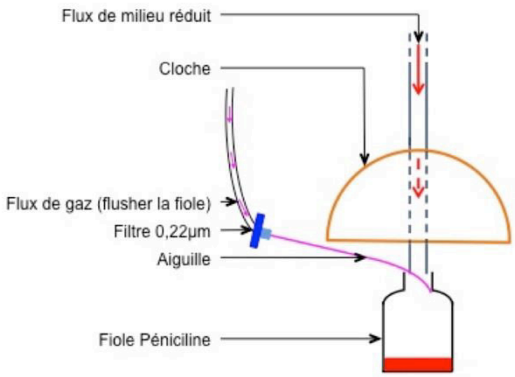


Figure synthétique sur la disposition de la fiole et de l'aiguille pour flusher

Fin de la manipulation :

- Sceller les fioles à l'aide de capsules d'aluminium
- Si nécessaire, ajouter une surpression d'azote dans les fioles à l'aide de la rampe à gaz
- Stocker les fioles remplies de milieu à 4°C
- Ranger les fioles non utilisées et les bouchons à l'endroit approprié

Pour le nettoyage, se référer à la partie « Nettoyage » de ce protocole.

N.B. : Au bout d'un certain temps, le tuyau de sortie de milieu peut devenir poreux. Il suffit juste de changer cette partie du tuyau. **(Ref. 910603, Dutscher)**

Appendix 4: Abstract and posters presented during the thesis

1) Poster presented during the "Journée des doctorants de l'EDSM" (EDSM PhD students' day), Brest, Nov. 2016:

Comparative genomics of a hyperthermophilic *Archaea*: *Thermococcus nautili*

2) Oral presentation presented at the 8th symposium of the "Association francophone d'écologie microbienne" (French-speaking association of microbial ecology), Camaret-sur-mer, Oct 2017:

Genomic diversity of closely related *Thermococcus* populations in deep-sea hydrothermal vent context

1) Poster presented during the "Journée des doctorants de l'EDSM" (EDSM PhD students' day), Brest, Nov. 2016

Génomique comparative d'une archée hyperthermophile: *Thermococcus nautili*

Damien Courtine[a,b,c], A. Murat Eren[d,e], Mohamed Jebbar[a,b,c], Myriam Georges[a,b,c], Karine Alain[a,b,c] et Loïs Maignien[a,b,c]

[a] Université de Bretagne Occidentale (UBO, UEB), Institut Universitaire Européen de la Mer (IUEM)—UMR 6197, Laboratoire de Microbiologie des Environnements Extrêmes (LM2E), Place Nicolas Copernic, F-29280 Plouzané, France
[b] CNRS, IUEM—UMR 6197, Laboratoire de Microbiologie des Environnements Extrêmes (LM2E), Place Nicolas Copernic, F-29280 Plouzané, France
[c] Ifremer, UMR 6197, Laboratoire de Microbiologie des Environnements Extrêmes (LM2E), Technopôle Pointe du diable, F-29280 Plouzané, France
[d] Josephine Bay Paul Center, Marine Biological Laboratory, Woods Hole, MA, United States
[e] Department of Medicine, The University of Chicago, Chicago, IL, United States

Les Archées et les Bactéries sont les êtres vivants parmi les plus abondants sur Terre. Ces organismes sont présents partout: dans les sols, dans les eaux, en symbiose avec des animaux ou des végétaux, etc. Les cheminées hydrothermales profondes (Figure 1) combinent des conditions extrêmes de haute température et de haute pression. Ces édifices profonds sont colonisées par de nombreux microorganismes extrêmophiles dont la résistance à ces conditions en fait des modèles particulièrement intéressants pour l'étude des mécanismes d'adaptation. L'objectif de ma thèse consiste à explorer la diversité génomique du genre *Thermococcus*, une archée hyperthermophile habituellement présente sur ces édifices hydrothermaux profonds (Figure 2).

Figure 1: Exemple de cheminée hydrothermale. Le fluide sortant de ces édifices a une température d'environ 200 à 300 °C alors que l'eau de mer est elle à 2°C. Cette différence brutale de température entraine la précipitation des éléments minéraux présents dans le fluide.

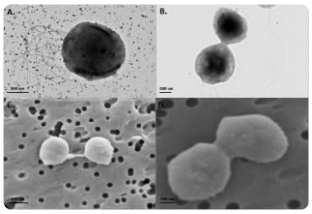


Figure 2: Photos d'un *Thermococcus*. A,B microscopie à transmission; C,D microscopie à balayage. Les *Thermococcus* sont des coques d'environ 1 µm de diamètre, ils ne supportent pas l'oxygène (anaérobie), et se nourrissent de sucres, protéines et lipides. Ils sont également capables de se déplacer grâce à leurs flagelles (A,B). (d'après Dalmasso *et al.*, 2016)

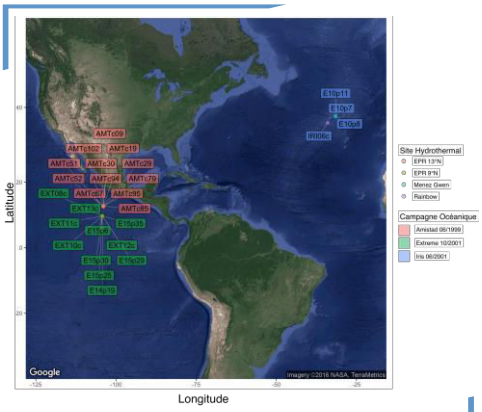


Figure 3: Lieux d'échantillonage des 23 souches apparentées à l'espèce *Thermococcus nautili*. Les isolats ont été récolté lors de 3 campagnes océanographiques, deux sur l'océan Pacifique et une sur l'océan Atlantique, totalisant 4 sites hydrothermaux.

Lors de ce projet, 48 isolats de *Thermococcus* ont été séquencés, ils appartiennent à deux populations distinctes. La première, proche de l'espèce *Thermococcus nautili*, comprend 27 isolats provenant des océans Atlantique et Pacifique (Figure 3). La seconde population est proche de l'espèce *Thermococcus sp. 4557* et compte 21 isolats. Actuellement, les génomes de 45 d'entre elles ont été séquencés et assemblés. Les résultats exposés ici sont préliminaires et portent sur 23 souches appartenant à la population '*Thermococcus nautili* '. Les isolats de cette population se trouvent être particulièrement proches. Ceci est visible sur l'arbre phylogénétique construit avec le marqueur ARNr 16S (Figure 4). Cette multitude d'origines est intéressante car chaque site hydrothermal est unique (géochimie, géographie). Or, si plusieurs isolats se révèlent être proches sur le plan génomique, mais proviennent de différents sites hydrothermaux, il y a une forte présomption d'adaptation de leur patrimoine génétique à leurs environnements respectifs.

Les génomes ont été annotés en utilisant Anvi'o (Eren et *al.*, 2006). La figure 5 représente les groupes de protéines propre à une souche de même que ceux partagés.

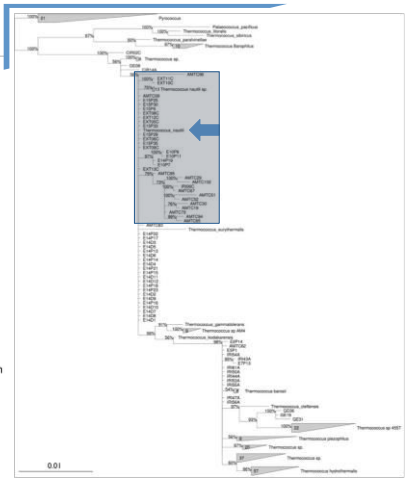


Figure 4: Arbre phylogénétique de l'ordre des *Thermococcales*. Cet ordre comprend les genres *Pyrococcus*, *Palaeococcus* et *Thermococcus*. La population de génomes utilisée dans la présente étude est située dans l'encadré bleu. L'arbre phylogénétique a été construit en utilisant les séquences du gène codant pour la petite sous-unité de l'ARN ribosomal (ARNr 16S) de chaque souche analysée ainsi que les souches référencées dans les bases de données.

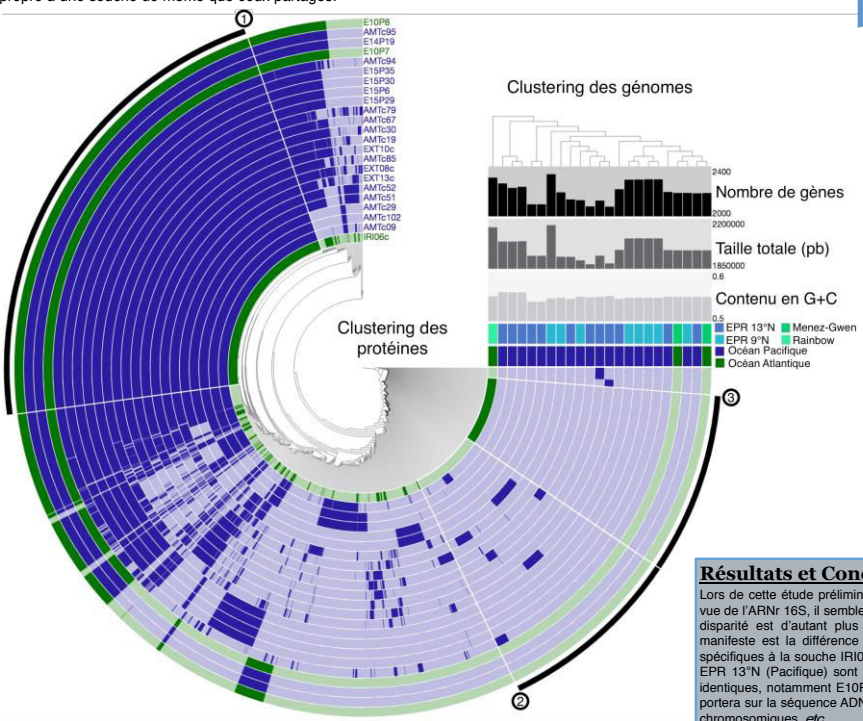


Figure 5: Comparaison des génomes réalisée avec Anvi'o (Eren et al., 2015). Chaque piste correspond à un isolat, et chaque trait correspond à un groupe de protéines. L'arbre au centre représente une hiérarchisation de ces groupes. Un groupe est composé de protéines similaires retrouvées dans un ou plusieurs génomes. En haut à droite sont présentées les différentes métadonnées des génomes et des isolats. 1) Génome-coeur ,c'est à dire l'ensemble des gènes partagés par tous les isolats étudiés. 2) Une partie des gènes accessoires présents uniquement au sein d'isolats du Pacifique. 3) Gènes spécifiques de l'isolat IRI06c, provenant de l'Atlantique.

Résultats et Conclusion:

Lors de cette étude préliminaire, bien que les souches étudiées soient très proches du point de vue de l'ARNr 16S, il semble qu'il y ait une certaine disparité en terme de contenu génique. Cette disparité est d'autant plus vraie lorsqu'on s'intéresse à la géographie des isolats. Le plus manifeste est la différence entre Atlantique/Pacifique avec environ 400 groupes de protéines spécifiques à la souche IRI06c. Les différences de contenu en gènes entre les sites EPR 9°N et EPR 13°N (Pacifique) sont moins flagrantes mais présentes. Certaines souches semblent être identiques, notamment E10P7/E10P8/AMTc95/E14P19. Pour celles-ci, le travail de comparaison portera sur la séquence ADN afin de mettre en évidence de fines variations, des réarrangements chromosomiques, *etc.*

Les perspectives sont nombreuses:

- Identification des fonctions métaboliques des groupes de protéines spécifiques pour chaque souche
- Mise en relation de la présence de gènes spécifiques chez une souche issue d'un site hydrothermal et de son rôle égologique associé
- Reconstruction de l'histoire évolutive de ces souches grâce à plusieurs jeux de données (protéines ribosomales, gènes en copie unique, *etc.*)

Références

Eren, A.M., et al., (2015). Anvi'o: an advanced analysis and visualization platform for 'omics data. PeerJ *3*, e1319.
Dalmasso, et al. Thermococcus piezophilus sp. nov., a novel hyperthermophilic and piezophilic archaeon with a broad pressure range for growth, isolated from a deepest hydrothermal vent at the Mid-Cayman Rise. Systematic and Applied Microbiology.

Ce travail a bénéficié d'une aide de l'Etat gérée par l'Agence Nationale de la Recherche au titre du programme « Investissements d'avenir » portant la référence ANR-10-LABX-19, ainsi que d'une aide de la Région Bretagne (allocation de recherche doctorale)

2) **Oral communication presented at the 8th symposium of the "Association Francophone d'Écologie Microbienne" (French-speaking association of microbial ecology), Camaret, Oct 2017**

Genomic diversity of closely related *Thermococcus* populations in deep-sea hydrothermal vent context

La définition de l'espèce microbienne admise repose sur le pourcentage de réassociation ADN-ADN entre souches. Une valeur supérieure à 70% détermine l'appartenance à la même espèce. Or ces valeurs seuils ne rendent pas toujours compte de la diversité fonctionnelle des populations. L'utilisation de séquençage massif permet désormais l'étude de nombreux génomes en parallèle. Ce travail se focalise sur l'étude des variations intra *vs* interspécifique au niveau génomique fonctionnel. Ceci dans l'objectif d'apporter de nouvelles pistes de réflexions sur une définition des espèces cohérente du point de vue écologique et taxonomique.

Une étude de génomique comparative a été menée sur deux clades qui appartiennent à un genre Archéen facilement isolable, le genre *Thermococcus*. Il est principalement retrouvé au niveau des environnements hydrothermaux océaniques profonds. Ces deux clades, représentant des espèces différentes, ont été sélectionnés dans une collection de 250 isolats de *Thermococcus* non caractérisés. Ces clades ont été retenus suivant deux critères: au moins une vingtaine d'isolats et différentes origines géographiques.

Les premiers résultats indiquent que l'isolement géographique des membres du premier clade est un facteur de spéciation et d'acquisition de différentes capacités métaboliques. Le second clade, dont un des membres est *Thermococcus nautili*, quant à lui, ne semble pas s'engager dans cette voie. Il apparaît plutôt un maintien de plusieurs phylotypes dans le même environnement, certainement par la colonisation de micro-niches spécifiques.

Genomic diversity of closely related *Thermococcus* populations in deep-sea hydrothermal vent context

Microbial species definition is based, among others, on the percentage of DNA-DNA relatedness between strains. A value greater than 70% determines the membership of the same species. These thresholds do not always reflect the functional diversity of populations. The use of next-generation sequencing now allows studying many genomes in parallel. This work focuses on the study of intra and interspecific variations at the functional genomic level. The aim is to provide new pieces of reflection on a coherent species definition from an ecological and taxonomic point of view.

A comparative genomic study was conducted on two clades that belong to an easily isolable archaeal genus: *Thermococcus*. It is mainly found in deep oceanic hydrothermal environments. These two clades, representing different species, were selected from a collection of 250 uncharacterized *Thermococcus* isolates. These clades were selected according to two criteria: at least twenty isolates and different geographical origins.

Within the first clade, 20 genomes from 9°N East Pacific Ridge (EPR), Rainbow in mid Atlantic ridge and Saint-Paul island (south Indian Ocean), that represent 5 species according to the actual microbial species definition. Here, the geographical isolation explains the genomic divergence, and thus speciation, across members of this clade. The second clade is composed of 27 genomes close to *Thermococcus nautili*. These 27 isolates come from two hydrothermal site along the EPR: 9°N EPR and 13°N EPR. They represent 6 species according to DNA-DNA relatedness. Based on the phylogeny of

single-copy core genes, these “species” or population co-occurred in both hydrothermal sites, suggesting colonization of different ecological niches. From the genomic comparative study, it appears that the main difference rely on the presence/absence of amino-acid biosynthesis pathway and transporters across the 6 populations. For example, the complete pathway for tryptophan biosynthesis is only present in 1 population (composed of 2 genomes). This maintenance of several phylotypes in the same environment may translate a recent or ongoing speciation event.

Synthèse de mes travaux

Ce travail de thèse s'inscrit dans un nouveau thème de recherche au LM2E. Ce volet d'étude à pour mission d'étudier l'écologie des microorganismes des environnements profonds en utilisant des approches haut-débit telles que le séquençage de génomes complets, des données de métagénomique et de métatranscriptomique, et également d'utiliser les banques d'amplicons (typiquement, les régions V4-V5 ou V6 du gène de l'ARNr 16S). Dans ce cadre écogénomique, je me suis intéressé à la diversité génomique d'isolats de *Thermococcus*, proches d'un point de vue phylogénétique. Nous souhaitions en effet tester certaines hypothèses sur les mécanismes qui influencent la diversification de ces génomes, et également mieux comprendre les pressions de sélection que peuvent subir de tels organismes ayant des niches écologiques restreintes, dans les environnements hydrothermaux marins profonds.

En début de thèse, l'objectif était tout d'abord de sélectionner une collection d'isolats adapté à notre question scientifique. Nous avons pour cela utilisé la collection de cultures disponible au laboratoire: UBO culture collection (UBOCC). La partie marine de la collection se compose d'environ 1300 isolats collectés à partir de différents échantillons marins collectés durant diverses des campagnes océanographiques. Parmi ces isolats, environ 300 étaient annotés comme appartenant à l'ordre des *Thermococcales*. Cet ordre est composé d'Archées hyperthermophiles principalement retrouvées en contexte hydrothermal marin profond. Cette assignation fût basée sur des critères morphologiques et culturaux, à savoir les isolats qui croissent en anaérobiose à une température d'environ 85°C, sur un milieu riche en matière organique, mobiles et ayant une morphologie coccoïde. Afin de confirmer et d'affiner cette première assignation, la séquence du gène codant pour l'ARN de la petite sous-unité du ribosome, c'est à dire l'ARNr 16S, a été utilisé comme marqueur. Ce marqueur, une fois séquencé, a

servi à construire un arbre phylogénétique afin de déduire une affiliation taxonomique au niveau du genre. Pour ce faire, l'ensemble des isolats ont été remis en culture et incubés à 80 ou 85°C pendant 16 à 18h. Puis l'ADN de chaque isolat a été extrait, puis la séquence du gène codant l'ARNr 16S, ainsi que la séquence située entre les gènes codant pour les ARNr 16S et 23S, appelée *Internal Transcribed Spacer* (ITS) ont été amplifiées par PCR. L'intérêt de séquencer l'ITS est d'avoir un marqueur supplémentaire pour construire un arbre phylogénétique plus robuste. Au total, pour chaque isolat, 3 séquençages selon la méthode de Sanger ont été nécessaires afin d'obtenir la séquence complète de chaque "16S-ITS". Les séquences ont été assemblées et leur qualité a été contrôlée. Au total, 273 séquences étaient complètes et prêtes pour passer à l'étape suivante, la construction de l'arbre phylogénétique. Des séquences de 16S-ITS des représentants des trois genres de *Thermococcales* (*Pyrococcus*, *Thermococcus*, et *Palaeococcus*) ont été téléchargés dans les bases de données publiques. Ces séquences ainsi que les 273 séquences obtenues lors de ce travail ont été alignées et un arbre a été construit en utilisant l'inférence bayésienne. La figure 20 représente une version simplifiée de cet arbre. Parmi les 273 isolats, 14 appartenait au genre *Pyrococcus* et les 259 autres étaient affiliés aux *Thermococcus* (Annexe 1). Enfin, à partir de cet arbre, deux groupes de génomes ont été sélectionnés selon les critères suivants : (i) plusieurs origines géographiques, (ii) plusieurs génomes pour une même origine géographique, (iii) groupe monophylétique si possible. En suivant ces critères, le premier groupe sélectionné comportait 21 isolats provenant de la dorsale Est-Pacifique 13°N (EPR 13°N), du champ hydrothermal Rainbow localisé sur la dorsale médio-Atlantique et de l'île Saint-Paul située au sud de l'océan Indien (Figure 21). Le second groupe renfermait 27 génomes proches de *Thermococcus nautili*, qui étaient originaires des sites EPR 9°N et EPR 13°N dans l'océan Pacifique (Figure 21).

La seconde partie de ma thèse a débuté par le séquençage des 48 génomes sélectionnés. Tous les ADN ont été extraits au laboratoire. Une moitié des génomes a directement été envoyée au Marine Biological Laboratory (MBL) à Woods Hole, USA, pour séquençage. Concernant la seconde moitié des génomes, les banques Illumina ont été préparées au laboratoire, puis envoyées au MBL pour séquençage. Nous avons choisi un séquençage sur Illumina MiSeq en paired-end reads 2x300 pb. J'ai ensuite assemblé les génomes à l'aide de CLC Genomics Workbench, en utilisant différentes tailles de *k-mer*. Au total 46 génomes ont été assemblés avec succès, dont 19 avec un seul contig circulaire, 15 avec un contig non circularisé et 13 génomes demeuraient fragmentés (2 à 57 contigs). Le séquençage a échoué pour un génome du group I (MC5), et il semble que le génome de l'isolat *Thermococcus* sp. E15P25 (groupe II) était contaminé, deux génomes semblant être présents, rendant sont utilisation impossible.

Dans un second temps, nous avons construit un arbre phylogénomique afin de mettre tous ces génomes dans un contexte évolutif. Cet arbre avait un second objectif, vérifier que ces génomes s'organisaient bien suivant les groupes définis avec les marqueurs ARNr 16S et ITS. Afin de construire une phylogénie solide, l'ensemble des génomes de *Thermococcales* disponibles a été utilisé. La phylogénie réalisée était basée sur les gènes du core-génome présents en simple copie, car c'est un jeu de donnée riche et défini sans *a priori* sur la fonction des gènes. Pour obtenir ce set de gènes, il a fallu réaliser une analyse de pangénomique, qui a permis de définir d'une part, l'ensemble des gènes partagés par tous les génomes étudiés (= core-génome), et d'autre part l'ensemble des autres gènes, soit le génome accessoire. L'union de ces deux catégories a formé le pangénome. Ce dernier a été établi avec anvi'o. Brièvement, anvi'o détecte les séquences codantes (CDS) dans tous les génomes, puis compare toutes les séquences elles. Ce résultat était ensuite fourni à MCL, un algorithme qui regroupe les gènes *via* une

approche graphique. Il en est ressorti une liste de groupes de gènes (PC, Protein Clusters). Dans un PC, si il y avait un gène de chaque génome, ce PC appartenait au core-génome. Lorsqu'un PC du core-génome contenait autant de gènes qu'il y a de génomes, le PC concerné appartenait aux « gènes du core-génome en copie unique » (SCGs). L'ensemble des autres PC représentait le génome accessoire. La figure 23 montre une représentation de ce pangénome et la distribution des PC accessoires. Après cette étape, les 602 SCGs ont été extraits, alignés, ajustés et concaténés. Cet alignement d'environ 92 000 positions non-redondantes nous a servi pour construire la phylogénie de ces 114 génomes, par la méthode du maximum de vraisemblance (Figure 24A). Dans cet arbre, les 21 isolats du groupe I formaient bien un groupe monophylétique. Deux génomes provenant des bases de données publiques venaient s'ajouter à ce groupe : *Thermococcus celericrescens* et *T.* sp 4557. Le groupe II était composé de 34 génomes : 25 séquencés dans ce projet, 1 génome de référence (*T. nautili*) et 8 génomes non-publiés fournis par le laboratoire MBGE de l'institut Pasteur. Ce groupe devait contenir 26 des isolats séquencés pour ce travail, mais la souche *T.* sp. IRI06c se situait ailleurs dans l'arbre. Après vérification, la séquence du gène de l'ARNr 16S du génome et celle obtenue lors de la première partie de la thèse différaient de 15 nucléotides. Ce génome a donc simplement été conservé pour le pangénome global, mais n'a pas été utilisé pour la suite du travail.

Cet arbre confirmant l'existence de ces deux groupes distincts, nous nous sommes interrogés sur la nature des paramètres pouvant expliquer l'organisation des isolats dans chaque groupe. Dans un premier temps, l'incidence de l'origine géographique a été analysée. En ce qui concerne le groupe I, ce facteur expliquait à lui seul l'organisation des génomes en clades (groupe ayant un ancêtre commun) dans cet arbre (Figure 24B). Les génomes du groupe II ne suivaient, quant à eux, pas cette tendance, l'origine

géographique n'expliquant pas leur organisation dans cet arbre. Le second paramètre qui a été étudié est la présence d'espèce(s) microbienne, *via* les deux métriques que sont l'ANI (identité nucléotidique moyenne entre les génomes) et l'hybridation ADN ADN (DDH) communément utilisées pour la délinéation des espèces. Le groupe I comptait 7 espèces selon ces deux métriques : une par origine géographique, sauf pour le site Rainbow, qui comptait 3 espèces sympatriques (Figure 25). Le groupe II était composé, quant à lui, de 6 espèces, toujours suivant l'ANI et la DDH (Figure 25).

La dernière étape de cette étude a consisté à identifier les gènes et les voies métaboliques spécifiques résultants des processus évolutifs de différentiation. Pour chaque groupe, un nouveau pangénome a été établi (Figures 26-27). À partir de là, l'objectif a ensuite été d'identifier les gènes spécifiques à chaque espèce et d'identifier les fonctions uniquement retrouvées dans ces gènes. En résumé, il y avait un nombre très variable de gènes spécifiques, de 1 à 336 suivant les espèces. De façon remarqualble, la moitié n'avait pas d'annotation. Parmi les gènes restants, beaucoup de fonctions étaient redondantes entre les espèces. Néanmoins, des fonctions uniques ont été identifiées : elles étaient majoritairement associées au métabolisme des acides-aminés, au métabolisme énergétique, au métabolisme des sucres ou encore au transport d'ions inorganiques. Ces résultats ont apportés des éléments d'information sur les pressions de sélections qui peuvent s'appliquer sur ces micro-organismes dans les environnements profonds. Ceci nous a également renseigné sur les métabolismes acquis ou perdus qui entrainent la formation de nouvelles espèces microbiennes dans l'environnement hydrothermal marin profond.

La dernière partie de cette thèse portait sur la dissémination des *Thermococcales* dans l'environnement. Pour ce faire, des métagénomes et métatranscriptomes provenant d'environnements hydrothermaux profonds et de sources chaudes terrestres ont été

alignés sur des génomes de *Thermococcales* (Figure 28). Dans l'ensemble, les génomes ont été retrouvés dans les métagénomes des fonds marins mais pas dans ceux des sources chaudes terrestres, la barrière entre ces deux environnements étant sans doute trop complexe à franchir par le simple fait de la dispersion aléatoire. Contrairement aux *Thermococcus*, les genres *Pyrococcus* et *Palaeococcus* ont été beaucoup moins détectés dans les métagénomes. Ceci laisse présager de niches écologiques plus restreintes et d'une faible abondance de ces Archées dans l'environnement. Enfin, certaines souches de *Thermococcus*, telles que *T. cleftensis,* semblaient quant à elles présenter une distribution plus comsmopolite (Pacifique Nord-Est et fosse des Caïmans en Atlantique), ce qui suggère qu'elles seraient capables de migrer sur de longues distances, alors que d'autres souches semblent ne rester qu'au niveau d'un seul site, ou qu'elle colonisent des niches au spectre écologique plus répandus. Le caractère novateur de cette dernière étude en outre était de donner des pistes de recherche afin d'isoler de nouveaux taxons intéressants, comme des *Palaeococcus*, faiblement représentés par rapport à l'immense majorité de *Thermococcus* isolés, caractérisés et séquencés.

Génomique comparative d'isolats phylogénétiquement proches appartenant au genre Thermococcus, une archée hyperthermophile

L'immense diversité génomique des microorganismes leur permet de vivre partout, même dans les environnements extrêmes tels que les sources hydrothermales profondes. Ces dernières, disséminées sur l'ensemble des fonds océaniques, sont un bon modèle pour étudier la biogéographie et la diversification des génomes. Une approche de génomique comparative a été employée sur des isolats du genre *Thermococcus* proches d'un point de vue évolutif. Ce travail visait à identifier des mécanismes ayant un rôle dans la diversification de ces génomes, et également d'identifier des gènes impliqués dans cette différenciation. A cette fin, deux groupes d'une vingtaine d'isolats ayant des origines géographiques diverses ont été sélectionnés et séquencés.
L'éloignement géographique résultant de la colonisation de nouveaux systèmes hydrothermaux semble être un facteur de diversification et de spéciation pour certains isolats. Cependant, lorsque les sites hydrothermaux sont relativement proches, il semblerait qu'un transfert de gènes entre les isolats soit toujours possible. Dans ce cas, l'adaptation à de nouvelles niches écologiques serait un facteur de la diversification des génomes. L'approche de génomique comparative a permis d'identifier des gènes spécifiques à certains sous-groupes, apparentés à des espèces. Ces gènes sont notamment impliqués dans les métabolismes des acides aminés, de production d'énergie et de transport d'ions inorganiques. Ceci reflète les pressions de sélections que peuvent subir ces organismes dans ces environnements hostiles à nombreuses formes de vie.

Mots clés : Génomique comparative, *Thermococcus*, Diversification, Génome, Hyperthermophile

Comparative genomics of closely related Thermococcus isolates, a genus of hyperthermophilic Archaea

The immense genomic diversity of microorganisms allows them to live everywhere, even in extreme environments such as deep hydrothermal vents. Scattered over the seabed, these are a good model for studying the biogeography and genomes diversification. A comparative genomics approach has been used on closely related isolates, of the genus *Thermococcus*. This work aimed at identifying mechanisms that have a role in the diversification of these genomes, and also to identify genes involved in this differentiation. For this purpose, two groups of about 20 isolates with different geographical origins were selected and sequenced.
The geographical isolation resulting from colonization of new hydrothermal systems is likely to be a diversification and speciation factor for some isolates. But when hydrothermal sites are relatively close, it would seem that gene transfer between isolates is still possible. In this case, adaptation to new ecological niches would be a factor contributing to the genomes diversification. The comparative genomics approach allowed highlighting genes specific to certain subgroups, related to species. These genes are involved in amino acid metabolism, energy production and the transport of inorganic ions. This reflects selection pressures that these organisms may experience in these environments, otherwise hostile to many forms of life.

Key words: Comparative genomics, *Thermococcus*, Diversification, Genome, Hyperthermophile