bioRxiv preprint doi: https://doi.org/10.1101/2020.12.19.423498; this version posted December 20, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

Genome evolution in bacteria isolated from million-year-old subseafloor sediments

William D. Orsi[1,2*], Tobias Magritsch[1], Sergio Vargas[1], Ömer K. Coskun[1], Aurele Vuillemin[1], Sebastian Höhna[1,2], Gert Wörheide[1,2,3], Steven D'Hondt[4], B. Jesse Shapiro[5,6,7], Paul Carini[8*]

[1]Department of Earth and Environmental Sciences, Paleontology & Geobiology, Ludwig-Maximilians-Universität München, Richard-Wagner-Strasse 10, 80333 Munich, Germany.

[2]GeoBio-CenterLMU, Ludwig-Maximilians-Universität München, Richard-Wagner-Strasse 10, 80333 Munich, Germany.

[3]SNSB- Bayerische Staatssammlung für Paläontologie und Geologie, Richard-Wagner-Strasse 10, 80333 Munich, Germany.

[4]Graduate School of Oceanography, University of Rhode Island, 215 South Ferry Road, 02882 Narragansett, USA.

[5]Department of Biological Sciences, University of Montreal, QC, Canada

[6]Department of Microbiology and Immunology, McGill University, QC, Canada

[7]McGill Genome Centre, Canada

[8]Department of Environmental Science, the BIO5 Institute, School of Plant Sciences, and the School of Comparative Animal & Biomedical Science, University of Arizona, Tucson, Arizona USA

Corresponding authors*:
Prof. Dr. William D. Orsi
Ludwig-Maximilians-Universität München, Department of Earth and Environmental Sciences, Paleontology & Geobiology, Richard-Wagner-Strasse 10, 80333 Munich, Germany.

Corresponding authors*:
Dr. Paul Carini
Address: University of Arizona, Department of Environmental Science, School of Plant Science, BIO5 Institute. Tucson, Arizona 85721
E-Mail: paulcarini@arizona.edu
Phone/Fax: Phone: 520-621-1646

Running title: Genome evolution in subseafloor bacteria

Keywords: genome evolution; deep biosphere

bioRxiv preprint doi: https://doi.org/10.1101/2020.12.19.423498; this version posted December 20, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

Significance statement: In microbial populations that subsist in isolation from the surface world in deep subseafloor sediment over millions of years, ultra-slow metabolic rates caused by long term energy limitation are hypothesized to restrict the spread of newly evolved traits. It remains unknown whether genomic evolution occurs under these extreme conditions. Our findings demonstrate that genomes of cultivated bacterial strains from the genus *Thalassospira* isolated from million-year-old abyssal sediment exhibit greatly reduced levels of homologous recombination, elevated numbers of pseudogenes, and widespread evidence of relaxed purifying selection. Our findings show that the genome evolution of these anciently buried bacteria has proceeded in a manner dominated by genetic drift, whereby in small population sizes, and in the absence of homologous recombination, mutations became fixed into the population which has led to the emergence of new genotypes.

Abstract: Deep below the seafloor, microbial life subsists in isolation from the surface world under perpetual energy limitation. The extent to which subsurface microbes evolve and adapt to their subseafloor habitat is unclear, given their ultra-slow metabolic rates. Here we show that genomes of *Thalassospira* bacterial populations cultured from million-year-old subseafloor sediments evolve by point mutation, with a relatively low rate of homologous recombination and a high frequency of pseudogenes. Ratios of synonymous to non-synonymous mutation rates correlate with the accumulation of pseudogenes, consistent with a dominant role for genetic drift in the subseafloor genomes, but not in type strains of *Thalassospira* isolated from surface world habitats. The genome evolution of these anciently buried bacteria has apparently proceeded in a genetic drift-like manner, whereby under long-term isolation with reduced access to novel genetic material from neighbors, new mutations became fixed into the populations leading to the emergence of new genotypes.

Main text.

The subseafloor biosphere contains one-third of all bacterial cells on Earth totaling $>10^{29}$ cells, which subsist over geological timescales under perpetual energy limitation (*1*). Whether evolution and ecological differentiation occurs in microbial populations below the seafloor has remained controversial. It is generally agreed that extreme energy limitation restricts metabolic activity and growth (*2*), which are

bioRxiv preprint doi: https://doi.org/10.1101/2020.12.19.423498; this version posted December 20, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

necessary for new mutations to propagate through populations to foster ecological differentiation and speciation (*3*). A metagenomic analysis showed that energy limitation and reduced growth restricted the spread of new mutations through microbial communities over 5,000 years in the upper 10 meters of anoxic continental shelf sediments (*4*). Yet, energy-starved cells under experimental laboratory conditions may undergo adaptations that confer fitness under long-term energy limitation (*5, 6*). Moreover, bacteria surviving for up to half a million years in deep subsurface permafrost have been shown to actively repair their DNA (*7*)—a process that can tamper the rate of molecular evolution in the subsurface (*4*). Here, we used the genomes of bacteria isolated from million-year-old subseafloor abyssal clay sediments to investigate the nature of genome evolution in subseafloor bacteria that persist under extreme energy limitation over long timescales.

Newly acquired mutations of functional significance can sweep through relatively fast-growing bacteria in surface world habitats and influence ecological differentiation (*8*), but it is unclear whether this occurs in ancient subseafloor sediment given the comparably slow subseafloor bacterial biomass turnover rates that are estimated to be on thousand-year timescales (*9*). The metabolic rates of microbes persisting in deep-sea abyssal clay are amongst the lowest observed in the subseafloor biosphere, such that these sediments are often oxic through the entire sediment column to the underlying oceanic crust (*10*). Microbes inhabiting such abyssal clay subseafloor settings are characterized by ultra-slow respiration rates and live near the low-energy limit to life (*11*). We cored a 15 m sedimentary sequence of oxygenated abyssal clay at a water depth of 6,000 m in the North Atlantic where the average sedimentation rate is an estimated 1 meter per million years (*12*). A relatively slow drawdown of O_2 with increasing depth at this site (Fig. 1A) reflects oxidation of organic matter by aerobic microbes.

We isolated colony-forming bacteria on petri dishes following an 18-month incubation of sediment and sterile ^{18}O-labeled artificial seawater (Fig S1) from 3 and 6 meters below the seafloor (mbsf)(see Methods). Because the mean sedimentation rate is estimated to be on average 1 m million yr^{-1}, the age of the sediments from which these bacteria were enriched and isolated are estimated to be 3 and 6 million years old, respectively. The full length 16S rRNA gene sequences from the isolates had closest similarity (90-99% sequence identity) to *Thalassospira xiamenensis and Thalassospira lohafexi* previously isolated from marine sediments (*13,14*) and oligotrophic marine waters (*15*) (these previously isolated microbes and their related cultured relatives are referred to as 'type strains' herein).

Several lines of evidence indicate the *Thalassospira* isolated from the 3 and 6 mbsf sediment enrichments are endemic to the subseafloor abyssal clays and are not a contaminant from the water column or other sources. First, the V4 hypervariable region of the 16S rRNA gene sequences from the sediment slurry enriched *Thalassospira* cultures share >99% sequence identity with a phylotype (OTU6) previously identified from the *in-situ* community determined from the frozen samples (Figure 1C). This OTU became

bioRxiv preprint doi: https://doi.org/10.1101/2020.12.19.423498; this version posted December 20, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

^{18}O-labeled during the 18-month incubation (atomic ^{18}O labeling of DNA: 59%) in the presence of sterile ^{18}O-labeled seawater (Fig S1), a proxy for growing microbes (*16*), and had an estimated doubling time in the incubation of 36 ± 1.5 (mean ± SD) days. This phylotype consists of two amplicon sequence variants (ASVs) that cluster with the *Thalassospira* subseafloor isolates (Otu6_ASV1, Otu6_ASV2), respectively, and are distinguished by a single nucleotide polymorphism (SNP) that is conserved between the ASVs and the subseafloor isolates (Figure 1C). The *in-situ* concentrations of both ASVs have highest abundance (ca. 1,000 16S rRNA gene copies g^{-1} sediment) between 4 - 6 mbsf, and both *Thalassospira* ASVs were detected in the 3 and 6 mbsf sediment (Figure 1B). This shows that the *Thalassospira* strains isolated from the 3 and 6 mbsf sediment enrichments are derived from the same distinct 16S rRNA gene ASVs present within the *in-situ* communities. The long-term physical isolation of these isolates in the subseafloor (see 'sediment physical properties' in SI) subsisting under uninterrupted energy limitation within these ancient sediments provides an opportunity to investigate how the relative effects of recombination, nucleotide substitution, and gene decay have shaped the genomes of the cultivated subsurface *Thalassospira* strains since their burial in the deep-sea clay millions of years ago.

Genome statistics

We sequenced the genomes of ten *Thalassospira* isolates each from the 3 and 6 mbsf sediment enrichments using a hybrid assembly approach consisting of long-read Nanopore sequencing, corrected and polished via short-read Illumina technologies at >100x coverage. The mean genome completeness of these hybrid assemblies was estimated to be 99.7% ± 0.3%; (mean ± SD), with most being 100% complete and representing the complete chromosome (Table S1). The mean length of these new *Thalassospira* genomes was 4.71 ± 0.08 Mbp (mean ± SD, Fig. S2), with 4,567 ± 107 (mean ± SD) protein coding genes (Fig. S2), and were assembled to an average of 12 ± 2 (mean ± SD) contigs (Table S1). The genome size and number of protein encoding genes are similar to those observed within the existing *Thalassospira* type strains isolated from the surface world (Fig. S2).

Core genome phylogenomic analysis

The core genome phylogeny of existing *Thalassospira* type species, and the newly isolated subseafloor *Thalassospira* strains, consisted of 1,809 orthologous genes and revealed three clades of subseafloor *Thalassospira*. One clade shares 96-97% genome-wide average nucleotide identity (ANI), and 99.9% 16SrRNA gene sequence identity, with *T. xiamenensis* and *T. permensis* (Fig. S3). We named isolates in this clade *T. xiamenensis* strain 'Neogene', after the Neogene eon (2.8 – 23 mya) which covers both estimated ages of sediment (3 mya and 6 mya) from which the strains in this clade were isolated. The subseafloor genomes in this clade correspond to the 16S rRNA gene ASV1 detected in the *in situ* frozen

bioRxiv preprint doi: https://doi.org/10.1101/2020.12.19.423498; this version posted December 20, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

sediment core samples (Fig. 1B, C). A second clade contained three isolates from 6 mbsf sediment shared 97% ANI with *T. xiamenensis* and *T. permensis.* Since all isolates in this clade were recovered from ca. 6 mya sediment deposited during the Miocene eon (5.33 – 23 mya), we report them as *T. xiamenensis* strain 'Miocene' (Fig. S3). However, despite sharing 97% ANI in the core genome with *T. xiamenensis* and *T. permensis* (Fig. S3), *T. xiamenensis* strain 'Miocene' only shared 90% 16S rRNA gene sequence identity with these closest related type strains. The subseafloor genomes in this clade also correspond to ASV1 detected in the *in situ* frozen sediment core samples (Fig. 1B, C). A third clade contained subseafloor *Thalassospira* cultures isolated only from 3 mbsf sediment and shared 95-96% ANI with *T. lucentensis* and *T. lohafexi* (Fig. S3). The subseafloor genomes in this clade correspond to the 16S rRNA gene ASV2 detected in the *in situ* frozen sediment core samples (Fig. 1B, C). Based on the genetic distinctness of these isolates we consider them to be a new candidate species, according to recently provided criteria based on genome-wide ANI (*17*). Because all isolates of this third clade were recovered from 3 mya sediment, we propose the candidate name '*Candidatus* Thalassospira pliocenensis', named after the Pliocene age (2.58 – 5.33 mya) of the deep-sea clay sediments from which they were isolated. Pangenome analysis revealed that flexible genome content is conserved within each of the three subseafloor clades, further evidence that each clade represents a genetically distinct population (Fig. S4).

Roles of mutation and recombination

The ratio of nucleotide diversity originating from mutations versus homologous recombination (r/m) can be used to measure the relative effect of homologous recombination on the genetic diversification of populations, (*18*). Due to the physical isolation of individual bacterial cells, reduced cell concentrations, and the reduced availability of extracellular DNA for recombination in subseafloor sediments (*2*) we hypothesized that rates of homologous recombination in the subseafloor *Thalassospira* populations would be lower compared to the type strains. To test this, we used an established method (*19*) to calculate the relative rate of recombination to mutation (R/θ), the mean length of recombined DNA (δ), and the mean divergence of imported DNA (v) for branch tips (existent genomes) and internal nodes (ancestral states) in the *Thalassospira* core genome phylogeny, which allows for a calculation of r/m (r/m = $(R/\theta) * \delta * v$). This analysis showed that in the *Thalassospira* core genome, the r/m is approximately ten times lower in the existent subseafloor core genomes (r/m = 0.078) compared to the type strains (r/m = 0.71) (Table 1), indicating that homologous recombination plays a much lesser role in the diversification of the subseafloor strains. The r/m values of the subseafloor *Thalassospira* are furthermore anomalously low compared to free-living bacteria isolated from the surface world, which have r/m values that range from 0.1-64 (*18*). Concomitant with the ten-fold lower r/m values compared to the type strains (Table 1), the subseafloor

bioRxiv preprint doi: https://doi.org/10.1101/2020.12.19.423498; this version posted December 20, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

Thalassospira core genomes exhibit far fewer numbers of inferred imported DNA from recombination events compared to the *Thalassospira* type strains and the ancestral states of the last common ancestors (Fig. 2).

We considered the possibility that subseafloor *Thalassospira* populations may be those with the necessary traits to survive at the time of burial (*2*), and little (or no) genome evolution may occur thereafter (*4*). We looked for evidence of evolution by investigating pairwise substitution numbers in the subseafloor *Thalassospira* genomes. We identified 10's to 1000's of nucleotide differences (single nucleotide polymorphisms [SNPs]) within each subseafloor *Thalassospira* clade, suggesting that evolution by point mutation has occurred since burial (Fig. S5). The SNPs in subseafloor *Thalassospira* strains were present in a clade-specific manner (Fig. S6) and included genes with predicted annotations involved in flagellar motility (FlhB, FliO), transcription (TetR and Fis family transcriptional regulators), cell wall biogenesis (peptidase S41, peptidoglycan DD-metalloendoptidase M23), and transport and metabolism of amino acids and carbohydrates (Fig. S6). Alternatively, the nucleotide diversity we observed in the distinct subsurface *Thalassospira* clades may have been present at the time of burial. However, the subseafloor *Thalassospira* genomes have widely different nucleotide diversity when compared to the predicted ancestral states of the last common ancestor with the type strains (Fig. 2). Thus, it is likely the clade-specific pairwise substitutions accrued during the time below the seafloor.

We considered the possibility that these substitutions occurred within the subseafloor populations during the culture enrichment process. However, the doubling times of the *Thalassospira* (OTU6) in the incubation measured with qPCR (36 ± 1.5 days; mean ± SD) indicate an estimated maximum of 15 doublings over the incubation (see SI). Given the number of generations, and the mutation rate for bacterial genomes (*20*) we calculated less than a single mutation would be expected to arise in the subseafloor *Thalassospira* genomes over the 15 generations that occurred during 18 month enrichment (4.7 Mbp in the *Thalassospira* genome × 1× 10^{-9} mutations bp^{-1} $generation^{-1}$ × 15 generations = 0.07 mutations). This is insufficient to explain the observed interpopulation nucleotide diversity between the subseafloor genomes which is 952 ± 177 nucleotide differences Mbp^{-1} (Fig. S5). Thus, the inter-population nucleotide diversity of the subseafloor strains likely arose during their long-term subsistence in the ancient sediments and is not the result of evolution during the laboratory incubation.

Substitutions and pseudogenes are fixed in subseafloor populations

The greatly reduced role homologous recombination in the subseafloor *Thalassospira* genomes (Table 1, Fig. 2) coincided with higher numbers of pseudogenes (non-functional parts of the genome that resemble functional genes), and non-synonymous substitutions (substitutions that alter the amnio acid sequence of a protein). We identified 47.9 ± 8.57 (mean ± SD) pseudogenes in the genomes of subseafloor

bioRxiv preprint doi: https://doi.org/10.1101/2020.12.19.423498; this version posted December 20, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

Thalassospira isolates, which is significantly higher than the number of pseudogenes identified in the type strains (22.1 ± 5.52 pseudogenes [mean ± SD]; Table 1) (two-sided T-test: P=1.5E-10). Similarly, we observed a modest but significant elevation of genome-wide nonsynonymous to synonymous substitution rates (dN/dS) in the genomes of the subseafloor *Thalassospira* cultures (0.035 ± 0.006; mean ± SD), relative to the type strains (0.022 ±0.012; mean ± SD) (two-sided t-test P=0.0002; Table 1 and Fig. 3). Like the SNPs, the composition of pseudogenes occurs in a clade-specific manner (ANOSIM: P = 0.001) (Fig. S6-S8). Compared to the type strains, the predicted annotations of the pseudogenes in the subseafloor *Thalassospira* genomes are skewed towards those involved in transcription, energy conservation, amino acid and carbohydrate metabolism, and flagellar motility (Fig. S7). Subseafloor genomes have significantly higher numbers of pseudogenes involved in flagellar biosynthesis (FliN, FliK, FlhO) compared to the type strains (Fig. S8).

In microbial genomes, non-essential genes under relaxed purifying selection are more susceptible to mutation-driven decay into pseudogenes (*21*). In small populations, mildly deleterious nonsynonymous substitutions and pseudogenes are fixed into the population by chance through genetic drift (*22*). Thus, key genomic signatures associated with drift-related evolutionary processes are elevated dN/dS ratios and a high proportion of pseudogenes (*23,24*). In the absence of recombination, the effects of nonsynonymous or slightly deleterious mutations compound with each generation because descendants carry the mutational burden of the parent generation (*23,24*).

The elevated and correlated genome-wide pseudogene count and dN/dS ratios (Fig. 3) in the subsurface *Thalassospira* isolates indicates that upon burial the population size became restricted relative to surface-world *Thalassospira*. Thus, the burial in sediments likely had two significant effects. First, the reduced energy availability—in the form of reduced quality and quantity of reactive organic carbon (*25*)—limited the environmental carrying capacity, and thus the population size, in these ancient subseafloor clays. Second, because the population size was limited, the chance for recombination events was also reduced due to infrequent cell-cell contact stemming from lower cell abundances and subseafloor *Thalassospira* cells were less likely to encounter genetically diverse recombination partners that might introduce genetic diversity into the population. The reduction of homologous recombination resulted in mutation becoming the dominant driver of evolution in these subseafloor *Thalassospira* strains as highlighted by the order of magnitude lower r/m values in the genomes of subseafloor *Thalassospira* (Table 1). Thus, our observations in subsurface *Thalassospira* populations are consistent with relaxed purifying selection in the absence of recombination in a small microbial population. Although we see elevated dN/dS ratios and an accumulation of pseudogenes across the core genomes of subsurface *Thalassospira*, some functions appear to be more prone to gene decay than others. For example, genes predicted to be involved in flagellar motility were present in both the SNP (*fliO*, *flhB*, *flgH*; Fig. S5) and pseudogene (*fliN*, *fliK*, *flhO*; Fig. S8) analysis,

bioRxiv preprint doi: https://doi.org/10.1101/2020.12.19.423498; this version posted December 20, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

suggesting physical restriction and energetic limitation has rendered motility as a superfluous function in the highly compacted ancient deep-sea clay.

Lack of genome reduction despite energy limitation

The concerted action of mutation and decay into nonfunctional pseudogenes can culminate in the loss of chromosomal DNA through a proposed deletional bias in microbial genomes (*26*). We did not observe evidence of genome reduction in the subseafloor *Thalassospira* genomes that might be expected in small, recombination-limited bacterial populations (*21*). We speculate that ultra-slow generation times of subsurface *Thalassospira* due to perpetual energy limitation has prevented sufficient generations to pass for the cells to lose superfluous DNA or genome reduction is balanced by the influx of new DNA by lateral gene transfer into the flexible genome (*27*). Alternatively, our selection of genomes and the inferences derived may be biased by our analysis of genomes from culturable microbes and genome reduction is often associated with complex growth requirements or unculturability (*28*).

Outlook:

Free-living bacterial populations may engage in high rates of homologous recombination (*8*), which may allow them to differentiate ecologically and purge deleterious mutations. However, the nearly complete lack of recombination in the subseafloor *Thalassospira* genomes analyzed here shows that they have evolved under a different regime akin to endosymbiotic bacteria, which lack homologous recombination and are thus subject to genetic drift whereby deleterious mutations become fixed (*21*). In the absence of recombination, the so-called 'Muller's Ratchet' eventually leads to the extinction of endosymbiotic bacterial lineages (*21*). Our genomes show similar signs of evolution, namely greatly reduced recombination and elevated dN/dS ratios and pseudogene numbers (Table 1). However, dN/dS ratios observed in the subseafloor *Thalassospira* genomes are lower than those seen in endosymbiotic bacteria (*24*) and genome reduction was absent. Thus, it appears that the subseafloor genomes are in a middle state, perhaps "one click" in the Mullers Ratchet, represented by a single burial event followed by a stable but low population size, in contrast to the repeated population bottlenecks experienced by endosymbionts at each insect generation.

Subseafloor communities are relatively small and become smaller and more physically isolated with increasing depth reaching the ultra-low concentrations of one cell per cubic centimeter of sediment at 2,500 meters below seafloor (*29*). Because genetic drift has a stronger effect on populations with small population sizes (*21*,*24*), microbes in the deep biosphere with small population sizes experiencing reduced homologous recombination may be particularly prone to genetic drift-mediated evolution. For example, a metagenomic analysis showed that an anaerobic subseafloor population displayed elevated

bioRxiv preprint doi: https://doi.org/10.1101/2020.12.19.423498; this version posted December 20, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

dN/dS ratios *(4)*. Because the subseafloor biosphere contains one-third of all bacterial cells on Earth (*1*), our findings suggest drift-like evolutionary processes in the absence of homologous recombination may be much more widely distributed in nature than previously thought. Future assessments of homologous recombination and drift in single cell genomes from uncultured lineages of bacteria and archaea that comprise most subsurface energy limited communities (*30)* could be used to assess how widespread this evolutionary mechanism is within the subsurface biosphere.

Acknowledgements. This work was supported primarily by the Deutsche Forschungsgemeinschaft (DFG) project OR 417/1-1 granted to W.D.O. Publication of the manuscript was supported by the LMU Mentoring Program. The expedition was funded by the US National Science Foundation through grant NSF-OCE-1433150 to S.D., who led the expedition. This is Center for Dark Energy Biosphere Investigations (C-DEBI) publication number XXX. A portion of this work was performed as part of the LMU Masters Program "Geobiology and Paleobiology" (MGAP). P.C. was supported by the University of Arizona's Technology and Research Initiative Fund (the Water, Environmental, and Energy Solutions initiative).

Author contributions. W.D.O., P.C., J.S., and G.W. conceived the work and experimental approach. W.D.O., T.M., O.K.C., S.V., P.C., S.H., and A.V. contributed to the laboratory and bioinformatics analyses and experimental work. S.D. obtained the samples during the KN223 R/V Knorr oceanographic expedition. All authors discussed and wrote the manuscript and commented on the paper.

Competing interests. The authors declare that they have no competing interests.

Supplemental information. Supplemental Table 1, Supplemental Figures S1-S8, References 31-49.

Data and materials availability. Data are publicly available through NCBI BioProject PRJNA473406. The 16S data are available in SRA BioSample accessions SAMN10929403 to SAMN10929517. Figures and output files from the pangenomic analysis in Anvio are available online through FigShare (https://figshare.com/s/06ba1287a00ab01a1ee). Additional data related to this paper may be requested from the authors.

bioRxiv preprint doi: https://doi.org/10.1101/2020.12.19.423498; this version posted December 20, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

References

1. J. Kallmeyer, R. Pockalny, R. R. Adhikari, D. C. Smith, S. D'Hondt, Global distribution of microbial abundance and biomass in subseafloor sediment. *Proc Natl Acad Sci U S A* **109**, 16213-16216 (2012).
2. M. A. Lever *et al.*, Life under extreme energy limitation: a synthesis of laboratory- and field-based investigations. *FEMS Microbiol Rev* **39**, 688-728 (2015).
3. O. X. Cordero, M. F. Polz, Explaining microbial genomic diversity in light of evolutionary ecology. *Nat Rev Microbiol* **12**, 263-273 (2014).
4. P. Starnawski *et al.*, Microbial community assembly and evolution in subseafloor sediment. *Proc Natl Acad Sci U S A* **114**, 2940-2945 (2017).
5. S. E. Finkel, Long-term survival during stationary phase: evolution and the GASP phenotype. *Nat Rev Microbiol* **4**, 113-120 (2006).
6. L. M. Wick, H. Weilenmann, T. Egli, The apparent clock-like evolution of Escherichia coli in glucose-limited chemostats is reproducible at large but not at small population sizes and can be explained with Monod kinetics. *Microbiology (Reading)* **148**, 2889-2902 (2002).
7. S. S. Johnson *et al.*, Ancient bacteria show evidence of DNA repair. *Proc Natl Acad Sci U S A* **104**, 14401-14405 (2007).
8. B. J. Shapiro *et al.*, Population genomics of early events in the ecological differentiation of bacteria. *Science* **336**, 48-51 (2012).
9. B. A. Lomstein, A. T. Langerhuus, S. D'Hondt, B. B. Jorgensen, A. J. Spivack, Endospore abundance, microbial growth and necromass turnover in deep sub-seafloor sediment. *NAture* **484**, 101-104 (2012).
10. H. Roy *et al.*, Aerobic microbial respiration in 86-million-year-old deep-sea red clay. *Science* **336**, 922-925 (2012).
11. S. D'Hondt *et al.*, Presence of oxygen and aerobic communities from sea floor to basement in deep-sea sediments. *Nature Geoscience* **8**, 299-304 (2015).
12. A. Vuillemin *et al.*, Archaea dominate oxic subseafloor communities over multimillion-year time scales. *Sci Adv* **5**, eaaw4108 (2019).
13. M. Li, S. Yang, Q. Lai, Z. Shao, Draft Genome Sequence of Thalassospira xiamenensis Strain MCCC 1A03042(T). *Genome Announc* **5**, (2017).
14. C. Liu, Y. Wu, L. Li, Y. Ma, Z. Shao, Thalassospira xiamenensis sp. nov. and Thalassospira profundimaris sp. nov. *Int J Syst Evol Microbiol* **57**, 316-320 (2007).
15. A. Hutz, K. Schubert, J. Overmann, Thalassospira sp. isolated from the oligotrophic eastern Mediterranean Sea exhibits chemotaxis toward inorganic phosphate during starvation. *Appl Environ Microbiol* **77**, 4412-4421 (2011).
16. B. A. Hungate *et al.*, Quantitative microbial ecology through stable isotope probing. *Appl Environ Microbiol* **81**, 7570-7581 (2015).

bioRxiv preprint doi: https://doi.org/10.1101/2020.12.19.423498; this version posted December 20, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

17. L. M. Rodriguez-R, K. Konstantinidis, Bypassing cultivation to identify bacterial species. *Microbe* **9**, 111-118 (2014).
18. M. Vos, X. Didelot, A comparison of homologous recombination rates in bacteria and archaea. *ISME J* **3**, 199-208 (2009).
19. X. Didelot, D. J. Wilson, ClonalFrameML: efficient inference of recombination in whole bacterial genomes. *PLoS Comput Biol* **11**, e1004041 (2015).
20. J. W. Drake, B. Charlesworth, D. Charlesworth, J. F. Crow, Rates of spontaneous mutation. *Genetics* **148**, 1667-1686 (1998).
21. J. P. McCutcheon, N. A. Moran, Extreme genome reduction in symbiotic bacteria. *Nat Rev Microbiol* **10**, 13-26 (2011).
22. H. Ochman, L. M. Davalos, The nature and dynamics of bacterial genomes. *Science* **311**, 1730-1733 (2006).
23. N. A. Moran, Accelerated evolution and Muller's rachet in endosymbiotic bacteria. *Proc Natl Acad Sci U S A* **93**, 2873-2878 (1996).
24. C. H. Kuo, N. A. Moran, H. Ochman, The consequences of genetic drift for bacterial genome complexity. *Genome Res* **19**, 1450-1454 (2009).
25. E. R. Estes *et al*., Persistent organic matter in oxic subseafloor sediment. *Nature Geoscience* **12**, 126-131 (2019).
26. A. Mira, H. Ochman, N. A. Moran, Deletional bias and the evolution of bacterial genomes. *Trends Genet* **17**, 589-596 (2001).
27. Nelson-Sathi *et al*., Origins of major archaeal clades correspond to gene acquisitions from bacteria. *Nature* **517**, 77-80 (2015).
28. P. Carini, A "Cultural" Renaissance: Genomics Breathes New Life into an Old Craft. *mSystems* **4**, (2019).
29. F. Inagaki *et al.*, Exploring deep microbial life in coal-bearing sediment down to ~2.5 km below the ocean floor. *Science* **349**, 420-424 (2015).
30. L. Solden, K. Lloyd, K. Wrighton, The bright side of microbial dark matter: lessons learned from the uncultivated majority. *Curr Opin Microbiol* **31**, 217-226 (2016).

bioRxiv preprint doi: https://doi.org/10.1101/2020.12.19.423498; this version posted December 20, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

Figure legends

Figure 1: Isolated subsurface *Thalassospira* are most abundant at 3-7 mbsf and distinct from related type strains isolated from overlying water and sediments. (A) Vertical profile of total 16S rRNA gene concentrations determined via qPCR (squares), and oxygen concentrations (circles). 16S rRNA gene concentration points are the average abundances of three technical qPCR replicates with ranges shown with error bars. (B) qPCR-normalized average concentrations of the *Thalassospira* affiliated 'Otu6' ASVs. Error bars are ranges from three technical qPCR replicates. Asterisks mark the depths for the long term ^{18}O-water incubation experiments, enrichments, and cultivation. (C) Maximum likelihood (PhyML) phylogenetic analysis of the *Thalassospira* 16S rRNA gene ASVs (V4 hypervariable region), together with subseafloor and type strain *Thalassospira* 16S rRNA gene V4 regions. The presence of the SNP is displayed. Black, grey, and white dots at the nodes represent >90%, >70%, >50% bootstrap support, respectively.

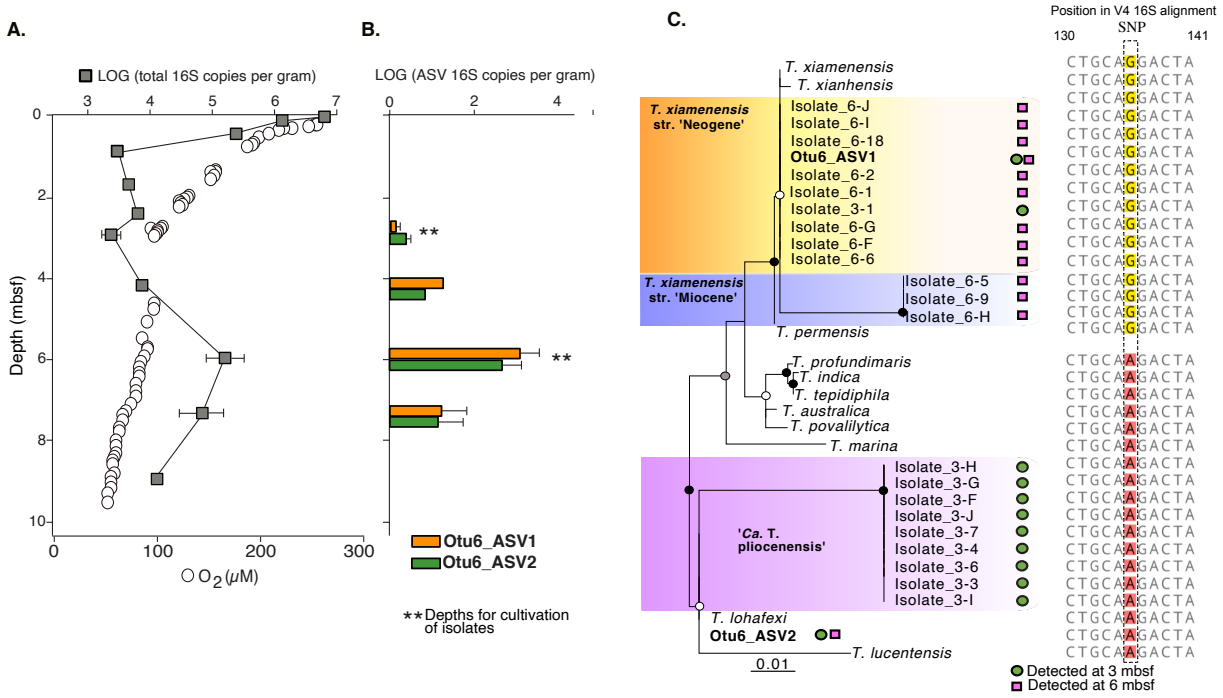
Figure 2: Recombination in the conserved core genome is limited in subseafloor *Thalassospira* populations. The maximum likelihood (PhyML) phylogenetic tree is based on a concatenated alignment of 1,809 genes conserved across all *Thalassospira* genomes ('core genes'). Black circles on nodes represent bootstrap values >95%. The position of recombination events in the core genome are represented by dark blue dots. Positions of low nucleotide diversity and no recombination events in the core genome are shown in light blue. Nucleotide diversity at specific sites in the core genome are illustrated with a color gradient (white: less diversity, orange: more diversity). Histograms on the right display the total number of recombination events (imports) in each genome sequence, and ancestral state reconstructions (internal nodes), as detected by ClonalFrameML (*19*).

Figure 3: The number of pseudogenes and dN/dS ratios are elevated and correlated in Subseafloor *Thalassospira* populations. Subseafloor genomes accumulate more pseudogenes as a function of increasing dN/dS ratios compared to the type strains. Linear regressions for type strains, subseafloor strains, and all strains are displayed.

Table 1. The contribution of recombination and mutation to nucleotide diversity in subseafloor populations. The results from ClonalFrameML (*19)* analysis used to calculate the relative contributions of recombination and mutation in the core genome (r/m). R/θ: the relative rate of recombination compared to mutation, δ: the average length of recombined (imported) DNA, v: mean divergence of imported DNA. Also displayed are the average number of pseudogenes and dN/dS ratios (+/- standard deviation). ** two sided T-test: P = 0.005. *****two sided T-test: P = 0.000001.

bioRxiv preprint doi: https://doi.org/10.1101/2020.12.19.423498; this version posted December 20, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

Figure 1

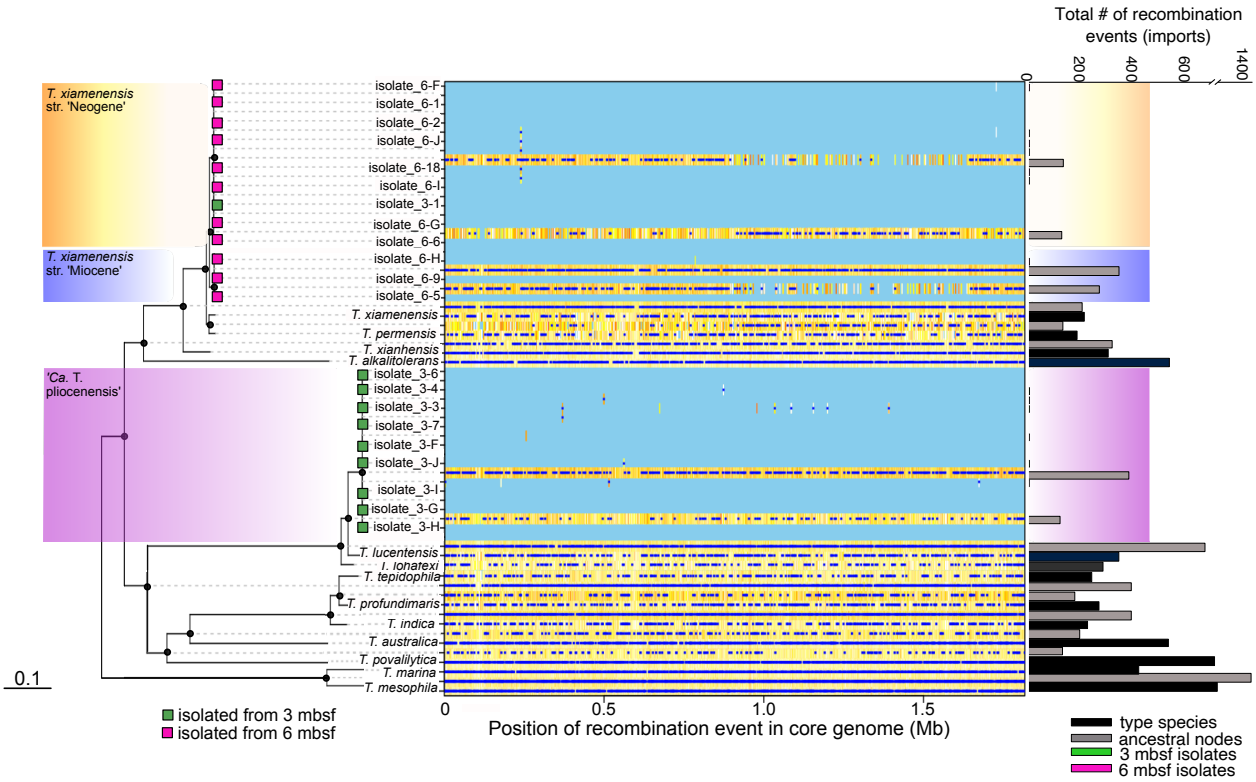


bioRxiv preprint doi: https://doi.org/10.1101/2020.12.19.423498; this version posted December 20, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

Table 1

Group	# strains	*R/θ*	*δ*	*v*	r/m	# pseudogenes	dN/dS
All subseafloor and type strains	34	0.053	244	0.055	0.71	37 (+/- 8)	0.025 (+/- 0.011)
Type strains	13	0.04	333	0.053	0.71	21 (+/- 3)	0.022 (+/- 0.01)
Subseafloor	21	0.006	500	0.026	0.078	48 (+/- 4)****	0.038 (+/- 0.007)**

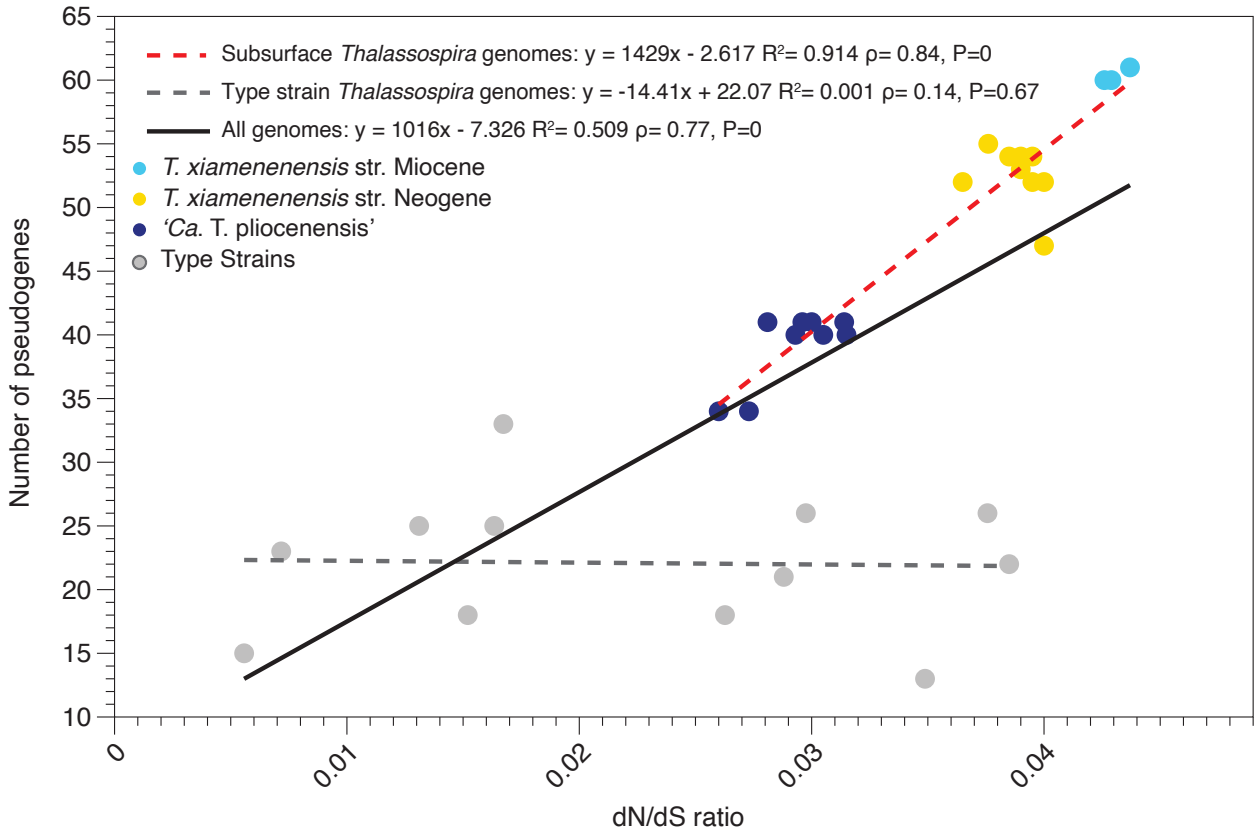
Figure 2



bioRxiv preprint doi: https://doi.org/10.1101/2020.12.19.423498; this version posted December 20, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

449

450 **Figure 3**



451

452

453

454

455

456

457

458

459

bioRxiv preprint doi: https://doi.org/10.1101/2020.12.19.423498; this version posted December 20, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

Supplemental Information

Genome evolution in bacteria isolated from million-year-old subseafloor sediments

William D. Orsi[1,2*], Tobias Magritsch[1], Sergio Vargas[1], Ömer K. Coskun[1], Aurele Vuillemin[1], Sebastian Höhna[1,2], Gert Wörheide[1,2,3], Steven D'Hondt[4], B. Jesse Shapiro[5,6,7], Paul Carini[8*]

[1]Department of Earth and Environmental Sciences, Paleontology & Geobiology, Ludwig-Maximilians-Universität München, Richard-Wagner-Strasse 10, 80333 Munich, Germany.

[2]GeoBio-CenterLMU, Ludwig-Maximilians-Universität München, Richard-Wagner-Strasse 10, 80333 Munich, Germany.

[3]SNSB- Bayerische Staatssammlung für Paläontologie und Geologie, Richard-Wagner-Strasse 10, 80333 Munich, Germany.

[4]Graduate School of Oceanography, University of Rhode Island, 215 South Ferry Road, 02882 Narragansett, USA.

[5]Department of Biological Sciences, University of Montreal, QC, Canada

[6]Department of Microbiology and Immunology, McGill University, QC, Canada

[7]McGill Genome Centre, Canada

[8]Department of Environmental Science, the BIO5 Institute, School of Plant Sciences, and the School of Comparative Animal & Biomedical Science, University of Arizona, Tucson, Arizona USA

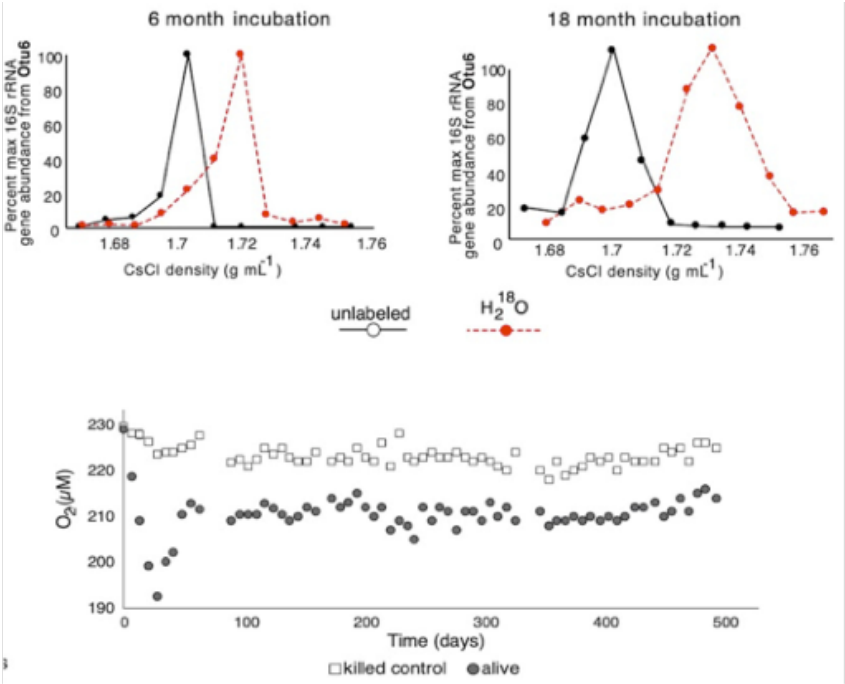
bioRxiv preprint doi: https://doi.org/10.1101/2020.12.19.423498; this version posted December 20, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

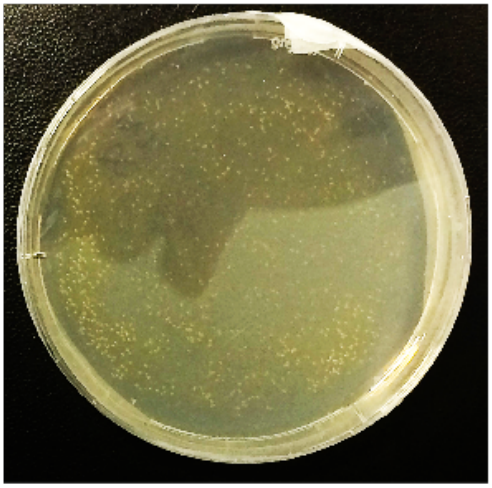
Group	isolation depth	isolate	N50 contig size (bp)	size (Mb)	# contigs	GC content	completeness (CheckM)	Contamination (CheckM)	predicted genome size (Mb)	# PEGs	Predicted #f PEGs	% core genome
Ca. T. pliocenensis'	3 mbsf	3.4	649,568	4.7	13	53.5	100%	0%	4.7	4463	4463	40.5
		3.7	1,392,694	4.73	13	53.5	100%	0%	4.73	4463	4463	40.5
		3.F	1,049,738	4.64	13	53.5	100%	0%	4.64	4463	4463	40.5
		3.G	1,232,970	4.64	13	53.5	100%	0%	4.64	4463	4463	40.5
		3.H	994,209	4.64	13	53.5	100%	0%	4.64	4463	4463	40.5
		3.I	735,219	4.64	11	53.5	100%	0%	4.64	4466	4466	40.5
		3.J	1,232,974	4.64	11	53.5	100%	0%	4.64	4466	4466	40.5
		3.3	556,527	4.64	13	53.5	100%	0%	4.64	4463	4463	40.5
		3.6	1,047,050	4.63	13	53.5	100%	0%	4.63	4463	4463	40.5
T. xiamenensis 'Neogene'	6 mbsf	6.1	823,074	4.74	10	54.8	99.50%	0%	4.76	4651	4674	38.9
		6.2	671,859	4.59	9	54.8	97%	0%	4.73	4481	4620	40.4
		6.6	1,045,808	4.74	12	54.8	100%	0%	4.74	4622	4622	39.1
		6.18	1,220,195	4.75	14	54.8	100%	0%	4.75	4651	4651	38.9
		6.F	1,372,017	4.74	9	54.8	99.50%	0%	4.76	4637	4660	39.0
		6.G	1,372,000	4.73	7	54.8	99.50%	0%	4.75	4636	4659	39.0
		6.I	1,156,923	4.75	11	54.8	100%	0%	4.75	4651	4651	38.9
		6.J	1,392,823	4.64	8	54.8	98%	0%	4.76	4555	4672	39.7
	3 mbsf	3.1	1,245,526	4.75	12	54.8	100%	0%	4.75	4673	4673	38.7
T. xiamenensis 'Miocene'	6 mbsf	6.5	484,976	4.87	16	54.7	100%	0%	4.87	4730	4730	38.2
		6.9	484,974	4.87	15	54.7	100%	0%	4.87	4728	4728	38.3
		6.H	484,664	4.87	16	54.6	100%	0%	4.87	4732	4732	38.2

Table S1. Genome summary statistics.

bioRxiv preprint doi: https://doi.org/10.1101/2020.12.19.423498; this version posted December 20, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

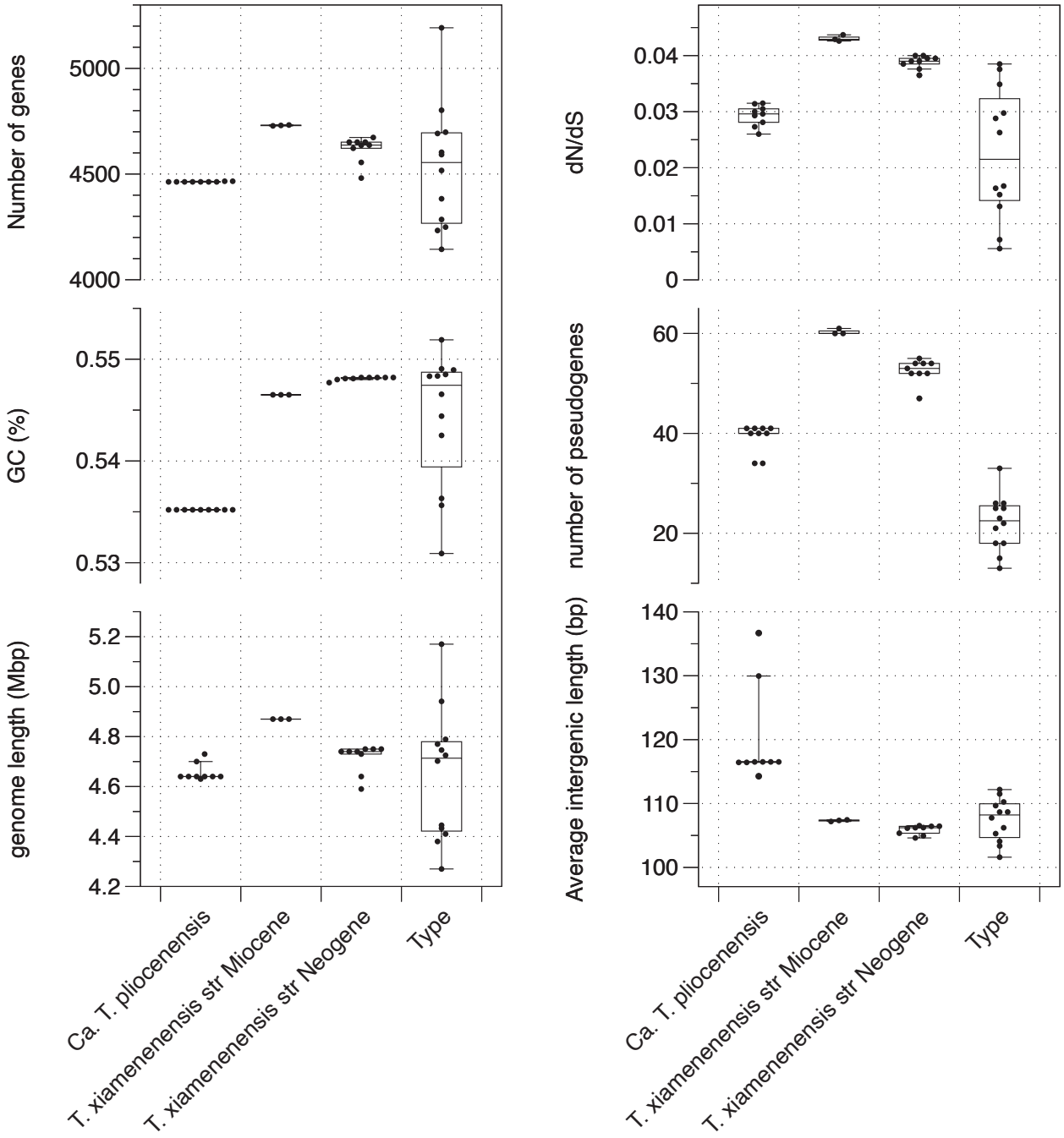
Figure S1. Top panel: ^{18}O-labeling of 16S rRNA genes from the *Thalassospira* OTU6 (see Figure 1), after 7 and 18 months of incubation with ^{18}O-labeled water from the 3 mbsf sediment (data originally published in Vuillemin et al., 2019). **Middle panel:** Oxygen consumption over time in the 18 month slurry from the 3 mbsf incubation (filled circles), and slurries containing labeled water and autoclaved sediment (killed control). **Bottom photo**: cultivation of colony forming bacteria on solid media after the 18 month incubation of sediment slurries in sterile ^{18}O-labeled artificial sea water. No bacterial colonies formed on petri dishes that were inoculated with the killed control slurries.





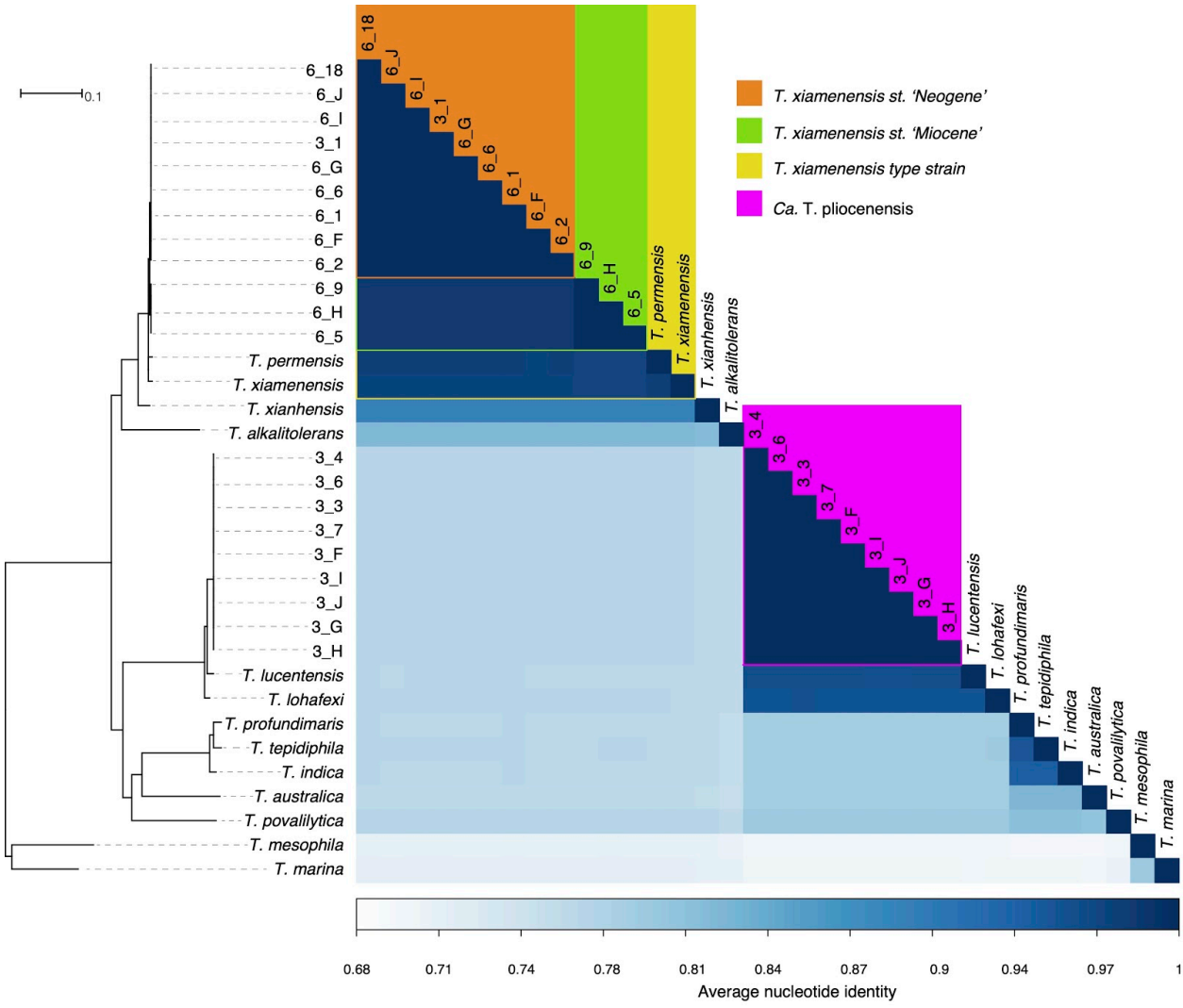
bioRxiv preprint doi: https://doi.org/10.1101/2020.12.19.423498; this version posted December 20, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

Figure S2: Summary of genome properties for *Thalassospira* strains used in this study. Points are derived from the analysis of existing genome sequences (for "Type" strains), and new high-quality draft genomes sequenced as part of this study.



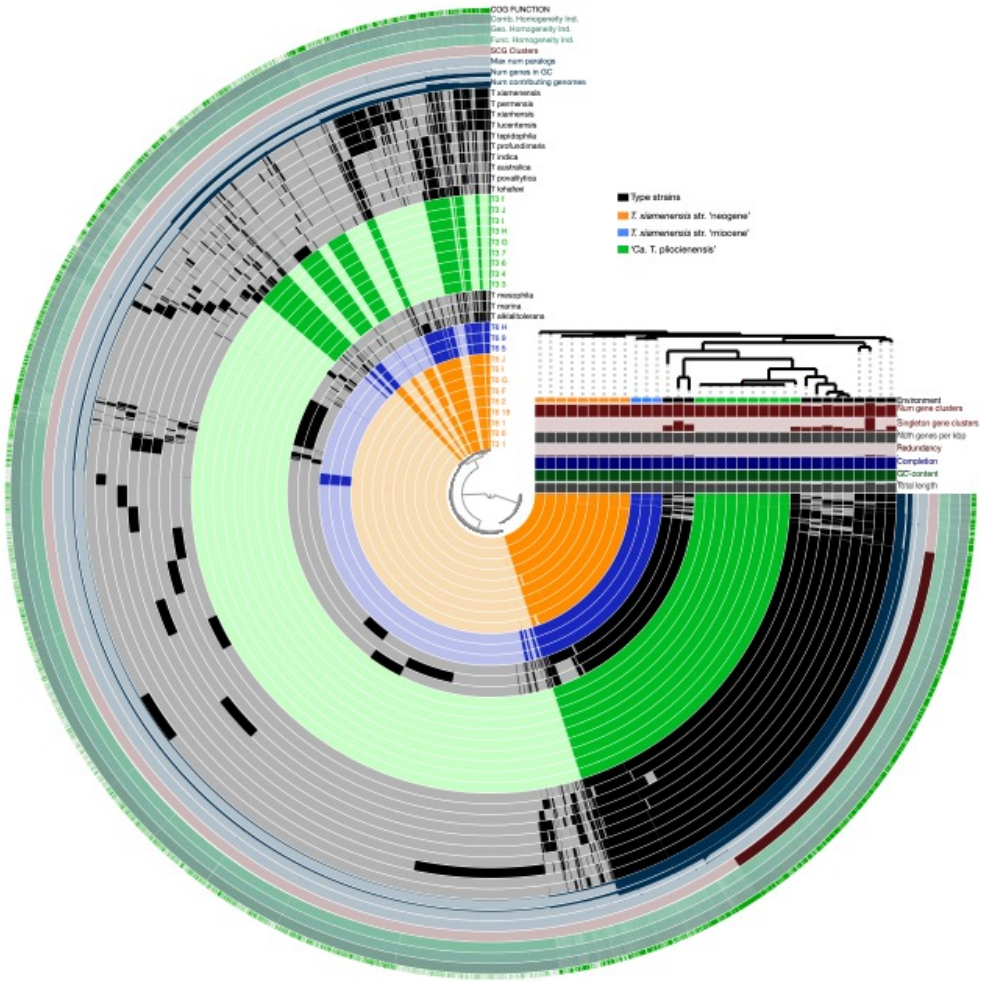
bioRxiv preprint doi: https://doi.org/10.1101/2020.12.19.423498; this version posted December 20, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

Figure S3. Average nucleotide identity (ANI) and the core genome phylogeny. The tree is based on maximum likelihood and a concatenated alignment of 1,809 core genes



bioRxiv preprint doi: https://doi.org/10.1101/2020.12.19.423498; this version posted December 20, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

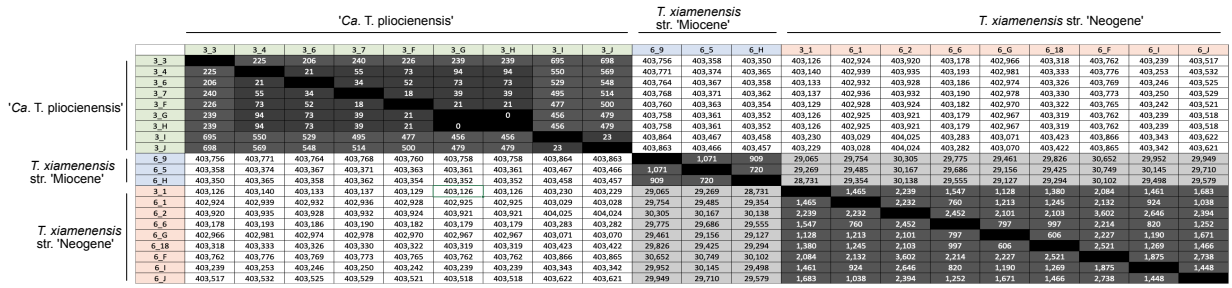
Figure S4: Pangenome analysis of all *Thalassospira* genomes included in the study. The internal dendrogram is a UPGMA based on the presence/absence of shared gene orthologs. Black bars in the first (inner) 34 circles show the occurrence of gene clusters in the genome of *Thalassospira* species. Grey areas and light colors in the circles represent gene clusters that were detected in the corresponding genome. The next eight bars show statistics for the pangenome analysis of each individual gene cluster (inner circle to outer circle) # contributing genomes: # of genomes that has a hit in a gene cluster, (GC), max # paralogs, single copy gene clusters (SCG), Functional Homogeneity index, Geometric homogeneity index, combined homogeneity index, presence of a COG functional assignment. The categories on the right side (below dendrogram) show the totals per genome for # of gene clusters found in each genome, Num genes per kbp: Number of genes per kilobase pairs of genome. Redundancy: Multiple occurrence of single copy genes in a genome, Completeness: Calculated from the occurrences of single copy gene set in a genome



bioRxiv preprint doi: https://doi.org/10.1101/2020.12.19.423498; this version posted December 20, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

568

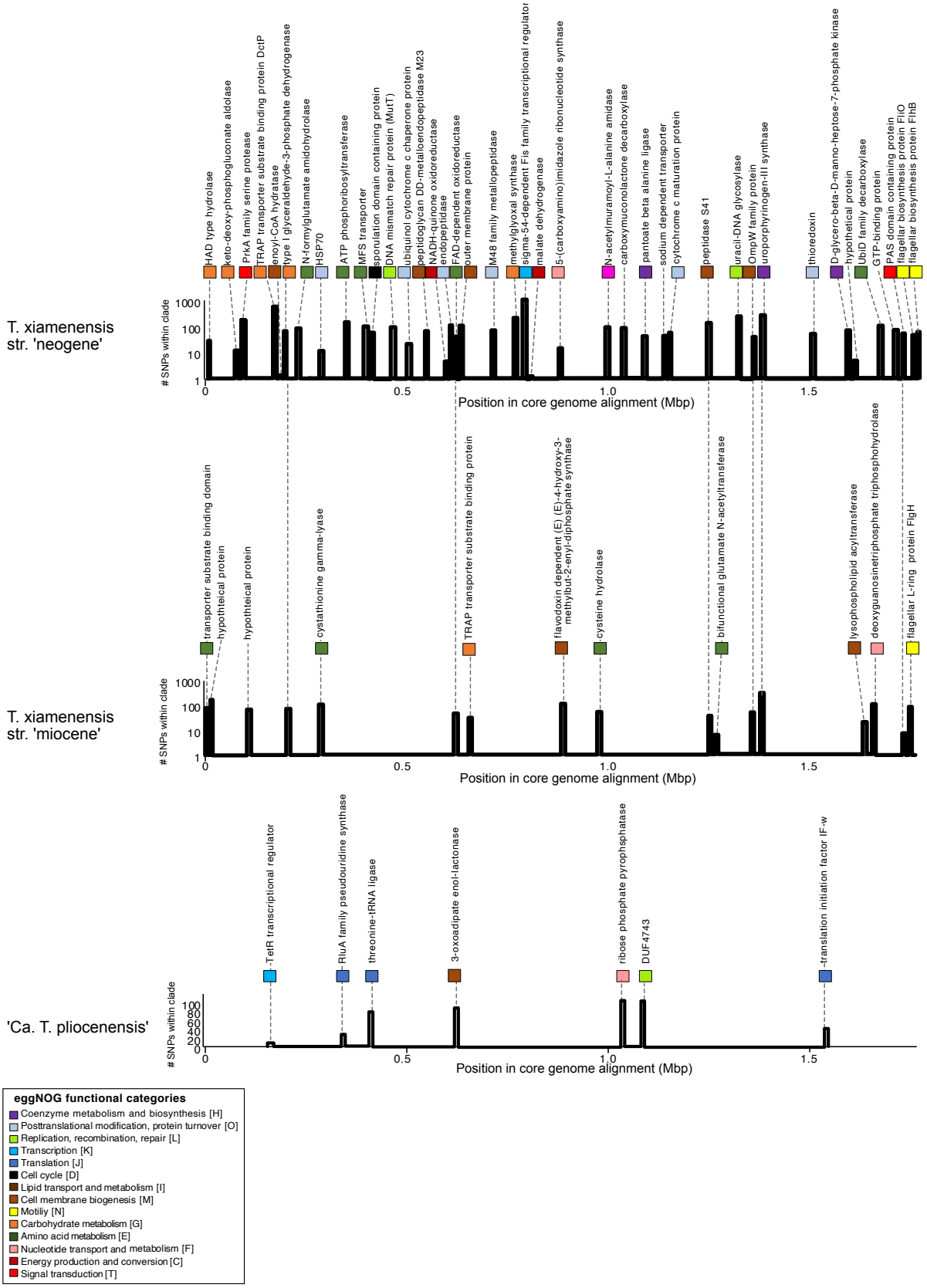
569 **Figure S5.** The number of SNPs between pairs of subseafloor *Thalassospira* genomes.



		'Ca. T. pliocienensis'									*T. xiamenensis* str. 'Miocene'			*T. xiamenensis* str. 'Neogene'								
		3_3	3_4	3_6	3_7	3_F	3_G	3_H	3_I	3_J	6_9	6_5	6_H	3_1	6_1	6_2	6_6	6_G	6_18	6_F	6_I	6_J
'Ca. T. pliocienensis'	3_3		225	206	240	226	239	239	695	698	403,756	403,358	403,350	403,126	402,924	403,920	403,178	402,966	403,318	403,762	403,239	403,517
	3_4	225		21	55	73	94	94	550	569	403,771	403,374	403,365	403,140	402,939	403,935	403,193	402,981	403,333	403,776	403,253	403,532
	3_6	206	21		34	52	73	73	529	548	403,764	403,367	403,358	403,133	402,932	403,928	403,186	402,974	403,326	403,769	403,246	403,525
	3_7	240	55	34		18	39	39	495	514	403,768	403,371	403,362	403,137	402,936	403,932	403,190	402,978	403,330	403,773	403,250	403,529
	3_F	226	73	52	18		21	21	477	500	403,760	403,363	403,354	403,129	402,928	403,924	403,182	402,970	403,322	403,765	403,242	403,521
	3_G	239	94	73	39	21		0	456	479	403,758	403,361	403,352	403,126	402,925	403,921	403,179	402,967	403,319	403,762	403,239	403,518
	3_H	239	94	73	39	21	0		456	479	403,758	403,361	403,352	403,126	402,925	403,921	403,179	402,967	403,319	403,762	403,239	403,518
	3_I	695	550	529	495	477	456	456		23	403,864	403,467	403,458	403,230	403,029	404,025	403,283	403,071	403,423	403,866	403,343	403,622
	3_J	698	569	548	514	500	479	479	23		403,863	403,466	403,457	403,229	403,028	404,024	403,282	403,070	403,422	403,865	403,342	403,621
T. xiamenensis str. 'Miocene'	6_9	403,756	403,771	403,764	403,768	403,760	403,758	403,758	403,864	403,863		1,071	909	29,065	29,754	30,305	29,775	29,461	29,826	30,652	29,952	29,949
	6_5	403,358	403,374	403,367	403,371	403,363	403,361	403,361	403,467	403,466	1,071		720	29,269	29,485	30,167	29,686	29,156	29,425	30,749	30,145	29,710
	6_H	403,350	403,365	403,358	403,362	403,354	403,352	403,352	403,458	403,457	909	720		28,731	29,354	30,138	29,555	29,127	29,294	30,102	29,498	29,579
T. xiamenensis str. 'Neogene'	3_1	403,126	403,140	403,133	403,137	403,129	403,126	403,126	403,230	403,229	29,065	29,269	28,731		1,465	2,239	1,547	1,128	1,380	2,084	1,461	1,683
	6_1	402,924	402,939	402,932	402,936	402,928	402,925	402,925	403,029	403,028	29,754	29,485	29,354	1,465		2,232	760	1,213	1,245	2,132	924	1,038
	6_2	403,920	403,935	403,928	403,932	403,924	403,921	403,921	404,025	404,024	30,305	30,167	30,138	2,239	2,232		2,452	2,101	2,103	3,602	2,646	2,394
	6_6	403,178	403,193	403,186	403,190	403,182	403,179	403,179	403,283	403,282	29,775	29,686	29,555	1,547	760	2,452		797	997	2,214	820	1,252
	6_G	402,966	402,981	402,974	402,978	402,970	402,967	402,967	403,071	403,070	29,461	29,156	29,127	1,128	1,213	2,101	797		606	2,227	1,190	1,671
	6_18	403,318	403,333	403,326	403,330	403,322	403,319	403,319	403,423	403,422	29,826	29,425	29,294	1,380	1,245	2,103	997	606		2,521	1,269	1,466
	6_F	403,762	403,776	403,769	403,773	403,765	403,762	403,762	403,866	403,865	30,652	30,749	30,102	2,084	2,132	3,602	2,214	2,227	2,521		1,875	2,738
	6_I	403,239	403,253	403,246	403,250	403,242	403,239	403,239	403,343	403,342	29,952	30,145	29,498	1,461	924	2,646	820	1,190	1,269	1,875		1,448
	6_J	403,517	403,532	403,525	403,529	403,521	403,518	403,518	403,622	403,621	29,949	29,710	29,579	1,683	1,038	2,394	1,252	1,671	1,466	2,738	1,448	

bioRxiv preprint doi: https://doi.org/10.1101/2020.12.19.423498; this version posted December 20, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

570 **Figure S6.** The number of interpopulation SNPs at different positions in the core genome
571 alignment, for each of the three subseafloor populations. The gene annotations to the
572 corresponding regions are shown.



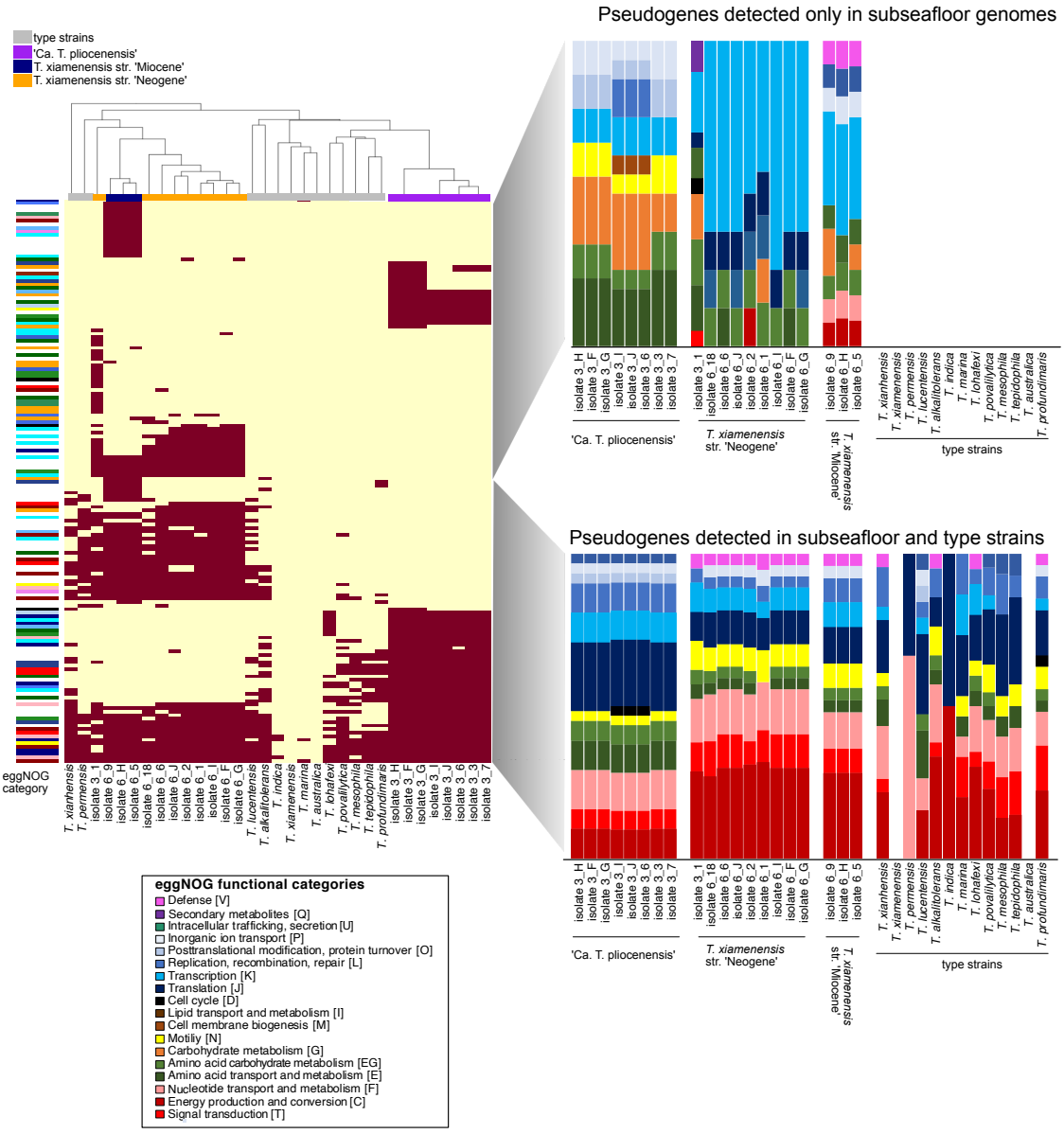
bioRxiv preprint doi: https://doi.org/10.1101/2020.12.19.423498; this version posted December 20, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

573 **Figure S7:** The heatmap shows the presence/absence of pseudogenes in the subseafloor genomes
574 and the presence of these pseudogenes in type strains. Functional annotation (against eggNOG)
575 of pseudogenes found in the subseafloor genomes only are compared to functional annotations of
576 pseudogenes found in both the subseafloor and type strains.

577

578

579

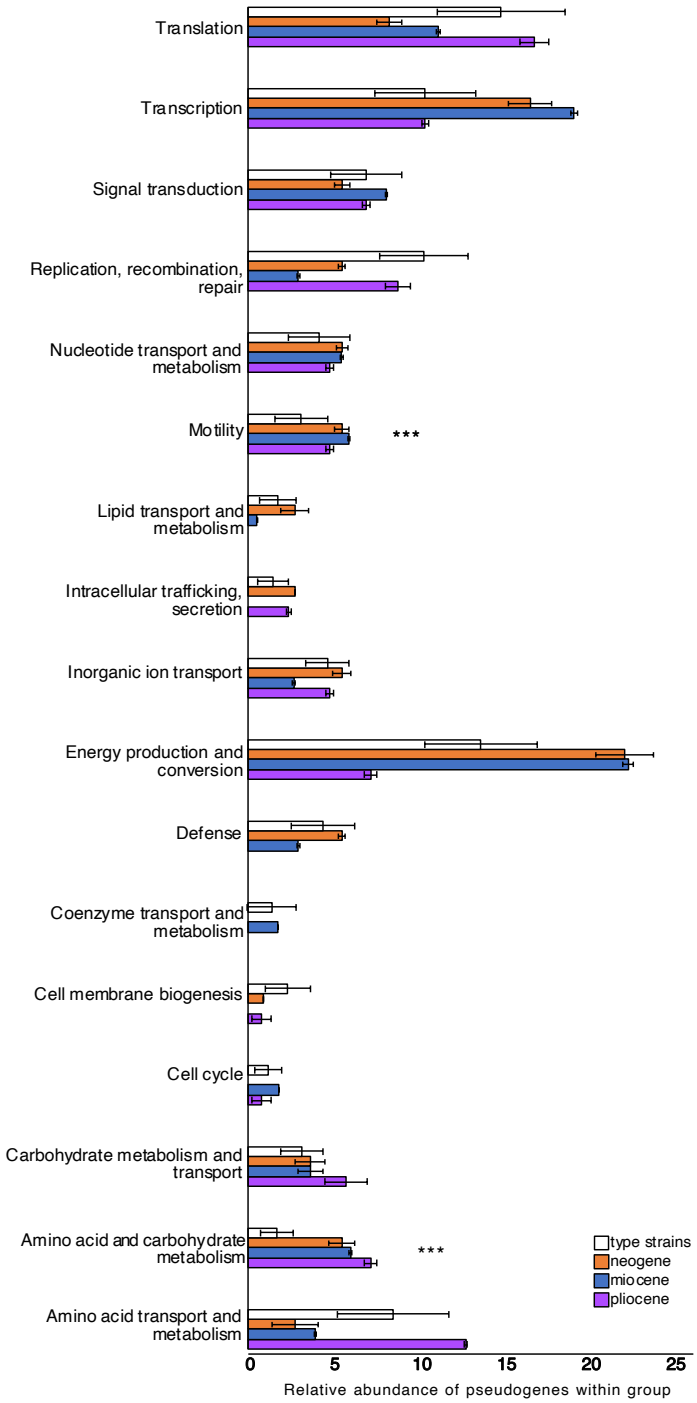


580

bioRxiv preprint doi: https://doi.org/10.1101/2020.12.19.423498; this version posted December 20, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

581 **Figure S8:** Histograms showing the average relative abundance of functional categories in
582 pseudogenes found within each of the three subseafloor populations, compared to the type
583 strains. The error bars represent standard deviations, and asterisks indicate functional categories
584 of pseudogenes that were significantly higher in the subseafloor genomes compared to the type
585 strain genomes (two sided T-Test, $P<0.05$).

586



bioRxiv preprint doi: https://doi.org/10.1101/2020.12.19.423498; this version posted December 20, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

Materials and Methods

Sampling, pore water chemistry, sediment properties. All samples were taken during Expedition KN223 of the *R/V Knorr* in the North Atlantic, from 26 October to 3 December 2014. At site 11 (22°47.0′ N, 56°31.0′ W, water depth ~5600 m) via a long core piston-coring device (~28 m). Additional details of sampling are published elsewhere (*11, 13*). Dissolved oxygen concentrations in the core sections were measured with optical O_2 sensors from the equilibrated core sections and measured with needle-shaped optical O_2 sensors (optodes) (PreSens, Regensburg, Germany) as described previously (*11, 13*). The dissolved O_2 data from Expedition KN223 are archived and available online in the Integrated Earth Data Applications (IEDA) database (http://www.iedadata.org/doi?id=100519).

Deep sea clay sediment particles have a grain size of <0.2 μm, and thus the pore space between the clay particles is smaller than that of a bacterial cell, limiting the movement of bacteria through pore space in the clays. Bioturbation can vertically redistribute cells within marine sediments, but bioturbation is restricted to the upper 0.5 meters of sediment (*31*), and thus cannot vertically transport sediment surface material and microbes to depths of 3 and 6 mbsf. Barring the possibility of vertical redistribution via bioturbation, and considering the mean sedimentation rate of 1 m per million years, it can be concluded that the bacterial cultures obtained from sediments collected at 3 and 6 mbsf have been physically isolated from the surface world for millions of years.

DNA extraction, qPCR, 16S rRNA gene sequencing. DNA extractions, qPCR, and 16S rRNA gene sequencing were performed previously and described in Vuillemin et al (*12*). In brief, subcores were sampled aseptically with sterile syringes were subsampled aseptically in an ultraviolet (UV)–sterilized DNA/RNA clean HEPA- filtered laminar flow hood. DNA extraction was extracted from 10 g of sediment transferred into 50 ml of Lysing Matrix E tubes (MP Biomedicals) containing silica glass beads and homogenized for 40 s at 6 m/s using a FastPrep-24 5G homogenizer (MP Biomedicals) in the presence of 15 ml of preheated (65°C) sterile- filtered extraction buffer [76 volume % 1 M NaPO4 (pH 8), 15 vol- ume % 200 proof ethanol, 8 volume % MoBio's lysis buffer solution C1, and 1 volume % SDS]. The samples were incubated at 99°C for 2 min and frozen overnight at −20°C, thawed, and frozen again at −20°C overnight, followed by additional incubation at 99°C for 2 min and a second homogenization using the settings described above. After the second homogenization, the samples were centrifuged for 15 min, and the supernatants were concentrated to a volume of 100 ml using 50-kDa Amicon centrifugal filters (Millipore). Coextracted PCR-inhibiting humic acids and other compounds

bioRxiv preprint doi: https://doi.org/10.1101/2020.12.19.423498; this version posted December 20, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

were removed from the concentrated extract using the PowerClean Pro DNA Cleanup Kit (MoBio). Extraction blanks were performed alongside the samples to assess laboratory contamination during the extraction process.

DNA was quantified fluorometrically using a Qubit with a double- stranded DNA high-sensitivity kit (Life Technologies). qPCR was performed using the custom primer dual indexed approach that targets the V4 hypervariable region of the 16S rRNA gene using updated 16S rRNA gene primers 515F/806R (515F, 5′-GTGYCAG- CMGCCGCGGTAA-3′; 806R, GGACTACNVGGGTWTCTAAT) (*32*). Barcoded V4 hypervariable regions of amplified 16S rRNA genes were sequenced on an Illumina MiniSeq following an established protocol (*33*). Bioinformatic processing of these previously published sequence data is described by Vuillemin et al (*12*) in detail.

Long term incubation set up. Prior to setting up the incubations, the subcores were sampled with sterile syringes using the sample aseptic technique used for the DNA extraction. For each sample depth, seven grams of abyssal clay was placed into sterile 20-mL glass flasks and incubated with 4 mL of sterile artificial seawater composed of either $H_2{}^{18}O$ (97% atomic enrichment) or unlabeled artificial seawater. Vials were crimp sealed, with an oxygenated headspace of approximately 10 mL, and incubated at 8 °C. The artificial seawater was different from the porewater at depth because there was no added nitrate, but there was also no added ammonia which should be similar to the *in situ* conditions where ammonia is generally below detection (*12*). Oxygen was measured continuously throughout the incubations using non-invasive fiberoptic measurements as described previously (*12*). Small fluctuations in the oxygen measurements in the killed control, and experimental incubations, were likely due to temperature fluctuations of the incubator itself (±1˚C), since the non-invasive fiber optic oxygen sensor spots are temperature sensitive (*12*). Oxygen consumption was detectable over 18 months in slurries consisting of sediment and sterile artificial seawater (Fig S1), suggesting the presence of actively respiring microbes.

We used qSIP to measure the atom % ^{18}O-enrichment of actively growing microbial taxa as described previously (*12*). In brief, after 7 and 18 months incubations DNA was extracted and subjected to Cesium Chloride (CsCl) density gradient centrifugation. The same 16S 515F/806R primers (described above) were used in qPCR (described above) to determine density shifts in the peak DNA of buoyant density (BD) for each incubation. 16S rRNA gene amplicons from each fraction resulting from the density gradient fractionation were Illumina sequenced as described previously (*12*). To identify contaminants that may have entered during the

bioRxiv preprint doi: https://doi.org/10.1101/2020.12.19.423498; this version posted December 20, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

fractionation process, we also included in the sequencing run extraction blanks from the SIP fractionation. OTUs containing sequences from extraction blanks were removed. Excess atm% ^{18}O-enrichment was calculated for each OTU (including OTU6, corresponding to the subseafloor *Thalassospira*) according to the equations for quantifying per OTU atomic enrichment.

The number of doublings for the *Thalassospira* OTU (OTU_6) detected at the 18 month timepoint was calculated using qPCR normalized relative abundance of the 16S rRNA genes at T0 and 18 months. The number of doublings was divided by the total number of days incubated to calculate doubling times in days.

Enrichments, cultivation, and sub-cultivation. After the 18 months of incubation in sterile ^{18}O-labeled artificial seawater, 25 µL of slurry was plated onto solid media (10 mg/mL yeast extract and 8 mg/mL agar in artificial seawater), and after 2 days incubated in the dark at room temperature, abundant colonies were observed growing on the surface of the petri dishes (Fig S1). No colonies were observed to grow on control petri dishes that received 25 µL of ^{18}O-labeled artificial seawater slurry incubated for 18 months using starting material from autoclaved sediment (killed controls). This indicated that the colony forming bacteria were from the sediments themselves and not due to contamination that was introduced during the experimental set up of the incubations. We attempted to culture chemoheterotrophic microbes directly from the collected sediment samples using the same conditions, but no colony forming units were observed on the petri dishes even after several months of incubation. Thus, a long term incubation of the sediments at 8 °C simply in the presence of added water apparently stimulated the activity of many subseafloor bacteria to a point at which they were then able to grow on the surface of a petri dish.

Ten colonies were picked from petri dishes containing colonies from the 3 and 6 mbsf slurries. These colonies were streaked onto new petri dishes, and a single colony was picked from these newly streaked bacteria and grown in sterile liquid media (10 mg/mL yeast extract in artificial seawater). Single colonies were then grown up in liquid media, a portion used for DNA extraction and genome sequencing, and the remaining volume frozen as glycerol stocks.

Assessing the possibility for genome evolution during the 18 month enrichment.

Since bacteria can evolve on lab experimental timescales (*5,6*), we considered the possibility that all diversification and evolution happened during the 1.5 year enrichment. Using the qPCR-based estimate for doubling time of the subseafloor Thalassospira OTU (OTU_6) in the incubation which was 36 (± 1.5)

bioRxiv preprint doi: https://doi.org/10.1101/2020.12.19.423498; this version posted December 20, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

683 days, the number of doublings with this rate over this time period would be approximately 15. According
684 to "Drake's rule" (*20*), bacteria experience on average one mutation per 300 genomes replicated, and thus
685 the amount of nucleotide diversity (hundreds to thousands of mutations : Fig S5) that could be accumulated
686 during the incubation is insufficient to explain the observed diversity between the three subseafloor
687 populations. We thus conclude that the inter-population nucleotide diversity is the result of new mutations
688 that were acquired after they were buried.
689
690 **Genome sequencing, de novo assembly, and annotation.** DNA was extracted from the isolates grown in
691 liquid culture until the end of exponential phase as described above. After reaching stationary phase,
692 cultures were pelleted via centrifugation and the supernatant was decanted. The cell pellets were
693 resuspended in a preheated (65°C) sterile filtered extraction buffer [76 volume % 1 M NaPO4 (pH 8), 15
694 volume %200 proof ethanol, 8 volume %MoBio lysis buffer solution C1,and 1 volume % SDS], and added
695 to lysing matrix E tubes (MP Biomedicals) containing silica glass beads and homogenized for 40 s at 6 m/s
696 using a FastPrep-24 5G homogenizer (MPBiomedicals). The samples were centrifuged for 15 min, and the
697 dissolved high molecular weight DNA in the supernatant was concentrated to a volume of 100 µL using
698 50-kDa Amicon centrifugal filters (Millipore). The concentrated extract was cleaned of proteins and other
699 non-genomic DNA organic matter using the PowerClean Pro DNA Cleanup Kit (MoBio). Extraction blanks
700 were performed alongside the samples to assess laboratory contamination during the extraction process.
701 Genomic libraries were prepared using the Nextera XT DNA Library Prep Kit (Illumina). Quality control
702 and quantification of the libraries were obtained on an Agilent 2100 Bioanalyzer System using the High
703 Sensitivity DNA reagents and DNA chips (Agilent Genomics). Metagenomic libraries were diluted to 1
704 nM using the Select-a-Size DNA Clean and Concentrator MagBead Kit (Zymo Research) and pooled for
705 further sequencing on the Illumina MiniSeq platform. Genomic libraries were sequenced to a depth of ca.
706 100x coverage using a high-output paired end 2 x 150 sequencing regent kit (Illumina).
707 In addition to Illumina sequencing, the high molecular weight genomic DNA was sequenced using the
708 NanoPore MinION. Sequencing libraries for the MinION were prepared using the Ligation Sequencing kit
709 (Oxford NanoPore Technologies), according to the manufacturers instructions. Barcoded libraries were
710 sequenced on the MinION using a Flongle R9 flow cell, base-called and demultiplexed using the MinIT
711 (Oxford NanoPore Technologies).
712 A hybrid assembly was performed using both the short (Illumina) and long (NanoPore) read
713 sequencing data using Unicycler (v.0.4.0), which uses *de novo* assembled Illumina data from SPADES to
714 polish the *de novo* assembled contigs obtained from NanoPore data using RACON (*34*). The combined
715 assemblies of Illumina and NanoPore data resulted in a relative low number of contigs (9-12 per genome),

bioRxiv preprint doi: https://doi.org/10.1101/2020.12.19.423498; this version posted December 20, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

and a predicted genome completeness of 100% of nearly all genomes (Table S1). Genome completeness was determined using CheckM (*35*). Genomes were annotated using RASTk (*36*).

Core genome phylogenetic analyses. The core genome was defined as the set of orthologous genes which were shared in all subseafloor and extant *Thalassospira* genomes. Orthologous genes were defined as those sharing >30% amino acid similarity to the collective suite of genes encoded within the type strain *Thalassospira xiamenensis* M-5. *T. xiamenensis* M-5 was chosen as the reference genome for this purpose, because it is the only publicly available genome of a cultivated *Thalassospira* isolate that is completely closed and represents a single chromosome and a 190 Kb plasmid (*14*). A total of 1,809 orthologous genes were identified that are encoded by all *Thalassospira* strains that had >30% sequence similarity to genes encoded within the *T. xiamenensis* M-5 genome. Each of these 1,809 genes were individually aligned between all *Thalassospira* strains using MUSCLE (*37*), and the individual 1,809 alignments were then concatenated into a single core genome alignment for the subsequent phylogenomic analysis (ClonalFrameML, HyPhy, aBRSEL) using Geneious Prime (version 2019.2.1). After concatenation of all core genes, the total size of the core genome alignment was 1,817,073 nucleotide characters, and 34 taxa (21 subseafloor strains, and 13 type strain taxa). A Maximum-Likelihood phylogeny was created using PhyML (*38*) with a GTR model of evolution and 100 bootstrap replicates, which was implemented within SeaView (*39*). The resulting phylogenetic tree and the concatenated core genome alignment were used as inputs for subsequent ClonalFrameML and dN/dS analyses.

The contributions of mutations and recombination to the genomic diversity in the concatenated core genome alignment, the number of recombination events (imports) per genome, and the positions of recombination hot spots, was investigated using ClonalFrameML (*19*). Nucleotides unaffected by recombination are referred to as unimported and nucleotides subject to recombination are referred to as imported (*19*). ClonalFrameML provides the relative rate of recombination to mutation (R/Theta), the mean length of recombined DNA (Delta), and the mean divergence of imported DNA (Nu). These results were used to calculate the relative contribution of recombination versus mutation to the overall genomic diversity (r/m), using the following formula r/m = (R/Theta) * Delta * Nu. ClonalFrameML was performed in three separate runs containing a core genome alignment that contained (1) all genomes, (2) only the subseafloor genomes, and (3) only the type strains. The resulting r/m values from these three groups (presented in Table 1) were then used to interpret the relative importance of mutations compared to recombination, in the separate groups (e.g., type strains versus subseafloor strains). We acknowledge that because the dataset contains genomes covering the diversity of a single bacterial genus (*Thalassospira*), the only detectable recombination most events are from donors from the species under study, so that the main source of recombination is not external (*19*).

bioRxiv preprint doi: https://doi.org/10.1101/2020.12.19.423498; this version posted December 20, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

In addition to calculating site and rates of recombination in the core genome, ClonalFrameML also estimates the ancestral sequences at internal nodes of the clonal genealogy, and any missing base calls in the observed sequences. The reconstruction of ancestral sequence states is performed using maximum likelihood and the ClonalFrame model can be thought of as a hidden Markov model (HMM) when the ancestral and descendant genomes for each branch of the clonal genealogy have been observed or reconstructed (*19*). The hidden state of the HMM records whether each nucleotide was subject to recombination or not on the branch connecting the two genomes. We acknowledge that drawing inference under the resulting ancestral recombination graph is a notoriously complex statistical problem (*19*). Instead, here we use ClonalFrameML only to assess within-group recombination (e.g., between species within the genus *Thalassospira*), and thus our analysis cannot assess the influence of external recombination (from species outside the genus *Thalassospira*).

The ratio of non-synonymous (dN) to synonymous (dS) mutations in the core genome alignment (global ω ratio) was estimated using HyPhy v2.2.4 (*40*), and applying the adaptive branch-site random effects likelihood (aBSREL) approach (*41*) to all branches in all subfamilies. Because of the high similarity of the subseafloor genomes, aBSREL was run multiple times using the core genome alignment with only one representative of the nearly identical subseafloor genomes included in each separate run. For this, one representative genome of *T. xiamenensis* strain 'Neogene', *T xiamenensis* strain 'Miocene, and 'Ca. T. pliocenensis' were included together with all other *Thalassospira* type strain genomes in each aBSREL run. Then, the aBSREL run was repeated with the same type strains but different subseafloor genomes from those same three clades, until dN/dS estimates for all subseafloor genomes were obtained.

Pangenome analysis. All subseafloor and extant *Thalassospira* genomes were analyzed in Anvi'o v6.2 using pangenome workflow (*42*). Briefly, each genome was converted into an anvi'o contigs database. Genes were functionally annotated using eggnog v5.0 (*43*) with eggNOG-mapper (*44*) and imported back to each genome's anvi'o contig database. Genome storages were generated using 'anvi-gen-genomes-storages' and 'anvi-pan-genome' was deployed with parameters '--min-bit 0.5' (*45*), '--mcl-inflation' 10 (*46*), and the flag '--use-ncbi-blast' (*47*). The anvi'o pan database and summary of gene clusters stored in FIGshare (https://figshare.com/s/06ba1287a00ab01a1ee).

Identifying pseudogenes. We estimated the number of pseudogenes within the genomes using two programs, Psi-Phi (*48*) and DFAST (*49*). Psi-Phi uses a conservative criterion considering a pseudogene only when it lost >20% of its original length, and enhances pseudogene recognition among closely related

bioRxiv preprint doi: https://doi.org/10.1101/2020.12.19.423498; this version posted December 20, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

strains both in annotated regions by identifying incorrectly annotated open reading frames (ORFs) and in intergenic regions by detecting new pseudogenes (*48*). Psi-Phi classifies pseudogenes as either identified pseudogenes and those as being possible, but potentially not pseudogenes. To be conservative, we only considered genes identified as pseudogenes from Psi-Phi and did not consider those flagged as 'potential pseudogenes'. As a second check of pseudogene content, we searched genomes for pseudogenes using DFAST (*49*). The estimated number of pseudogenes per genome was then taken as an average of the numbers detected both using both methods (Psi-Phi and DFAST). On average, Psi-Phi identified a higher number of pseudogenes per genome (57 ±10) compared to DFAST (32 ±4), but the variation between methods for the same genome was consistent (average variation =27, standard deviation of averages = 7). This minimal variation between individual genomes indicates that biases inherent to the pseudogene prediction methods affected the different genomes equally, and thus allow for a pseudogene comparison between the genomes.

Supplemental references

31. L. R. Teal, M. T. Bulling, E. R. Parker, M. Solan, Global patterns of bioturbation intensity and mixed depth of marine soft sediments. *Aquatic Biology* **2**, 207-218 (2008)
32. A. E. Parada, D. M. Needham, J. A. Fuhrman, Every base matters: assessing small subunit rRNA primers for marine microbiomes with mock communities, time series and global field samples. *Environ Microbiol* **18**, 1403-1414 (2016).
33. M. Pichler *et al.*, A 16S rRNA gene sequencing and analysis protocol for the Illumina MiniSeq platform. *Microbiologyopen*, e00611 (2018).
34. R. R. Wick, L. M. Judd, C. L. Gorrie, K. E. Holt, Unicycler: Resolving bacterial genome assemblies from short and long sequencing reads. *PLoS Comput Biol* **13**, e1005595 (2017).
35. D. H. Parks, M. Imelfort, C. T. Skennerton, P. Hugenholtz, G. W. Tyson, CheckM: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. *Genome Res* **25**, 1043-1055 (2015).
36. R. Overbeek *et al.*, The SEED and the Rapid Annotation of microbial genomes using Subsystems Technology (RAST). *Nucleic Acids Res* **42**, D206-214 (2014).
37. R. C. Edgar, MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nulceic Acids Res* **32**, 1792-1797 (2004).
38. S. Guindon, F. Lethiec, P. Duroux, O. Gascuel, PHYML Online--a web server for fast maximum likelihood-based phylogenetic inference. *Nucleic Acids Res* **33**, W557-559 (2005).

bioRxiv preprint doi: https://doi.org/10.1101/2020.12.19.423498; this version posted December 20, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

39. M. Gouy, S. Guindon, O. Gascuel, SeaView version 4: A multiplatform graphical user interface for sequence alignment and phylogenetic tree building. *Mol Biol Evol* **27**, 221-224.
40. S. L. Pond, S. D. Frost, S. V. Muse, HyPhy: hypothesis testing using phylogenies. *Bioinformatics* **21**, 676-679 (2005).
41. M. D. Smith *et al.*, Less is more: an adaptive branch-site random effects model for efficient detection of episodic diversifying selection. *Mol Biol Evol* **32**, 1342-1353 (2015).
42. A. M. Eren *et a*l., Anvi'o: an advanced analysis and visualization platform for 'omics data. *PeerJ* **3**, e1319 (2015).
43. J. Huerta-Cepas *et a*l., eggNOG 5.0: a hierarchical, functionally and phylogenetically annotated orthology resource based on 5090 organisms and 2502 viruses. *Nucleic Acids Res* **47**, D209-D314 (2018).
44. J. Huerta-Cepas *et a*l., Fast Genome-Wide Functional Annotation through Orthology Assignment by eggNOG-Mapper. *Mol Biol Evol* **34**, 2115-2122 (2017).
45. M. N. Benedict et al., ITEP: and integrated toolkit for exploration of microbial pan-genomes. *BMC Genomics* **15**, 12.1186/1471-2164-15-8 (2014).
46. S. van Dongen, C. Abreu-Goodger, Using MCL to extract clusters from networks. In "Bacterial Molecular Networks: Methods and Protocols" Eds. J. van Helden, A. Toussaint, D. Thieffry. Springer New York, New York, NY. Pages 281-295.
47. S. F. Altschul *et al.,* Basic local alignment search tool*. J Mol Biol* **215**, 403-410 (1990).
48. E. Lerat, H. Ochman, Psi-Phi: exploring the outer limits of bacterial pseudogenes. *Genome Res* **14**, 2273-2278 (2004).
49. Y. Tanizawa, T. Fujisawa, Y. Nakamura, DFAST: a flexible prokaryotic genome annotation pipeline for faster genome publication. *Bioinformatics* **34**, 1037-1039 (2018).