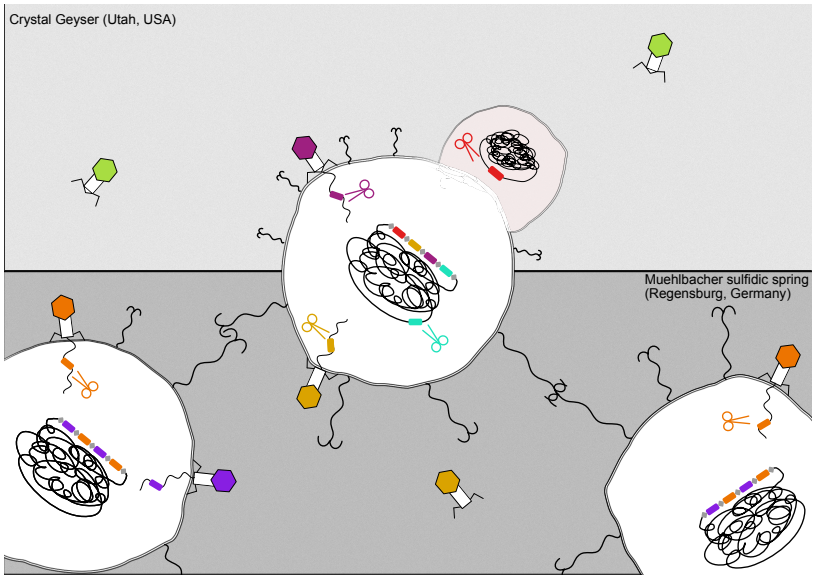


CRISPR-mediated interference and transcriptomic complexity of Altiarchaea interacting with symbiotic archaea in the deep biosphere

DISSERTATION

Zur Erlangung des Doktorgrades der Naturwissenschaften (Dr. rer. nat.)

der Fakultät Chemie der

UNIVERSITÄT DUISBURG ESSEN

Offen im Denken

Universität Duisburg-Essen

vorgelegt von

Sarah Eßer

2023

Title image: Schematic image of *Candidatus* Altiarchaea (*Ca.*) clustered regularly interspaced short palindromic repeat (CRISPR) interference within Crystal Geyser (Utah, USA; top) and Muehlbacher sulfidic spring (Regensburg, Germany). Different CRISPR interferences are indicated by colored boxes, scissors, and respectively colored targets. Top: *Ca.* Altiarchaeum crystalense targeted by viruses and attached to its symbiont *Ca.* Huberiarchaeum crystalense. Bottom: Ca. Altiarchaeum hamiconexum interconnected biofilm cells heavily targeted by viruses. © Sarah Eßer

Das Promotionsgesuch wurde eingereicht am:		06. Juni 2023
Disputation erforlgte am:		23. Januar 2024
Diese Arbeit wurde angeleitet von:		Prof. Dr. Alexander Probst
Prüfungsausschuss:	Vorsitzender:	Prof. Dr. Sebastian Schlücker
1.	Gutachter:	Prof. Dr. Alexander Probst
2.	Gutachterin:	Prof. Dr. Bettina Siebers
3.	Gutachter:	Prof. Dr. Folker Meyer

_____

Unterschrift: Sarah Eßer

CRISPR-mediated interference and transcriptomic complexity of Altiarchaea interacting with symbiotic archaea in the deep biosphere

DISSERTATION

For achieving the doctoral degree of natural sciences (Dr. rer. nat.)

At the faculty of Chemistry at the University of Duisburg-Essen

by

Sarah Eßer

2023

Keywords: deep subsurface, aquatic ecosystems, DPANN archaea, archaeal symbiosis, CRISPR systems

DOI: 10.17185/duepublico/81937
URN: urn:nbn:de:hbz:465-20240617-130249-2

The dissertation was performed under the auspices of

Prof. Dr. Alexander J. Probst (PhD supervisor)

at the faculty of Chemistry under fulfillment of all guidelines according to the

Promotionsordnung of the faculty

“Eine wissenschaftliche Entdeckung ist nie die Arbeit von nur einer Person.”

Louis Pasteur (1822-1895)

“In den Wissenschaften ist viel Gewiss, sobald man sich von den Ausnahmen nicht irre machen lässt und die Probleme zu ehren weiß.”

Johann Wolfgang von Goethe (1749 – 1832)

“Was dem einzelnen nicht möglich ist, das schaffen viele.”

Friedrich Wilhelm Raiffeisen (1818 – 1888)

Acknowledgements

First and foremost, I would like to thank Prof. Dr. Alexander J. Probst for giving me the opportunity to take over this interesting part of the project and the support during the time of the thesis. Thank you for believing in my person, knowledge, and strength. In addition, I would like to thank the DFG (Deutsche Forschungsgemeinschaft) for funding this project.

Second, I would like to thank the GEMs, or I would rather say good friends, that always supported me in good times and hard times. Thank you for all the gossip, the helpful discussions, the tipsy evenings, and the laughter, which I will always remember. I hope we all stay in contact or really open the dreamt of company to be our own bosses. Especially, I would like to thank Sabrina, Abi, Indra, and Sophie for being such good friends, for helping me stay afloat when the workload seemed to overwhelm me, for mental support whenever needed, and for the coffee, wine, and gin. Thank you, Ines, and Indra, for the office support and the tissues whenever they were needed, as well as for the open ear and the amazing “out of context” laughter. Also, I would like to thank the rest of the GEMs for great parties, game nights, funny conferences, and an awesome time.

I would like to say a huge **THANKS** to my family and friends. Thank you for accepting my weirdness, my ups and downs and for always giving me the strength that I needed to reach my goal. A special thank you to my mum who made it possible for me to study whatever I wanted and for always having an open ear for me. Dominik, thank you that you were my anchor when I came home and for giving me the freedom to reach my aim. An enormous thank you to everyone who helped me correcting my thesis.

And just to express how thankful I am and because I might have not mentioned it often enough:

THANK YOU

List of publications

In the following paragraph all publications as first author, co-first author, and co-author are listed which are either in submission, review or published. At the current date (06. June 2023) these include two first-author papers and eight co-author papers.

First author paper:

1. **S. P. Esser**, J. Rahlff, W. Zhao, M. Predl, J. Plewka, K. Sures, F. Wimmer, J. Lee, P. S. Adam, J. McGonigle, V. Turzynski, I. Banas, K. Schwank, M. Krupovic, T. L. V. Bornemann, P. A. Figueroa-Gonzalez, J. Jarett, T. Rattei, Y. Amano, I. K. Blaby, J.-F. Cheng, W. J. Brazelton, C. L. Beisel, T. Woyke, Y. Zhang, A. J. Probst, A CRISPR-mediated symbiosis of uncultivated archaea predicted from sequencing data. *In Press at Nature Microbiology* (2023).
   (Now available as: **S. P. Esser**, J. Rahlff, W. Zhao, M. Predl, J. Plewka, K. Sures, F. Wimmer, J. Lee, P. S. Adam, J. McGonigle, V. Turzynski, I. Banas, K. Schwank, M. Krupovic, T. L. V. Bornemann, P. A. Figueroa-Gonzalez, J. Jarett, T. Rattei, Y. Amano, I. K. Blaby, J.-F. Cheng, W. J. Brazelton, C. L. Beisel, T. Woyke, Y. Zhang, A. J. Probst, A predicted CRISPR-mediated symbiosis between uncultivated archaea. *Nature Microbiology* **8**, 1619–1633 (2023). Date modified: 11th March 2024)

2. **S. P. Esser**, V. Turzynski, J. Plewka, C. Moore, I. Banas, J. Lee, T. Woyke, A. J. Probst, Differential expression of core metabolic functions in DPANN Archaea of distinct subsurface ecosystems. *Prepared for submission to ISME communications* (2023).
   (Now available as: **S. P. Esser**, V. Turzynski, J. Plewka, C. J. Moore, I. Banas, A. R. Soares, J. Lee, T. Woyke, A. J. Probst, Differential expression of core metabolic functions in *Candidatus* Altiarchaeum inhabiting distinct subsurface ecosystems. *bioRxiv*, 2023.11.20.567779 (2023). Date modified: 11th March 2024)

Co-author papers:

3. Turzynski, V., Griesdorn, L., Moraru, C., Soares, A.R., Simon, S.A., Stach, T.L., Rahlff, J., **Esser, S.P.**, Probst, A.J. Virus-Host Dynamics in Archaeal Groundwater Biofilms and the Associated Bacterial Community Composition. *Viruses* 15. (2023)

4. Bornemann, T.L.V., **Esser, S.P.**, Stach, T.L., Burg, T., Probst, A.J. uBin – a manual refining tool for genomes from metagenomes. *Environmental Microbiology* 1-7 (2023)

5. L. Listmann, C. Peters, J. Rahlff, **S. P. Esser**, E. Schaum, Coevolutionary patterns shown in *Ostreococcus*-virus system from the Western Baltic Sea in freshly isolated hosts and viruses. *bioRxiv* (2023).

6. S. Ninck, T. Klaus, T. V. Kochetkova, **S. P. Esser**, L. Sewald, F. Kaschani, C. Bräsen, A. J. Probst, I. V. Kublanov, B. Siebers, M. Kaiser, Environmental activity-based protein profiling for function-driven enzyme discovery from natural discovery from natural communities. *bioRxiv* (2022)

7. J. Rahlff, **S. P. Esser**, J. Plewka, M. E. Heinrichs, A. Soares, C. Scarchilli, P. Grigioni, H. Wex, H.-A. Giebel, A. J. Probst, Heads in the clouds: marine viruses disperse bidirectionally along the natural water cycle. *bioRxiv* (2022)

8. Monsees Indra, Turzynski Victoria, **Esser Sarah P.**, Soares André, Timmermann Lara I., Weidenbach Katrin, Banas Jarno, Kloster Michael, Beszteri Bánk, Schmitz Ruth A., Probst Alexander J., Label-Free Raman Microspectroscopy for Identifying Prokaryotic Virocells. *mSystems*. **7**, e01505-21 (2022).

9. J. Rahlff, V. Turzynski, **S. P. Esser**, I. Monsees, T. L. V. Bornemann, P. A. Figueroa-Gonzalez, F. Schulz, T. Woyke, A. Klingl, C. Moraru, A. J. Probst, Lytic archaeal viruses infect abundant primary producers in Earth's crust. *Nature Communications*. **12**, 4642 (2021).

<u>As a member of the "RNA discovery consortium":</u>

10. U. Neri, Y. I. Wolf, S. Roux, A. P. Camargo, B. Lee, D. Kazlauskas, I. M. Chen, N. Ivanova, L. Zeigler Allen, D. Paez-Espino, D. A. Bryant, D. Bhaya, A. B. Narrowe, A. J. Probst, A.

Sczyrba, A. Kohler, A. Séguin, A. Shade, B. J. Campbell, B. D. Lindahl, B. K. Reese, B. M. Roque, C. DeRito, C. Averill, D. Cullen, D. A. C. Beck, D. A. Walsh, D. M. Ward, D. Wu, E. Eloe-Fadrosh, E. L. Brodie, E. B. Young, E. A. Lilleskov, F. J. Castillo, F. M. Martin, G. R. LeCleir, G. T. Attwood, H. Cadillo-Quiroz, H. M. Simon, I. Hewson, I. V. Grigoriev, J. M. Tiedje, J. K. Jansson, J. Lee, J. S. VanderGheynst, J. Dangl, J. S. Bowman, J. L. Blanchard, J. L. Bowen, J. Xu, J. F. Banfield, J. W. Deming, J. E. Kostka, J. M. Gladden, J. Z. Rapp, J. Sharpe, K. D. McMahon, K. K. Treseder, K. D. Bidle, K. C. Wrighton, K. Thamatrakoln, K. Nusslein, L. K. Meredith, L. Ramirez, M. Buee, M. Huntemann, M. G. Kalyuzhnaya, M. P. Waldrop, M. B. Sullivan, M. O. Schrenk, M. Hess, M. A. Vega, M. A. O'Malley, M. Medina, N. E. Gilbert, N. Delherbe, O. U. Mason, P. Dijkstra, P. F. Chuckran, P. Baldrian, P. Constant, R. Stepanauskas, R. A. Daly, R. Lamendella, R. J. Gruninger, R. M. McKay, S. Hylander, S. L. Lebeis, **S. P. Esser**, S. G. Acinas, S. S. Wilhelm, S. W. Singer, S. S. Tringe, T. Woyke, T. B. K. Reddy, T. H. Bell, T. Mock, T. McAllister, V. Thiel, V. J. Denef, W.-T. Liu, W. Martens-Habbena, X.-J. Allen Liu, Z. S. Cooper, Z. Wang, M. Krupovic, V. V. Dolja, N. C. Kyrpides, E. V. Koonin, U. Gophna, Expansion of the global RNA virome reveals diverse clades of bacteriophages. *Cell*. (2022)

Conference Proceedings

During the time of the PhD thesis, in total three poster (Vereinigung allgemeiner und angewandter Mikrobiologie 2020, Viruses of Microbes 2022 and International Society of Microbial Ecology 2022), one talk at "Vereinigung allgemeiner und angewandter Mikrobiologie" in 2021 was contributed to conferences. Additionally, the EMBO course (European Molecular Biology Organization) "Integrated multi-omic analyses of microbial communities" was visited in April 2022 as participant, where I presented a poster.

Table of Content

1. Abstract

The microbial communities of the deep subsurface are as complex and diverse as the variety of ecosystems present. By using culture-independent techniques, several branches of the tree of life have been phylogenetically extended and therefore shedding light on the microbial dark matter. Aquatic environments, such as groundwater and marine ecosystems, the archaeal branch gained a broad range of new representatives, predominantly within the DPANN branch. *Candidatus* Altiarchaeum (*Ca.*) is a globally widespread DPANN archaeon found in, *e.g.*, geysers and subsurface aquifers of diverse chemical properties and with varying bio-processible compounds. Previous metagenomic analyses revealed a CRISPR system, known as immune systems of microorganisms, encoded in population genomes of *Ca.* Altiarchaeum as well as an association with an episymbiont of the DPANN branch *Ca.* Huberiarchaeum. Yet the type and complexity of the CRISPR systems of *Ca.* Altiarchaeum and the respective targets were unknown. This thesis aims to analyze the potential targets of the CRISPR systems and to compare expression rates of conserved core metabolic functions in ecosystems around the globe in order to delineate the focus of metabolic functions of *Ca.* Altiarchaeum. The analyses of the CRISPR interference with spacers derived from the hosts' CRISPR array revealed targeting of viral sequences, the hosts' own genome and genomic regions of the episymbiont. By mapping metagenomic reads to the targeted sites, decreases in coverage indicated that the CRISPR system cleaves the genomic DNA of the host via self-targeting and the episymbiont within gene coding regions. Transferring the reported results to publicly available archaeal genomes on NCBI revealed that spacer acquisition of genomic DNA from divergent archaeal lineages might be an widespread phenomenon, particularly within extreme subsurface environments. The results suggest that CRISPR targeting alters symbiotic relationship and might result in shifts from parasitic to mutualistic behaviors based on necessary metabolic complementation between archaea. Moreover, the differential expression rates of core metabolic functions from *Ca.* Altiarchaea were analyzed in ecosystems with and without its episymbiont. Although the compared ecosystems have significantly different chemical compositions and available bio-processable molecules, the shared core metabolic functions suggest that *Ca.* Altiarchaea are in one ecosystem influenced by viral attacks and, in the other, affected by environmental stressors. The combined results

indicate that *Ca.* Altiarchaea, as a player in carbon fixation in the deep subsurface, is affected by viral attacks and symbiotic relationship dependent on the ecosystems. Compared to other known DPANN archaea–archaea host-symbiont associations, the symbiosis of *Ca.* Altiarchaea and *Ca.* Huberiarchaea provides new insights into microbial interactions via CRISPR-Cas interference. In general, this thesis sheds light onto the potential diversity of CRISPR interference in DPANN archaea of aquatic subsurface ecosystems and expands the knowledge about uncultivated archaeal symbiotic partners. Furthermore, the comparison of expression rates of core metabolic functions of the DPANN archaeon *Ca.* Altiarchaeum gives an insight into the variety of the archaeal metabolism dependent on the aquatic subsurface ecosystems.

2. Introduction

2.1. The hidden microbial majority

Within the Earth's emergence microorganisms populated many variable ecosystems. Microorganisms are small organisms, which are visible under the microscope and are monocellular. Prokaryotes, consisting of bacteria and archaea, are the second largest entity on Earth, with about 10^{29} - 10^{30} cells and they can be found in various ecosystems (Flemming and Wuertz, 2019). They are only outnumbered by viruses, virus-like particles, and mobile genetic elements (MGEs) with approx. 10^{31} entities (Hendrix et al., 1999; Mushegian, 2020). As early studies, conducted before the rise of sequencing and bioinformatic approaches, focused on cultures of microorganisms from ecosystems, these extensive hidden microbial communities, also known as 'microbial dark matter' (Marcy et al., 2007; Rinke et al., 2013), were a surprising find. The complexity of prokaryotic communities was underestimated for a long time, as most prokaryotes (up to 99%) are not-yet cultivated, and therefore, the community compositions in ecosystems as well as the complex interactions of different microbes were unknown (Head et al., 1998; Solden et al., 2016). Phylogenetic relationships within the eukaryotic (which range from unicellular Eukarya, to complex organisms, such as animals and plants), bacterial and archaeal branch were uncovered using the highly conserved 16S ribosomal RNA (16S rRNA) gene (Maidak et al., 1997; Woese et al., 1990; Woese and Fox, 1977). To provide an identifier, in cases when culture-independent approaches were used to classify a species, the term *Candidatus* was introduced. These *Candidatus* species are distributed across the tree of life depending on their phylogenetic assignment and their gene or protein composition (Brown et al., 2015; Méheust et al., 2019; Oren, 2021). The most prominent tree of life, which sheds light on the phylogenetic relationships of microorganisms was presented by Woese, Kandler and Whellis in 1990 (Woese et al., 1990), proposed with prior studies of Woese and Fox in 1977 (Woese and Fox, 1977). Their tree of life was presented as a three-domain tree separating bacteria, archaea, and eukaryotes (Woese et al., 1990; Woese and Fox, 1977). Before the introduction of the three-domain tree, Whittaker described a five-domain tree divided by protists, monera (bacteria), plants, fungi, and animals (Whittaker, 1969). With progressive technologies and approaches, the tree of life evolved into a two-domain tree which consisted of a bacterial and

an archaeal branch with eukaryotes originating within the latter (reviewed in Koonin, 2015 and Pace, 2009, 2006), although this is discussed in the literature (Forterre, 2013).

Specifically, the archaeal branch was for a long time only known to consist of extremophiles, e.g., methanogens, thermophiles, and halophiles (Fox et al., 1977; Magrum et al., 1978; Woese et al., 1978), but analyses of 16S rRNA genes within various ecosystems diversified the knowledge of archaea within moderate ecosystems (DeLong, 1998). It was shown that archaea are present in numerous ecosystems such as marine sediments, polar seas, soils, and human guts (*e.g.*, Bintrim et al., 1997; Deming, 2002; Gaci et al., 2014; Hinrichs et al., 1999) but can also be involved in nutrient cycling for plants (reviewed in Jung et al., 2020). Archaea generally influence the biogeochemical cycles. Studying these can hence give insights into global carbon and nutrient cycles as well as the respective influence of microorganisms (reviewed in Offre et al., 2013).

Studies focusing on shedding light into the microbial dark matter or in microbial communities of harsh environments are profiting from technological progress of sequencing techniques and bioinformatical analyses.

2.2. Modern resolution of microbial communities

The sequencing of desoxyribonucleic acid (DNA) and ribonucleic acid (RNA) is a significant element in resolving microbial communities in different ecosystems. Even complex ecosystems can be depicted with this technique. Chain terminated Sanger sequencing was the first DNA sequencing technique that could decipher complex DNA fragments. It was even used to decode the first whole genome, namely the bacteriophage phiX174 infecting *Escherichia coli* (*E. coli*) with a genome size of 5,375 nucleotide bases (Sanger et al., 1977, reviewed in Men et al., 2008). After Sanger sequencing, which was the laboratory standard for multiple decades, the necessity of high throughput sequencing methods grew with the higher complexity of samples. As a result, new sequencing techniques like Illumina sequencing and Nanopore sequencing were developed (reviewed in Kchouk et al., 2017). Notably, Illumina sequencing resulted in the affordable and quick decoding of DNA of complex microbial communities since multiple small DNA fragments can be deciphered in parallel via reverse identification chemistry, which detects fluorophore removal (reviewed by Ari and Arikan, 2016; Bentley et al., 2008).

Along with this fast evolution in sequencing techniques, the attention was rapidly directed to manual and automized bioinformatic tools that could handle large amounts of sequencing data and carry out different types of analyses (collected information in Abdurakhmonov, 2016; Mount, 2004). The term bioinformatics is attributed to interdisciplinary sciences, such as mathematical/statistical approaches, computational analyses combined with biological information and was introduced around the turn of the millennium (Luscombe et al., 2001). With the help of bioinformatics, not only complex DNA and RNA datasets can be analyzed, also proteins and metabolites can diversify the current knowledge about microbial interactions, communities, and single cells with the help of bioinformatics. These interdisciplinary methods alongside statistical approaches are combined in the term "multi-omics" (Figure 2.2.1; Krassowski et al., 2020).

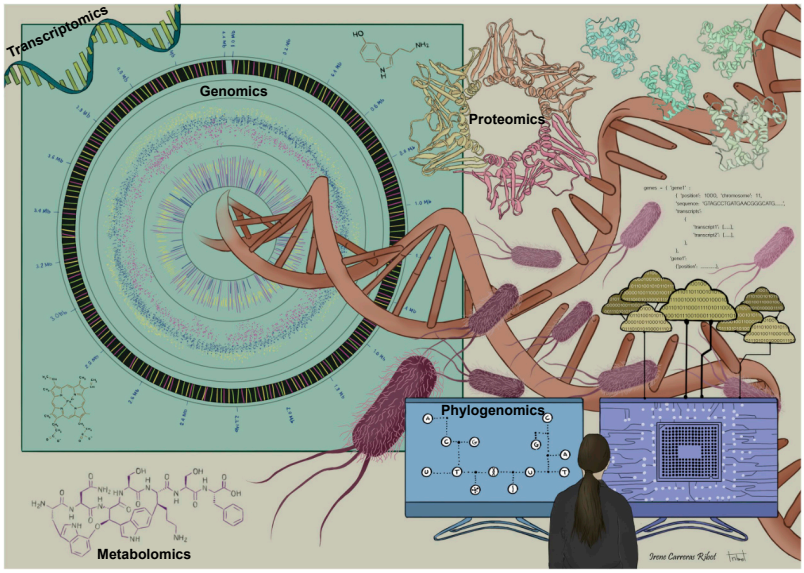


Figure 2.2.1 | *Merged view on multi-omics (modified from Krassowski et al., 2020). In this figure the different analyses that are included in the term are depicted. They range from genomics and transcriptomics, as well as metabolomics, phylogenomics and proteomics, to statistical approaches that can verify and falsify the results.*

2.2.1. Metagenomics and Metatranscriptomics

Metagenomic analyses describe complex communities of microorganisms from environments with varying environmental conditions (reviewed in Zhou et al., 2015). DNA and RNA extractions of microbial communities, the baseline for a fruitful sequencing and metagenomic analysis, aim to be non-selective to a specific organism and are usually adapted to the ecosystem, *e.g.*, to volcanic environments (reviewed in Herrera and Cockell, 2007). Sequencing DNA from environmental or complex microbial communities results in

metagenomes that represent the respective environment. In addition, sequencing data can be used to address different ecological questions, such as the microbial community composition or the phylogenetic variation (reviewed in Zhou et al., 2015).

To achieve and maintain high-quality datasets, sequencing reads should be quality-filtered to remove sequencing adapters and read fragments of low quality (Eren et al., 2013; Kunin et al., 2010). After assembling the quality-filtered metagenomic dataset, scaffolds or contigs can be binned into metagenome-derived genomes (MAGs), viral scaffolds, or fragments that could not be assigned to a bin (reviewed in Navgire et al., 2022 and Yang et al., 2021). Binning produces reliable insights into the functionality of population genomes of microorganisms and therefore portrays a cross-section of a microbial species within an ecosystem. Binning is carried out by multiple factor validations, depending on the tool used, and afterwards completed by curating and quality assessment of the genome (reviewed in Yang et al., 2021). The diversity of microbial communities can also be assessed with single-cell amplified genomes (SAGs) (Nawy, 2014; Zong et al., 2012). The advantage of SAGs over MAGs is that only one single cell is used to prepare a DNA library, and the results depict one cell's metabolic capacity instead of the microbial community. Single-cell genomics is favorable while studying a specific strain or species of microbes (Navin et al., 2011; Nawy, 2014; Shapiro et al., 2013; Zong et al., 2012).

Nevertheless, not only can genomic information be derived from microbial communities, single cells, or environmental microbial communities, also metatranscriptomes can be analyzed by RNA extractions (reviewed in Shakya et al., 2019) that are translated to cDNA prior to sequencing and combined with rRNA depletion (O'Neil et al., 2013). The latter is necessary as the ribosomal RNA is transcribed in higher copy numbers (up to 15 gene copies per genome) than the remaining encoded genes (Pei et al., 2010), and therefore, uses the sequencing depth more efficiently (O'Neil et al., 2013). For metatranscriptomes (cDNA-sequencing) a quality-filtering of sequencing read is, in contrast to metagenomic raw reads, not inevitably necessary as it does not influence the amount of reads assigned to the respective metagenomic sequence to calculate expression profiles (Liao and Shi, 2020). Therefore, these metatranscriptomes can be used to identify the transcribed metabolic capacity of the community. To assign the transcriptomic reads to genomes derived from metagenomes, the quality-filtered reads can be mapped to the metagenomic-derived

genomes. This approach results in coverage estimations that can be traced back to the identification of actively transcribed genes within the community or microorganism. The power of metatranscriptomics is hence of interest, especially when looking at the metabolic capacity of microbial dark matter.

2.2.2. Metaproteomics, meta-metabolomics, and phylogenomics

The collective information of multi-omics resolve metabolic, genomic, and phylogenetic information of microbial communities in distinct ecosystems revealing unknown complexities of microorganisms on Earth (collected information in Agrawal and Verma, 2021). By definition, metaproteomics describes the analyses and quantification of the protein complexity in the environment (Wilmes and Bond, 2004). Metabolomics, also known as metabonomics, refer to the metabolites or byproducts analyzable from environmental samples (Marchesi and Ravel, 2015; Nicholson, 2006; Nicholson et al., 1999). Metaproteomic datasets help researchers to collect information not only about the genetic potential of a community but also about the proteins actively transcribed and present in an environmental sample (Maron et al., 2007; Tringe et al., 2005; Wilmes and Bond, 2006, 2004). The information extractable from analyzing the byproducts and metabolites present in an environment or culture can indicate the current metabolic capacity of the microbes (Fiehn, 2002). Further, the analysis can indicate if environmental stressors or changes in the microbial community impact the overall reaction, such as carbon cycling or the availability of bio-processable products (Aguiar-Pulido et al., 2016). While meta-genomics, -proteomics, -bolomics/bonomics, and -transcriptomic focus on cell-specific complexity, phylogenomics concentrate on the phylogenetic complexity of microbial populations by comparing taxonomical assignments based on gene alignments (reviewed by Delsuc et al., 2005 and Philippe et al., 2005).

In general, a combination of all “omic” approaches can give a complex insight into environmental microbial communities and are under constant development within the latest research approaches. These culture independent methods are specifically useful for environments that diverge from the average and are therefore called extreme environments.

2.3. Extreme environments

Unlike macroorganisms, microorganisms can thrive under extreme environmental conditions. By definition, every ecosystem in which humans cannot survive over an extended period of time is an extreme environment (Chénard and Lauro, 2017; Gómez, 2011). Varying environmental factors are, among others, temperature, pressure, and salinity. In addition, high concentrations of different bio-processable molecules and elements can stand out in an extreme environment. Furthermore, a combination of different extreme conditions still allows microorganisms to thrive where 'higher' lifeforms would be unable to survive. Examples of such an extreme environment with multiple extreme factors are deep-sea hydrothermal vents, *e.g.,* black or white smokers (Jannasch and Mottl, 1985; Tivey, 2014). These hydrothermal vents occur characteristically in volcanically active regions and at oceanic spreading ridges (Baker and German, 2004), can have temperatures of up to 405 degrees Celsius, high concentrations of sulfur, carbon dioxide or iron, and due to their depth, are high-pressure systems (reviewed by Martin et al., 2008). The high energetic potential supports hydrothermal vents or sediments in being a favorable ecosystem for diverse microbial communities (*e.g.* Baross and Hoffman, 1985; Corliss et al., 1979; reviewed by Martin et al., 2008; Spiess et al., 1980). Another example of hydrothermal vents and sediments with highly complex eukaryotic and prokaryotic populations are the vents at the Guaymas Basin, located at the deep oceanic seabed in the Gulf of California (Dhillon et al., 2003; Dombrowski et al., 2018; Edgcomb et al., 2002; Teske et al., 2002). Prokaryotes, such as members of the genus *Sulfurimonas* and the ANME group (anaerobic methanotrophic archaea), within the Guaymas Basin are metabolically complex and can anaerobically oxidize sulfur and methane respectively (Dhillon et al., 2003; Teske et al., 2021).

While hydrothermal vents and sediments are examples of oceanic, deep subsurface environments, similar environments exist for the terrestrial subsurface, such as geysers and springs fed by hot or cold aquifers. Due to their high pressure and primarily challenging composition of bio-processable molecules, these count as extreme environment. Examples for terrestrial hot springs are alkaline hot springs, located in the Yellowstone National Park (USA), which are dominated by thermophilic bacteria of the genus *Thermus* that thrive in environments at 75°C (De León et al., 2013). Despite the extreme conditions for microbial communities in these ecosystems, changes in the abundance of microbes and metabolic

capacities (De León et al., 2013) indicate active microbial adaptations to the environment. The Crystal Geyser (Utah, USA) and the Geyser Andernach (Andernach, Germany) are examples of cold water-fed terrestrial geysers with complex microbial communities (Bornemann et al., 2022; Emerson et al., 2016; Probst et al., 2018). Despite the adaptations to environmental factors, microbes can also adapt to viral attacks and hinder viral induced cell lysis by using clustered regularly interspaced short palindromic repeat (CRISPR) systems as microbial immune systems (reviewed by Horvath and Barrangou, 2010).

2.4. CRISPR systems

CRISPR systems are colloquially defined as immune systems of prokaryotes against invading mobile genetic elements (MGEs), such as viruses, phages, and retroviruses (Barrangou et al., 2007b; Garneau et al., 2010; reviewed by Horvath and Barrangou, 2010). Approximately 85% of all archaea and 40% of all bacteria encode CRISPR systems in their genome with varying complexity (reviewed by Makarova et al., 2020). However, until the research teams led by Mojica and Jansen further described these systems, they were for many years only known as repetitive nucleic acid regions within a complex gene-coding genomic region (Jansen et al., 2002; Mojica et al., 2000, 2005). One of the first known prokaryotes with an identified repetitive sequence system, later revealed to be a CRISPR system, was *E. coli* K12 in 1987 (Ishino et al., 1987).

2.4.1. Function and structure of CRISPR systems

The function of a CRISPR system can be divided into three stages: (1) acquisition/adaptation, (2) expression of the CRISPR locus, and (3) interference (reviewed in Horvath and Barrangou, 2010; Rath et al., 2015). The first phase, the adaptation, describes the process of acquiring short fragments (26-62 bp), namely spacers, of DNA from invading mobile genetic elements (Grissa et al., 2007; Rath et al., 2015). These spacers can be used in bioinformatic analyses to explore infection histories in microbial populations by comparing the spacer sequences to protospacer sequences (the protospacer is equal to the same nucleotide sequence as a spacer but within the MGE) (reviewed by Bhaya et al., 2011). Spacers are acquired in between the two inverted repetitive nucleotide sequences with a length between 21-48 bp, namely repeats (Bolotin et al., 2005; Grissa et al., 2007; Jansen et

al., 2002; Rath et al., 2015). Alternating modules consisting of repeat and spacer are named CRISPR arrays. Since CRISPR arrays allow for a lead back to the microorganism and the MGE, they can be used to reveal infection histories within the ecosystem (reviewed in Westra et al., 2014, 2012). CRISPR arrays are flanked by an AT-rich leader sequence, which on the one hand, signals the adaptation of the CRISPR system to invading MGEs, and, on the other hand, encode promotors for the CRISPR transcription (Alkhnbashi et al., 2016; Erdmann and Garrett, 2012; Lillestøl et al., 2006; Yosef et al., 2012). Thereby the acquisition of new spacers was reported to be at the leader-proximal end, which means that the order of spacers within an array is a blueprint of the infection history (Barrangou et al., 2007b; Deveau et al., 2008). The position of the acquired spacers is also decisive for the success of the immune response of the prokaryote, as the expression of spacers close to the leader sequence is significantly higher than the expression of spacers closer to the CRISPR arrays end (McGinn and Marraffini, 2016). Interestingly, the position of the protospacers within plasmids or viruses does not seem to be affected by any kind of selectivity (Yosef et al., 2012).

The second phase of CRISPR functionality portrays the gene expression of CRISPR associated *cas* genes into crRNA that partake in different roles within the cell defense process (reviewed by Horvath and Barrangou, 2010). As the CRISPR gene locus is transcribed into multiple crRNA fragments, each of these carries a spacer and a part of the palindromic repeat (Garrett et al., 2011). The cascade is defined as the CRISPR-associated complex of anti-viral defense and is a multi-subunit protein complex involved in the interference phase of the CRISPR defense mechanism (reviewed by Bhaya et al., 2011).

The interference with invading DNA or RNA describes the last phase of the CRISPR system function. This phase includes directing the expressed *cas* cassette to the invading genetic material and the respective cutting of these (Marraffini and Sontheimer, 2010a; Reeks et al., 2013). A successful interference of the CRISPR system is only possible if the microorganism was able to adapt a spacer previously (McGinn and Marraffini, 2016).

2.4.2. Classification of CRISPR systems

CRISPR systems are divided into two classes based on the functionality and structure of the respective systems (Koonin et al., 2017; Makarova et al., 2015, 2011; Makarova and Koonin, 2015). Depending on the *cas* genes encoded in the CRISPR loci, CRISPR systems are

currently categorized into six types (Type I-VI) with multiple subcategories. At the same time, the CRISPR types I, III, and IV are assigned to CRISPR class 1, and the types II, V, and VI are grouped within CRISPR class 2 (Koonin et al., 2017).

Within the CRISPR types of class 1, which is preeminent in prokaryotes, CRISPR type I is the most complex (Koonin and Makarova, 2022; Makarova et al., 2018, 2015, 2011). CRISPR type I divides into the subtypes I-A to I-F and type I-U, whereby type I-F has two variants. Of these CRISPR subtypes, the CRISPR system I-E, found in *E. coli*, is one of the most studied systems. The type I-E consists of *cas1*, *cas2*, *cas3*, *cse1*, *cse2*, *cas7*, *cas5*, and *cas6e* (Jore et al., 2011; Swarts et al., 2012). While the *cas1* and *cas2* genes are involved in the adaptation phase, and the *cas3* gene is responsible for cleaving foreign DNA, the other *cas* genes are building the CRISPR cascade within the interference stage (Jore et al., 2011; Terns and Terns, 2011).

In the CRISPR system of class 2, the CRISPR type V is known for its complexity, whereby the *cas12* and *cas14* enzymes seem to be the drivers of the diversification of this CRISPR system (Yan et al., 2019). A study by Yan and collaborators in 2019 identified a total of 17 phylogenetically distinct CRISPR type V systems, which divides into four branches (Yan et al., 2019).

While the CRISPR associated *cas1* and *cas2* enzymes are preeminent in nearly all CRISPR system types and utilize the adaptation phase, other *cas* genes, which are variable dependent on the CRISPR type, build complexes for the interference phase of cell defense (Brouns et al., 2008, reviewed in Barrangou et al., 2007; Bhaya et al., 2011). Whereas in most eukaryotes, *cas9* enzymes are responsible for cleaving foreign DNA/RNA (reviewed by Doudna and Charpentier, 2014; Ran, 2014), in many prokaryotic CRISPR systems, this role is carried out by *cas3* genes (reviewed in He et al., 2020).

2.4.3. Applications of the CRISPR system

CRISPR systems do not only affect microbial communities, nutrient cycles, and the success of viral infections (Andersson and Banfield, 2008; Heidelberg et al., 2009; Tyson and Banfield, 2008), but researchers from different fields are also using them to genetically modify crops, microorganisms, and human genetic material. However, the latter is highly debated in regards to ethics and human rights (Brokowski and Adli, 2019).

The CRISPR system I-E can be used to genetically modify, *e.g.*, bacteriophages. A study of Kiro, Shitrit, and Qimron in 2014 used this CRISPR system to delete the nucleotide kinase encoding gene in the T7 phage that infects *E. coli*, illustrating efficient engineering of phage genomes (Kiro et al., 2014). The researchers constructed three plasmids according to Brouns et al., 2008, whereby two encode the *cas* genes of the CRISPR type I-B and one plasmid encodes for spacers matching specific protospacer regions of the bacteriophage T7 (Brouns et al., 2008b; Kiro et al., 2014). With an active and functional expression of the CRISPR system, the targeted nucleotide kinase was deleted in the phage genome (Kiro et al., 2014).

Not only do microorganisms constantly adapt to invading MGEs, but viruses and phages also encode, *e.g.*, anti-CRISPR proteins to escape active CRISPR systems (Pawluk et al., 2018). These anti-CRISPR systems were first found in a *Pseudomonas spp.* phage, although the CRISPR system of the host inhabited the spacer matching to protospacer regions of the phage (Bondy-Denomy et al., 2012). Anti-CRISPR proteins can prevent the binding of the cleavage protein, namely *cas3*, or convert the CRISPR systems to a transcriptional repressor (Bondy-Denomy et al., 2015). This inactivation of the CRISPR systems of microorganisms led to the understanding that viruses and phages can cope with and prevent from extinction by adapting to the prokaryotic immune system.

2.5. DPANN archaea

Archaea are divided into several branches according to their phylogeny. Based on the name of the first five archaea, which were first discovered, in this branch, *Ca.* Diapherotrites, *Ca.* Parvarchaeota, *Ca.* Aenigmarchaeota, Nanoarchaeota, and *Ca.* Nanohaloarchaeota, the DPANN archaea were named (Rinke et al., 2013). Until today several other archaea were reported to belong to this branch based on phylogenetic analyses (*i.e.*, Castelle et al., 2018; Castelle and Banfield, 2018; Chen et al., 2018; Liu et al., 2018). Within phylogenomic studies, DPANN archaea are proposed to originate and diverge from other archaea early in times (reviewed by Spang et al., 2017). The first cultivated member of the DPANN archaea was *Nanoarchaeum equitans* (*N. equitans*), an ectosymbiont to *Ignicoccus hospitalis* (*I. hospitalis*), whereby *N. equitans* is entirely dependent on the presence of its host to flourish (Huber et al., 2002). DPANN archaea have little metabolic capacity with significant gaps in the core metabolism, which is why most of them are known to be obligate symbiotic partners (Castelle

et al., 2018; Castelle and Banfield, 2018). Although the first DPANN archaea described was the cultivated Nanoarchaeum equitans (Huber et al., 2002), the branch is dominated by candidate archaea identified with culture-independent methods (Castelle et al., 2018). Members of the DPANN archaea were found to be very abundant in aquatic systems. Not only surface waters but also groundwater systems are ecosystems dominated by DPANN archaea (Castelle et al., 2015; He et al., 2021). While most DPANN archaea are obligate symbionts, some lineages, such as *Ca.* Diapherotrites are proposed to be free-living microorganisms based on metagenomic analyses of their metabolic capacity (Youssef et al., 2015). Pacearchaeota and Woesearchaota, also members of the DPANN archaea, are abundant in high altitude surface waters and can dominate these with up to 46 % and 38 %, respectively, and DPANN can therefore dominate aquatic ecosystems (Ortiz-Alvarez and Casamayor, 2016).

2.5.1. Symbiosis of DPANN archaea

As mentioned above, DPANN archaea were proposed to be obligate symbionts, with the first in co-culture detected representative being *N. equitans* (Huber et al., 2002). Symbiosis is a collective term for all symbiotic life forms, including parasitism, mutualism, and commensalism. A characteristic of symbiotic relationships is the dependency, disregard the kind of dependency, of the symbiont on the hosts' metabolites or metabolic pathways. Exemplarily, *N. equitans* encodes a total of only 552 genes encoded in its 490 kb genome (coding density of 95%) and is therefore only cultivatable in co-culture with its host *I. hospitalis* (Waters et al., 2003). None of these genes encode for chemolithoautotrophy, which is the energy production source by reducing sulfur based on hydrogen as an electron donor within *Ignicoccus* species (Huber et al., 2000). Next to the limited encoded energy production of *N. equitans*, the genome lacks essential genes for functional cell organelles and processes, *e.g.*, biosynthesis of nucleotides, co-factors, amino acids, and lipids (Waters et al., 2003). To transfer these necessary metabolites from the host to the symbiont, the genome of *N. equitans* encodes for multiple antiporters, cell appendages, and secretory signal proteins (Huber et al., 2003; Waters et al., 2003).

Next to the prominent symbiotic example of *Nanoarchaeum equitans* and *Ignicoccus hospitalis*, another example of a symbiont-host relationship is *Ca.* Micrarchaea and *Metallosphaera* sp. AS-7. *Ca.* Micrarchaea, as members of the DPANN superphylum, were first

detected within an acid mine drainage (Archaeal Richmond Mine Acidophilic Nanoorganisms = ARMAN) (Baker et al., 2006, 2010; Sakai et al., 2022). Interestingly, transmission electron microscopy (TEM) revealed that ARMAN archaea, which include *Ca.* Micrarchaea and *Ca.* Parvarchaea, have a cell size ranging from 193 to 299 nm and are therefore smaller than Nanoarchaea, which have a mean cell size of 400 nm (Baker et al., 2006; Huber et al., 2002). Although the cell size of *Ca.* Micrarchaea is smaller than the cells of Nanoarchaea, Micrarchaea encode, on average, 912 genes with a genome size between 0.64 – 1.08 Mb (coding density of ~90.4%) (Chen et al., 2018). Host-switching experiments revealed that *Ca.* Micrarchaea can not only live in co-culture with *Metallosphaera* sp. AS-7, a member of the *Sulfolobales* order, but also with members of the order *Metallosphaera*, *Acidianus*, and *Saccharolobus* (Sakai et al., 2022), which is, until this date, a unique behavior for DPANN symbionts. Comparable to *N. equitans*, genomes of *Ca.* Micrarchaea lack genes for most of the amino acid biosynthetic pathways, nucleotides and they do not encode for carbon fixation pathways (Chen et al., 2018). Within analysis of the encoded metabolic capacities of the representative genomes of *Ca.* Micrarchaea unveiled horizontal gene transfer between the potential host of the order Sulfolobales (Sakai et al., 2022).

2.6. *Candidatus* Altiarchaeota

Ca. Altiarchaea is a DPANN archaeon that can live independently from symbiosis due to its increased genome size of approximately 1.5 Mb, compared to other DPANN archaea, and their metabolic capacity (Bird et al., 2016; Bornemann et al., 2022; Probst et al., 2014). Altiarchaea were first described as a string-of-pearls in the Sippenhauer Moor (SM) and, based on 16S rRNA analyses, first called *SM1 archaea* due to the lack of other representatives in public databases (Rudolph et al., 2001). These string-of-pearls are macroscopically visible year-round and are white biofilms with a variable number of cells. The ecosystem has a constant temperature of 10 °C, high sulfur concentrations (up to 1.2 mg/L), and a pH of 6.5. In addition, the oxygen concentration was determined to be low at 1.4 mg/L, which is favorable for the anaerobic nature of *Ca.* Altiarchaeum (Rudolph et al., 2001). With fluorescence *in-situ* hybridization (FISH) approaches, these string-of-pearls were determined to be biofilm structures composed of cocci-shaped *Ca.* Altiarchaea surrounded by sulfide-reducing bacteria of the family *Thiothrix* (Moissl et al., 2002; Rudolph et al., 2001). Close to

the Sippenhauer Moor, at an approximate distance of 30 km, the Muehlbacher sulfidic spring Isling (MSI) was also found to inhabit archaea of the same genus, determined by phylogenetic analyses, FISH imaging, and identical restriction fragment length polymorphism (RFLP) patterns (Henneberger et al., 2006; Rudolph et al., 2004). Within the tree of life, *Ca.* Altiarchaea are assigned to the branch of Euryarchaeota within the DPANN superphylum but are distinct from methanogens (Moissl et al., 2002; Rudolph et al., 2001).

Compared to the natural springs in SM, MSI is a metal-mantled drilling whole, drilled in 1925, releasing 5.5 m^3 water per hour from a depth of 36.5 m (Henneberger et al., 2006; Probst et al., 2014). The cells within elevated biofilm flocks, which also consist of *Ca.* Altiarchaea, appear to be in closer proximity to each other than within the spring-of-pearls in SM. The researchers around Huber (2006) uncovered, via FISH images, that the biofilms in MSI have a purity of 95 %, with the remaining 5 % belonging to rod-shaped bacteria (Henneberger et al., 2006). Later this finding was confirmed by adjusted PhyloChip analyses, which resulted in an increase of 2130 % in the abundance of *Ca.* Altiarchaea in the biofilm compared to the spring water (Probst et al., 2013). *Ca.* Altiarchaea from both environments (SM and MSI) have a G/C content ranging between 34.5 and 35 mol% (Henneberger et al., 2006; Moissl et al., 2003).

2.6.1. Global distribution of the phylum *Ca.* Altiarchaeota

Sequencing-based analyses of worldwide distributed deep subsurface aquifers revealed that organisms of the phylum *Ca.* Altiarchaeota can have different core metabolisms. Therefore generated genomes are sorted into the Alti1 and Alti2 clade, showing a strong phylogenetic distribution pattern alongside the continent found in (Bornemann et al., 2022). In general, genomes of *Ca.* Altiarchaea were found in samples originating from, *e.g.*, Japan (Hernsdorf et al., 2017), the United States of America (Probst et al., 2014), and Italy (Bornemann et al., 2022; Hamilton et al., 2015). The following thesis focuses only on the Alti1 clade.

2.6.2. Physiology and metabolic capacity of *Ca.* Altiarchaea hamiconexum

Interestingly, despite their phylogenetically determined branching within the DPANN archaea, *Ca.* Altiarchaea of SM and MSI, no indication of the coenzyme F_{420} was found, resulting in no fluorescence emission typical for methanogens (Moissl et al., 2003). The

coenzyme F_{420} is an H_2-sensible cytoplasmic enzyme involved in the redox reaction of deazaflavin within the methanogenesis that is excited at a wavelength of 420 nm and that emits a blue fluorescence signal (Cheeseman et al., 1972; Eirich et al., 1979, 1978). Further studies related to the core metabolic potential of *Ca.* Altiarchaeum hamiconexum, the dominant species within MSI, unveiled that the core metabolism is based on an adaptation of the Wood-Ljungdahl pathway (Probst et al., 2014). This pathway was found to be independent of light, enabling the organism to live autotrophically within subsurface ecosystems, and is reliant on reductive acetyl-CoA pathways to fix carbon within cells (reviewed by Ljungdhal, 1986; Wood, 1991). As this pathway is very energy efficient, one ATP can be used for the production of one mol pyruvate (reviewed in Berg et al., 2010) and might favor the dominance of Altiarchaea in various ecosystems (Probst et al., 2017). Moreover, the genomes of the *Ca.* Altiarchaea genus encode transport systems to channel activated sugars outside of the cell, supporting the biofilm formation and CRISPR systems (Probst et al., 2014).

Focusing on the physiology of biofilm-forming *Ca.* Altiarchaea populations, cell appendages are visible within the ecosystems SM and MSI. Within the study of the researchers around Moissl, it was found that these pili-like structures, later called *hami* (Rudolph et al., 2004), are digestible by proteinase K, which is an indicator that these pili are composed of proteins (Moissl et al., 2003). These hook-like protein structures interconnect Altiarchaea cells and are essential in forming the ecosystem's biofilm (Moissl et al., 2005; Rudolph et al., 2004). The Altiarchaea *hami* proteins have a protein weight of approximately 120 kDa (Kilo Dalton), a length of up to 5 µm, and a diameter of approx. 7 - 8 nm (Moissl et al., 2005; reviewed by Probst and Moissl-Eichinger, 2015). These *hami* are a unique finding in subsurface ecosystems.

In an immunofluorescence-based study of *Ca.* Altiarchaeum hamiconexum, the FtsZ protein was labeled, which is involved in the early steps of the cell division process and building the midcell ring (Bi and Lutkenhaus, 1991; Den Blaauwen Tanneke et al., 1999), indicating that the cells within biofilms are dividing and replicating (Probst et al., 2014). Furthermore, within FISH images the DAPI fluorescence signal, which labels DNA fragments, a strong intrinsic polar orientation of genomic DNA within the cells is visible, even without labeling the FtsZ cell division proteins (Rahlff et al., 2021).

Sequencing-based studies focusing on the genomic variability of Altiarchaea genomes unveiled that a core genome, including a CRISPR system, is preserved within the Alti-1 clade (Bornemann et al., 2022; Rahlff et al., 2021). Although the adaptation to invading DNA, namely the variance of spacer sequences, is highly variable between the ecosystems, the direct repeat sequence of the CRISPR system I-B is highly conserved within the Altiarchaea population across the globe (Rahlff et al., 2021). Within studies by Rahlff *et al*. in 2021, a second CRISPR system of type III, which was not predicted in the subtype, was proposed for *Ca.* Altiarchaeum hamiconexum in different ecosystems such as MSI and HURL. These CRISPR systems' complexity support the high infection rates predicted within MSI (Rahlff et al., 2021; Turzynski et al., 2023). Interestingly, one of the most dominant viruses infecting *Ca.* Altiarchaeum hamiconexum is a lytic virus with a genome size of 8.9 kb. The prediction of lytic archaeal viruses within the deep continental subsurface was the first of its kind in 2021 (Rahlff et al., 2021). Previous research predicted lysogeny as a dominant form of viral lifestyle and therefore, the piggy-back-the-winner theorem was the prevailing concept of viral lifestyle in the deep subsurface (Anderson et al., 2011).

2.6.3. Altiarchaea and its DPANN symbiont *Ca.* Huberiarchaea

In a publication by Probst *et al.* in 2018, a symbiont named *Ca.* Huberiarchaeum was proposed to belong to *Ca.* Altiarchaeum crystalense. The symbiotic prediction was based on metagenomic abundance correlations for 25 samples from one sampling campaign in 2015 (Probst et al., 2018). In contrast to the extended metabolic capacity of *Ca.* Altiarchaeum, the genome of *Ca.* Huberiarchaeum crystalense encodes not all essential pathways for cell functionality, e.g., not all amino acid synthesis pathways are encoded in the genome (Probst et al., 2018). A follow-up study on the potential exchange of metabolites between the two DPANN archaea hypothesized the possibility of cytoplasmic contact (Schwank et al., 2019). Other discussed pathways for metabolite exchange could be cysteine-rich large surface proteins or transport by membrane-enclosed vesicles (Probst et al., 2018; Schwank et al., 2019). Exemplary *Ca.* Huberiarchaeum is not able to synthesize the energy-related metabolites $NAD(P)^+$ or ferredoxin that would be needed in several synthesis pathways encoded in the genome. Next to the near-complete encoded RNA synthetase pathways, the

lipid biosynthesis, essential to form a cell membrane, is also nearly fully encoded within the *Ca.* Huberiarchaeum population genome (Probst et al., 2018; Schwank et al., 2019).

2.7. Knowledge gap

While there is no confirmed symbiotic relationship within DPANN archaea (intra-branch) reported yet, this thesis aims to explain the proposed symbiotic relationship of *Ca.* Altiarchaeum crystalense and *Ca.* Huberiarchaeum crystalense within Crystal Geyser. In addition, this relationship was kept under surveillance within other ecosystems with variable chemical and environmental conditions. Not only the metabolic dependencies of *Ca.* Huberiarchaeum on its host *Ca.* Altiarchaeum also the relevance of the intercellular defense systems of *Ca.* Altiarchaeum, namely the CRISPR system, and its relevance within the symbiotic relationship should be highlighted. Although self-targeting of CRISPR systems were observed in previous publications the report of spacers targeting a symbiotic partner is especially within DPANN archaea a new finding. This questions which kind of symbiotic relationship is prevalent for the two DPANN archaea. In addition, the analysis of CRISPR activity would allow conclusions about the nutrient supply and carbon fixation within the ecosystem by lysis of cells.

Furthermore, this thesis aims to shed light on the metatranscriptomic expression of the core metabolism of *Ca.* Altiarchaeum within different ecosystems. As *Ca.* Altiarchaeum is found to be free-living, and in symbiotic relationship with its episymbiont *Ca.* Huberiarchaeum in several ecosystems but biofilm-forming and without metagenomic and fluorescent evidence of the episymbiont in other ecosystems, the influence of environmental conditions and the presence/absence of the episymbiont on the metabolism of the host *Ca.* Altiarchaeum will be evaluated in this thesis.

3. Publications

3.1. Overview

The following chapter lists the two first author papers (first in press at Nature Microbiology, second in manuscript version) achieved during the duration of the PhD (~3.5 years, starting Nov. 2019). Both publications aim to enlighten different aspects of the metabolic functionality of *Ca.* Altiarchaeum in ecosystems with varying environmental conditions. In addition, the publications shed light into the symbiotic relationship of two DPANN archaea in the deep terrestrial subsurface. The aims of the two different work packages are listed individually below.

3.1.1. Aim and scope of the publication: '**A CRISPR-mediated symbiosis between uncultivated archaea predicted from sequencing data'**

Archaea of the DPANN archaea are widespread across the globe, inhabit diverse ecosystems encode a variety of CRISPR systems even in the deep subsurface. While the CRISPR interactions against viruses is a well-known concept the targeting of a symbiotic partner within the DPANN branch is rather unknown. The publication '**A CRISPR-mediated symbiosis between uncultivated archaea predicted from sequencing data**' addresses the question if the conserved CRISPR systems of *Ca.* Altiarchaea targets viruses, performs self-targeting and/or targeting of the symbiont. Furthermore, the metabolic dependency of the host-symbiont (*Ca.* Altiarchaea-*Ca.* Huberiarchaea) affected by the targeting of the respective genomes is determined. This publication also includes a discussion about the symbiotic nature of intra-branch DPANN archaea. The publication sheds light onto the DPANN archaea branch using state of the art metagenomic, metatranscriptomic analyses and metabolic modeling.

3.1.2. Aim and scope of the publication: '**Differential expression of core metabolic functions in DPANN Archaea of distinct subsurface ecosystems**'

The calculation of expression profiles is a major indicator if microbial communities are either actively replicating or if microbial communities are healthy. As *Ca.* Altiarchaeum is a widespread DPANN archaea found in association in presence and absence of the symbiont, the effects of the symbiont on the metabolism of *Ca.* Altiarchaea sheds a light onto the metabolic differences within different ecosystems. This analysis is provided in the publication

‘**Differential expression of core metabolic functions in DPANN Archaea of distinct subsurface ecosystems**’ Moreover, the calculation of the expression profiles of the core metabolism of *Ca.* Altiarchaea within the distinct ecosystem should give an insight into the possible reasons of the changing morphological appearance of the Altiarchaea population (free-living archaea against biofilm-forming archaea).

Publication 1: Now available as:
Esser, S. P. *et al.* A predicted CRISPR-mediated symbiosis between uncultivated archaea. *Nature Microbiology* **8**, 1619–1633 (2023).

3.2. Publication 1: **A CRISPR-mediated symbiosis between uncultivated archaea predicted from sequencing data**

Sarah P. Esser[1,2,*], Janina Rahlff[2,*,†], Weishu Zhao[3,††], Michael Predl[4,5], Julia Plewka[1,2], Katharina Sures[1,2], Franziska Wimmer[6], Janey Lee[7], Panagiotis S. Adam[2], Julia McGonigle[8], Victoria Turzynski[1,2], Indra Banas[1,2], Katrin Schwank[2,†††], Mart Krupovic[9], Till L. V. Bornemann[1,2], Perla Abigail Figueroa-Gonzalez[1,2], Jessica Jarett[7], Thomas Rattei[4,5], Yuki Amano[10], Ian K. Blaby[7], Jan-Fang Cheng[7], William J. Brazelton[8], Chase L. Beisel[6,11], Tanja Woyke[7], Ying Zhang[3], and Alexander J. Probst[1,2,12,#]

[1]Environmental Metagenomics, Research Center One Health of the University Alliance Ruhr, Faculty of Chemistry, University of Duisburg-Essen, 45151 Essen

[2]Group for Aquatic Microbial Ecology, Environmental Microbiology and Biotechnology, University of Duisburg-Essen, 45141 Essen

[3]Department of Cell and Molecular Biology, College of the Environment and Life Sciences, University of Rhode Island, Kingston, RI 02881, USA

[4]Computational Systems Biology, Centre for Microbiology and Environmental Systems Science, University of Vienna, 1030 Vienna, Austria

[5]Doctoral School in Microbiology and Environmental Science, University of Vienna, 1030 Vienna, Austria

[6]Helmholtz Institute for RNA-based Infection Research (HIRI), Helmholtz-Centre for Infection Research (HZI), 97080 Würzburg, Germany

[7]DOE Joint Genome Institute, Lawrence Berkeley National Laboratory, One Cyclotron Rd, Berkeley, CA, 94720, USA

[8]School of Biological Sciences, University of Utah, Salt Lake City, Utah, USA

[9]Institut Pasteur, Université Paris Cité, CNRS UMR6047, Archaeal Virology Unit, 75015 Paris, France

[10]Nuclear Fuel Cycle Engineering Laboratories, Japan Atomic Energy Agency, Tokai, Ibaraki, 319-1194 Japan

[11]Medical faculty, University of Würzburg, 97080 Würzburg Germany

Publication 1: Now available as:
Esser, S. P. *et al.* A predicted CRISPR-mediated symbiosis between uncultivated archaea. *Nature Microbiology* **8**, 1619–1633 (2023).

[12]Centre of Water and Environmental Research (ZWU), University of Duisburg-Essen, 45141 Essen, Germany

[*]authors contributed equally

[†]present address: Centre for Ecology and Evolution in Microbial Model Systems (EEMiS), Department of Biology and Environmental Science, Linnaeus University, Kalmar, Sweden

[††]present address: Shanghai Jiao Tong University, School of Life Sciences and Biotechnology, International Center for Deep Life Investigation (IC-DLI), Shanghai Jiao Tong University, Shanghai, 200240, China

[†††]present address: University of Regensburg, Biochemistry III, 93053 Regensburg, Germany

[#]corresponding author: alexander.probst@uni-due.de

Publication 1: Now available as:
Esser, S. P. *et al.* A predicted CRISPR-mediated symbiosis between uncultivated archaea. *Nature Microbiology* **8**, 1619–1633 (2023).

SUMMARY

CRISPR-Cas systems defend prokaryotic cells from invasive DNA of viruses, plasmids, and other mobile genetic elements (Horvath and Barrangou, 2010). Capitalizing on multi-omics approaches, we show here that the CRISPR systems of uncultivated archaea also target chromosomal DNA of episymbionts with thousands of spacers over a period of six years and in two distinct ecosystems. A comprehensive analysis of CRISPR-Cas-based infection histories revealed that uncultivated, deep-subsurface archaeal primary producers might defend themselves from archaeal episymbionts of the DPANN superphylum, some of which are known to fuse their cytoplasms with their host (Baker et al., 2010; Comolli and Banfield, 2014; Hamm et al., 2019; Heimerl et al., 2017; Schwank et al., 2019). Using single-cell genomic and metagenomic data, we show that host CRISPR spacer of a CRISPR system I-B match putative essential genes of its episymbiont. Transcriptomics provided evidence for the expression of specific episymbiont-targeting spacers and suggests along with varying spacer compositions over six years an active CRISPR-Cas system *in situ*. However, genome-scale modeling of metabolic interactions between two deep subsurface host-symbiont systems revealed that host cells could also benefit from their symbionts via metabolic complementation. We propose that these uncultivated archaeal episymbionts are either parasitic or mutualistic depending on the respective genotype of the host. By expanding our analysis to 7,012 archaeal genomes, we suggest that CRISPR-Cas targeting of genomes associated with symbiotic archaea evolved independently in various archaeal lineages and might be a widespread phenomenon.

MAIN

Clustered regularly interspaced short palindromic repeats associated systems (CRISPR-Cas) facilitate adaptive prokaryotic immunity via cleavage of mobile genetic elements (MGEs), e.g., viruses and plasmids (Garneau et al., 2010). CRISPR *loci* consist of a series of direct repeat (DR) sequences interspaced by short variable fragments, *i.e.*, spacers, flanked by *cas* genes. Upon exposure to novel MGEs, short DNA fragments from these invaders are incorporated into the CRISPR array as spacers. The spacers are then used as templates to form CRISPR RNAs (crRNAs) that guide effector Cas nucleases to complementary nucleic acid sequences. Spacer

sequences can also be used to study infection histories *in silico* based on matches to protospacers, corresponding nucleic acid regions in the MGE (Andersson and Banfield, 2008).

CRISPR systems exhibit remarkable diversity and functional plasticity including roles in non-defensive functions (reviewed by Koonin and Makarova, 2022). Six main types of CRISPR-Cas systems have been described, including different subtypes, *e.g.*, type I-A to I-F, depending on signature genes and their arrangements (Makarova et al., 2020, 2015). Target identification in type I and II systems is dependent on the recognition of a short protospacer-adjacent motif (PAM) in the target DNA sequence, which elicits cleavage and clearance of the MGE's protospacer. Rather than relying on a defined PAM for target recognition, other CRISPR systems (*e.g.* type III) generally evaluate the extent of hybridization between the flanking portions of the crRNA (called protospacer-flanking sequence) and the target (Maniv et al., 2016; Marraffini and Sontheimer, 2010b). While CRISPR-Cas systems are widely distributed, they are more common in archaea (in ~85% of genomes) than in bacteria (in ~40% of genomes; reviewed by Makarova et al., 2020).

Branching from the archaeal tree of life, the DPANN superphylum including *i.e.*, Diapherotrites, Parvarchaeota, Aenigmarchaeota, Nanoarchaeota and Nanohaloarchaeota, and several other recently proposed phyla (Dombrowski et al., 2019; Rinke et al., 2013), comprises a vast collection of microorganisms remarkably small in size and enigmatic due to the scarcity of cultivated representatives (Castelle et al., 2018; Sakai et al., 2022). Insights into the physiological characteristics of DPANN archaea arise primarily from detailed analyses of co-cultivation with amenable microorganisms (Huber et al., 2002; Jahn et al., 2008) and/or imaging of environmental samples (Schwank et al., 2019). These inferences, along with the limited metabolic potential contained in their comparatively small genomes, suggest that most DPANN archaea exist as (epi-)symbionts of other archaea (Hamm et al., 2019; Huber et al., 2002; Jarett et al., 2018; Munson-McGee et al., 2015; Wurch et al., 2016) or even as intracellular symbionts (Hamm et al., 2023). The independent and autotrophic *Candidatus* Altiarchaeum sp. is host organism to another uncultivated DPANN archaeon, *Candidatus* Huberiarchaeum crystalense (Probst et al., 2018; Schwank et al., 2019). As previous evidence suggested that certain DPANN archaea can fuse their cytoplasm with that of their hosts (Baker et al., 2010; Comolli and Banfield, 2014; Hamm et al., 2019; Heimerl et al., 2017; Schwank et

al., 2019) and even exchange enzymes (Heimerl et al., 2017), we investigated the symbiotic nature of *Ca.* Altiarchaeum and *Ca.* Huberiarchaeum using metagenomics and metabolic modeling in two independent subsurface ecosystems (Fig. 3.2.1A). Based on our results, we suggest that CRISPR-Cas systems play an integral role in mediating archaeal host-DPANN interactions.

Type I CRISPR-Cas system targets genomes of archaeal episymbionts

Two subsurface ecosystems separated by 8,255 km (Fig. 3.2.1A) and derived from different geological formations (Hernsdorf et al., 2017; Probst et al., 2018)*, i.e.*, a Wingate Sandstone-hosted aquifer of the Colorado Plateau at ~ 350 m depth (Crystal Geyser (CG), Utah, USA) (Emerson et al., 2016; Probst et al., 2018, 2017, 2014) and a diatomaceous/siliceous mudstone-hosted aquifer of the Horonobe Underground Research Laboratory (HURL, Hokkaido, Japan) (Hernsdorf et al., 2017) at ~ 250 m depth, were dominated by two species of *Ca.* Altiarchaea (up to 24.5% and 51.6% of the community, respectively). *Ca.* Altiarchaea dominated both ecosystems, and we show their association with cells of *Ca.* Huberiarchaea, their DPANN episymbiont, using species-specific fluorescence *in-situ* hybridization (FISH) (Fig. 3.2.1A; Supplementary Methods). The *Ca.* Altiarchaea genomes retrieved from CG and HURL were shown to encode a I-B CRISPR system and an abundant CRISPR array, which could not be assigned to a specific *cas* gene casette, as has been reported for other *Ca.* Altiarchaea species (Probst et al., 2014; Rahlff et al., 2021). Confidence in assigning the CRISPR-Cas system to its correct metagenome-assembled genome (MAG; Altiarchaea genomes n=1; Table S1-S2) derives from the exceedingly high abundance of *Ca.* Altiarchaea genome fragments in the CG samples (Fig. S3.2.1.1). In addition, within 219 single-cell amplified genomes (SAGs; Altiarchaea SAGs n=7; Table S3) from CG, only *Ca.* Altiarchaea bore the corresponding consensus DR sequence (see Extended Data Fig. 3.2.1 for additional correlation-based evidence), which were remarkably well-conserved across ecosystems (Rahlff et al., 2021) (Fig. 3.2.1B).

Analyses of spacers from *Ca.* Altiarchaeum crystalense detected in 66 CG metagenomes over six years of surveillance (1.07 Tbps of sequencing data, Table S1) revealed 297,531 distinct spacer clusters (Fig. 3.2.1B), indicative of a complex CRISPR spacer repertoire

system for this organism (Figs. S3.2.1.2 and S3.2.1.3). Within these metagenomes, CRISPR type I-B spacers matched the protospacers of 64 viral DNA sequences corresponding to 14 distinct viral genus clusters (Fig. 3.2.1C, Table S6, Extended Data Fig. 3.2.2, Fig. S3.2.1.4-S3.2.1.6, details in Supplementary Results). The PAM sequence 5'-TTN-3' was identified on viral targets matched by type I-B spacers (Fig. S3.2.1.7). However, we were unable to experimentally confirm this PAM using an established PAM assay (Wimmer et al., 2022) or to assess GFP repression (Marshall et al., 2018) using the 5'-TTN-3' PAM in a cell-free transcription-translation (TXTL) system (Wimmer et al., 2022) (Supplementary Material and Methods, Section 3.2.1), likely due to the divergent settings (including temperature) of the host environment compared to those used in the established assay (Wimmer et al., 2022).

The finding that all virus-matched spacers detected in the exhaustive CG survey derived from the CRISPR I-B system and the ubiquitous nature of this system in *Ca.* Altiarchaea worldwide (Rahlff et al., 2021) suggests that the I-B system serves as a primary line of defense against viruses infecting these archaea. A substantial fraction of the spacers matched microbial genomes, including those of *Ca.* A. crystalense, *i.e.*, its own genome (self-targeting, up to 2.9% in sample CG16) and of its episymbiont *Ca.* Huberiarchaeum crystalense (up to 2.8% in sample CG08; Fig. 3.2.1C-3.2.1D, Fig. 3.2.2A-D). The relative proportion of spacers matching the episymbiont was greater than that matching the host genome (Fig. 3.2.1C-D), indicative of biased acquisition, negative selection of self-targeting spacers or a positive selection for spacers from the episymbiont genome. The positions of these spacers in the CRISPR I-B array encoded in an altiarchaeal SAG suggest that these spacers prevailed in the system for extended periods (Fig. 3.2.2D). While 17% of the protospacers self-targeted through the I-B system showed a significant decrease in metagenomic coverage compared to untargeted scaffold regions (bootstrapped Wilcoxon paired signed rank test, target sites = 196, FDR-corrected p-value < 0.05), 30% of the I-B protospacers in *Ca.* H. crystalense genomes showed a significant drop in coverage, suggesting *in situ* targeting of the episymbiont in CG (bootstrapped Wilcoxon signed rank test, target sites = 73, FDR-corrected p-value < 0.05; Supplementary Results, Section 3.2.1, Extended Data Fig. 3.2.3). The coverage of the significantly different targeted regions, compared to the average coverage of the scaffold decreases in *Ca.* A. crystalense and *Ca.* H. crystalense by 10.74% (median) and 36.99%

(median), respectively (bootstrapped Wilcoxon signed rank test, n=990, FDR-corrected p-value < 0.05; details in Supplementary Results, Section 3.2., Table S5). Supporting this difference, the PAM sequence detected next to the protospacers in *Ca.* H. crystalense was identical to that of the virus-targeting PAM sequence (Fig. S3.2.1.7). Coverage drops as observed herein could also arise from misassemblies, regions excised in subpopulations, or elevated SNPs resulting in low recruitment of reads.

In contrast to the conserved PAM in the episymbiont and the viruses, the self-targeted protospacer regions were not associated with the 5'-TTN-3' PAM (Fig. S3.2.17). As shown for other microbial communities, self-targeting can result in cell suicide (reviewed in Heussler and O'Toole, 2016) or transcriptional regulation (Stern et al., 2010) of genes influencing the fitness of the microbial population and can thus reduce the strain variation within an ecosystem (Aklujkar and Lovley, 2010). However, the lack of the PAM, the essential motif for successful targeting of DNA by CRISPR system type I (reviewed in Bhaya et al., 2011) in the population genomes of *Ca.* Altiarchaea, might on the one hand prevent subpopulations of *Ca.* Altiarchaea from cell death by autoimmunity. On the other hand, the correct PAM could still lead to cell death in subpopulations, given that the PAM has not been silenced by mutations. Based on the overall results from metagenomics and metatranscriptomics we suggest that CRISPR-Cas systems may function similarly against viral DNA and chromosomal DNA of episymbionts.

Previous investigations, which were based on either species-specific FISH or electron microscopy, indicate that many DPANN archaea (including *Ca.* A. crystalense and its episymbiont) fuse their cytoplasms (Baker et al., 2010; Comolli and Banfield, 2014; Hamm et al., 2019; Heimerl et al., 2017; Schwank et al., 2019). This direct interaction of the host's and the symbiont's cytoplasms, and a potentially predatory nature of the symbiont (Probst et al., 2018; Schwank et al., 2019), likely underlie the evolution of a direct assault on the episymbiont's genome by the Altiarchaeota CRISPR system (Fig. 3.2.1A). To this end, we annotated genes of *Ca.* H. crystalense targeted by *Ca.* A. crystalense's CRISPR type I-B system and identified several hypothetical proteins, proteins lacking annotation, and non-coding genomic regions (these categories sum up to 98.25%). Targeted genes included a CTP synthase and a DNA methyltransferase N-4/N-6 domain protein (Fig. 3.2.2C, Table S5).

Methyltransferases protect DNA against cleavage by restriction enzymes (Wilson, 1991). Inactivation of such a methyltransferase might increase vulnerability of the episymbiont towards enzymatic cleavage by the host.

CRISPR-Cas targeting of similar episymbiont genomes in two independent subsurface ecosystems

Targeting of episymbiont's genomes by altiarchaeotal CRISPR spacers was also observed in the HURL ecosystem. In contrast to *Ca.* A. crystalense's CRISPR-Cas I-B dependent targeting of *Ca.* H. crystalense genomes in the CG environment, *Ca.* Altiarchaeum horonobense found within the HURL ecosystem appeared to employ CRISPR spacers of an unassigned array (*i.e.*, no *cas* genes in direct vicinity could be detected due to genome fragmentation but the DR sequence is identical to type III CRISPR-Cas systems of other Altiarchaea (Rahlff et al., 2021)) to potentially ward off *Ca.* Huberiarchaeum julieae episymbionts and viral invaders (Fig. 3.2.1D). While spacers of this unassigned array targeting the *Ca.* H. julieae's genome exhibited greater diversity compared to the self-targeting counterparts of *Ca.* A. horonobense's (Fig. 3.2.1D), their relative abundance in the metagenome was nearly two-fold lower (Fig. 3.2.1E). The ecosystem-specific involvement of CRISPR-Cas I-B along with the unassigned array targeting of the episymbiont genomes in two distinct subsurface ecosystems seemingly indicates an independent evolution of defense against intruding DNA, which aligns with previous investigations that demonstrated a strict biogeography of Altiarchaea core genomes and site-specific evolution (Bornemann et al., 2022). Given the site-specific evolution of *Ca.* Altiarchaea an alternative explanation for the acquisition of spacers against foreign chromosomal DNA might be avoidance of spoilage of the host chromosome by intruding genes (horizontal gene transfer). In Haloarchaea, such a mechanism has been shown to indirectly control for unwanted horizontal gene transfer between strains of the same genus (Turgeman-Grott et al., 2019).

Spacers of the unassigned CRISPR array were detected in much greater diversity than those of CRISPR I-B systems at the HURL site (Fig. 3.2.1D-E). While spacers of the unassigned array of CG-derived *Ca.* A. crystalense self-target chromosomal gene sequences, the spacers of the unassigned array of *Ca.* A. horonobense's self-target intergenic regions (Fig. 3.2.1C,

Table S7). Notably, it has been demonstrated in haloarchaea that self-targeting does not necessarily lead to cell suicide (Stachler et al., 2017). Assuming that the CRISPR-Cas interference is associated with a defense against the symbiont, a plethora of spacers present at CG might effectively repress the symbiont (host:symbiont = 11:1 based on metagenomic read mapping), while a lower abundance of spacers targeting *Ca.* H. julieae at HURL was associated with a higher presence of episymbionts (host:symbiont = 6:1).

Episymbionts complement metabolic demands originating from self-targeting of the host

We applied genome-scale metabolic modeling to examine the different symbiotic interactions of *Ca.* Altiarchaea and *Ca.* Huberiarchaea implicated by variations in the CRISPR systems and host-symbiont ratios of the two ecosystems analyzed. MAGs (ten genomes of *Ca.* A. crystalense, ten of *Ca.* H. crystalense, one of *Ca.* A. horonobense, one of *Ca.* H. julieae), SAGs (seven of *Ca.* A. crystalense and one of *Ca.* H. crystalense) and transcriptomic data (extracted spacers of samples CG05, CG08 and CG16 from 2015) from CG and HURL environments were used to render genome-scale metabolic reconstructions. Although we applied thorough manual data curation(Emerson et al., 2016; Probst et al., 2018), the genomes were fairly fragmented (average $N50_{host/CG}$ = 8067.24, average $N50_{symbtiont/CG}$ = 14983.73, average $N50_{host/HURL}$ = 3604, average $N50_{symbtiont/HURL}$ = 4115), and missing information due to fragmentation or binning errors cannot be excluded.

A consensus model was created for each ecosystem to cogently summarize and compare the metabolic capacities of *Ca.* Altiarchaeum and *Ca.* Huberiarchaeum, and constraint-based modeling of these metabolic networks facilitated an assessment of host-symbiont metabolic complementarity (Fig. 3.2.2E-F, Fig. S3.2.1.8). Models of the CG and HURL environment both revealed a significant reliance of *Ca.* Huberiarchaea upon its host's metabolism yet little to no dependency of the host upon the metabolism of *Ca.* Huberiarchaea. For example, glucose, amino acids, vitamins, and energy carrying compounds like adenosine triphosphate (ATP) were transferred from *Ca.* Altiarchaeum to *Ca.* Huberiarchaeum in both models (Table S8-S11; details in Supplementary Results, Section 3.2.1), supporting the notion that *Ca.* Altiarchaeum is a primary producer, while *Ca.* Huberiarchaeum relies on its host for carbon and energy sources (Schwank et al., 2019).

Analyses of CG and HURL host-symbiont relationships also revealed highly variable metabolic collaborations between episymbionts and their hosts. In the CG ecosystem, a deoxycytidylate monophosphate (dCMP) deaminase was absent in *Ca.* A. crystalense but present in *Ca.* H. crystalense. This gene is essential to reach a non-zero biomass for *Ca. A.* crystalense in the model (see Supplementary Results), suggesting a collaborative effort of synthesizing pyrimidine (Fig. S3.2.1.8C). Similarly, HURL-borne *Ca.* A. horonobense genomes lacked deoxythymidine monophosphate (dTMP) synthase genes, while these genes were present in the genomes of *Ca.* H. julieae – once again implicating collaboration, namely in folate biosynthesis (Fig. S3.2.1.8C; Fig. 3.2.3). At HURL, the self-targeting of genes in *Ca.* Altiarchaea did not impact the host's dependency on the symbiont's metabolism within both CRISPR systems (Fig. S3.2.1.8, Supplementary Results, Section 3.2.1). At CG, however, eliminating the functions of genes self-targeted by the I-B system in metabolic models exposed additional modes of complementing *Ca.* A. crystalense's metabolic demands by *Ca.* Huberarchaea via lysyl-tRNA synthetases and phenylalanyl-tRNA synthetases (Fig. 3.2.2D-E, Fig. 3.2.3, and Fig. S3.2.1.8A, and C-F). The respective protein sequences were not horizontally transferred between *Ca.* Altiarchaea and *Ca.* Huberiarchaea based on phylogenetic analyses; instead, the phenylalanyl-tRNA synthetase of *Ca.* Huberiarchaeum can be traced back with strong confidence to *Ca.* Woesarcheaota and *Ca.* Pacearchaeota (Supplementary Data).

While the protospacers of *Ca.* Altiarchaea viruses and *Ca.* H. crystalense harbored a definitive PAM (5'-TTN-3'associated with other I-B systems (Vink et al., 2021), no such clear motif was detected in the host protospacers. Here, the second base of the putative PAM region, exhibited a four-fold greater than the average single nucleotide polymorphism (SNP) rate of genes (Fig. S3.2.1.9; details in Supplementary Results, Section 3.2.1). Mutations in the PAM region diverging from the 5'-TTN-3' motif would prevent self-targeting at least for parts of the altiarchaeal population (Pyenson et al., 2017) and thus protect the host chromosome from CRISPR-Cas-mediated cleavage. In our model, removal of self-targeting would lessen the metabolic dependence on the symbiont and enable subpopulations of *Ca.* Altiarchaea to flourish more independently. The missing PAM sequence for self-targeting spacers and the increased SNP-rate in such regions compared to those targeting the episymbiont suggest that the population of *Ca.* Altiarchaea is adapting to escape the dependency of the symbiont.

Considering that acquisition of self-targeting spacers is a stochastic process (Chabas et al., 2022), escape mutations or deletions within the essential targeted genes could have detrimental effects on the cell viability due to the deficits in the corresponding metabolic activities resulting in cell suicide (reviewed in Heussler and O'Toole, 2016). Episymbionts could provide a temporary relief to the host cell by complementing the metabolic deficiency, becoming a bona fide symbiont, at least until the metabolic autonomy of the host is reestablished. We thus hypothesize that interactions between hosts and episymbiont depend on the genotype of the host and can consequently be either mutualistic or parasitic. However, cultivation of the host-symbiont system along with establishing a genetic system to modify the host genome are necessary to test this hypothesis.

Applicability of findings to decipher inter-phylum interactions of other symbiotic archaea

To facilitate the overlay of our findings on other potential archaeal host-DPANN episymbiont relationships, we analyzed CRISPR spacer matches between all archaeal genomes publicly available in NCBI's GenBank (7,012 genomes: state May, 2021, Table S4). After having extracted 106,641 spacer sequences, 39,875 distinct spacer-to-protospacer matches across all genomes were detected. Few contigs carrying CRISPR arrays (*e.g.,* for *Ca.* Micrarchaeum) also contained taxonomic hallmark genes, such as those coding for DNA-directed RNA polymerase subunit or ribosomal proteins, which provided additional confidence for the correct assignment of spacers to the fragmented public MAGs. The spacer hits accounted for both self-targeting and interspecies spacer interactions (Extended Data Fig. 3.2.4). Network analyses showed the genomes of the DPANN *Ca.* Aenigmarchaeota and *Ca.* Altiarchaeota (Fig. 3.2.4), as well as *Sulfolobus*, *Methanosarcina, Haloferax,* and *Halobacterium* spp. forming large clusters resulting from a wealth of interspecies hits and/or self-targeting (Extended Data Fig. 3.2.4), which was also previously shown for other archaea (Brodt et al., 2011; Turgeman-Grott et al., 2019). Well-established DPANN-host co-cultures, *e.g. Ignicoccus hospitalis* and *Nanoarchaeum equitans* (Paper et al., 2007), did not exhibit CRISPR-Cas-derived targeting to either of the symbionts in our archaeal genome dataset.

Particularly for the hydrothermal system of Guaymas Basin, Gulf of California (Dombrowski et al., 2018), our approach enabled the *a priori* prediction of DPANN-host

interactions based on CRISPR-Cas genome targeting (Fig. 3.2.4). Analyses of the spacer-protospacer matches from the read data of Guaymas Basin revealed frequent targeting (160 spacer-protospacer matches) of *Ca.* Aenigmarchaeota by *Ca.* Bathyarchaeota. Genes targeted by these spacer-protospacer matches, *e.g.,* encode for the LamGL domain-containing protein, which is *inter alia* responsible for the binding of sulfated glycolipids (Hohenester, 2019; Hohenester and Yurchenco, 2013), and the ribonucleoside triphosphate reductase, amenable for catalysis of the conversion of ribonucleotides into deoxyribonucleotides (Benner et al., 1989).

Another novel host-DPANN interaction unveiled by these analyses involves *Ca.* Micrarchaeota spacers matching a Thermoprotei archaeon, with both of these genomes arising from the same ecosystem but a few centimeters apart in depth (Dombrowski et al., 2018). Comparing those targeted gene-encoding regions to the targeted genetic regions in *Ca.* Huberiarchaeum by spacers of *Ca.* Altiarchaeum (CTP synthase and DNA methylase, see above), no acquisition pattern of spacers directed against genomic regions that encode specific functions could be detected. Overall, these findings suggest spacer-protospacer matches are a useful tool for identifying *in-silico* host-symbiont interactions of uncultivated archaea based on metagenomic analyses.

The findings discussed here demonstrate that archaeal CRISPR-Cas systems acquire resistance not only to genomes of foreign MGEs (Garneau et al., 2010) and closely related species (Stern et al., 2010) but also to archaea of other phyla, particularly episymbionts belonging to the DPANN superphylum. Our results suggest that CRISPR-Cas-mediated adaptive immunity might lead to complex interactions between the host and symbiont at the population level, possibly drawing the host into maintaining a collaborative relationship with the symbiont due to balancing the self-targeting nature of the host's CRISPR system and the potential defense against the episymbiont. Based on our results from single-cell genomic data, metagenomes, and metatranscriptomes, we suggest that a double-edged sword drives the evolution of microbial populations, *i.e.*, CRISPR-Cas-mediated defenses likely render a major fraction of the DPANN episymbiont population truly parasitic, while the remainder seem to support the host in a mutualistic fashion.

Publication 1: Now available as:
Esser, S. P. *et al.* A predicted CRISPR-mediated symbiosis between uncultivated archaea. *Nature Microbiology* **8**, 1619–1633 (2023).

ACKNOWLEDGEMENTS

This effort was funded by the Ministerium für Kultur und Wissenschaft des Landes Nordrhein-Westfalen ("Nachwuchsgruppe Dr. Alexander Probst") and the German Science Foundation under project NOVAC (grant number DFG PR1603/2-1) and through SPP 2141 (grant number DFG BE 6703/1-1). Genome-scale metabolic modeling was supported by the National Science Foundation under Grant No. 1553211. The Ministry of Economy, Trade and Industry of Japan funded a part of the work as "The project for validating assessment methodology in geological disposal system" (2019 FY, Grant Number: JPJ007597). The work (proposal: 10.46936/10.25585/60000800) conducted by the U.S. Department of Energy Joint Genome Institute (https://ror.org/04xm1d337), a DOE Office of Science User Facility, is supported by the Office of Science of the U.S. Department of Energy operated under Contract No. DE-AC02-05CH11231. P.S.A is supported by a postdoctoral fellowship from the Alexander von Humboldt Foundation. J.P. was supported by Lundin Energy Norway AS within the framework of the GeneOil Project. J.R. was supported by the German Science Foundation (grant number RA3432/1-1, project number 446702140). Support by the German Federal Ministry of Education and Research within the project "MultiKulti" (BMBF funding code: 161L0285E) is acknowledged. We thank Ken Dreger for exemplary server administration, and Bettina Siebers, Ivan Berg, Jillian F. Banfield and Benjamin Meyer for insightful discussion.

CONFLICT OF INTEREST

The authors declare no conflict of interest.

METHODS

Data availability. Metagenomic datasets generated from the Crystal Geyser (CG) (Emerson et al., 2016; Probst et al., 2018) ecosystem (Utah, USA) in 2009, 2014, and 2015 (n = 66), and the Horonobe Underground Research Laboratory (HURL) (Hernsdorf et al., 2017) (Hokkaido, Japan) environment (n = 2) were downloaded from the NCBI' Sequence Read Archive (SRA) in April 2019 (Table S1). SAGs generated in a previous study (Probst et al., 2018) (n = 219) were retrieved from the JGI's Integrated Microbial Genomes and Microbiomes database (Chen et al., 2019) (Table S3). The metagenome-derived genomes of *Ca.* A. crystalense and *Ca.* H. crystalense from CG are publicly accessible from NCBI (accession numbers in Table S2). The genomes of *Ca.* A. horonobense and *Ca.* H. julieae from HURL were newly reconstructed in this investigation (Table 2). All unpublished genomes used in this study are available in a Figshare folder (10.6084/m9.figshare.22339555).

Metagenome assembly and genome reconstruction. For all metagenomic datasets of CG and HURL, quality filtering and trimming of reads was done using BBduk (https://github.com/BioInfoTools/BBMap/blob/master/sh/bbduk.sh) and Sickle (Joshi and Fass, 2011). The MetaSPAdes (Nurk et al., 2017) (version 3.10) and Bowtie2 (Langmead and Salzberg, 2012) utilities (--sensitive) were applied to assemble reads and estimate coverage, respectively. Scaffolds < 1 kbp were excluded from further analysis. The interactive uBin (Bornemann et al., 2023) software was used to segregate the genomes of *Ca.* Altiarchaeum horonobense and *Ca.* Huberiarchaeum julieae based on %GC content, taxonomy, and coverage information. Previously published *Ca.* Huberiarchaeum genomes generated from each of the CG and HURL environments were used as probes to identify respective scaffolds at the protein level (≥80% similarity).

Phylogeny of Altiarchaeum and Huberiarchaeum. A reference dataset spanning the diversity of 176 archaeal genomes was used to place *Ca.* Huberarchaeota and *Ca.* Altiarchaeota phylogenetically. To avoid redundancy, all genomes annotated as *Ca.* Altiarchaeota on NCBI (June 2019), previously published *Ca.* Altiarchaeota genomes (Bornemann et al., 2022), and one representative genome from *Ca.* Altiarchaeum and *Ca.* Huberiarchaeum were consolidated for this work. Individual homology searches were executed across these datasets, using HMMER 3.2.1 (Eddy, 2011) with the Phylosift (Darling et al., 2014) marker

HMM profiles and an e-value cutoff of 1 x 10^{-5}. All DNA sequences were aligned with MUSCLE v3.8.31 (Edgar, 2004) (default parameters) and manually curated to fuse fragmented genes and remove distant homologs and paralogous copies. One *Ca.* Altiarchaeota genome (GCA_003663105) was likely contaminated and thus removed from the final alignments. Sequence sets resulting from each of the four datasets were fused together (36 single-gene datasets; one of the 37 Phylosift marker genes (DNGNGWU00035) was omitted due to many missing taxa), realigned as before, trimmed with BMGE (BLOSUM30)(Criscuolo and Gribaldo, 2010), and concatenated into one supermatrix (200 taxa; 5,974 amino acid positions). Phylogenies were reconstructed with IQ-TREE 2 (Minh et al., 2020) (v2.0.5), first using ModelFinder (Kalyaanamoorthy et al., 2017), then using that phylogeny as a guide, with the PMSF model (Wang et al., 2017) (LG+C60+F+G). Branch supports were calculated using 1,000 ultrafast bootstrap (Hoang et al., 2017) and 1,000 SH-aLRT (Guindon et al., 2010) replicates and the aBayes (Anisimova et al., 2011) test and trees were visualized in iTOL (Letunic and Bork, 2019) (version 5).

RNA extraction and metatranscriptomic sequencing. Samples for transcriptomics were collected along with DNA samples as previously published (Probst et al., 2018). For the samples CG05, CG08 and CG16 we filtered approx. 189, 151, and 151 L of geyser-erupted water, respectively. The MoBio PowerMax Soil DNA kit, now re-branded as the Qiagen DNeasy PowerMax Soil kit (Qiagen, Germantown, MD), was used to perform all metagenomic RNA extractions. Filters were aseptically cut into pieces, and 20 mL of lysis buffer from the kit was added for removal of cells from the filters. The manufacturer's alternative protocol, entitled "Alternative PowerMax Protocol for Isolation of RNA and DNA from Low Biomass Soil with Low Humics" was adjusted as follows: briefly, 10 mL of Bead Solution was added to the thawed filter and vortexed at maximum speed for 5 minutes to remove cells. The cell solution was transferred to a bead tube, and 5 ml of phenol:chloroform:isoamyl alcohol (25:24:1), pH 6.6, was added and homogenized by vortexing for 10 minutes. The manufacturer's protocol was followed thereafter. The metagenomic RNA extracts underwent DNase treatment using the Qiagen DNase Max kit (Qiagen), following manufacturer's standard protocol. Quality control and quantification of all RNA extracts were performed using the Agilent Bioanalyzer RNA 6000 Nano kit (Agilent, Santa Clara, CA). Sequencing libraries were created using the

Illumina TruSeq Stranded mRNA Library Prep Kit, following manufacturer's protocol (Illumina, San Diego, CA). Libraries were sequenced on the Illumina HiSeq 2500 platform (Illumina).

CRISPR system extraction and viral sequence determination. The CRISPR systems of 18 distinct *Ca.* Altiarchaeum crystalense genomes (Probst et al., 2018, 2017) (Table S2) and one *Ca.* Altiarchaeum horonobense genome (Hernsdorf et al., 2017) (Table S2) were extracted with CRISPRCasFinder (Couvin et al., 2018) (ver. 1.2), and annotated *cas* genes were used to identify CRISPR-Cas cassettes. Two resulting consensus DR sequences were used as input for MetaCRAST (Moller and Liang, 2017) (-d 1 -c 1 -a 1 -h -r), analysis of metagenomic reads, metatranscriptomic reads, and single cell genome reads. Only spacers having adjacent repeat sequences bearing 100% similarity with the respective read were considered. All spacers shorter than 24 bps, longer than 57 bps, or harboring homopolymers of six or more identical bases in a row were excluded. Spacers were clustered to 97% nucleotide identity using CD-hit (Fu et al., 2012) and respective centroid sequences were used in downstream analyses.

To check if spacers were biased towards matching genome transcripts, the orientation of the CRISPR array was confirmed on all available *Ca*. Altiarchaeum crystalense genomes to identify the forward strand that corresponds to CRISPR-RNA by using CRISPRDirection2.0 with default settings (Biswas et al., 2014). To avoid false positive predictions of self-targeting and episymbiont targeting, we masked prophage region, predicted by VirSorter (Roux et al., 2015) (category 1-3, 4-6) and transposon regions, predicted by ISEScan (Xie and Tang, 2017). The spacers from this analysis were blasted (nucleotide blast, bidirectional [default setting] and unidirectional [-strand plus] on the forward strand) against the CDS data of 18 *Ca.* Altiarchaeum crystalense genomes (including seven SAGs), eleven *Ca.* Huberiarchaeum crystalense (including 2 SAGs), one genome of *Ca.* Altiarchaeum horonobense and *Ca.* Huberiarchaeum julieae, respectively. All unpublished viral genomes used in this study are deposited in the Figshare folder (10.6084/m9.figshare.22738568).

Detection, dereplication, and genomic analyses of DNA viruses were performed as described previously (Rahlff et al., 2021). Please see supplementary methods for details.

Metabolic modeling. Consensus models of *Ca.* Altiarchaeum to *Ca.* Huberiarchaeum were constructed based on collections of MAGs and SAGs from CG (20 MAGs, 8 SAGs) and HURL (2 MAGs) (Table S2) to capture the metabolic potential of each population. Candidate genes

were first identified based on a pangenome analysis, which was performed following ortholog identification using a bidirectional best hit approach (Zhang and Sievert, 2014). All representative genes from the MAGs or SAGs of a given ecosystem served as candidates for that ecosystem's metabolic reconstruction. Complementary metabolic characteristics were identified between *Ca.* Altiarchaeum and *Ca.* Huberiarchaeum via a *fastgapfill* implementation in the PSAMM software package (Dufault-Thompson et al., 2018; Steffensen et al., 2016). Simulations targeted the growth optimization of *Ca.* Altiarchaeum while applying the metabolic functions of *Ca.* Huberiarchaeum as a reference, which facilitated the identification of *Ca.* Huberiarchaeum-encoded complementary functions essential for *Ca.* Altiarchaeum - and vice versa. Combined *Ca.* Altiarchaeum and *Ca.* Huberiarchaeum metabolic models were formulated with exchange constraints representative of environmental *in situ* geochemical measurements corresponding to either CG or HURL (Table S9 and S10). Comparative analyses based on computational simulations were carried out both in the presence and absence of targeted functions. This enabled the identification of changes in metabolite transfer and/or metabolic collaboration between *Ca.* Altiarchaeum and *Ca.* Huberiarchaeum (Fig. S3.2.1.8) upon targeting specific genes with spacers.

PAM analysis of *Ca.* Altiarchaeota, *Ca.* Huberarchaeota, and viruses. Applying CRISPRTarget (Biswas et al., 2013) (accessed in June 2020) with default settings, protospacer adjacent motifs (PAMs) were identified within the genomes of *Ca*. Altiarchaeota, *Ca.* Huberarchaeota, and viruses using spacers bearing 80% sequence similarity. CRISPRTarget results were screened with WebLogo (Crooks et al., 2004; Schneider and Stephens, 1990) in batches of 10,000 8-bps sequences.

SNP analysis. To identify *Ca.* Altiarchaeum crystalense SNPs, reads from samples CG05, CG08, and CG16 (samples for which also transcriptomic datasets were available, and which were used in the metabolic modeling) were aligned to nine different MAGs (Table S2) and analyzed individually by using BBMap (https://sourceforge.net/projects/bbmap/) (default parameters). SNPs were predicted using the VarScan (Koboldt et al., 2012) pileup2snp command (v2.4.3; default settings) with observations and coverage thresholds set to a minimum of two and eight, respectively. SNPs bearing the reference allele 'N' were excluded if all base called reads showed this 'N'.

CRISPR-Cas interactions across archaeal diversity. All archaeal genomes housed in the publicly accessible NCBI database (May 2021; Table S4) were screened for viral sequence contaminants using VirSorter (Roux et al., 2015) (default settings), and all respective hits, as well as annotated plasmids, were excluded from consideration. The CRISPRCasFinder (Couvin et al., 2018) utility was used to extract spacers, DR, and *cas* genes from each genome individually with the help of the *cas* gene database (-ArchaCas). All CRISPR arrays detected were masked in their respective genomes to avoid false positives, and spacers were filtered for homopolymers and sequence length as described above. All spacer sequences were queried (Altschul et al., 1990) against all archaeal genomes to an 80% nucleotide similarity threshold, and interactions between genomes based on CRISPR spacer matches were visualized in Cytoscape (Shannon et al., 2003). The taxonomy of each genome was pulled from the NCBI taxonomy database and in single cases validated using Genome Taxonomy Database (Chaumeil et al., 2020; Parks et al., 2020, 2018)(GTDB-Tk classify, version v0.3.3, database r89). To avoid false positive predictions of self-targeting and episymbiont targeting, we masked prophage region, predicted by VirSorter (Roux et al., 2015) (category 1-3, 4-6).

CRISPR binding, cleavage and PAM assay. For details, please refer to the Supplementary Methods.

Authorship contributions

SPE and AJP performed genome-resolved metagenomics, while SPE and JR performed viromics. JR analyzed viral genomes with input from MK. CRISPR-Cas analyses were done by SPE, JR and AJP. SNP analysis was performed by MP and TR. Genome-scale modeling was conducted by WZ and YZ with input from SPE, PAFG, and AJP. Phylogenomic analyses were carried out by PSA. TLVB provided bioinformatic assistance, and KSch and VT performed microscopy and initial metabolic analyses. JM and WB re-sampled Crystal Geyser and, JL, TW, and AJP conducted RNA extraction and sequencing, and SPE analyzed transcriptomes. FW and CB performed binding, cleavage and PAM assays and JL, JJ, YA, TW, and AJP generated/provided raw data. KS and SPE analyzed the archaeal CRISPRCas interactions from published NCBI archaeal genomes. SPE, JR, WZ, YZ, and AJP wrote the manuscript with input from all authors.

Publication 1: Now available as:
Esser, S. P. *et al.* A predicted CRISPR-mediated symbiosis between uncultivated archaea. *Nature Microbiology* **8**, 1619–1633 (2023).

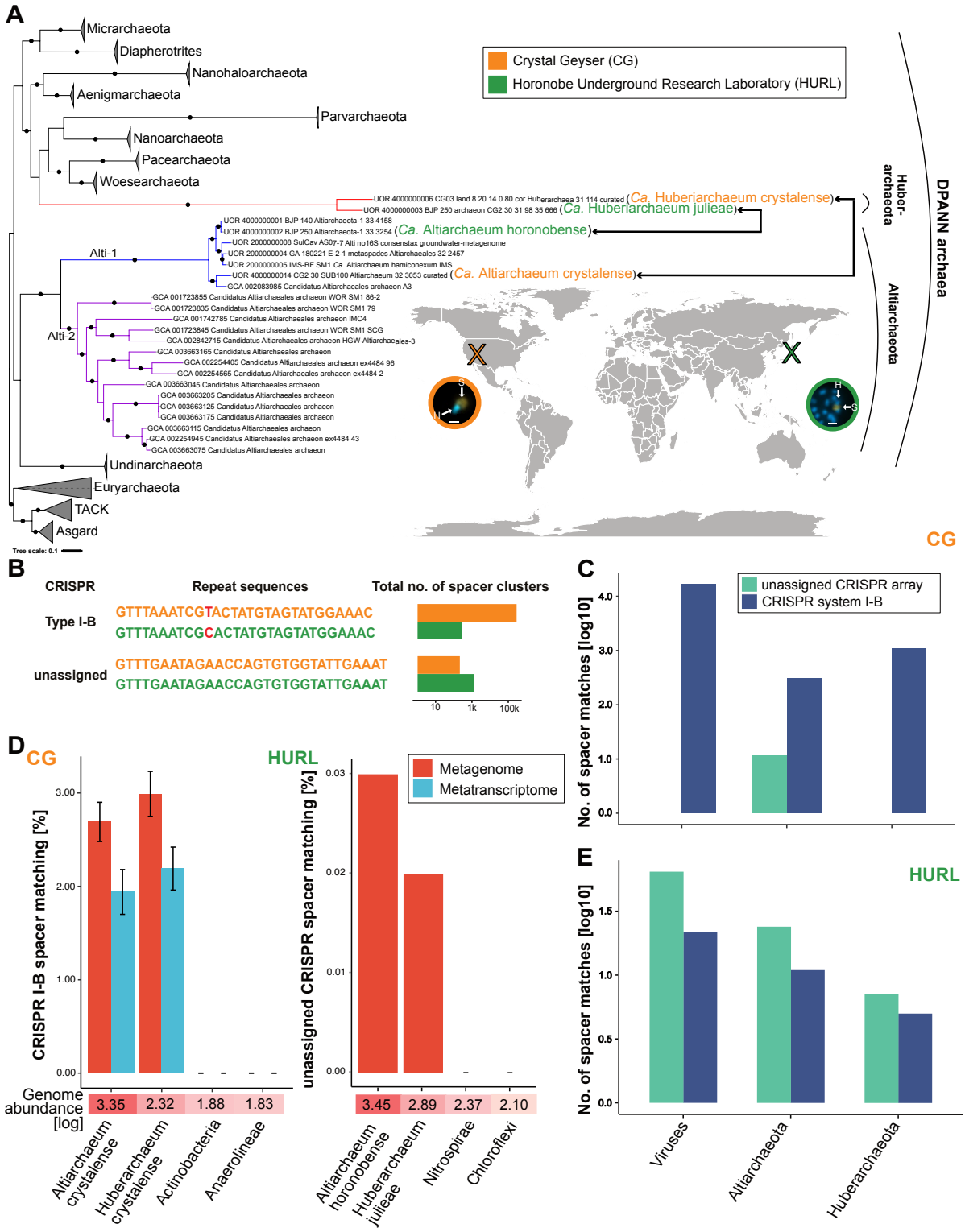


Figure 3.2.1. Phylogenetic positioning of Ca. Altiarchaeota and Ca. Huberarchaeota, sampling locations, FISH analysis, and CRISPR-Cas targets | (A) *Phylogenetic tree of archaea highlighting Candidatus Altiarchaeum and Candidatus Huberiarchaeum of the sampling*

locations Crystal Geyser (CG, Utah, USA, orange) and Horonobe Underground Research Laboratory (HURL, Hokkaido, Japan, green). Fluorescence pictures are highlighted with circles colored according to the sapling site and show Ca. Altiarchaeum (blue; H - host) as host and its episymbiont Ca. Huberiarchaeum (orange; S - symbiont) in the respective ecosystems. Scale bar 1 µm. ***(B)*** *Ca. Altiarchaeota CRISPR systems, their associated direct repeat (DR) sequences, and the number of spacer clusters arising from the two sampling sites. DR sequences (presented in forward orientation) are conserved within the unassigned CRISPR system between CG and HURL with the exception of a point mutation in the CRISPR system I-B DR (red). Spacer clusters were calculated at 97% nucleotide identity. (**C**) Logarithmic number of centroid spacers derived from spacer clusters matching 64 extracted viral sequences (total number of spacer matches: 0 of unassigned CRISPR system and 16561 of CRISPR system IB), 17 binned genomes of Ca. Altiarchaeum crystalense (total number of spacer matches: 115 of unassigned CRISPR system and 1,311 of CRISPR system IB) and 11 binned genomes of Ca. Huberiarchaeum crystalense (total number of spacer matches: 0 of unassigned CRISPR system and 1,445 of CRISPR system IB) originating from the CG site (Table S2). Spacers were derived from the complete 66-sample metagenomic dataset. (**D**) Percentage of CRISPR system I-B spacer cluster abundances matching to organisms that were previously detected in this ecosystem at the CG site. Listed are the logarithmic genome abundances of the respective organisms. Error bars denote the standard deviation of the abundance of matching spacer clusters for samples CG05, CG08, and CG16 of the year 2015. These were displayed because also transcriptomic data was available. The dataset of HURL is referring to one metagenome, as no other data was available. (Means and standard deviation: CG Altiarchaeum crystalense: 2.69 ± 0.21, 1.93 ± 0.24; Huberiarchaeum crystalense: 2.99 ± 0.23, 2.19 ± 0.23; HURL Altiarchaeum horonobense: 0.029; Huberiarchaeum julieae: 0.019) (**E**) Logarithmic number of centroid spacers derived from spacer clusters matching extracted viral sequences (total number of spacer matches: 64 of unassigned CRISPR system and 22 of CRISPR system IB), two binned genomes of Ca. Altiarchaeum horonobense (total number of spacer matches: 19 of unassigned CRISPR system and 2 of CRISPR system IB) and one binned genome of Ca. Huberiarchaeum julieae (total number of spacer matches: 7 of unassigned CRISPR system and*

0 of CRISPR system IB) originating from the HURL site. Spacers were derived from one metagenomic dataset.

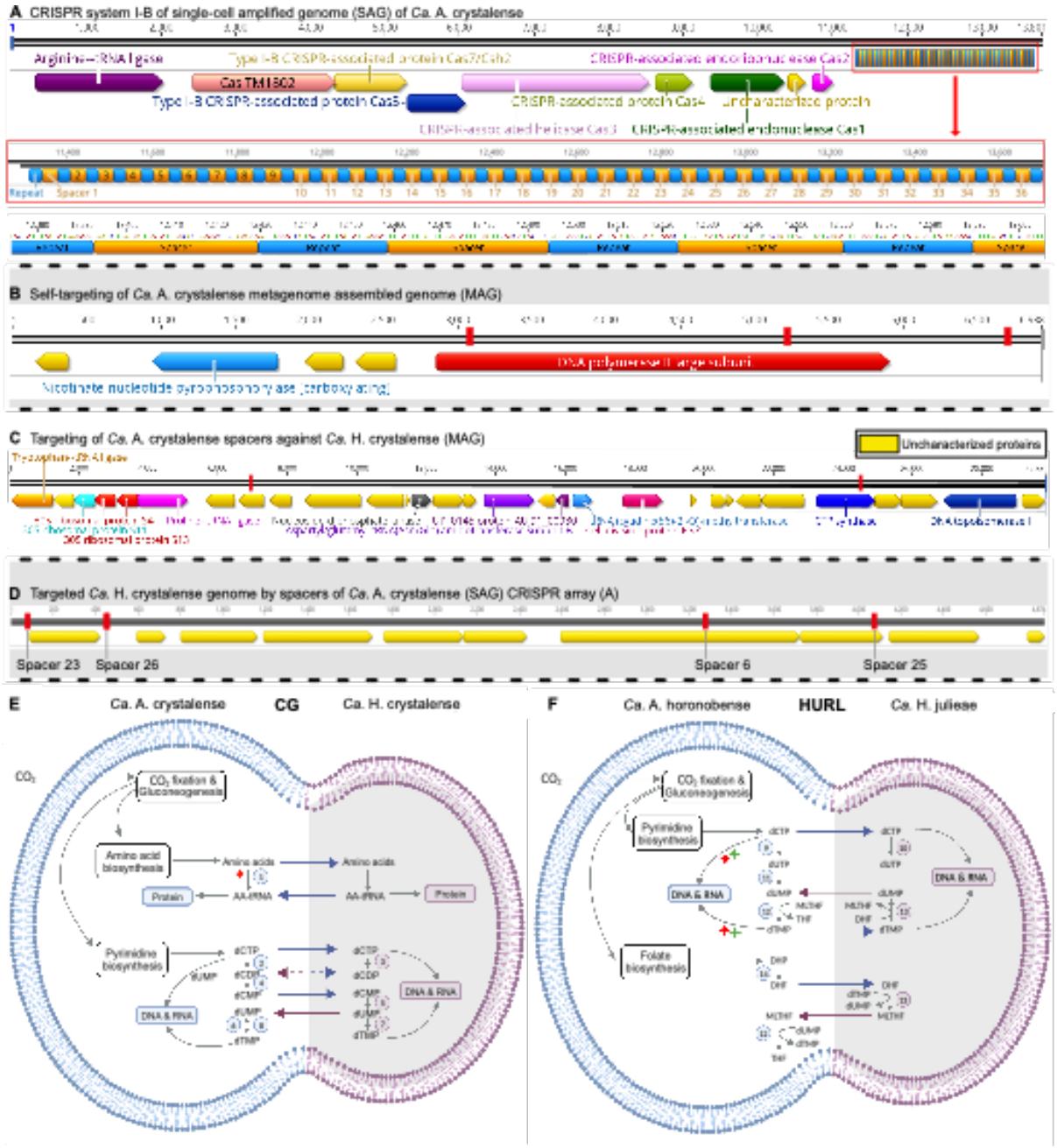
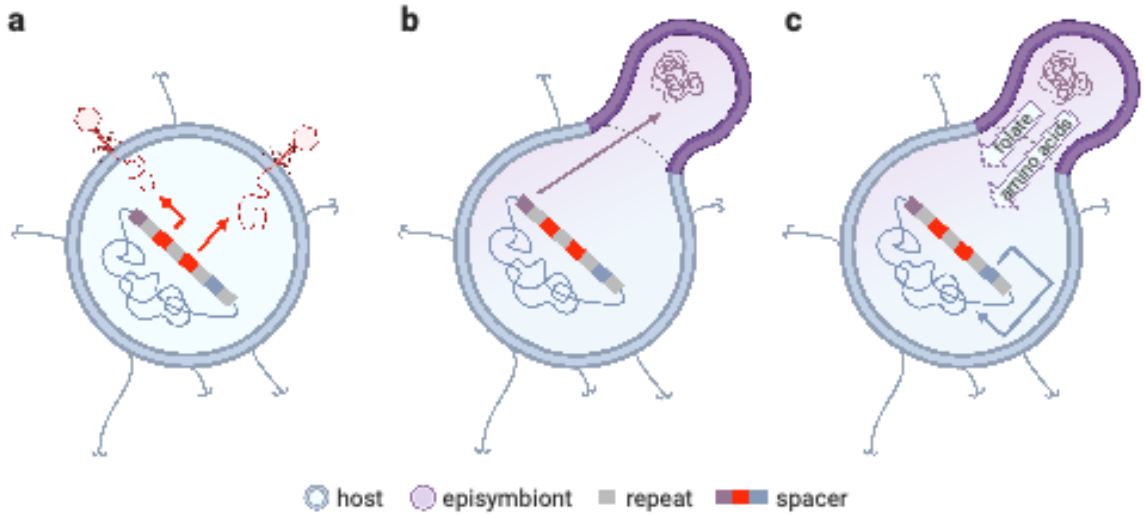


Figure 3.2.2. Example of Ca. Altiarchaeota CRISPR-Cas type IB loci, gene targets on host and episymbiont genomes, and metabolic interactions between Ca. Altiarchaea and Ca. Huberiarchaea as inferred from genome-scaled metabolic modeling | (A) Example of *CRISPR system I-B locus of Ca. A. crystalense with assembled CRISPR array from a single amplified genome (accession no. 1088571). Red box highlights the analysed CRISPR array bearing the*

repeat sequence GTTTAAATCGTACTATGTAGTATGGAAAC and its respective spacers within the array. ***(B)*** *Example of a Ca. A. crystalense DNA polymerase II large subunit locus self-targeted by altiarchaeotal spacers extracted from metagenomes (accession no. 2786546692). Red boxes on the genomic region highlight spacer matching regions. Yellow genes are annotated as uncharacterized proteins.* ***(C)*** *Example of a Ca. H. crystalense genome (accession no. 2785510793) partially matched by Ca. A. crystalense spacers at the genetic loci of the 30S ribosomal protein S11, CTP synthase, and an uncharacterized protein.* ***(D)*** *Example of a Ca. H. crystalense SAG (accession no. 1088571) partially matched by Ca. A. crystalense spacers at the genetic loci of uncharcterized proteins.* ***(E)*** *Metabolic interactions between Ca. Altiarchaeum and Ca. Huberiarchaeum in Crystal Geyser, CG (17 genomes of Ca. A. crystalense and eleven of Ca. H. crystalense, and spacers extracted from transcriptomes) and in* ***(F)*** *Horonobe Underground Research Laboratory, HURL (one genome of A. horonobense and one Ca. H. julieae). Solid arrows denote exchanges of putative essential metabolites between Ca. Altiarchaeum and Ca. Huberiarchaeum. Dashed arrows indicate exchange of metabolites that are only required when CRISPR spacers attack certain target genes (type I-B labeled with red diamonds and the unassigned type labeled with green diamonds).* ***(E-F)*** *While most compounds were produced by Ca. Altiarchaeota, the production of dUMP requires an essential gene, ⑤-dCMP deaminase (EC 3.5.4.12), in Huberiarchaea. Circled numbers indicate key enzymes involved in symbiotic metabolic interactions at CG: ①–Phenylalanyl-tRNA synthetase (EC 6.1.1.20), Lysyl-tRNA synthetase (EC 6.1.1.6); ②, ③–(d)NDP kinase (EC 2.7.4.6); ④–dCMP kinase (EC 2.7.4.25); ⑤–dCMP deaminase (EC 3.5.4.12); ⑥, ⑦–dTMP synthase (EC 2.1.1.45); ⑧–FAD-dependent dTMP synthase (EC 2.1.1.148). The production of tetrahydrofolate (THF) requires an essential gene encoded by Ca. Huberiarchaeum julieae, ⑬–dTMP synthase (EC 2.1.1.45). Circled numbers denote key enzymes involved in the symbiotic metabolic interactions at HURL: ⑬–dTMP synthase (EC 2.1.1.45); ⑫–FAD-dependent dTMP synthase (EC 2.1.1.148); ⑨, ⑩–dCTP deaminase (EC 3.5.4.13); ⑪–dUTPase (EC 3.6.1.23); and ⑭–Dihydrofolate synthase (EC 6.3.2.12).*



a) Viral targeting. **b)** Targeting of episymbiont. **c)** Self-targeting and respective metabolic complementation.

Figure 3.2.3. Illustration of the newly discovered functionality of CRISPR-Cas systems within Ca. Altiarchaea | ***a. Viral targeting:*** *CRISPR-Cas system targets the genomes of MGEs that infect the cell (current state of knowledge).* ***b. Targeting of episymbiont:*** *CRISPR-Cas system targets the genome of the episymbiont Ca. Huberiarchaeum to defend against the parasite.* ***c. Self-targeting and respective metabolic complementation:*** *Self-targeting of CRISPR-Cas in Altiarchaeota mediates metabolic patchiness, which is complemented by the episymbiont metabolism, leading to mutualism. Please note, that this mutualism might be limited to a subset of organisms in the host population. Arrows symbolize spacer-protospacer interactions. The Figure was created with Biorender.com.*

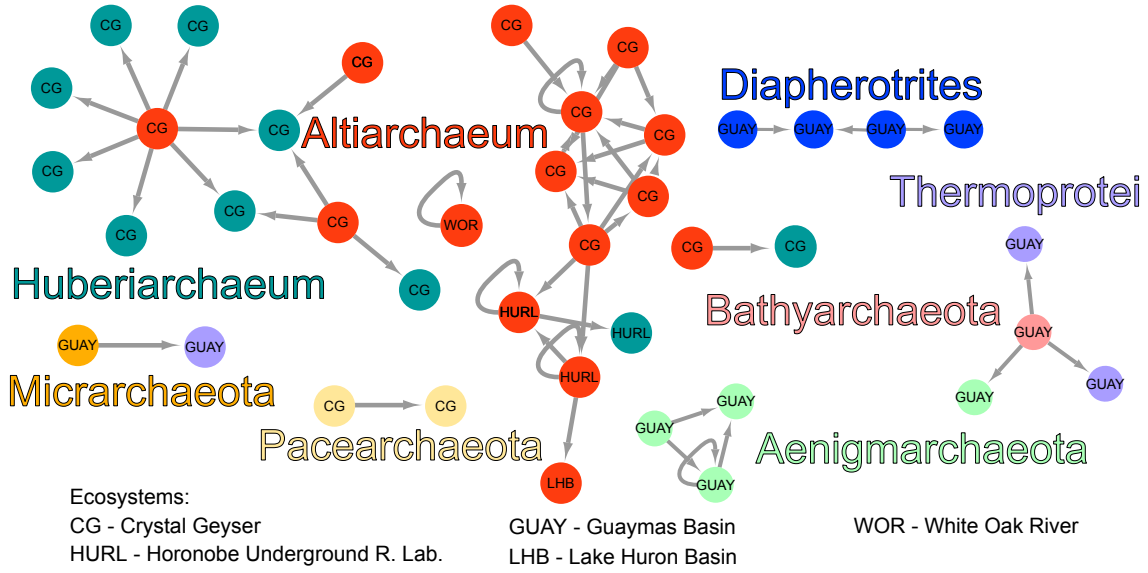
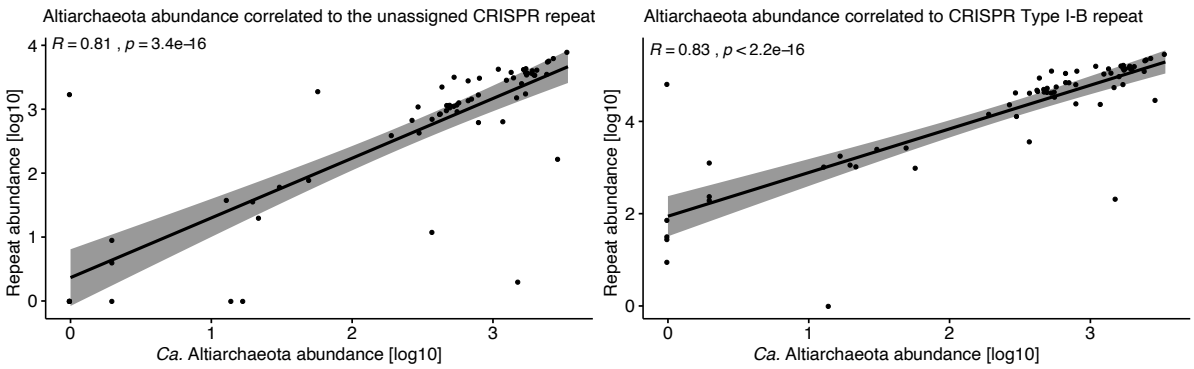
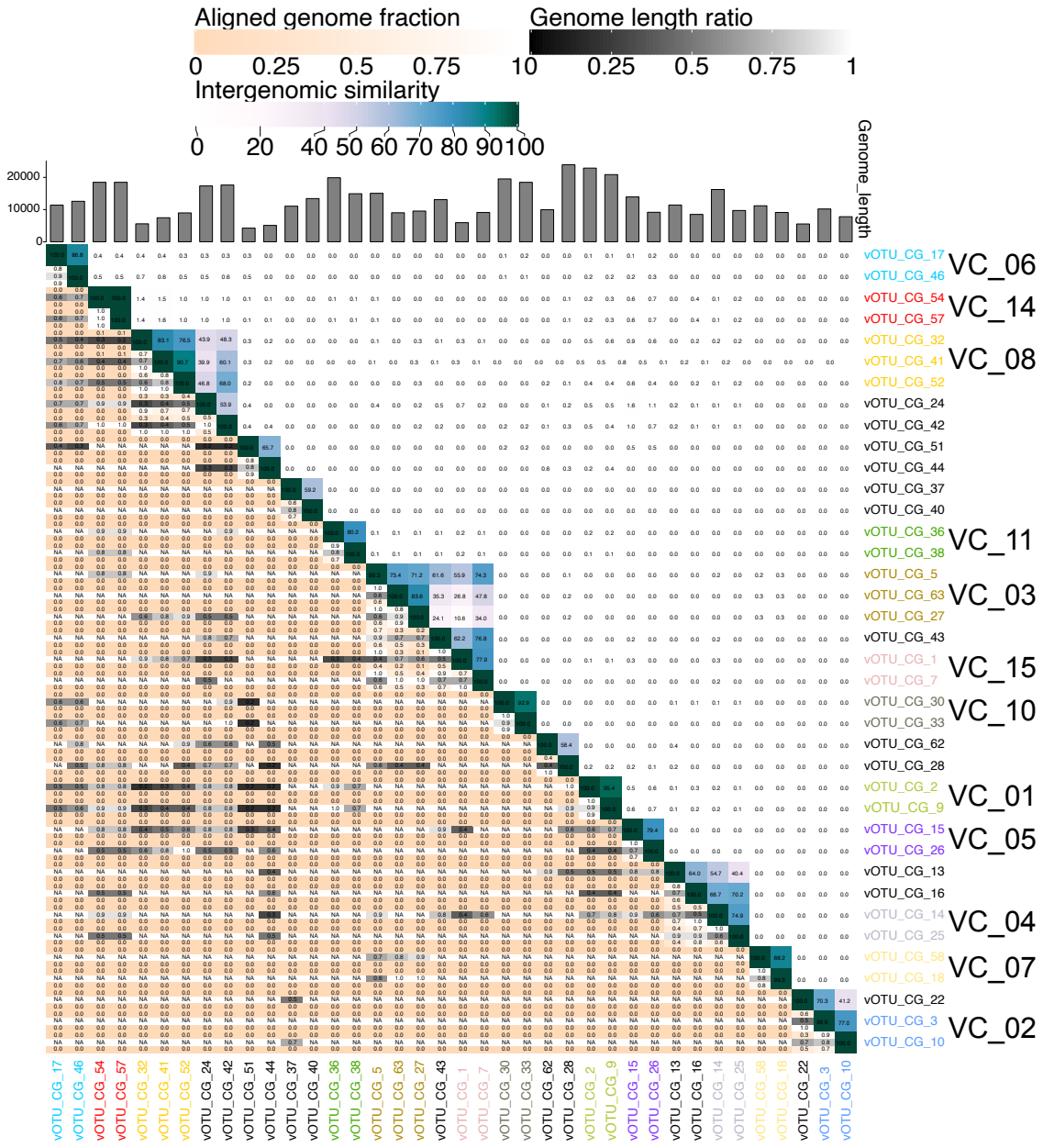


Figure 3.2.4. Directed spacer interaction of DPANN archaea derived from the analysis of 7,012 publicly available archaeal genomes | *Nodes correspond to archaeal genomes. Boomerang and linear grey arrows indicate self-targeting and non-self (including interspecies) targeting spacers, respectively. With the exception of Thermoprotei and Bathyarchaeota, all of the archaea pictured belong to the DPANN superphylum. Colors represent the phylogenetic affiliation of genomes. Genomes of Ca. Altiarchaeum and Ca. Huberiarchaeum derives primarily from CG. Genomes coded according to their corresponding ecosystem: CG - Crystal Geyser* (Emerson et al., 2016; Probst et al., 2018, 2014)*; LHB - Lake Huron Basin* (Sharrar et al., 2017)*; WOR - White Oak River* (Bird et al., 2016)*; GUAY - Guaymas Basin* (Dombrowski et al., 2018)*; HURL - Horonobe Underground Research Laboratory* (Hernsdorf et al., 2017).



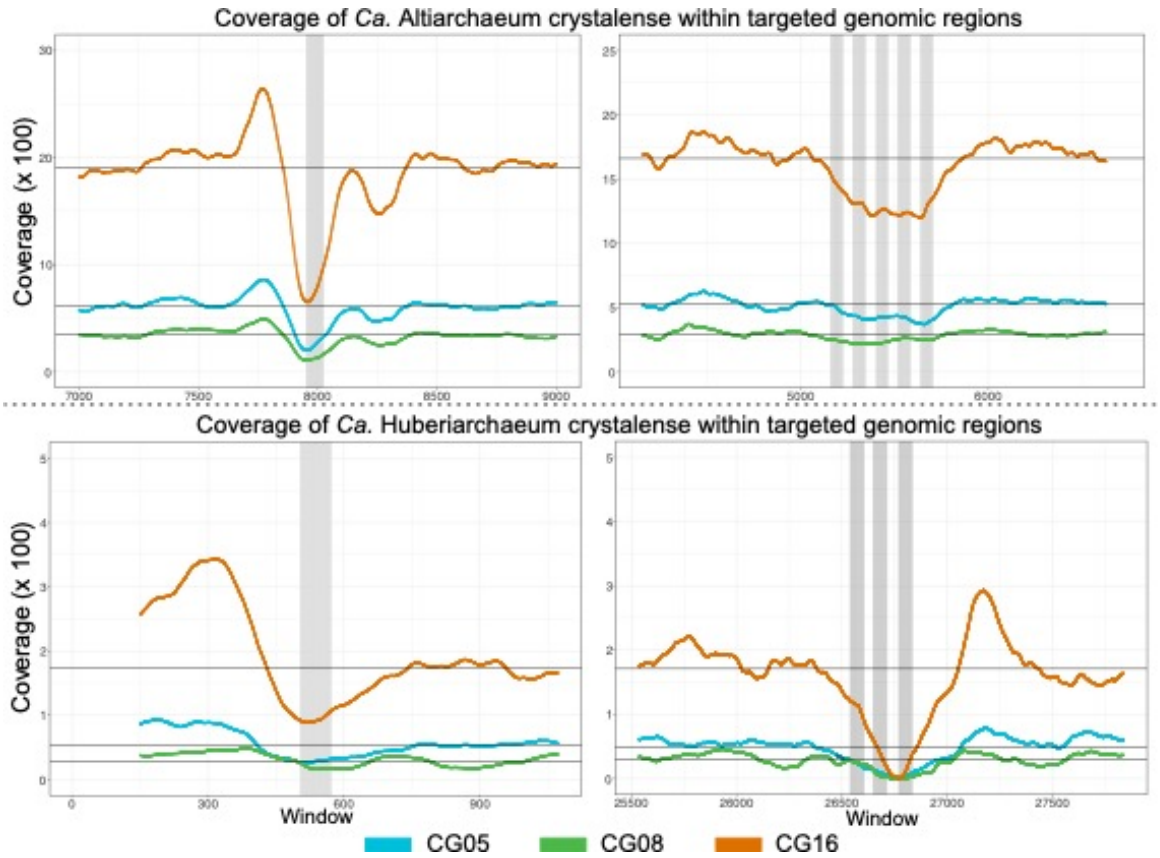
Extended Data Fig. 3.2.1 | Correlation of repeat abundance and abundance of Ca. Altiarchaea genomes. *Spearman rank correlation of logarithmic abundances of Ca. A.*

crystalense and logarithmic abundances of repeat sequences of the unassigned CRISPR array and the I-B CRISPR system in metagenomes from CG (n=66). The grey area depicts all samples within a confidence interval of 0.95. Visualization was performed with R (Posit team, 2022; Team, 2013) (version 3.6.1).



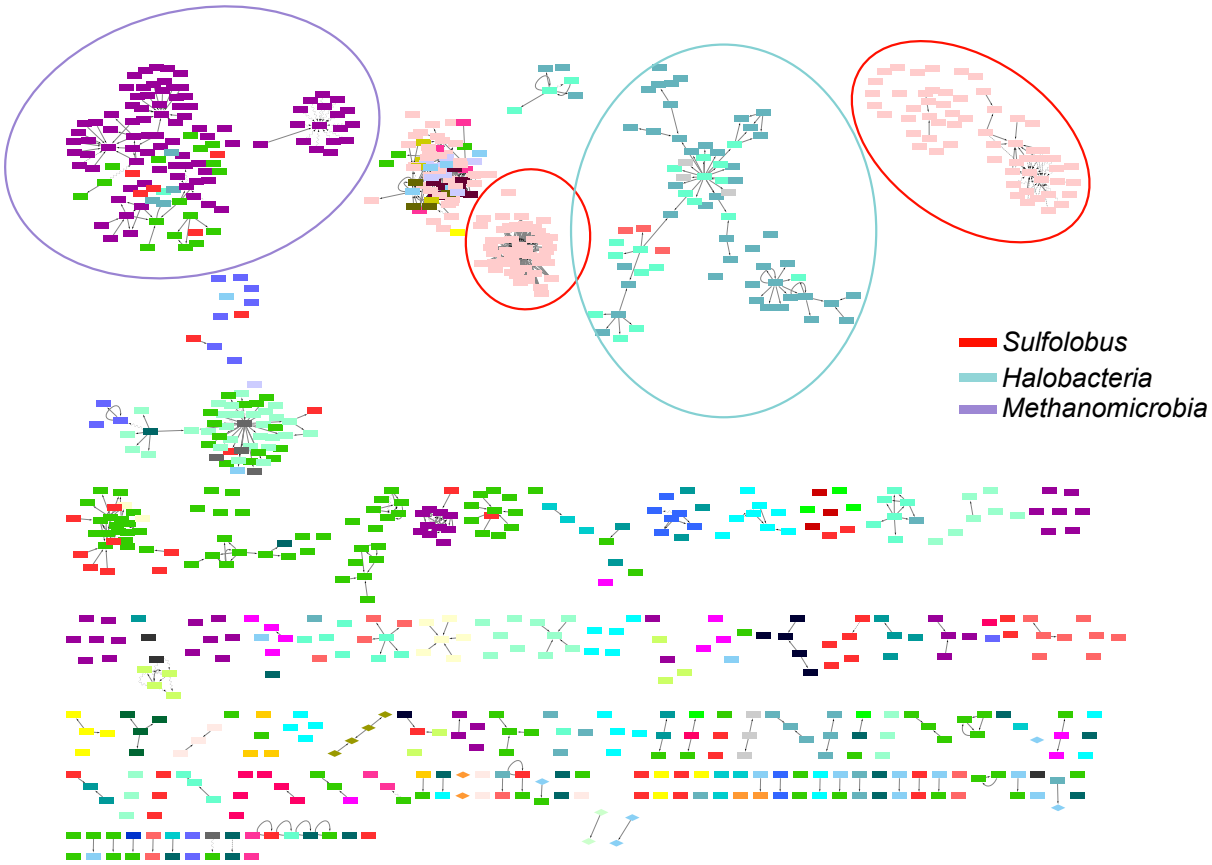
Extended Data Fig. 3.2.2 | Viral clusters predicted by VIRIDIC*(Moraru et al., 2020)***.** *Heatmap showing intergenomic similarity for viral scaffolds of viral clusters (VC_XY) and some singletons (black). Coloring of viral OTUs (vOTUs) according to Table S6. VC_09, _12, _13*

determined by the other tools were not found by VIRIDIC. Only scaffolds with intergenomic similarity of >10 between two viral scaffolds are shown.



Extended Data Fig. 3.2.3 | Coverage analyses of scaffolds targeted by spacers from Ca. Altiarchaea. *Coverage changes within targeted regions by CRISPR system IB of Ca. Altiarchaeum and Ca. Huberiarchaeum based on metagenomic read mapping. The vertically grey marked regions are spacer targeted regions of either Ca. Altiarchaeum or Ca. Huberiarchaeum, whereby the horizontally dark grey lines are showing the average coverage of the scaffold. The colored graphs show the coverage across the spacer targeted region of three samples from the minor eruption phase, where Ca. Altiarchaeum is the most abundant organism (Fig. S3.2.1.1).*

Publication 1: Now available as:
Esser, S. P. *et al.* A predicted CRISPR-mediated symbiosis between uncultivated archaea. *Nature Microbiology* **8**, 1619–1633 (2023).



Extended Data Fig. 3.2.4 | Spacer targeting analyses of publicly available archaeal genomes. *Directed spacer analysis of 7,012 publicly available archaeal genomes (Table S4) shows large clusters of spacers targeting at species level. The targeting spacers (edges) of the genomes Sulfolobus, Methanomicrobia and Halobacterium (nodes) form large clusters performing self-targeting or targeting other genomes of the same family. The clustering was illustrated with Cytoscape (Shannon et al., 2003) (version 3.9.1). Please note that targeting within the same genus might limit the interspecies recombination, as demonstrated in haloarchaea (Turgeman-Grott et al., 2019), or reflect the presence of multiple conserved genomic regions between the genomes.*

Publication 1: Now available as:
Esser, S. P. *et al.* A predicted CRISPR-mediated symbiosis between uncultivated archaea. *Nature Microbiology* **8**, 1619–1633 (2023).

3.2.1 Supplementary Information for **A CRISPR-mediated symbiosis between uncultivated archaea predicted from sequencing data**

Sarah P. Esser[1,2,*], Janina Rahlff[2,*,†], Weishu Zhao[3,††], Michael Predl[4,5], Julia Plewka[1,2], Katharina Sures[1,2], Franziska Wimmer[6], Janey Lee[7], Panagiotis S. Adam[2], Julia McGonigle[8], Victoria Turzynski[1,2], Indra Banas[1,2], Katrin Schwank[2,†††], Mart Krupovic[9], Till L. V. Bornemann[1,2], Perla Abigail Figueroa-Gonzalez[1,2], Jessica Jarett[7], Thomas Rattei[4,5], Yuki Amano[10], Ian K. Blaby[7], Jan-Fang Cheng[7], William J. Brazelton[8], Chase L. Beisel[6,11], Tanja Woyke[7], Ying Zhang[3], and Alexander J. Probst[1,2,12,#]

[1]Environmental Metagenomics, Research Center One Health of the University Alliance Ruhr, Faculty of Chemistry, University of Duisburg-Essen, 45151 Essen

[2]Group for Aquatic Microbial Ecology, Environmental Microbiology and Biotechnology, University of Duisburg-Essen, 45141 Essen

[3]Department of Cell and Molecular Biology, College of the Environment and Life Sciences, University of Rhode Island, Kingston, RI 02881, USA

[4]Computational Systems Biology, Centre for Microbiology and Environmental Systems Science, University of Vienna, 1030 Vienna, Austria

[5]Doctoral School in Microbiology and Environmental Science, University of Vienna, 1030 Vienna, Austria

[6]Helmholtz Institute for RNA-based Infection Research (HIRI), Helmholtz-Centre for Infection Research (HZI), 97080 Würzburg, Germany

[7]DOE Joint Genome Institute, Lawrence Berkeley National Laboratory, One Cyclotron Rd, Berkeley, CA, 94720, USA

[8]School of Biological Sciences, University of Utah, Salt Lake City, Utah, USA

[9]Institut Pasteur, Université Paris Cité, CNRS UMR6047, Archaeal Virology Unit, 75015 Paris, France

[10]Nuclear Fuel Cycle Engineering Laboratories, Japan Atomic Energy Agency, Tokai, Ibaraki, 319-1194 Japan

[11]Medical faculty, University of Würzburg, 97080 Würzburg Germany

Publication 1: Now available as:
Esser, S. P. *et al.* A predicted CRISPR-mediated symbiosis between uncultivated archaea. *Nature Microbiology* **8**, 1619–1633 (2023).

[12]Centre of Water and Environmental Research (ZWU), University of Duisburg-Essen, 45141 Essen, Germany

[*]authors contributed equally

[†]present address: Centre for Ecology and Evolution in Microbial Model Systems (EEMiS), Department of Biology and Environmental Science, Linnaeus University, Kalmar, Sweden

[††]present address: Shanghai Jiao Tong University, School of Life Sciences and Biotechnology, International Center for Deep Life Investigation (IC-DLI), Shanghai Jiao Tong University, Shanghai, 200240, China

[†††]present address: University of Regensburg, Biochemistry III, 93053 Regensburg, Germany

[#]corresponding author: alexander.probst@uni-due.de

Content:

1. Supplementary Methods

Detection, dereplication and analysis of DNA viral scaffolds

Assembled metagenomes were used to extract and predict viral and putatively viral sequences as previously performed (Rahlff et al., 2021). In brief, predicted viral operational taxonomic units (vOTUs) >3kb were dereplicated via usearch (Edgar, 2010) at 95% nucleotide identity resulting in centroid sequences for downstream analysis. VOTUs were identified via blastn (Altschul et al., 1990) (--short, filtering for 80% similarity) of Clustered regularly interspaced short palindromic repeat (CRISPR)-derived spacers against centroid vOTUs. Completeness and origin (host, viral, unclassified) of vOTUs was assessed using CheckV v.0.4.0 (Nayfach et al., 2021). Clustering of viral sequences with a recent viral Refseq database (Cook et al., 2021) (release July 2022). and previously detected Altiarchaeota-targeting viruses

(Rahlff et al., 2021) was performed using vConTACT2 (Bin Jang et al., 2019; Bolduc et al., 2017) v.0.11.3, VICTOR (Meier-Kolthoff and Göker, 2017) (using nucleic acid sequences) and VIRIDIC (Moraru et al., 2020) under default settings and for calculating intergenomic similarities. In VICTOR, all pairwise comparisons of the nucleotide sequences were conducted using the Genome-BLAST Distance Phylogeny (Meier-Kolthoff et al., 2013) (GBDP) method under settings recommended for prokaryotic viruses (Meier-Kolthoff and Göker, 2017). The resulting intergenomic distances from VICTOR were used to infer a balanced minimum evolution tree with branch support via FastME including Subtree Pruning and Regrafting post-processing (Lefort et al., 2015) for the distance formula D0. Branch support was inferred from 100 pseudo-bootstrap replicates each. Trees were rooted at the midpoint (Farris, 1972). Visualization of viral clusters identified with vConTACT2 in conjunction with the viral RefSeq database was performed using Cytoscape v.3.9.02 (Shannon et al., 2003). In addition, a circular proteomic tree with viral genomes using the Virus-Host DB: RefSeq release 217 was build using ViPTree version 3.5. (Nishimura et al., 2017) . Within ViPTree, dsDNA was selected as nucleic acid type and “any host” chosen as host category.

Sample preparation for Fluorescence *in situ* hybridization (FISH)

Groundwater for FISH analysis was sampled to visualize the Altiarchaea-Huberarchaea relationship within the Crystal Geyser (CG) and Horonobe Underground Research Laboratory (HURL) environment. Water from CG was sampled onto a 0.2 µm filter with a syringe filter holder until the filter started clogging and afterwards fixed by slowly pressing 3% formaldehyde (Thermo Fisher Scientific, MA, USA) through the filter to exchange the sample water with fixative. Fixation was performed for one hour in the dark. Within the filter holder, a washing step with 3x 20 mL Phosphate Buffered Saline (PBS) (conc. 1 v/v%) was done, followed by alternating washing and incubation with ethanol with 50, 70 and 100% (v/v)% for 10 minutes at room temperature. The filter holder was opened in a sterile environment, and the filter was stored in a petri dish with the biofilm facing upwards and then air dried for 10 minutes. Filter samples for FISH from CG of the sampling campaign in 2021 were covered and stored in RNA*later*™ (Invitrogen by Thermo Fisher Scientific; Ref: AM7021).

Imaging of FISH samples

FISH was performed according to Schwank et al. (Schwank et al., 2019) with the following modifications. DAPI (4',6-Diamidino-2-Phenylindole) was used at concentrations of 4 µg per mL without dilution in the washing buffer. Visualization was performed with an Axio Imager M2m epifluorescence microscope (X-Cite XYLIS Broad Spectrum LED Illumination System, Excelitas) equipped with an Axio Cam MRm and a Zen 3.4 Pro software (version 3.4.91.00000) (Carl Zeiss Microscopy GmbH, Jena, Germany). Imaging was carried out by using the 110x/1.3 oil objective EC-Plan NEOFLUAR (Carl Zeiss Microscopy GmbH) and three different filter sets (Carl Zeiss): 49 DAPI for imaging *Ca.* A. crystalense/horonobense cells and *Ca.* H. crystalense/julieae cells, 43 Cy3 for the detection of *Ca.* Huberiarchaea signals, and 09 for achieving 16S rRNA signals of *Ca.* Altiarchaea. The FISH images are shared within a Figshare folder (10.6084/m9.figshare.22739849).

Synthesis of *cas* genes derived from *Ca.* Altiarchaeota MAGs

The CRISPR-Cas gene cassette (Cas1, Cas2, Cas3, Cas4, Cas5, Cas8b) of one single-cell amplified genomes (SAG) of *Ca.* Altiarchaeum was used in gene synthesis. The Cas6 gene was annotated in two other SAGs of *Ca.* Altiarchaeota, once with 438 and 468 amino acids respectively. To synthesize these genes, the sequences were first codon optimized using the BOOST design software v.1.3.9 (Oberortner et al., 2017) and an *E. coli* codon frequency table. The synthetic DNA fragments were obtained from Twist Bioscience, CA, USA, which were later PCR amplified and cloned into the NcoI and XhoI sites of the pET28b vector using the NEBuilder HiFi Assembly kit (E2621X, New England BioLabs). The PCR was performed using the KAPA HiFi HotStart ReadyMix (Roche Sequencing, Pleasanton, CA, USA) according to the manufacturer recommended cycling protocol. The sequences of the refactored *cas* genes were verified by Pacific Bioscience sequencing. The synthetic building blocks and PCR primer sequences are listed in Table S13.

CRISPR-Cas activity assay in TXTL

The activity of the *Ca.* Altiarchaeota type I-B CRISPR-Cas system was tested in a cell-free transcription-translation (TXTL) system. Circular or linear DNA constructs that were added to

a TXTL reaction were transcribed and translated, and RNAs and proteins were produced (Garamella et al., 2016). The reaction conditions of the TXTL reactions performed here were adapted from Wimmer et al. (Wimmer et al., 2022). A deGFP reporter plasmid was generated with Site Directed Mutagenesis (SDM) using p70a_deGFP_PacI(Wimmer et al., 2022) as backbone and introducing a TTTTC motif 12 nucleotides upstream of the p70a promoter driving the deGFP expression. The TTTTC motif was used as a putative PAM sequence because this motif was found next to a sequence matching a type I-B spacer (see main text). Constructs encoding single spacer arrays driven by the constitutive promoter J23119 contained either a spacer targeting the p70a promoter region of the reporter plasmid or a non-targeting spacer. These constructs were generated by Golden Gate adding spacer sequences in a plasmid which contained two repeat sequences interspaced by two BbsI restriction sites. The construct p70a-T7RNAP (Garamella et al., 2016) encoding the T7 RNA polymerase and Isopropyl β-D-1-thiogalactopyranoside (IPTG; Carl Roth, Karlsruhe, Germany) was added to the TXTL reaction to ensure expression of the *cas* genes. Two Master mixes containing plasmids encoding for Cascade-forming *cas* proteins were prepared using the stoichiometry Cas8b1-Cas77-Cas51-Cas61, namely one for the 245 and the 268 amino acids long Cas6. A volume of 3 µL TXTL reaction were prepared in Costar 3357 96-well V-bottom plates (Corning, NY, USA) with Costar 2080 cover mats (Corning) using the liquid handling machine Echo525 (Beckman Coulter, Brea, CA, USA) including the following components: 2.25 µL myTXTL Sigma 70 MasterMix (Arbor Biosciences, MI, USA), 0.2 nM p70a-T7RNAP, 0.5 mM IPTG, 3 nM Cascade Master mix, 1 nM Cas3 plasmid, and 1 nM targeting or non-targeting spacer plasmid. After a 4 h pre-incubation period at 29°C to allow the ribonucleoprotein complex of Cascade and crRNA to form, 1 nM deGFP reporter plasmid containing the TTTTC motif was added to the TXTL reactions. The reactions were incubated at 29°C for additional 16 h while measuring deGFP expression with BioTek Synergy H1 plate reader (BioTek, Winooski, VT, USA) at 485/528 nm excitation/emission (Shin and Noireaux, 2012). Targeting spacer-mediated binding of the Cascade complex to the target region in the deGFP driving promoter or target plasmid degradation by Cas3 would lead to inhibition of deGFP production. The non-targeting spacer does not affect deGFP production and was used as a control. The fluorescence background values were measured with reactions containing solely myTXTL Sigma 70

MasterMix and nuclease-free H_2O and were subtracted from the endpoint deGFP values of the TXTL reactions. Significance between deGFP values derived from the non-targeting and targeting samples was calculated with Welch's t-test. All results showed a p-value > 0.05 and were therefore seen as non-significant. Hence, we concluded that the type I-B systems do not exhibit binding or degradation activity under the tested conditions. This could be due to the conditions used here not reflecting the conditions at the sampling site of *Ca.* Altiarchaeota, or the motif TTTTC being a non-recognized PAM. All reactions were performed in triplicates.

PAM assay in TXTL (PAM-DETECT)

To reveal the PAM diversity recognized by the type I-B system of *Ca.* Altiarchaeota, PAM-DETECT (PAM DETermination with Enrichment-based Cell-free TXTL) was performed. A detailed protocol can be found in Wimmer et al. (Wimmer et al., 2022). A plasmid containing a PAM library of five randomized nucleotides was used as a target plasmid. A single spacer array plasmid is constructed as mentioned above harboring a spacer targeting the target plasmid adjacent to the randomized nucleotides. Upon recognition of a PAM sequence, the Cascade complex binds to its target and thereby covers a PacI recognition site included in the target region. Cascade-bound target plasmids are protected from PacI digestion leading to an enrichment of recognized PAMs, detected by next-generation-sequencing (NGS; specified below). Separate 6 µL TXTL reactions were prepared containing one or the other Cascade Master mix mentioned above. TXTL reactions contained: 4.5 µL myTXTL Sigma 70 Master mix, 0.2 nM pET28a_T7RNAP (Wimmer et al., 2022), 0.5 mM IPTG, 3 nM Cascade Master mix, 1 nM targeting spacer plasmid (targeting PAM library plasmid) and 1 nM PAM library plasmid (pPAM_library)(Wimmer et al., 2022). After incubation at 29°C for 16 h, the TXTL samples were diluted 1:400 in nuclease-free H_2O. A volume of 500 µL of the dilution was digested with 0.09 units μL^{-1} PacI (NEB) in 1x CutSmart Buffer (NEB) at 37°C for 1 h. A "non-digested" control was prepared using 500 µL of the dilution and adding nuclease-free H_2O instead of PacI. PacI was inactivated at 65°C for 20 min and proteins were digested with 0.05 mg mL^{-1} Proteinase K (Cytiva, Marlborough, MA, USA) at 45°C for 1 h. Proteinase K was inactivated at 95°C for 5

min and remaining plasmids were extracted with standard ethanol precipitation. To prepare sequencing libraries, Illumina adapters with unique dual indices were added in two amplifications steps using KAPA HiFi HotStart Library Amplification Kit (KAPA Biosystems, Wilmington, MA, USA) and purification by AMPure XP (Beckman Coulter) after every amplification step. A volume of 15 µL of the ethanol-purified samples was used in a 50 µL PCR reaction with 19 cycles to add Illumina sequencing primer sites. The flow cell binding sequence was added in the second PCR reaction using 1 ng purified amplicons generated with the first PCR in a 50 µL reaction and 18 cycles. NGS was performed on an Illumina NovaSeq 6000 sequencer with 50 bp paired-end reads and 2.0 million reads per sample. PAM wheels were generated according to Leenay et al. (Leenay et al., 2016) and Ondov et al. (Ondov et al., 2011) and are not depicted here as no PAM enrichment was observed. Absence of PAM enrichment might be due to the reaction conditions of PAM-DETECT deviating from the conditions at the sampling site of *Ca.* Altiarchaeota. PAM-DETECT assays were performed in duplicates.

Naming of archaeal genomes

Except for *Ca.* Huberiarchaeum crystalense, all host and episymbiont species were previously only classified at the genus level or – in case of the episymbiont from the HURL ecosystem – not classified at all. Using established average nucleotide identity (ANI) and Average Amino Acid Identity (AAI) cutoffs along with phylogenetic analyses (Fig. 3.2.1 and Supplementary Data 1), we established the host-symbiont pairs as *Ca.* Altiarchaeum crystalense and *Ca.* Huberiarchaeum crystalense from the CG ecosystem (named after the ecosystem Crystal Geyser) and *Ca.* Altiarchaeum horonobense (named after the sampling site Horonobe) and *Ca.* Huberiarchaeum julieae (named after subsurface microbiologist Julie Huber).

Overview of models for *Ca.* Altiarchaeota and *Ca.* Huberarchaeota host-symbiont interaction based on genomic information

To infer metabolic interactions, genome scale metabolic reconstructions of *Ca.* Altiarchaeum crystalense/horonobense and *Ca.* Huberiarchaeum crystalense/julieae (see accession numbers Table S2) were based on MAGs and SAGs identified from CG (AltiCG-HuberCG

model) and HURL (AltiHURL-HuberHURL model). Relative abundance (based on coverage) of the respective genomes was calculated via read mapping with Bowtie2(Langmead and Salzberg, 2012) (--sensitive mode). The AltiCG-HuberCG model included 515 genes of *Ca.* A. crystalense and 88 genes of *Ca.* H. crystalense, associated with 477 and 125 reactions, respectively (Table S8). The AltiHURL-HuberHURL model included 388 *Ca.* A. horonobense and 78 *Ca.* H. julieae genes, associated with 495 and 128 reactions, respectively (Table S9). Each model contained two compartments (one for *Ca.* Altiarchaeum and one for *Ca.* Huberiarchaeum), with either restricted or unlimited metabolite exchanges between the two compartments to model the metabolite availability upon cytoplasmic fusion of the two organisms.

Metabolic model reconstruction

The genome-scale metabolic models of AltiCG-HuberCG and AltiHURL-HuberHURL were represented in a YAML format following conventions defined by the PSAMM software package (Dufault-Thompson et al., 2018; Steffensen et al., 2016). Details of the model are represented in Tables S8 - S11. Metabolic simulations were performed with PSAMM version 1.0 using the IBM ILOG CPLEX Optimizer version 12.7.1.0 (https://www.ibm.com/products/ilog-cplex-optimization-studio). The CG model was based on the prediction of metabolic pathways using combined annotation of all MAGs and SAGs identified from this and a prior study (Schwank et al., 2019). Protein sequences annotated from the individual MAGs and SAGs were clustered at 100% amino acid identity using CD-HIT (Fu et al., 2012; Li and Godzik, 2006), followed by a pangenome analysis to capture metabolic capacities represented by the entire population. Automated metabolic reconstruction was performed based on ortholog mapping to (i) existing models of other archaeal strains, i.e. *Pyrococcus furiosus*, *Thermococcus eurythermalis*, *Methanosarcina barkeri* and *Methanococcus maripaludis* (Gonnerman et al., 2013; Goyal et al., 2014), and (ii) public databases, such as the Kyoto Encyclopedia of Genes and Genomes (Kanehisa et al., 2017) (KEGG), EggNOG (Huerta-Cepas et al., 2019) and Transporter Classification Database (Saier et al., 2016) (TCDB). Extensive manual curations were carried out following the automated reconstruction to integrate prior annotations of *Ca.* Altiarchaeum's and *Ca.*

Huberiarchaeum's metabolism (Probst et al., 2018; Schwank et al., 2019), as well as latest biochemical evidence of enzymatic functions in archaeal organisms (Tables S8 and S9). Overall, literature evidence was assigned to 137 reactions in the model for AltiCG-HuberCG and 144 reactions in the model for AltiHURL-HuberHURL through homologous mapping to experimentally verified enzymes.

The biomass equations of *Ca.* Altiarchaeum and *Ca.* Huberiarchaeum were individually formulated in both models following a standard procedure (Table S8 and S9). The biosynthesis of macromolecules (e.g., DNA, RNA, protein, and lipids) were defined to account for the mM composition of each building block in assembling 1 g of a given component and the associated energy cost. The stoichiometry of DNA and RNA biosynthesis was derived based on the average composition of nucleotides in the genomes and coding genes, respectively. The energy cost for DNA and RNA synthesis was estimated as 2 mM of ATP per millimole of nucleotides according to the mechanism of polynucleotide biosynthesis (Neidhardt et al., 1990). The stoichiometry of protein biosynthesis was calculated based on the average composition of amino acids in the corresponding proteome, and the associated energy cost was estimated based on the mechanism of protein synthesis (Nelson et al., 2004), where one ATP was consumed for each tRNA charging, and two GTPs were consumed for extending one amino acid to a growing peptide chain. The tRNA charging equations were represented separately for each amino acid. The stoichiometry of lipid biosynthesis was formulated based on experimental measurements of the weight compositions of core lipids and header groups of *Ca.* Altiarchaeum or *Ca.* Huberiarchaeum of the respective system (Probst et al., 2018). Following the definition of macromolecular synthesis functions, the biomass equations of *Ca.* A. crystalense, *Ca.* H. crystalense, *Ca.* A. horonobense, and *Ca.* H. julieae were formulated to represent the gram composition of DNA, RNA, proteins, lipids, and vitamins in 1 g of cell dry weight (gDW). The CG- and HURL-specific *Ca.* Altiarchaeum and *Ca.* Huberiarchaeum biomass were then combined based on an estimation of their relative abundance in the respective ecosystems using the metagenomic data. Specifically, the combined Altiarchaeum-Huberiarchaeum biomass has a relative Huberiarchaeum:Altiarchaeum ratio between 0.06 and 0.12 in the CG system, and a ratio of 0.205 in the HURL system (as estimated via stringent read mapping, see above).

Simulation of the *Ca.* Altiarchaeum – *Ca.* Huberiarchaeum metabolism was formulated with exchange constraints that represent the corresponding *in situ* geochemical measurements in the CG (Probst et al., 2018) and HURL (Hernsdorf et al., 2017). These geochemical measurements included the ion concentrations in porewater and the compositions of headspace gas (Table S8 - S11). Some measurements, e.g., CO_2 and H_2 at the CG site, were not available, but the compounds were required for biomass production in the *Ca.* Altiarchaeum – *Ca.* Huberiarchaeum system, and thus they were added to the exchange without implicit constraints. To simulate the fusion of the cytoplasm between *Ca.* Altiarchaeum and *Ca.* Huberiarchaeum, unlimited metabolite exchange was introduced to allow the free transfer of all small-molecular metabolites (excluding macromolecules, such as DNA, RNA, protein, lipids, and biomass) between the *Ca.* Altiarchaeum and *Ca.* Huberiarchaeum cell compartments. Metabolic gaps in the production of biomass components by *Ca.* Altiarchaea were identified using the PSAMM *fluxcheck* and *gapcheck* functions (Dufault-Thompson et al., 2018; Steffensen et al., 2016) Candidate gap-filling reactions for unblocking each biomass component were identified using the PSAMM *fastgapfill* implementation with the KEGG reaction database (Kanehisa et al., 2017) as a reference, and subsequently curated before being incorporated into the models. A total of 17 gap-filling reactions were included in the *Ca.* Altiarchaea compartment of both CG and HURL models, including functions in the citrate cycle, amino acids-, lipids-, and cofactor-biosynthesis. The overall stoichiometric consistency, formula and charge balance of the model were validated using the PSAMM *masscheck*, *formulacheck*, and *chargecheck* functions (Dufault-Thompson et al., 2018; Steffensen et al., 2016). The exchange reactions, compound sources or sinks, biomass equations and reactions involving compounds with undefined group R or X were excluded from formula and charge checks but instead manually inspected to ensure proper formulation.

To identify complementary metabolic processes between *Ca.* Altiarchaea and *Ca.* Huberiarchaea, the PSAMM *fastgapfill* implementation (Dufault-Thompson et al., 2018; Steffensen et al., 2016) was applied to optimize the *Ca.* Altiarchaeota biomass while using all metabolic reactions in the *Ca.* Huberiarchaea compartment as a reference database, and vice versa, using corresponding models for CG or HURL. A list of metabolic reactions, including

metabolite exchange functions between *Ca.* Altiarchaea and *Ca.* Huberiarchaea, was identified from this automated gap filling procedure to reveal the potential metabolic interactions between the two archaea at each site. The predicted complementary functions were subsequently confirmed by showing that the removal of any interactions would lead to a non-viable *Ca.* Altiarchaeota or *Ca.* Huberiarchaea (biomass production is zero), suggesting that these metabolite exchanges reflect minimal essential interactions between *Ca.* Altiarchaea and *Ca.* Huberiarchaea of a given site (Table S8 and S9). Genes corresponding to the CRISPR type I-B and the unassigned CRISPR array spacer targeting in both CG and HURL systems were mapped to the metabolic reconstructions of AltiCG-HuberCG and AltiHURL-HuberHURL, respectively, for the identification of putative targets for simulating the metabolic influences of CRISPR spacers attacking (Table S10 and S11). To identify changes in the *Ca.* Altiarachaea – *Ca.* Huberiarchaea metabolic collaboration when considering attacks of respective genes by CRISPR-Cas systems, comparisons were made between the exchange unlimited model (where all metabolites (with the exception of macromolecules) were allowed to transfer freely between *Ca.* Altiarchaeum and *Ca.* Huberiarchaeum) and the exchange limited model (where only the complementary metabolites were allowed to transfer between *Ca.* Altiarchaeum and *Ca.* Huberiarchaeum). Flux variability analysis (FVA) was applied to the optimization of the combined Altiarchaeum-Huberiarchaeum biomass in the limited or unlimited models. Pathways that are required for complementing the effect of CRISPR spacer attacks were identified by comparing the FVA results of the limited and unlimited models. If the deletion of a spacer attacked gene would result in a zero-biomass flux in the limited model while a non-zero biomass flux in the unlimited model, a complementary pathway to the corresponding gene deletion was explored by identifying the enabling functions in the unlimited model. Note that the FVA was performed in PSAMM using the CPLEX Optimizer version 12.7.1.0, a zero range is defined as any fluxes within $1E^{-6}$ from zero.

Phylogenetic reconstruction of gene transfers

For the phylogenies of lysine and phenylalanine (subunit B) tRNA synthetases, the protein sequences inferred from both genes from *Ca.* Altiarchaeum hamiconexum and *Ca.* Huberiarchaeum crystalense were used for homology searches against local databases of

1808 archaeal and 25118 bacterial genomes (all genomes of the respective domain on NCBI as of 2019.06.01 dereplicated at species level) with DIAMOND v2.0.15.153 (Buchfink et al., 2014). The maximum number of target sequences (-k 400) was determined by trying different numbers (100, 200, 400, 800, 1000, 0/all), aligning with MAFFT FFT-NS-2 (v7.505) and running a preliminary phylogeny (BioNJ or PhyML without tree topology optimization) in Seaview version 5.0.4 (Gouy et al., 2021). We picked the number that we deemed to give a reasonable view of the origin of each sequence, without including too many divergent homologs or increasing the downstream computational load too much. The original query sequences were added to the set of hits and aligned with MAFFT E-INS-I. The datasets were curated semi-manually (https://github.com/ProbstLab/Adam_Kolyfetis_2021_methanogenesis/blob/master/fuse_sequences.py) to fuse fragmented sequences, realigned as before, and trimmed with BMGE (Criscuolo and Gribaldo, 2010) (BLOSUM30). Phylogenies were reconstructed with IQ-TREE 2 (Minh et al., 2020) using ModelFinder (Kalyaanamoorthy et al., 2017) for the model selection and branch supports calculated using 1000 ultrafast bootstrap (Hoang et al., 2017) and 1000 SH-aLRT replicates.

Sliding window for coverage analysis of regions targeted by CRISPR spacers

Variations in coverage over the genomes were investigated to deduce possible negative selection at targeted sites. Targeted scaffolds from individual genomes were mapped back to the raw reads (from sample CG05, CG08 and CG16) with Bowtie2 (Langmead and Salzberg, 2012) with default settings. Mappings were filtered to remove hits with more than three mismatches using SAMtools (Li et al., 2009). Genomecov from BEDtools was used to calculate coverage per position (Quinlan and Hall, 2010). The first and last 150 base pairs of each scaffold, and possible transposons and viruses were masked by setting the breadth to zero. Mean breadth from sliding windows of 35 base pairs were calculated. In addition, all position with a coverage lower than 10 were excluded. The median coverage of each scaffold (δ) serves to differentiate high and low breadth. Wilcoxon signed rank tests (standard function R (Team, 2013)) were performed between targeted regions of a scaffold and the same amount

of randomly drawn non-targeted windows from the same scaffold. Random sampling and the test were repeated 1000 times for each scaffold.

2. Supplementary Results

Analysis of Altiarchaeota-viral interactions reveal new viral clades in the CG ecosystem

In total, 64 predicted vOTUs (vOTU_CG_1 – vOTU_CG_64) were extracted and compared to previously published altiarchaeotal vOTUs Altivir_1 - Altivir_8 from different ecosystems (Rahlff et al., 2021) (Fig S5). Many of the singletons and outliers as determined by vConTACT2 were identified to be of viral origin using CheckV (Nayfach et al., 2021) (Table S6). In total 41 fragments were assigned as viral, one as partly viral and host, and 17 as complete host fragment. The five residual sequences remained unassigned. The completeness of the extracted viral sequences varied between 2% and 56.1%, according to CheckV and under consideration of the most adequate method for completeness determination (AAI-based or HMM-based, Table S6). This underlines how fragmented these vOTUs are. The GC content and coding densities on vOTUs ranged between 18.7-65.1% (median=34.3%) and 51.1-98.7% (median=92.9%), respectively (Table S6). All vOTUs were classified as lytic viruses being affiliated to VirSorter(Roux et al., 2015) category 1-3, supporting previous evidence for a predominance of lytic infections of *Ca.* Altiarchaeota in subsurface ecosystems (Rahlff et al., 2021).

Both VICTOR and vConTACT2 v.0.11.3 resulted in 14 VCs, each of them corresponding to an individual genus. Both tools revealed 40 and 42 new viral clades (including unclustered singletons and outliers, which likely represent new viruses), respectively. Many VCs were further supported using VIRIDIC and therein a genus and species threshold of 70% and 95% (Extended Data Fig. 3.2.2), respectively. Here, we detected 11 genus-based VCs with two of them also forming species clusters based on intergenomic comparisons (Extended Data Fig. 3.2.2). Including unclustered singletons, VIRIDIC revealed 46 new viral clades, and all tools showed profound overlap for VC detection (Table S6). However, it should be noted that all viruses from all metagenomes were dereplicated at 95% nucleotide identity to reduce the dataset. Clustering of *Ca.* Altiarchaeota associated vOTUs, determined by spacer-protospacer matches against vOTUs from VirSorter, with a recent database (release July 2022) revealed no clustering with any previously known virus from the database or previously described *Ca.* Altiarchaeota viruses (Rahlff et al., 2021) at the genus level. Despite not being related at the

genus level, some of the vOTUs share protein clusters with known bacteriophages of different VCs (Figure S5, Table S6). Three of the vOTUs vOTU_CG_11, vOTU_CG_44, and vOTU_CG_51, were targeted by one, 350 and 295 spacers from transcriptomic data (samples CG05, CG08, and CG16), respectively. These matching spacers from the transcriptomes indicate active acquisition of spacers from these viruses. By constructing a proteomic tree (Fig. S3.2.1.6) together with RefSeq database viruses, several vOTUs with infection histories with *Ca.* Altiarchaea clustered with phages of the host group Pseudomonadota, Actinomycetota and Bacillota, but also viruses from Chordata and "other" hosts, while "others" included different archaeal hosts such as Thermoproteus, Pyrobaculum, Pyrococcus, or Thermococcus. The tree thus indicates that CRISPR spacer acquisition by Ca. Altiarchaea from vOTUs might occur from phylogenetically distant viruses, a phenomenon that has been observed for other ecosystems (Hwang et al., 2023).

Protospacer regions in host and episymbiont are associated with drops in scaffold coverage

A total of 17% of the protospacer regions (34 out of 196) of the CRISPR-Cas type I-B system that represent targeted genomic regions of the host (*Ca.* A. crystalense) showed significant decreases in metagenomic coverage compared to untargeted regions of the very same scaffold, which we determined using a sliding window algorithm (see Supplementary Methods) paired with a bootstrapped Wilcoxon paired signed rank test. A significance level of 5% ($p < 0.05$) was tested for each protospacer region and the p-values were FDR-adjusted.

A total of 30% of the protospacers (22 out of 73) of the CRISPR-Cas type I-B system that represent targeted genomic regions of the episymbiont (in *Ca.* H. crystalense) showed significant decreases in metagenomic coverage compared to untargeted regions of the very same scaffold, which we determined using a sliding window algorithm (see Supplementary Methods) paired with a bootstrapped Wilcoxon paired signed rank test. A significance level of 5% ($p < 0.05$) was tested for each protospacer region and the p-values were FDR-adjusted.

The coverage of the targeted regions, compared to the average coverage of the scaffold decreases in *Ca.* A. crystalense and *Ca.* H. crystalense by 30.13% (mean ± STD, STD=40.79%) and 32.3% (mean ±37.32%), respectively (bootstrapped Wilcoxon signed rank test, n=1000, where for n=990 FDR-corrected p-values were < 0.05, for median please see

main manuscript). In other words, 17% of the self-targeted host regions showed a drop of 30.13% in coverage, while in 30% of the episymbiont-target sites the coverage of the scaffold dropped by 32.3%. The high variation of coverage was expected as scaffold coverage was determined for a microbial population (rather than a discrete species) across multiple environmental samples with varying microbial composition (Probst et al., 2018). Overview of the most important percent values are presented in Table S5.

Complementary metabolism between *Ca.* Altiarchaea or *Ca.* Huberiarchaea

In the CG system, 32 metabolites were predicted to be transferred from *Ca.* A. crystalense to *Ca.* H. crystalense under minimum interactions between these two to produce non-zero biomass for both, including glucose, ATP, amino acids (all except for Asn, Cys, Phe, Tyr), reducing equivalents (i.e. NADH and reduced ferredoxin), formate, chorismate (precursor of aromatic amino acids biosynthesis), and vitamins (Table S10). Eight metabolites were predicted to be transferred from *Ca.* H. cystalense to *Ca.* A. crystalense (Fig. S3.2.1.8 A-C), including malate, oxidized compounds, and dUMP, which is the product of a dCMP deaminase reaction in *Ca.* H. crystalense that is required for completing the pyrimidine biosynthesis in *Ca.* A. crystalense (Fig. S3.2.1.8 C). Similarly, in the HURL system, 24 metabolites were predicted to be transferred from *Ca.* A. horonobense to *Ca.* H. julieae, including glucose, ATP, amino acids (all except for Asp, Asn, Cys, Phe, Tyr), formate, chorismate, dTMP, UTP, and vitamins. Seven metabolites were predicted to be transferred from *Ca.* H. julieae to *Ca.* A. horonobense, including 5,10-Methylenetetrahydrofolate (MLTHF), which is the product of a dTMP synthase function in *Ca*. H. julieae required for completing the folate biosynthesis pathway in *Ca.* A. horonobense (Fig. S3.2.1.8 D).

It is worth noting that certain functions in *Ca.* Huberiarchaea formed complementary pathways for the biosynthesis of essential biomass components in both the CG as well as in the HURL system. In the CG system, metabolic cooperation was observed in pyrimidine biosynthesis. A dUTPase (EC 3.6.1.23) was missing in *Ca.* A. crystalense, which blocks the biosynthesis of dUMP from dUTP, a pathway encoded in *Ca.* A. horonobense. As a result, the biosynthesis of dUMP in *Ca.* A. crystalense relied on a dCMP deaminase (EC 3.5.4.12), which was present in *Ca.* H. crystalense but absent in other *Ca.* Altiarchaea or *Ca.* Huberiarchaea

strains examined in this study. The biosynthesis of dCMP, a substrate for the dCMP deaminase, was absent in *Ca.* H. crystalense. Therefore, *Ca.* A. crystalense was responsible for dCMP biosynthesis via a dCMP kinase (EC 2.7.4.25), while the dCMP deaminase in *Ca.* H. crystalense was responsible for bridging the dCMP biosynthesis with the production of dTTP (Fig. S3.2.1.8 C). Interestingly, although the collaboration of pyrimidine biosynthesis was essential for AltiCG-HuberCG, alternative metabolites could be transferred between *Ca.* A. crystalense and *Ca.* H. crystalense due to the presence of redundant pathways (represented as dotted lines in Fig. S3.2.1.8), which contributed to the robustness of the predicted metabolic collaboration.

In the HURL system, metabolic cooperation was observed in folate biosynthesis. This was mediated by a dTMP synthase (EC 2.1.1.45) function, which was also linked to pyrimidine biosynthesis (Fig. S3.2.1.8 D). The dTMP synthase in *Ca*. H. julieae bridged a missing link in the tetrahydrofolate (THF) biosynthesis pathway of the *Ca.* A. horonobense by converting dihydrofolate (DHF) to 5,10-Methylenetetrahydrofolate (MLTHF). As a byproduct, the dTMP synthase in *Ca*. H. julieae also converted dTMP to dUMP. In collaboration with the FAD-dependent dTMP synthase (EC 2.1.1.148) in *Ca.* A. horonobense, the dTMP and dUMP were cycled while driving the biosynthesis of THF in the AltiHURL-HuberHURL system (Fig. S3.2.1.8 E). Interestingly, while the dTMP synthase was encoded in *Ca.* Altiarchaeum crystalense and *Ca.* H. crystalense, it was only encoded in *Ca*. H. julieae in the HURL system, indicating the observed coupling of folate biosynthesis could be unique to the HURL system.

Single nucleotide polymorphism (SNP) prediction within PAM sequences

Using the predicted spacer target locations and the SNP locations, SNP frequencies in vicinity of spacer targets were calculated for all positions inside the spacer targets relative to the 5’ end, as well as outside the spacer targets in a range of 2000 bases with location relative to the 5’ or 3’ end of the spacer respectively. Regression curves were calculated via locally estimated scatterplot smoothing (LOESS), as well as a fitted polynomial regression model of order three. Visualization was performed with R (Posit team, 2022; Team, 2013) (version 3.6.1). Targeted genes (coverage > 8) of 9 *Ca.* A. crystalense MAGs by spacers extracted from the three samples CG05, CG08, and CG16 were used to calculate the average SNP ratio of the

putative protospacer adjacent motif (PAM) sequence and the residual gene. A Spearman rank correlation test between the SNP frequency and the location relative to the spacers shows a significant negative correlation (exact values of correlation coefficient ρ and p-value, upstream (start) and downstream (end) of the spacer respectively: **CG05** end: ρ = -0.461, p-value < 2.2e-16, start: ρ = -0.564, p-value < 2.2e-16; **CG08** end: ρ = -0.148, p-value = 3.177e-11, start: ρ = -0.439, p-value < 2.2e-16; **CG16** end: ρ = -0.239, p-value < 2.2e-16, start: ρ = -0.663, p-value < 2.2e-16, for neighboring 2000 bases).

In the neighborhood of 500 bases, a significant negative correlation was found for samples CG05 and CG16 (CG05 end: ρ = -0.404, p-value < 2.2e-16, start: ρ = -0.225, p-value = 3.89e-7; CG16 end: ρ = -0.318, p-value = 3.028e-13, start: ρ = -0.174, p-value = 9.574e-5), but not for CG08, where the Spearman correlation is close to 0 and the p-value is > 0.05 (CG08 end: ρ = -0.064, p-value = 0.1558, start: ρ = 0.009, p-value = 0.8471). For the analysis of SNP frequency in the 500bp neighborhood outside of spacers, only self-targeted genes of *Ca.* A. crystalense genomes were selected. The total number of matching spacers and self-targeted genes in each sample were 154 genes and 396 spacers for CG05, 84 genes and 220 spacers for CG08, and 260 genes and 628 spacers for CG16.

In detail, the first base after the spacer matching region in the *Ca.* Altiarchaeum crystalense genome was analyzed separately for elevated SNP frequency. To test whether the mean SNP frequency at this position is higher than can be expected by chance, it is compared to randomly chosen spacer positions. In this procedure, for every spacer matching a stretch of a gene, a random position on this gene was chosen, leading to the same number of spacers and genes, as in the real samples. 100 of such randomized spacer position sets were generated for each sample and the mean SNP frequency of the first base relative to the spacer end was calculated. This analysis shows an increase in SNP frequency ranging from 22% to 69%, at the first base after real spacers compared to random draws. Utilizing the normal distribution of means, as indicated by the central limit theorem, a one-sided Z-test was chosen. The results show that the increase is statistically significant (Z-test) for samples CG08 (55% increase, p-value = 0.005) and CG16 (69% increase, p-value = 4.04e-9). Sample CG05 misses the threshold of significance with a p-value of 0.063 and a SNP frequency increase of 22% over the randomized data.

3. Supplementary Figures

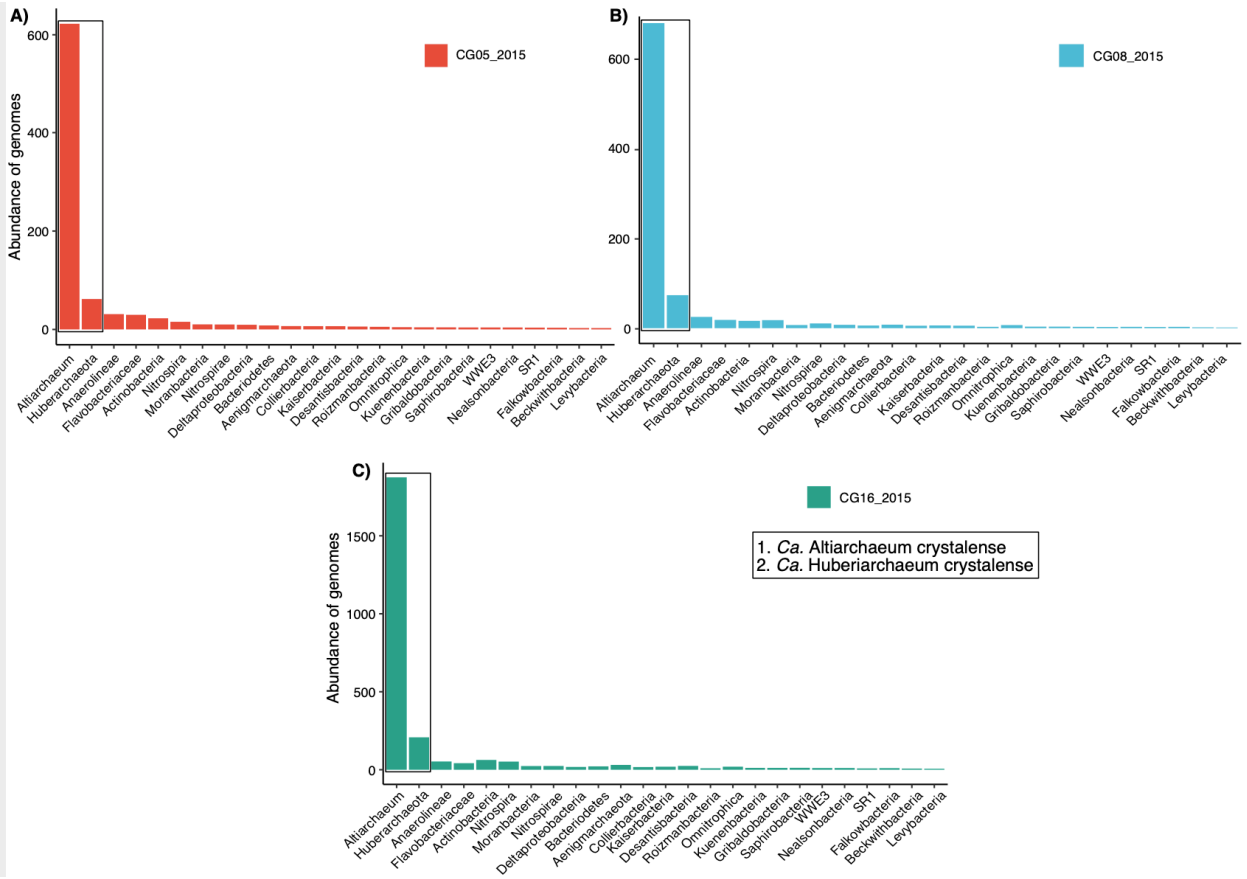


Figure S3.2.1.1 | Rank abundance curves based on relative abundance of genomes in samples A) CG05, B) CG008 and C) CG16 from 2015 (Probst et al., 2018). These three samples were selected due to the availability of corresponding transcriptomes, and because *Ca.* A. crystalense and *Ca.* H. crystalense were the most abundant organisms therein. Rank abundance curves are sorted based on sample CG05 for better comparability. Please note that the y-axes are differently scaled due to varying abundances of other organisms in the samples, and we only displayed organisms determined to occur in the minor eruption phase, i.e., in this subsurface ecosystem, of the Geyser (in total 25 organisms). The calculations of the rank abundances also visualize that the relative abundance ratio of *Ca.* A. crystalense and *Ca.* H. crystalense is constant over these samples, which underlines their host-symbiont relationship, and which was previously published in Schwank et al. (2018) for this ecosystem (Schwank et al., 2019).

Publication 1: Now available as:
Esser, S. P. *et al.* A predicted CRISPR-mediated symbiosis between uncultivated archaea. *Nature Microbiology* **8**, 1619–1633 (2023).

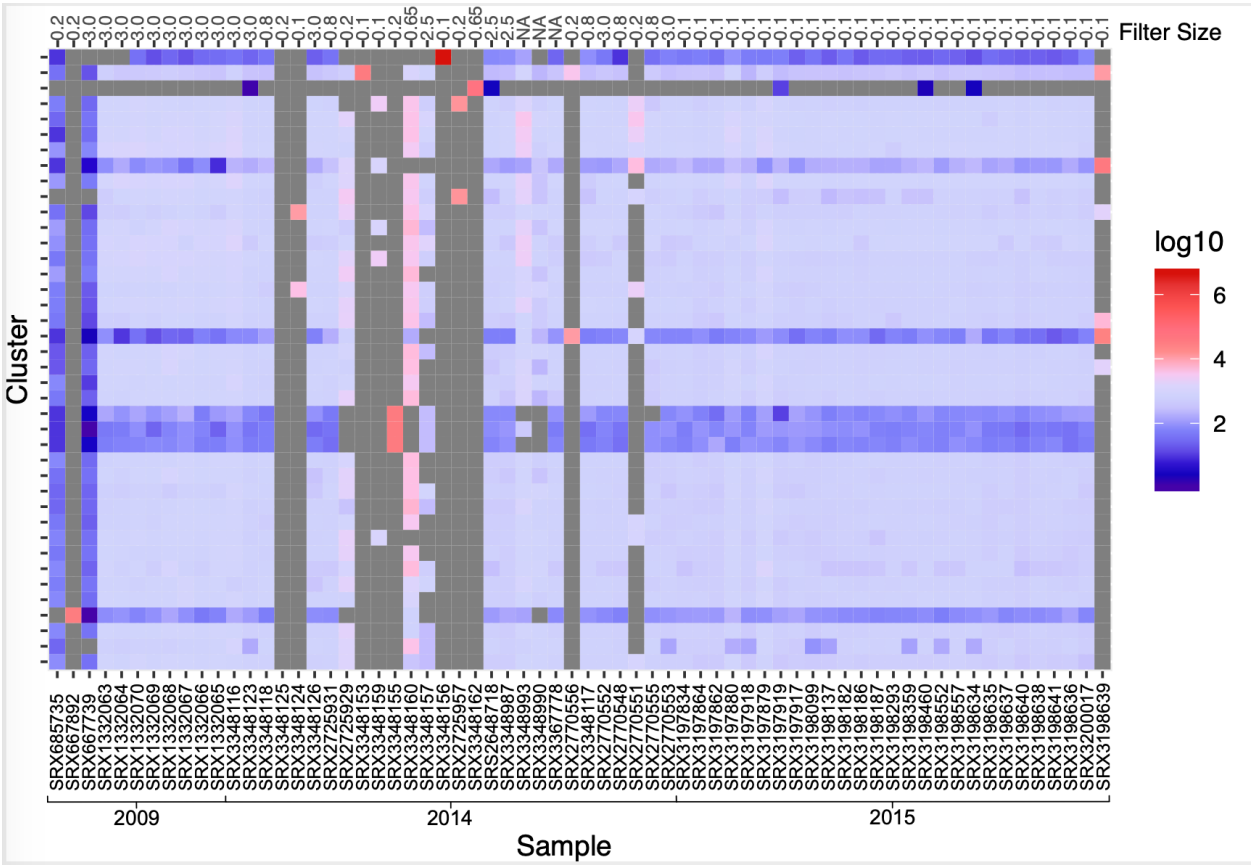


Figure S3.2.1.2 | Relative abundance of spacer clusters from CRISPR system type I-B in Crystal Geyser over a time span of six years and based on 66 metagenome samples (n=66), with x-axis showing the NCBI accession number of the respective sample (Table S1). Spacer clusters are normalized on *Ca.* A. crystalense genome abundances within each sample. The 40 most abundant spacer clusters (out of 297,531 spacer clusters) are depicted. Grey color indicates absence of spacer clusters in a metagenome; please note that some samples were retrieved from sequential filtration through different pore sizes and thus *Ca.* Altiarchaeota genomes were nearly absent in some samples. For sample details, please refer to Table S1. Visualization was performed with R (Posit team, 2022; Team, 2013) (version 3.6.1).

Publication 1: Now available as:
Esser, S. P. *et al.* A predicted CRISPR-mediated symbiosis between uncultivated archaea. *Nature Microbiology* **8**, 1619–1633 (2023).

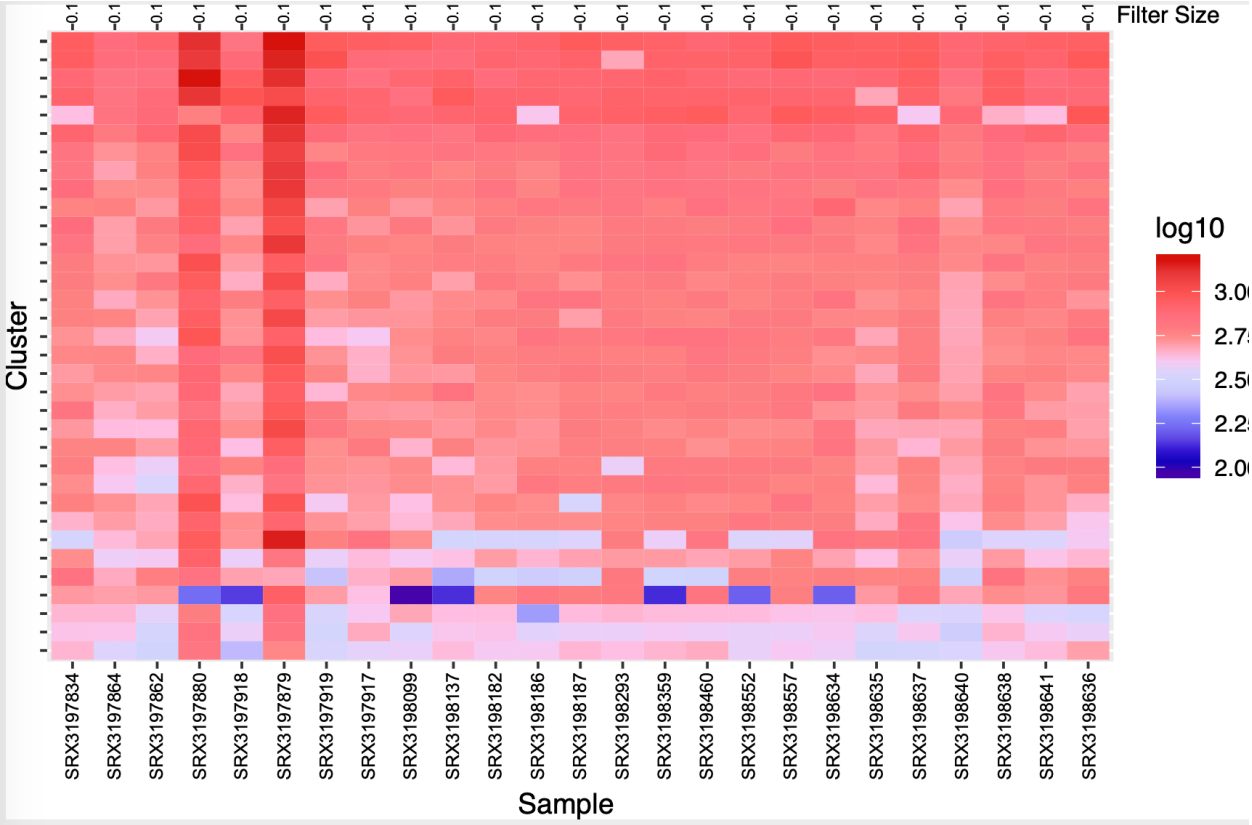


Figure S3.2.1.3 | Relative abundance of spacer clusters from CRISPR system type I-B in Crystal Geyser over one eruption phase in 2015 (n=25), with x-axis showing the NCBI accession number of the respective sample (Table S1). Spacer clusters are sorted in order of sampling. All PCR amplified metagenomic datasets are excluded(Probst et al., 2018). Spacer clusters are normalized on *Ca.* A. crystalense genome abundances within each sample. The 34 most abundant spacer clusters (out of 297,531 spacer clusters) are depicted. For sample details, please refer to Table S1. Visualization was performed with R (Posit team, 2022; Team, 2013) (version 3.6.1).

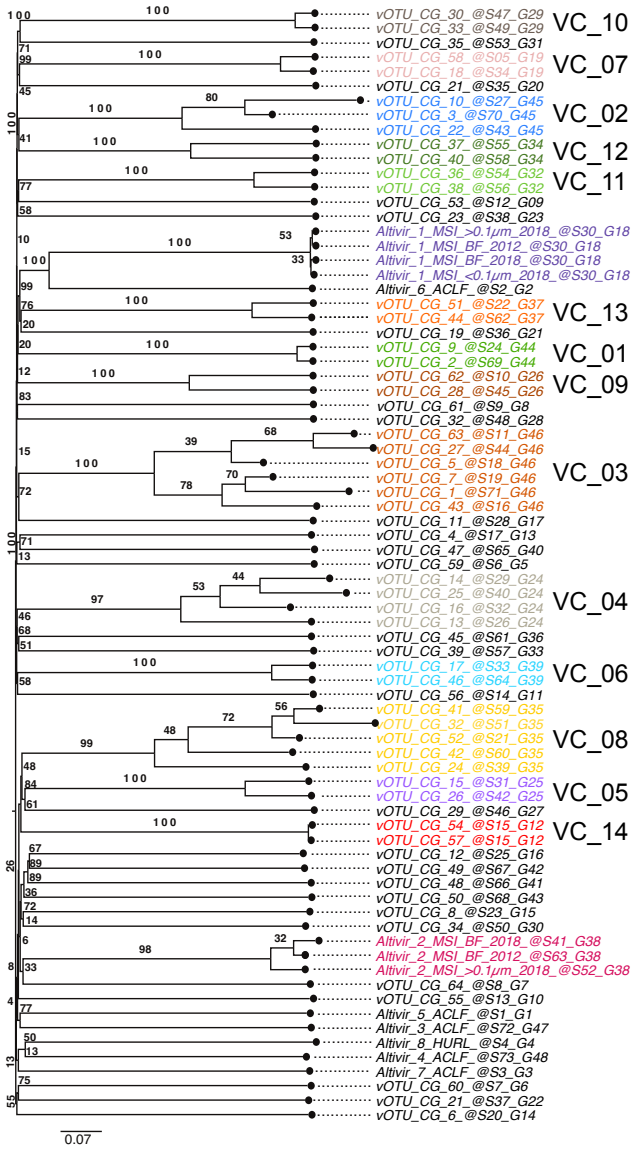


Figure S3.2.1.4 | Clustering of Crystal Geyser vOTUs with previously published altiarchaeotal vOTUs *Altivir_1 - Altivir_8* from other subsurface ecosystems (Rahlff et al., 2021) using VICTOR (Meier-Kolthoff and Göker, 2017). The Figure shows the phylogenomic GBDP (Meier-Kolthoff et al., 2013) tree inferred using the formula D0 yielding average support of 64%. OPTSIL clustering yielded 73 species clusters and 48 genus clusters including previously detected vOTUs (Rahlff et al., 2021). The numbers above branches are GBDP pseudo-bootstrap support values from 100 replications. All identical colored viruses are at least assigned to the same genus. Viral clusters (VC) refer to clusters mentioned in Table S6. Singletons are indicated in

black. VCs formed by Altivir_1_MSI and Altivir_2_MSI were described in Rahlff *et al* 2021 (Rahlff et al., 2021).

Publication 1: Now available as:
Esser, S. P. *et al.* A predicted CRISPR-mediated symbiosis between uncultivated archaea. *Nature Microbiology* **8**, 1619–1633 (2023).

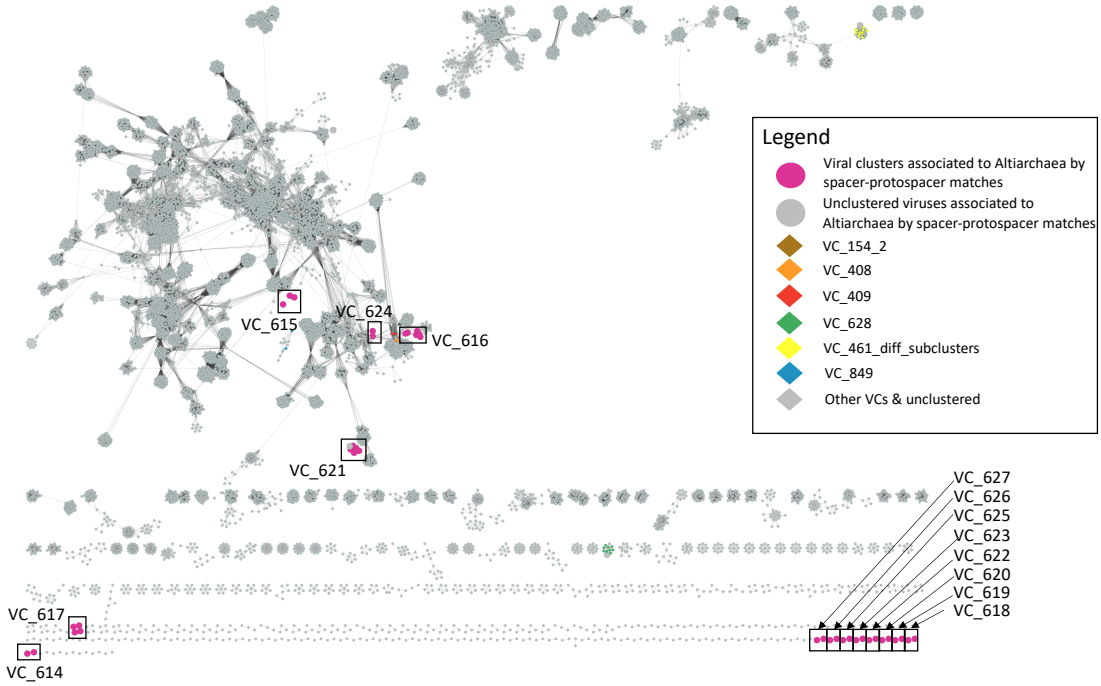


Figure S3.2.1.5 | Network of *Ca.* Altiarchaeum associated vOTUs identified by spacer-protospacer matches, and forming 14 viral clusters (VCs) and their relationships with viruses from a viral Refseq database (Cook et al., 2021) (release July 2022). The network shows interaction (edges) between viruses (nodes) based on shared protein clusters. Most vOTUs form genus clusters with other viruses associated to *Ca.* Altiarchaeum (pink enlarged circles) and cluster with database viruses from VC_145_2, VC_408, VC_409, VC_849. Some unclustered vOTUs (outliers in vConTACT2, grey enlarged circles) were related to viruses from different subclusters of VC_461, VC_628, or other outliers (grey diamonds) of the reference database. For further information about correspondence of VC designation with VC designations in Fig. S3.2.1.4 & Extended Data Fig. 3.2.2 as well as members of the VCs clustering with viruses at higher rank than genus level please refer to Table S6. Data were visualized using Cytoscape v.3.9.0. (Shannon et al., 2003).

Publication 1: Now available as:
Esser, S. P. *et al.* A predicted CRISPR-mediated symbiosis between uncultivated archaea. *Nature Microbiology* **8**, 1619–1633 (2023).

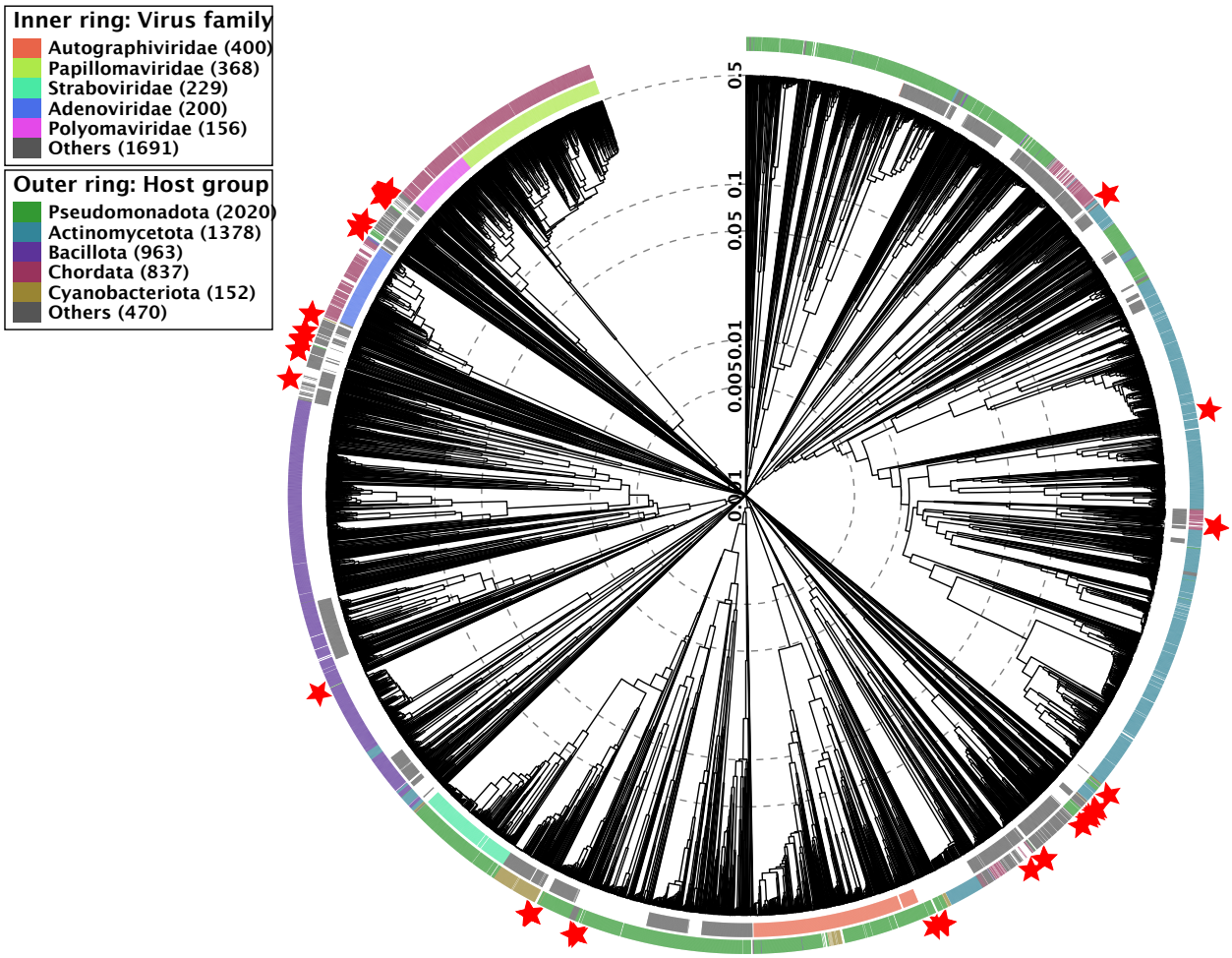


Figure S3.2.1.6 | Circular viral proteomic tree designed by ViPTree (Nishimura et al., 2017). Positioning of vOTUs from this study is indicated by red stars in the tree.

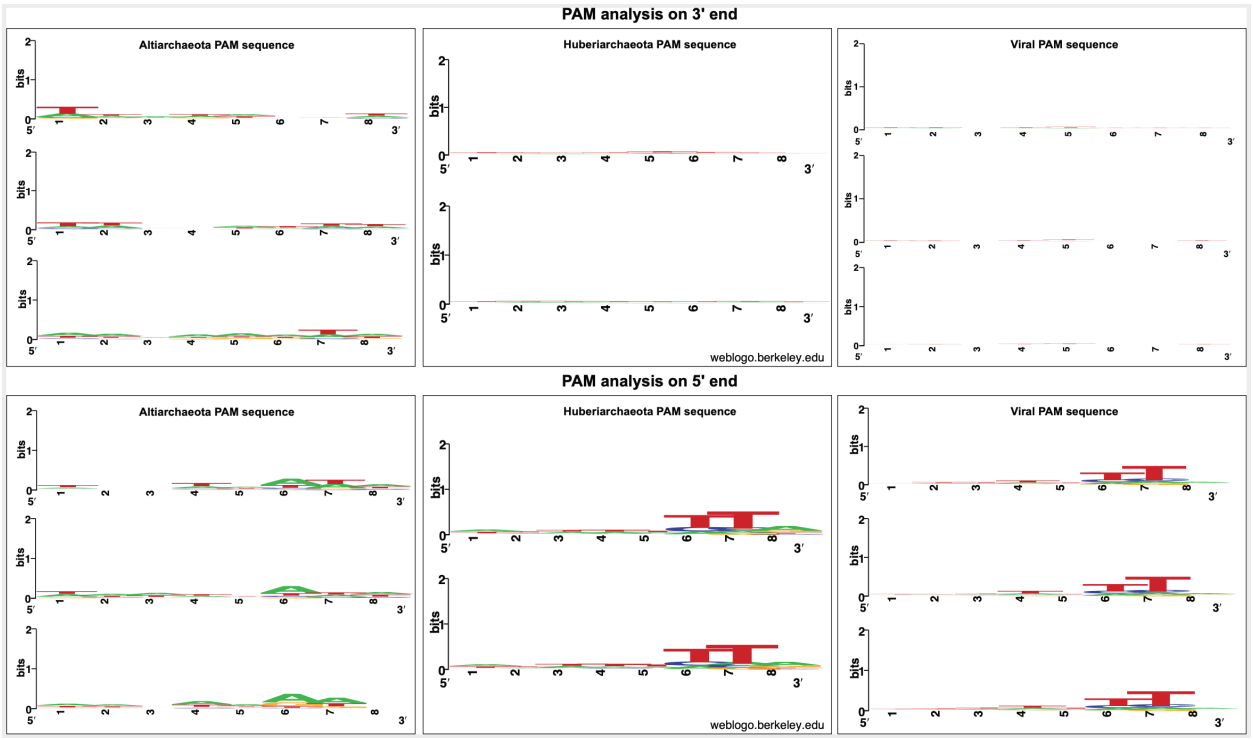
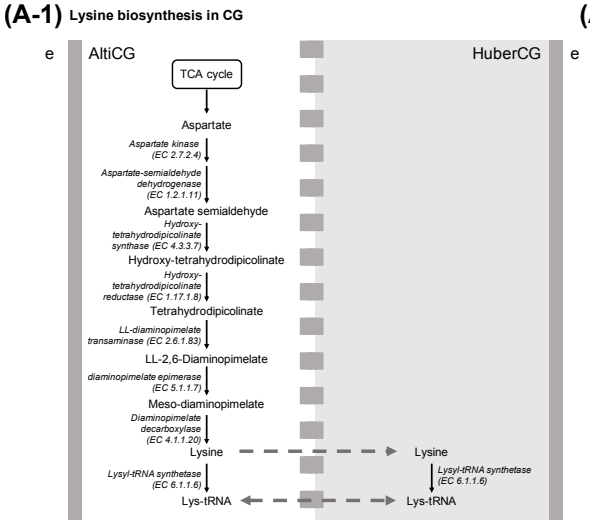
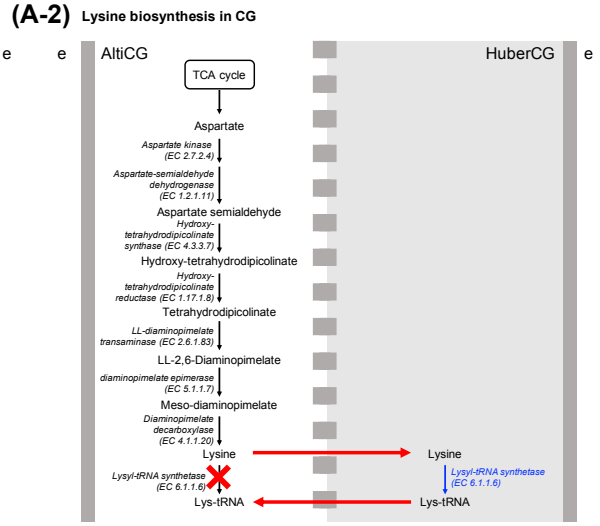
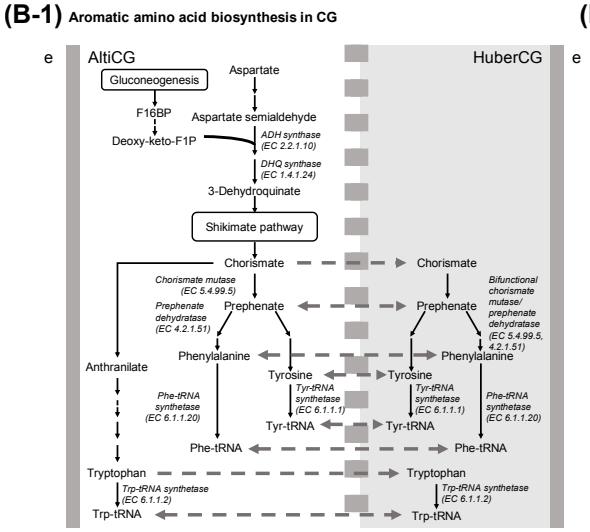


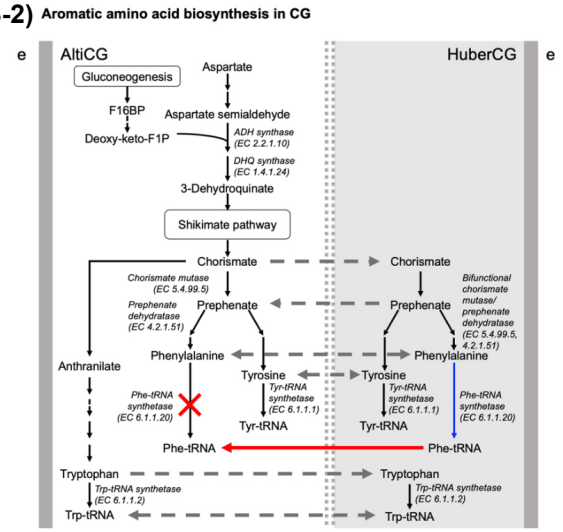
Figure S3.2.1.7 | Protospacer adjacent motifs (PAM) of *Ca.* Altiarchaea, *Ca.* Huberiarchaea and all vOTUs from CG for CRISPR system type I-B analyzed with CRISPRTarget (Biswas et al., 2013) (accessed in June 2020) and WebLogo (Crooks et al., 2004; Schneider and Stephens, 1990) (accessed in June 2020). For the analysis of PAM sequences all matching spacers with 80% similarity were used. As WebLogo has a maximal input of 10,000 spacers for one sequence design, the analysis to create a PAM sequence logo had to be split in two to three batches. We found 23,208, 14,479, and 29,044 spacer matches against 18 *Ca.* A. crystalense genomes, eleven *Ca.* H. crystalense genomes and 64 vOTUs, respectively, with a cutoff score >19 in CRISPRTarget for the identification of a PAM. A 5' consensus PAM of the sequence TTN was found for *Ca.* H. crystalense and the vOTUs, but none was detected for *Ca.* A. crystalense. Contrasting the analysis with spacers from CRISPR system type I-B, the identification of a PAM sequence for the unassigned CRISPR array resulted in no consensus motif sequence (data not shown).

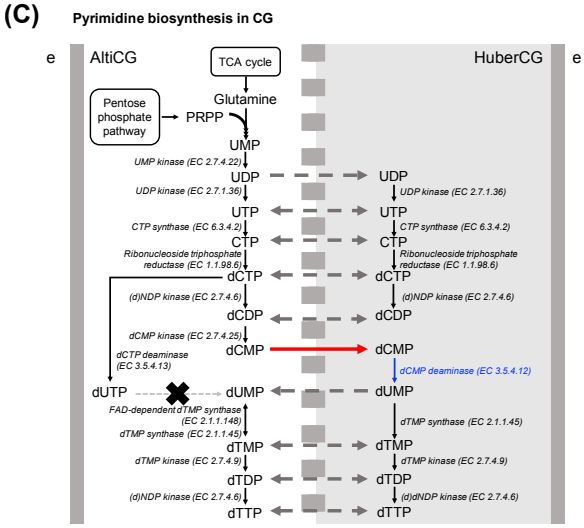
Publication 1: Now available as:
Esser, S. P. *et al.* A predicted CRISPR-mediated symbiosis between uncultivated archaea. *Nature Microbiology* **8**, 1619–1633 (2023).











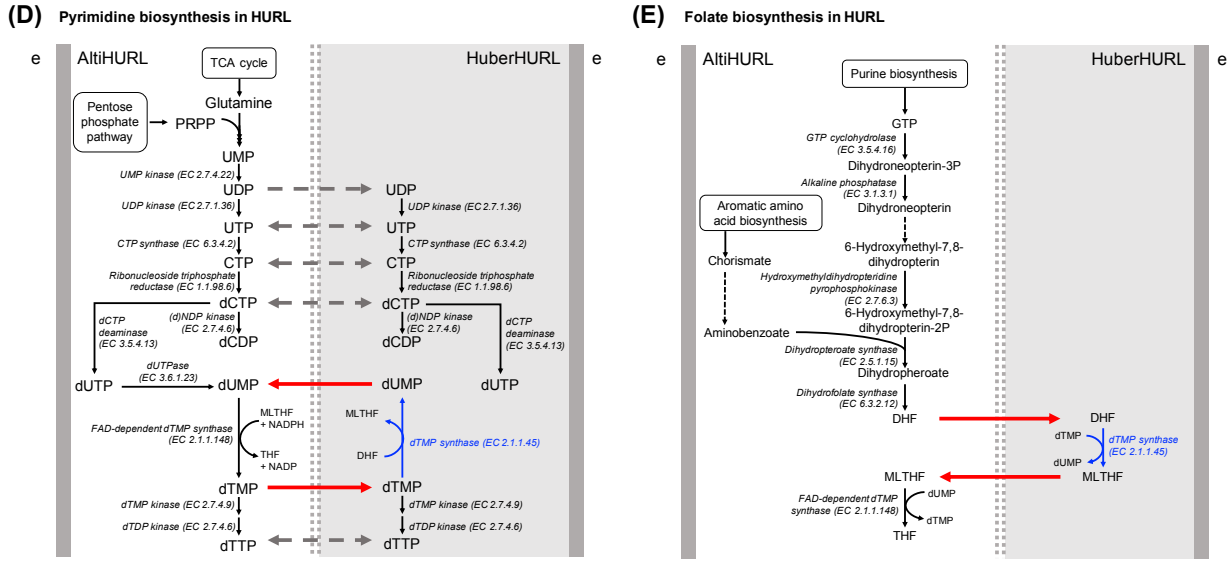


Figure S3.2.1.8 | Complementary metabolic pathways between *Ca.* Altiarchaeum crystalense/horonobense and *Ca.* Huberiarchaeum crystalense/julieae during self-targeting of *Ca.* A. crystalense/horonobense. The dotted arrows symbolize exchange of metabolites between cells. Red arrows symbolize metabolic complementation between *Ca.* A. crystalense/horonobense and *Ca.* H. crystalense/julieae via exchange of metabolites. Blue arrows highlight genes/pathways in *Ca.* H. crystalense/julieae are necessarily involved in the metabolic interactions for this complementation. Depicted metabolic interactions without considering CRISPR-Cas interference within CG (**A-1 and B-1).** Red crosses indicate the potential deletion of metabolic functions due to attack of genes by self-targeting spacers within CG **(A-2, B-2, C-E)**. In HURL, the necessary interactions within the pyrimidine and folate biosynthesis are visualized **(D and E).**

Publication 1: Now available as:
Esser, S. P. *et al.* A predicted CRISPR-mediated symbiosis between uncultivated archaea. *Nature Microbiology* **8**, 1619–1633 (2023).

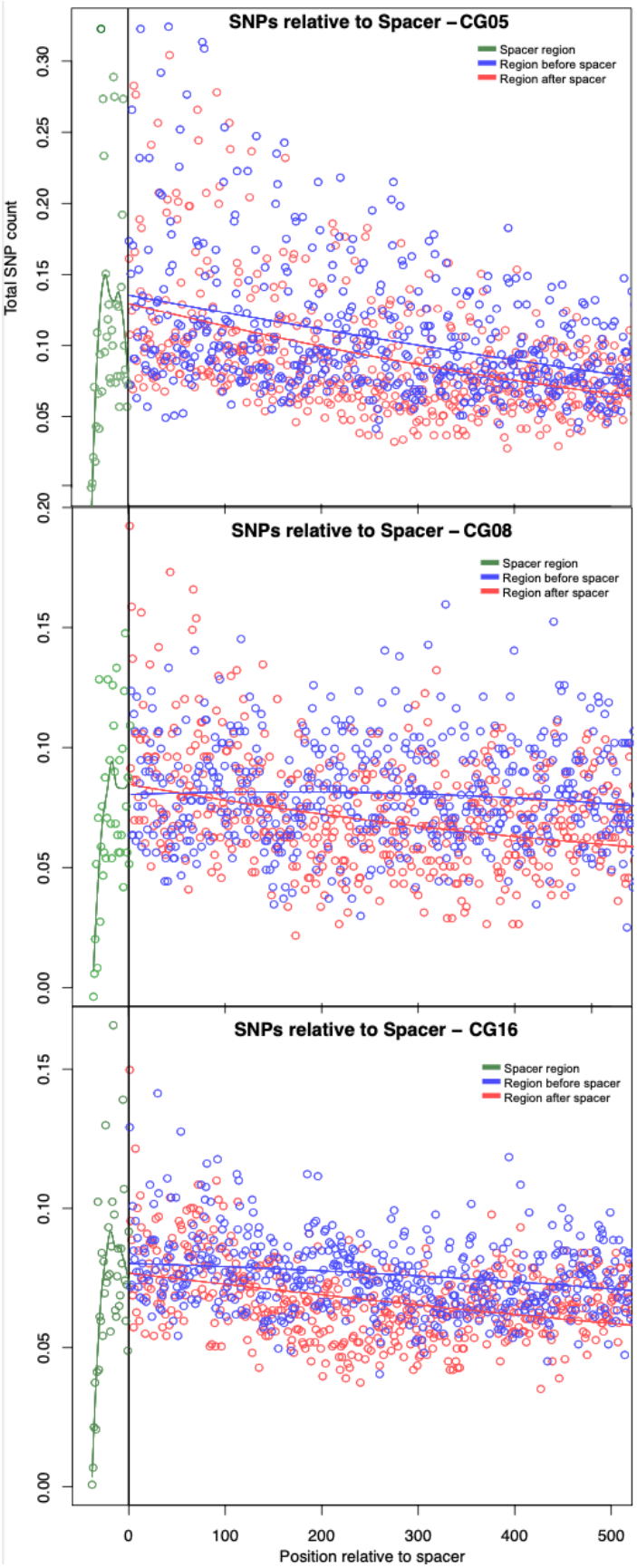


Figure S3.2.1.9 | Single nucleotide polymorphism (SNP) calling on all *Ca.* Altiarchaeum crystalense genes targeted by spacers of the CRISPR type I-B system (self-targeting). Depicted are the protospacer region (green), the bases upstream of the spacers (blue) and the bases downstream of the protospacer matching region (red). CRISPR spacer targets are oriented 3' to 5' with the 5' end being adjacent to the PAM motif.

Self-targeting of *Ca.* Altiarchaea (host)		
CRISPR system	**CG (percent spacer matches)**	**HURL (percent spacer matches)**
Type I-B	DNA forward: 42% vs. DNA reverse: 58%	DNA forward: 77% vs. DNA reverse: 23%
Unassigned array	DNA forward: 0% vs. DNA reverse: 100%	DNA forward: 84.5% vs. DNA reverse: 15.5%
Targeting of *Ca.* Huberiarchaea (episymbiont)		
CRISPR system	**CG (percent spacer matches)**	**HURL (percent spacer matches)**
Type I-B	DNA forward: 44% vs. DNA reverse: 56%	N.A.
Unassigned array	N.A.	DNA forward: 88.8% vs. DNA reverse: 11.2%

Supplementary Table S3.2.1.7 | Strand-specific analysis of targets of *Ca.* Altiarchaeum-derived spacers to determine the specificity of the two CRISPR-Cas systems. While CRISPR system type I-B is targeting the forward DNA strand in both ecosystems (for self-targeting and for targeting of the episymbiont), the unassigned CRISPR array seems to target solely the reverse strand of DNA (targeting threshold 90%) in CG and the forward strand of DNA in HURL. The analysis includes self-targeting spacers of *Ca.* Altiarchaea and spacers targeting *Ca.* Huberiarchaea genomes via matching to prodigal-predicted CDS (Hyatt et al., 2010). In total 16 *Ca.* A. crystalense genomes, 11 *Ca.* H. crystalense genomes, and one genomes of each *Ca.* A. horonobense and *Ca.* H. julieae were used for this analysis. Please note that the number of spacers matching to the genomes are different to the numbers in the main figure 1C and 1E, because only spacers matching in the CDS regions were analyzed here. Percentages describe the proportion of spacers matching the forward strand and/or the reverse strand, respectively. N.A. = not available (no spacers/insufficient number of spacers available for analysis).

Publication 1: Now available as:
Esser, S. P. *et al.* A predicted CRISPR-mediated symbiosis between uncultivated archaea. *Nature Microbiology* **8**, 1619–1633 (2023).

4. **List of Supplementary Tables and Data**

These tables are provided as separate files on the supporting CD within the folder Publication_1_Section_3.2.

Supplementary Table S1 | Overview of accessed metagenomes and metatranscriptomes. Accession numbers, read length, number of bases sequences, number of assembled sequences, number of extracted spacers for CRISPR system type I-B and the unassigned CRISPR array.

Supplementary Table S2 | Accession numbers for *Ca.* Altiarchaeum crystalense/horonobense and *Ca.* Huberiarchaeum crystalense/julieae genomes.

Supplementary Table S3 | Accession numbers and JGI IMG annotation of single amplified genomes (SAGs) from Crystal Geyser.

Supplementary Table S4 | Accession numbers of publicly available archaeal genomes accessed May 2021.

Supplementary Table S5 | Overview of annotated genes matched by CRISPR spacers derived from metagenomes of the samples CG05, CG08 and CG16 (Table S5.3) and of single amplified genomes. The table contains all genes of *Ca.* A. crystalense/horonobense (Table S5.1 and S5.4, respectively) and *Ca.* H. crystalense/julieae (Table S5.2 and Table S5.5, respectively) annotated with DIAMOND against the UniRef100 database and manual annotations of 101 genes matched by the spacers targeted by spacers. These datasets were used as input for genome-resolved metabolic modeling and single nucleotide polymorphism (SNP) analysis. Calculated coverage drops of targeted scaffolds in *Ca.* A. crystalense and *Ca.* H. crystalense (Table S5.6). Extended explanations of the individual data sheets are provided as content sheet in the document.

Supplementary Table S6 | Viral scaffolds targeting *Ca.* Altiarchaeum crystalense as determined by spacer-protospacer matches and application of different tools for virus detection(Rahlff et al., 2021). Table shows putative origin of scaffold region (viral, host, unclassified) as determined by CheckV(Nayfach et al., 2021), genome completeness, contig length, number of open reading frames (ORF), GC content, coding density, and viral genus and species clusters as determined using vConTACT2, VICTOR and VIRIDIC. Depending on the

clustering tool used, 40 to 46 new viral clades (=sum of different viral clusters (VC), singletons, outliers) could be identified. n.d.=none determined, ^{T} indicates that viruses that were targeted by CRISPR spacers from the transcriptome, ^{C} indicates circular viral scaffolds.

Supplementary Table S8 | Genome-scale model of Alti-Huber metabolism in the CG system. Listed are the used basic reactions (Table S8.1) and compounds (Table S8.2) for the metabolic modeling. In addition, Table S8.3 describes the genes involved in the reactions. Unlimited interactions (Table S8.4) describes the interactions between *Ca.* Altiarchaea and *Ca.* Huberiarchaea if compounds are freely moving between the cells and the limited interactions (Table S8.5) only allows the interactions needed to result a non-zero biomass production. Table S8.6 lists the chemical compounds measured (Probst et al., 2018) or assumed to calculate the metabolic model. The biomass objective functions for *Ca.* Altiarchaea and *Ca.* Huberiarchaea are listed in Table S8.7-S8.8. Extended explanations of the individual data sheets are provided as content sheet in the document.

Supplementary Table S9 | Genome-scale model of Alti-Huber metabolism in the HURL system. Listed are the used basic reactions (Table S9.1) and compounds (Table S9.2) for the metabolic modeling. In addition, Table S9.3 describes the genes involved in the reactions. Unlimited interactions (Table S9.4) describes the interactions between *Ca.* Altiarchaea and *Ca.* Huberiarchaea if compounds are freely moving between the cells and the limited interactions (Table S9.5) only allows the interactions needed to result a non-zero biomass production. Table S9.6 lists the chemical compounds measured (Hernsdorf et al., 2017) or assumed to calculate the metabolic model. The biomass objective functions for *Ca.* Altiarchaea and *Ca.* Huberiarchaea are listed in Table S9.7-S9.8. Extended explanations of the individual data sheets are provided as content sheet in the document.

Supplementary Table S10 | Results of flux variability analysis (FVA) on both wild-type and spacer-attacked Alti-Huber models in the CG system. Table S10.1 and Table S10.2 describe genes targeted by spacers and deleted genes within *Ca.* Altiarchaeum crystalense by self-targeting, respectively. The FVA simulation data of limited interactions and unlimited interactions of *Ca.* Altiarchaeum crystalense and *Ca.* Huberiarchaeum crystalense with no spacer to protospacer matches is listed in Table S10.3 and Table S10.4, respectively. Table S10.5 and Table S10.6 contains the FVA data while self-targeted genes are deleted with

unlimited interactions of *Ca.* Altiarchaeum crystalense and *Ca.* Huberiarchaeum crystalense, respectively. Extended explanations of the individual data sheets are provided as content sheet in the document.

Supplementary Table S11 | Results of flux variability analysis (FVA) on both wild-type and spacer-attacked Alti-Huber models in the HURL system. The FVA simulation data of limited interactions and unlimited interactions of *Ca.* Altiarchaeum horonobense and *Ca.* Huberiarchaeum julieae with no spacer to protospacer matches is listed in Table S11.1 and Table S11.2, respectively. Table S11.3 describe genes targeted by spacers and deleted genes within *Ca.* Altiarchaeum horonobense by self-targeting. Extended explanations of the individual data sheets are provided as content sheet in the document.

Supplementary Table S12 | CRISPR spacer matches of publicly available archaeal genomes in Guaymas Basin, and all spacer matched (above 3 matched per genome) within the NCBI public archaeal genomes.

Supplementary Table S13 | Amino Acid sequences of *cas* genes that were synthesized for the PAM and TXTL assay.

Supplementary Data | Collection of all phylogentic trees calculated for this study. (1) Phylogenetic tree of *Ca.* Altiarchaeum crystalense/horonobense and *Ca.* Huberiarchaeum crystalense/julieae located within the archaeal branch. (2) Phylogenetic analysis of the phenylalanine—tRNA synthetase of *Ca.* Huberiarchaeum crystalense located within the archaeal branch. Calculated to determine the closest relative within the archaeal branch. (3) Phylogenetic analysis of the phenylalanine—tRNA synthetase of *Ca.* Altiarchaeum crystalense located within the archaeal branch. Calculated to determine the closest relative within the archaeal branch. (4) Phylogenetic analysis of the lysine—tRNA synthetase of *Ca.* Huberiarchaeum crystalense located within the archaeal branch. Calculated to determine the closest relative within the archaeal branch. (5) Phylogenetic analysis of the lysine—tRNA synthetase of *Ca.* Altiarchaeum crystalense located within the archaeal branch. Calculated to determine the closest relative within the archaeal branch.

Publication 1: Now available as:
Esser, S. P. *et al.* A predicted CRISPR-mediated symbiosis between uncultivated archaea. *Nature Microbiology* **8**, 1619–1633 (2023).

Publication 2: Now available as: Esser, S. P. *et al.* Differential expression of core metabolic functions in *Candidatus* Altiarchaeum inhabiting distinct subsurface ecosystems. *bioRxiv* 2023.11.20.567779 (2023).

3.3. Publication 2: **Differential expression of core metabolic functions in DPANN Archaea of distinct subsurface ecosystems**

Sarah P. Esser[1,*], Victoria Turzynski[1], Julia Plewka[1], Carrie Moore[1], Indra Banas[1], André R. Soares[1], Janey Lee[2], Tanja Woyke[2], Alexander J. Probst[1,3,4,*]

[1] Environmental Metagenomics, Research Centre One Health Ruhr of the University Alliance Ruhr, Faculty of Chemistry, University Duisburg-Essen, 45141 Essen

[2] DOE Joint Genome Institute, One Cyclotron Rd, Berkeley, CA, 94720, USA

[3] Centre of Water and Environmental Research (ZWU), University of Duisburg-Essen, Universitätsstraße 5, 45141, Essen, Germany

[4] Center for Medical Biotechnology (ZMB), University of Duisburg-Essen, 45141 Essen, Germany

[*]co-corresponding authors: sarah.esser@uni-due.de, alexander.probst@uni-due.de

ISME communications: Brief communications

Abstract (max. 200 words):

Candidatus Altiarchaea are widespread across aquatic subsurface ecosystems and possess a highly conserved core genome (Bornemann et al., 2022), yet adaptations to different biotic and abiotic factors of the core genome based on gene expression remain unknown. Here, we investigated the metatranscriptome of two *Ca.* Altiarchaeum populations that thrive in ecosystems with substantial differences; while Crystal Geyser in USA is a high-CO_2 groundwater system, and *Ca.* Altiarchaeum crystalense co-occurs with the symbiont *Ca.* Huberiarchaeum crystalense, the Muehlbacher Schwefelquelle in Germany is an artesian spring high in sulfide concentration, where *Ca.* A. hamiconexum is heavily infected with viruses (Rahlff et al., 2021; Turzynski et al., 2023). We mapped metatranscriptomic reads against their genomes to analyze the expression profile of their core genomes. Out of 537 shared gene clusters, 331 were functionally annotated and 130 showed a significant

difference between the two sites. Main differences were related to cell defense like CRISPR-Cas, virus defense, replication, and transcription as well as energy and carbon metabolism. Our results demonstrate that altiarchaeal populations in the subsurface are highly adapted to their environment likely driven by other biological entities that temper with their core metabolism. We consequently posit that viruses and symbiotic interactions can be major energy sinks for organisms in the deep biosphere.

(Originality-Significance Statement:

Organisms of the uncultivated phylum *Ca.* Altiarchaeota are globally widespread and fulfill essential roles in carbon cycling, e.g., carbon fixation in the continental subsurface. Here, we show that the activity of organisms in the continental subsurface differ significantly depending on the geological and microbial setting of the ecosystem explaining many of the previously observed physiological traits of this organism group.)

Main (approx. 1200-1500) (no specific boundaries set):

The aquatic deep subsurface houses some of the most diverse and complex ecosystems on Earth, which vary in their chemical, physical and biological parameters and are therefore ecological niches for differently adapted microorganisms. DPANN (Diapherotrites, Parvarchaeota, Aenigmarchaeota, Nanohaloarchaeota and Nanoarchaeota) archaea are alongside with other archaea and bacteria found in various aquatic deep subsurface ecosystems such as marine and terrestrial geysers, lakes and boreholes (Castelle et al., 2015; Dombrowski et al., 2019; Huber et al., 2002; Momper et al., 2017; Ortiz-Alvarez and Casamayor, 2016). To cope with their little metabolic capacity DPANN archaea are known to be able to live in symbiosis with other archaea, with the best studied example of *Ignicoccus hospitalis* and *Nanoarchaeota equitans* (Huber et al., 2002; Jahn et al., 2008, 2004; Paper et al., 2007), but also recently described DPANN-host associations from the realm of Micrarchaeota (Sakai et al., 2022). One exception from this rule is *Ca.* Altiarchaeum, a geographically widespread genus of organisms that live freely as carbon fixing organism in the deep subsurface. Recent investigations of *Ca.* Altiarchaea showed site-specific genomic adaptations likely occurring due to horizontal gene transfer, yet these organisms harbor a

highly conserved core genome that follows a strict biogeographic pattern (Bornemann et al., 2022).

Using metagenomic abundance correlations and FISH imaging *Ca.* Altiarchaea were shown to be a host of the DPANN archaea *Ca.* Huberiarchaea in chemically and biologically distinct ecosystems Crystal Geyser (Probst et al., 2018, 2017; Schwank et al., 2019) (CG; Utah, USA) and the Horonobe Underground Research Laboratory (Hernsdorf et al., 2017) (HURL; Hokkaido, Japan). In contrast to these symbiotic systems, *Ca.* Altiarchaea was originally found to be a biofilm-forming archaeon in the subsurface without its symbiont in the Muehlbacher sulfidic spring (Moissl et al., 2002; Probst et al., 2014; Probst and Moissl-Eichinger, 2015; Rahlff et al., 2021) (MSI; Regensburg, Germany). However, little is known about the differences in the expression profile of the core genome of *Ca.* Altiarchaea across different ecosystems, particularly also as the symbiont *Ca*. Huberiarchaeon might highly influence host's metabolism (Publication 1, Section 3.2).

This study focuses on metatranscriptomic expression profiles of *Ca.* Altiarchaeum with its episymbiont *Ca.* Huberiarchaeum present in CG compared to MSI, where the symbiont is absent. We propose that the differences in the expression profile are not only influenced by the differing chemical composition of the two ecosystems, but also by the presence of the episymbiont in Crystal Geyser. The two sampling sites, MSI and CG, are located in Regensburg, Germany (N 48 59' 8.999''; O 12 7' 38.459'') and in Utah, USA (N 38 56' 18.125''; W 110 8' 7.389''), respectively, and also differ in chemical composition of bio-processable ions and other molecules in their groundwater (Probst et al., 2018; Rudolph et al., 2001) (summarized in Fig. 3.3.1). Particularly, the chemical composition of CG varies throughout the phases of eruption (Probst et al., 2018). When comparing the ion composition of MSI to CG's minor eruption phase that sources groundwater from the deepest intersected aquifer that is dominated by *Ca.* Altiarchaea (Probst et al., 2018; Rudolph et al., 2001), the former sampling site appears to be rather limited in ion and nutrient availability (Fig. 3.3.1). MSI as a sulfidic spring with high levels of sulfide, has a three-fold lower sulfate concentration than CG. Interestingly, also nitrate, sodium and potassium have a lower abundance in MSI compared to CG, which might influence the necessity to form biofilms for nutrient retention and filtration from groundwater. This is in agreement with fact, that *Ca.* Altiarchaea are

predominantly present as single cells in CG (Probst et al., 2018) (Fig. 3.3.1, Fig. S3.3.1.1) and almost purely as biofilms in MSI, where the cells interconnect with *hami* (Probst et al., 2014; Probst and Moissl-Eichinger, 2015; Rahlff et al., 2021) (Fig. 3.3.1, Fig. S3.3.1.2). In addition to these geochemical and ecophysiological differences of Altiarchaea in the two ecosystems, *Ca.* A. crystalense also shows the presence of an attached episymbiont, namely *Ca.* Huberiarchaeum crystalense, in CG (Fig. 3.3.1, Fig. S3.3.1.1-S3.3.1.2) (Probst et al., 2018; Schwank et al., 2019). By contrast, this episymbiont has not been detected in metagenomes from MSI (Probst et al., 2014; Rahlff et al., 2021; Publication 1, Section 3.2). Irrespective of these physiological differences and site-specific horizontal evolution of *Ca.* Altiarchaea, their core genome is heavily conserved across many deep subsurface sites (Bornemann et al., 2022) rendering it the ideal genus for studying differential gene expression with respect to different environmental conditions.

Predicted genes of previously published genomes from CG and MSI (Table S2) were clustered at 80% nucleotide similarity. Of the 537 shared gene clusters (Fig. 3.3.2A), 430 were assigned a functional annotation. The 107 remaining gene clusters were either not annotated (no hit in FunTaxDB 1.2) or annotated as uncharacterized proteins. From the 430 annotated gene clusters, 94 were within the first or last 200-bps of the respective scaffold causing irregularities in transcriptome mapping (underestimation of coverage). These genes were also excluded from downstream statistical analysis. The remaining 336 gene clusters were sorted according to their functional annotation, whereby the overall differential expression for most gene clusters in CG (~90.8%, n=305) was higher than in MSI (~9.2%, n=31).

Based on DESeq2 calculated Log2FoldChange (below -1.47 and above 1.44; FDR corrected p-values < 0.05), the overall regulation of genes revealed a significant upregulation of 76 genes in MSI compared to 54 genes in CG. Genes overexpressed in the latter ecosystem included Modification Methylase MboII, a Type II site-specific deoxyribonuclease, a FAD synthase, and an RNA synthase (Fig 2B). By contrast, MSI upregulated genes included ribosomal proteins (e.g., S11, S28e, L35Ae), a phage shock protein, a cell division protein FtsZ, and a Thermosome subunit (Fig. 3.3.2B).

Focusing on the difference in the expression profile of shared protein clusters (Fig. 3.3.2C, Fig. S3.3.1.3-S3.3.1.10) it was evident that the increase in expression of any given gene

that showed a significant difference between the two sites, was substantially greater in CG than in MSI (Fig. S3.3.1.11). Particularly, some genes related to replication (*e.g.*, DNA polymerases, Table S3-S4) and nutrient metabolism have an up to two-fold higher expression in CG than in MSI (Fig S3.3.1.5 and S3.3.1.8) indicating that *Ca.* A. crystalense is more active and replicating than *Ca.* A. hamiconexum when sampling the respective ecosystem. However, the expression of FtsZ, a protein involved in forming the septum of a dividing *Ca.* A. hamiconexum (Probst et al., 2014; Probst and Moissl-Eichinger, 2015; Rahlff et al., 2021) cell seems upregulated in MSI (Fig. 3.3.2C). The accumulation of this protein, which usually has similar concentrations in the cell irrespective of cell division, can be used as an indicator of a starting cell division process as it localizes at midcell early in the division process shown for *E. coli* (Den Blaauwen Tanneke et al., 1999). In addition, the elongation factor 1 alpha, which is included in the aminoacyl tRNA incorporation in archaea and eukaryotes (Xu et al., 2022), and the thermosome subunit, which represents the chaperonin family in archaea and is accordingly involved in the protein folding (Phipps et al., 1993, 1991), is up to two fold higher expressed in MSI then in CG. This can also be an indicator that the cell division in MSI is upregulated. Therefore, the upregulation of the FtsZ gene indicates a higher replication rate of *Ca.* Altiarchaeum hamiconexum in MSI and supports the visible diploidy of the cells in previously published and here shown FISH images (Henneberger et al., 2006; Probst et al., 2014) (Fig. 3.3.1). Comparing the function of the cellular replication genes upregulated in CG (*i.e.*, DNA polymerases) versus MSI (i.e., cell division process, it is evident that *Ca.* A. crystalense appears to heavily replicate the genome but somehow does not proceed the cell division cycle as corresponding cells in MSI. Previous cell biology studies on archaea cell division showed, that the depletion of the FtsZ proteins in *Haloferax volcanii* inhibit cell division, although DNA replication is still ongoing (Liao et al., 2021). We propose that although the DNA synthesis is very prominent in Altiarchaea from CG, their final cell division seems hampered, likely due to the presence of the symbiont *Ca.* H. crystalense or many different viruses in CG that show infection histories with this organism (Publication 1, Section 3.2).

Ca. A. hamiconexum has been described to be infected by at least two different viruses, one of which is a lytic virus (Rahlff et al., 2021; Turzynski et al., 2023). In agreement with these findings, our differential expression analysis revealed a significant increase of

phage shock proteins, which are a stress response when the membrane gets penetrated by invading MGEs (Brissette et al., 1990), in MSI compared to CG. By contrast, CRISPR Cas 5, which is a protein involved in the cascade building for the splicing mechanisms in CRISPR type I systems (reviewed by Hille and Charpentier, 2016) is upregulated in for *Ca.* A. crystalense (Fig. S3.3.1.2C), which shows a tremendous spacer variety over six years not only against MGEs but also against its episymbiont (Publication 1, Section 3.2). Consequently, we identified a specific adaptation to defense against lytic viruses and the episymbiont, respectively.

Beyond upregulation of the CRISPR system, the interaction of the episymbiont might also be responsible for the upregulation of other metabolic functions in *Ca.* A. crystalense. For example, multiple genes encoding for proteins involved in energy metabolism such as the quinolinate synthase, FO synthase subunit 2 and the FAD synthase (Fig. 3.3.2C, Fig. S3.3.1.5-S3.3.1.6) were significantly enriched in the CG transcriptome. Increased energy demands might stem from the highly active CRISPR Cas system which acquired hundreds of thousands of different spacers in the *Ca.* Altiarchaea population (Publication 1, Section 3.2). In addition, the polysaccharide biosynthesis was also found to be upregulated in CG, although we only seldomly found biofilms of *Ca.* Altiarchaea in CG compared to MSI (Fig. 3.3.1). This upregulation could either be related to the scavenging nature of the episymbiont (Schwank et al., 2019) or indicate an intrinsic tendency of the *Ca.* A. crystalense to form biofilms without success due to the turbulent geyser system (Probst et al., 2018).

In sum, the analyses of the differential gene expression profiles of two *Ca.* Altiarchaea populations from distinct geological settings are influenced by environmental factors and biological interactions. While the populations in MSI appear to be heavily influenced by viral attacks, the episymbiont in CG and the turbulence of the geyser system seem to upregulate the CRISPR system, the energy metabolism, and the biofilm formation. Consequently, this study contributes to a long-standing, DNA sequencing-based body of literature on the deep biosphere in general and on *Ca.* Altiarchaea in particular by leveraging metatranscriptomes of low-biomass deep subsurface ecosystems.

ACKNOWLEDGEMENTS

This effort was funded by the Ministerium für Kultur und Wissenschaft des Landes Nordrhein-Westfalen ("Nachwuchsgruppe Dr. Alexander Probst") and the German Science Foundation under project NOVAC (grant number DFG PR1603/2-1). J.P. was supported by Lundin Energy Norway AS within the framework of the GeneOil Project. We thank Christopher T Brown (UC Berkeley) to provide the image of Crystal Geyser, Ken Dreger for exemplary server administration, and Sabrina Eisfeld, Ines Pothmann and Maximilliane Ackers for administrative support.

CONFLICT OF INTEREST

The authors declare no conflict of interest.

Data availability statement

All metatranscriptomics datasets are published under the accession number SAMN14515498, SAMN14515403, SAMN14515402 for CG and for MSI, in the scope of this thesis available upon request or under the following server path (/data4/IMS_Transcriptome/raw.d/). The accession numbers of metagenomic assembled genomes are listed in Table S1.

Author Contribution

SPE performed RNA extraction for MSI and bioinformatical analysis of this study. VT did FISH imaging with assistance of IB. JL performed RNA extractions for CG. JP, CM and AJP helped with bioinformatical assistance. SPE and AJP conceptualized the study. The manuscript was written by SPE and AJP with input from all co-authors.

Publication 2: Now available as: Esser, S. P. *et al.* Differential expression of core metabolic functions in *Candidatus* Altiarchaeum inhabiting distinct subsurface ecosystems. *bioRxiv* 2023.11.20.567779 (2023).

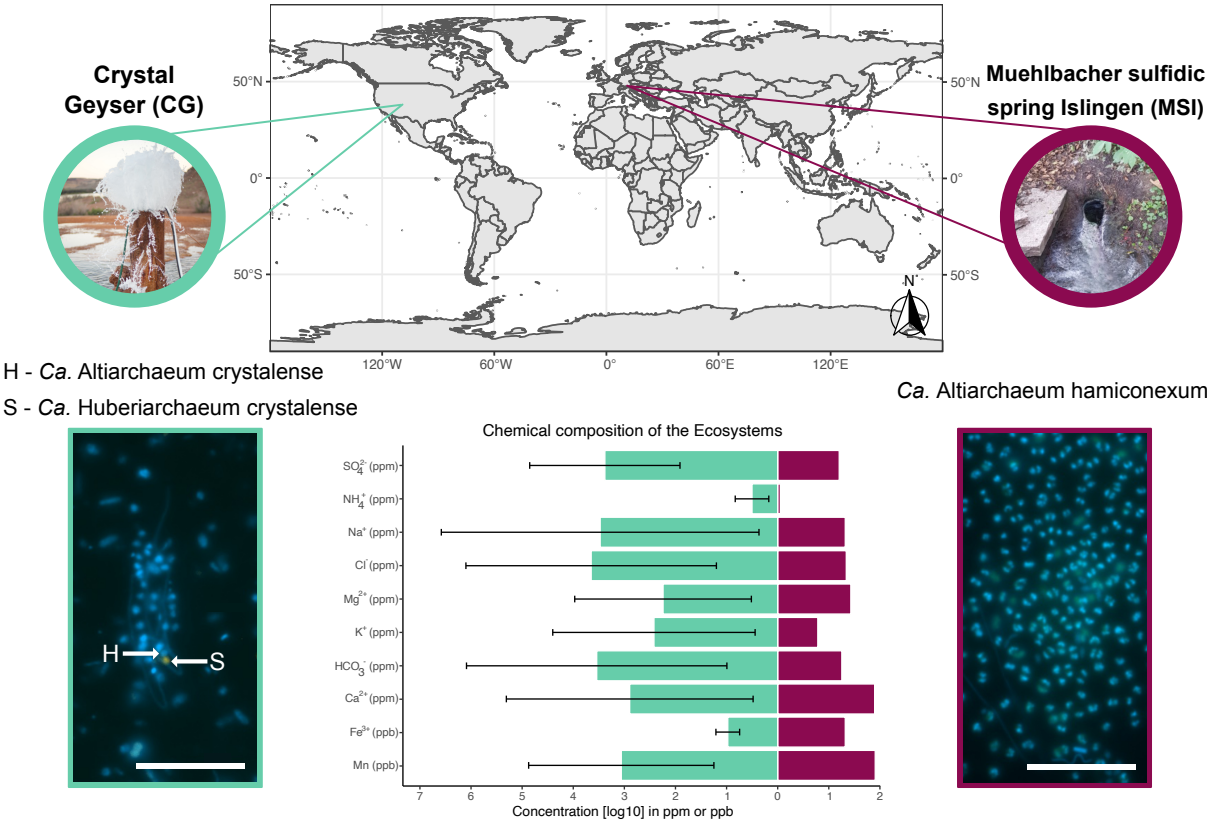


Fig. 3.3.1| World map showing the two sampling sites, Crystal Geyser (CG, blue, image taken by Christopher T. Brown) and Mühlbacher sulfidic spring (MSI, magenta). Differences in the chemical composition of MSI (n=1)[21]. and CG (n=3)[14] are shown for all nutrients and ions measured in both ecosystems Fluorescence *in-situ* hybridization (FISH) images show the host *Ca.* Altiarchaeum (SMARCH714 labeled with Atto448 (Rudolph et al., 2001)) in green (H) and the symbiont *Ca.* Huberiarchaeum (HUB1206 with Cy3 (Schwank et al., 2019)) in orange (S). Please note that the symbiont was not detected in MSI. Scale bar 10 μm.

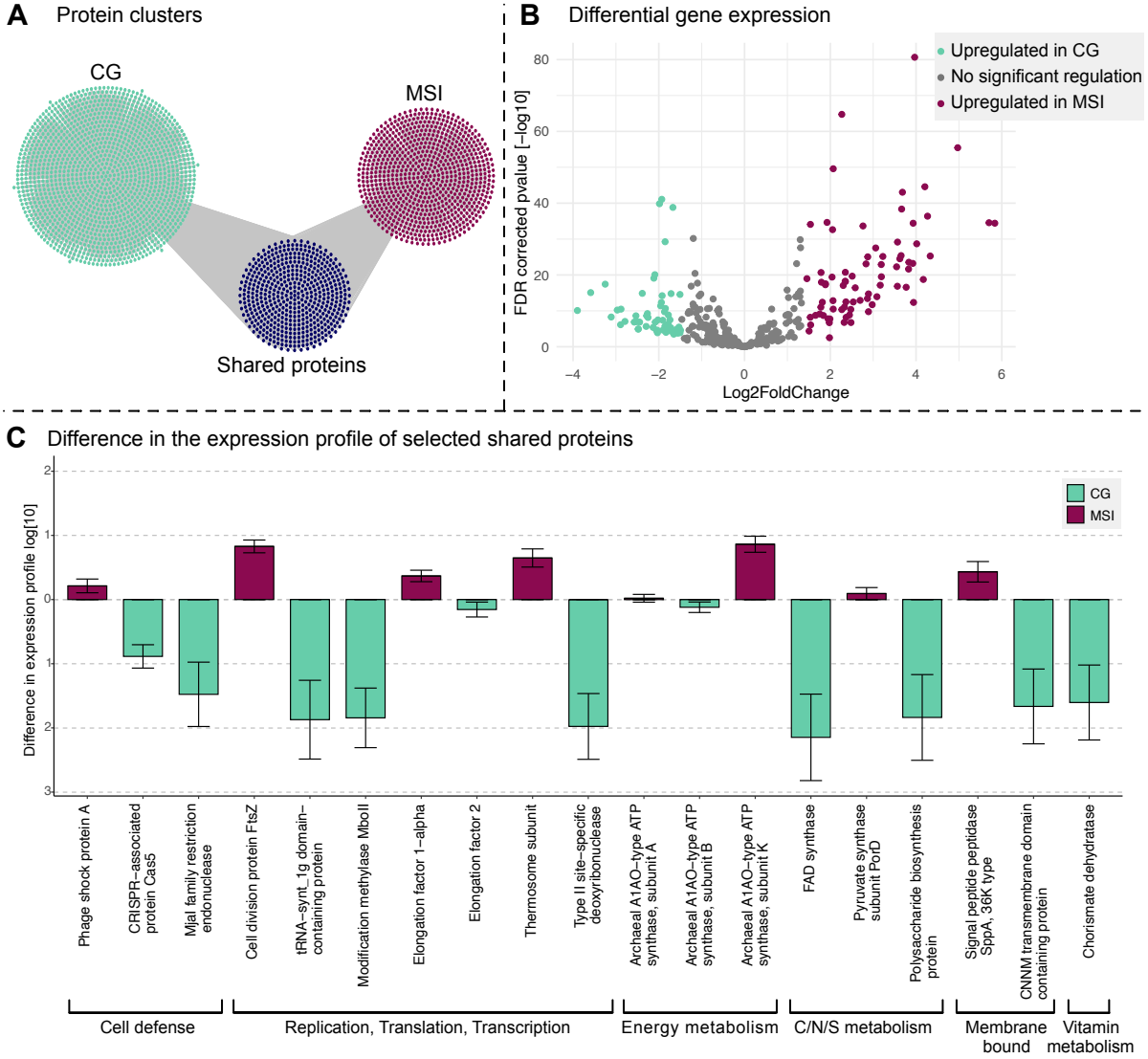


Fig. 3.3.2| Gene clusters, differential gene expression and differences in gene expression profile of two *Ca.* Altiarchaea populations. **A** Gene clusters (80% AA similarity) of CG (n = 1447) and MSI (n = 824) as well as shared protein clusters (n = 537). **B** Differential gene expression of MSI compared to CG. The count data was normalized based on the coverage of ten ribosomal proteins (see methods) and then evaluated with DESeq2 (Liu et al., 2021; Love et al., 2014) in R studio (Posit team, 2022; Team, 2013). **C** Difference graph of the expression profile for 19 shared gene clusters selected based on annotation, difference, and relationship to physiology/ecology (the values represent the difference in the expression of the gene clusters). All values, including mean expression rate, and standard deviations are listed in Table S3.

Publication 2: Now available as: Esser, S. P. *et al.* Differential expression of core metabolic functions in *Candidatus* Altiarchaeum inhabiting distinct subsurface ecosystems. *bioRxiv* 2023.11.20.567779 (2023).

3.3.1. Supplementary information for Differential expression of core metabolic functions in DPANN Archaea of distinct subsurface ecosystems

Sarah P. Esser[1,*], Victoria Turzynski[1], Julia Plewka[1], Carrie Moore[1], Indra Banas[1], André R. Soares[1], Janey Lee[2], Tanja Woyke[2], Alexander J. Probst[1,3,4,*]

[1] Environmental Metagenomics, Research Centre One Health Ruhr of the University Alliance Ruhr, Faculty of Chemistry, University Duisburg-Essen, 45141 Essen

[2] DOE Joint Genome Institute, One Cyclotron Rd, Berkeley, CA, 94720, USA

[3] Centre of Water and Environmental Research (ZWU), University of Duisburg-Essen, Universitätsstraße 5, 45141, Essen, Germany

[4] Center for Medical Biotechnology (ZMB), University of Duisburg-Essen, 45141 Essen, Germany

[*]co-corresponding authors: sarah.esser@uni-due.de, alexander.probst@uni-due.de

Content:

1. Methods

RNA extraction and sequencing

For RNA extraction from *Ca.* Altiarchaeota biofilms in MSI the biofilm flocks were harvested as previously described (Probst et al., 2013) and the flocks were directly frozen at - 80°C until further processing in the lab in November 2021. RNA was extracted with the RNeasy PowerBiofilm RNA extraction kit (Qiagen, Germany) according to the manufacturer's guideline. The extracted DNA was sequenced at the LCSB (Luxemburg) with 150bp paired-end Illumina technology. Prior to sequencing the rRNA was depleted and the RNA was transformed into cDNA.

Sequencing data from Crystal Geyser was retrieved from a previous study (Publication 1, Section 3.2) (minor eruption phase, samples CG05, CG08, and CG16), ensuring that *Ca.* Altiarchaeum crystalense is the most abundant organism in the samples. All accession numbers are listed in Table S1.

Coverage-based normalization of metatranscriptomes

After quality filtering with bbduk (https://github.com/BioInfoTools/BBMap/blob/master/sh/bbduk.sh) and sickle (Joshi and Fass, 2011), the metatranscriptomics reads were normalized by mapping (Langmead and Salzberg, 2012) reads against representative genomes of *Ca.* Altiarchaeum hamiconexum (MSI) and *Ca.* A. crystalense (CG) (see Table S1 for accession numbers). The mean coverage of ten house-keeping genes [30S ribosomal proteins: S4, S5, S7, S8e, S9, S10, S11, S12, S13, S15] were used to calculate the normalization factor of each metatranscriptome sample (per ecosystem n=3) [Table S2]. Prior to choosing these genes for normalization the position of the gene on the scaffold (not within the first/last 200 bp of the scaffold) and a stable coverage distribution across the gene was taken into consideration.

Clustering of genes and calculation of expression profile

Genes of the abovementioned *Ca.* A. hamiconexum and *Ca.* A. crystalense genomes were predicted with prodigal (Hyatt et al., 2010), and consecutively clustered with cdhit (Fu et al.,

2012; Li and Godzik, 2006) at 80% amino acid similarity. The genes from the two ecosystems sharing a cluster were annotated with the FunTaxDB (Bornemann et al., 2023) (version 1.2) database. All genes that either had no or an unclassified annotation and/or which started/ended within the first/last 200 bps of the scaffolds were discarded within the expression profile. Removing genes starting/ending within the first/last 200 bps of a scaffold was chosen to avoid mapping distortions resulting from inaccurate mapping at these scaffold regions.

The abovementioned normalization factors were used to determine the coverage differences introduced by the varying extraction and sequencing methods and the mean coverage with standard deviation of the shared gene clusters were calculated. The visualization was performed with ggplot2 (Wickham, 2016) in R studio (Posit team, 2022; Team, 2013) (version 2023.03.0+386). The data for evaluating the count data within the RNA-seq data was performed with DESeq2 implemented in R (Love et al., 2014). The threshold values of up- and downregulation of differential gene expression were calculated to the base expression of MSI according to Quackenbush 2002 (Quackenbush, 2002).

Fluorescence *in-situ* hybridization

For FISH of *Ca.* A. hamiconexum and *Ca.* A. crystalense cells filtered onto 0.2 μm PFE filters the 16S rRNA probe “SMARCH714” from Moissl et al. 2003 was used (Moissl et al., 2003), whereby the staining of *Ca.* Huberiarchaeum crystalense was performed with the HUB1206 probe from Schwank et al. 2019 (Schwank et al., 2019). Cells were also counterstained with DAPI. The samples from Crystal Geyser and ‘Mühlbacher Schwefelquelle’ were taken in August 2021 and February 2022, respectively, fixed on site with 3% (v/v%) formaldehyde and stored at – 80°C.

2. Supplementary Figures

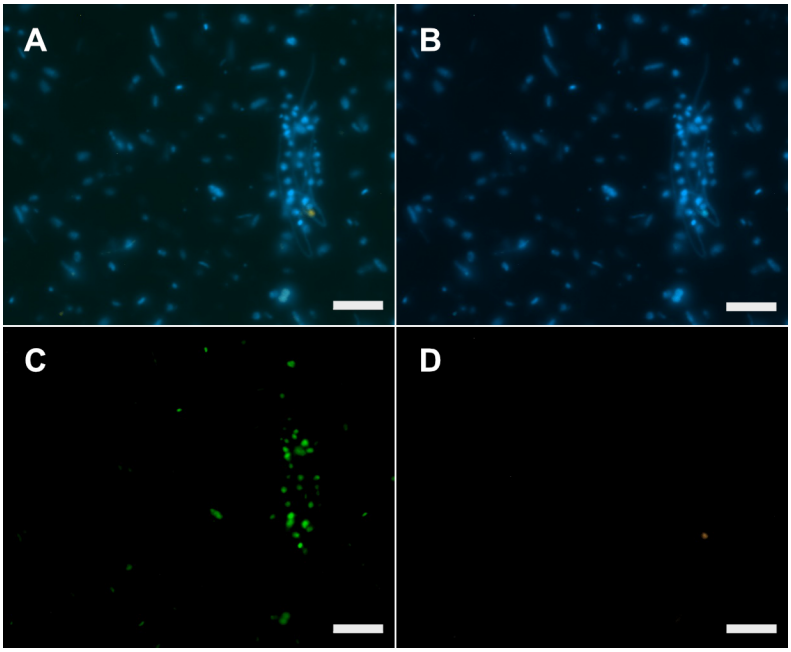


Figure S3.3.1.1| Fluorescence *in situ* hybridization images of *Ca.* Altiarchaeum crystalense and *Ca.* Huberiarchaeum crystalense in Crystal Geyser. **A** Merged FISH images as shown in Fig. 3.3.1. **B** DAPI staining of the sample tissue. **C** 16S rRNA fluorescence signal of *Ca.* Altiarchaeum with SMARCH714 probe with Atto488 labeling. **D** *Ca.* Huberiarchaeum crystalense with HUB1206 probe and Cy3 labeling. Scale bar 10 µm.

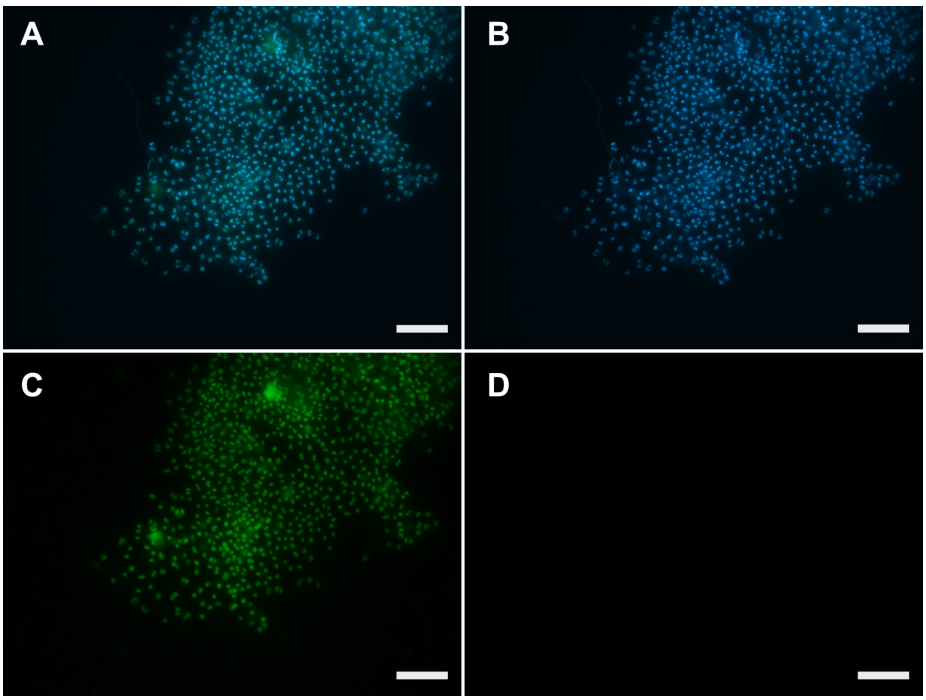


Figure S3.3.1.2| Fluorescence *in situ* hybridization images of *Ca.* Altiarchaeum crystalense and *Ca.* Huberiarchaeum crystalense in Crystal Geyser. **A** Merged FISH images as shown in Fig. 3.3.1. **B** DAPI staining of the sample tissue. **C** 16S rRNA fluorescence signal of *Ca.* Altiarchaeum with SMARCH714 probe with Atto488 labeling. **D** *Ca.* Huberiarchaeum crystalense with HUB1206 probe and Cy3 labeling. Scale bar 10 µm.

Publication 2: Now available as: Esser, S. P. *et al.* Differential expression of core metabolic functions in *Candidatus* Altiarchaeum inhabiting distinct subsurface ecosystems. *bioRxiv* 2023.11.20.567779 (2023).

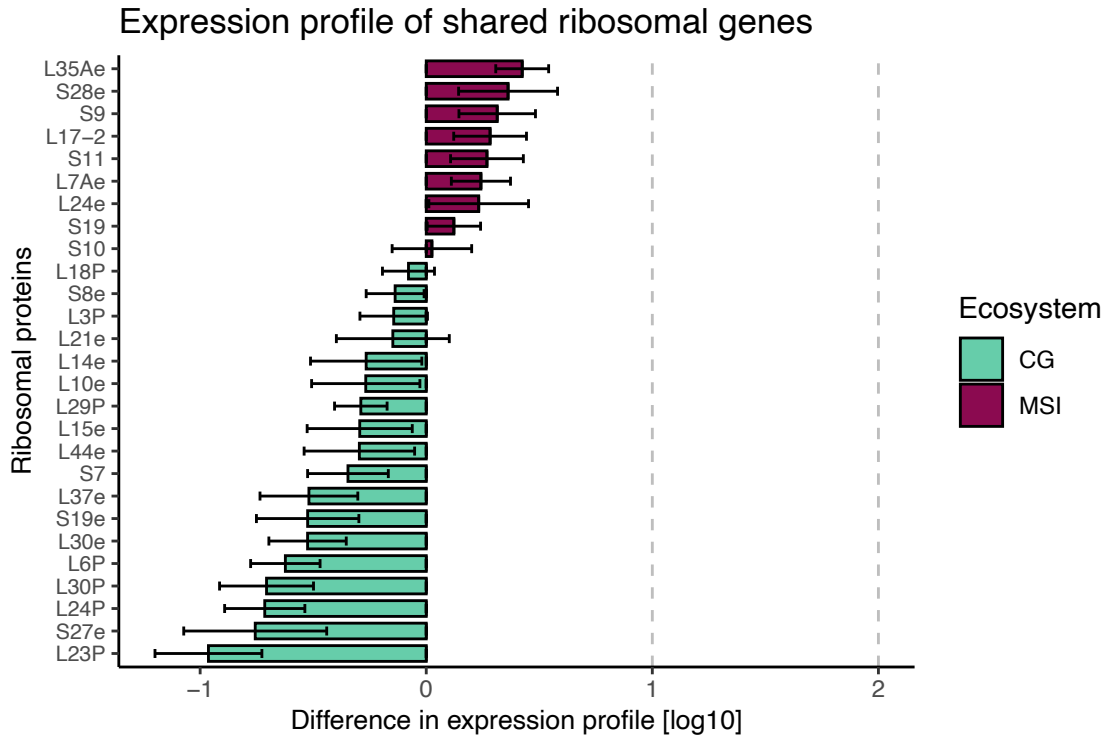


Figure S3.3.1.3| Difference in expression profile of ribosomal proteins derived from metatranscriptomics data of CG (blue) and MSI (magenta), respectively. Visualization was performed with ggplot2 (Wickham, 2016) in R studio (Posit team, 2022; Team, 2013).

Publication 2: Now available as: Esser, S. P. *et al.* Differential expression of core metabolic functions in *Candidatus* Altiarchaeum inhabiting distinct subsurface ecosystems. *bioRxiv* 2023.11.20.567779 (2023).

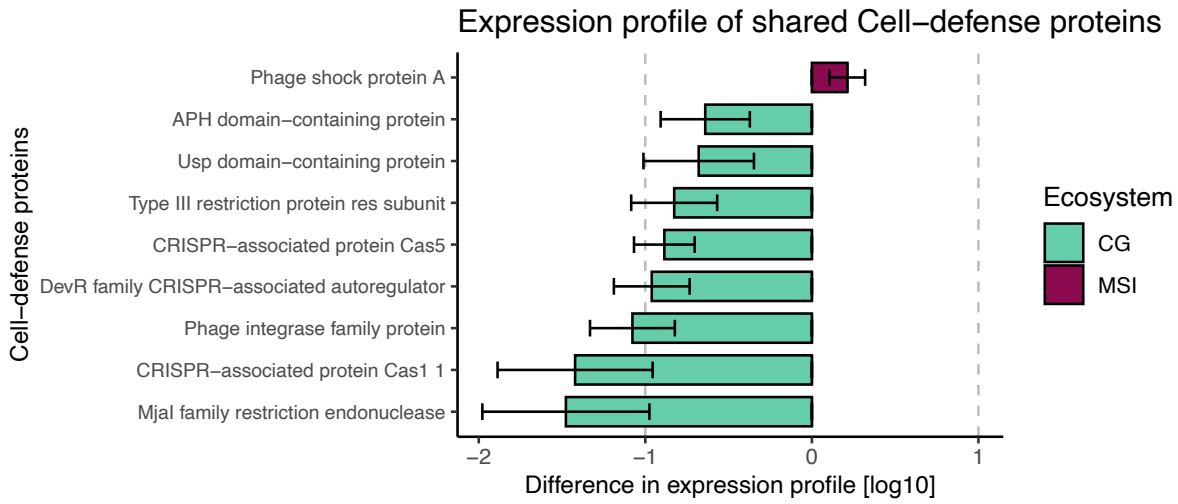


Figure S3.3.1.4| Difference in expression profile of proteins related to cell-defense and derived from metatranscriptomics data of CG (blue) and MSI (magenta), respectively. Visualization was performed with ggplot (Wickham, 2016) in R studio (Posit team, 2022; Team, 2013).

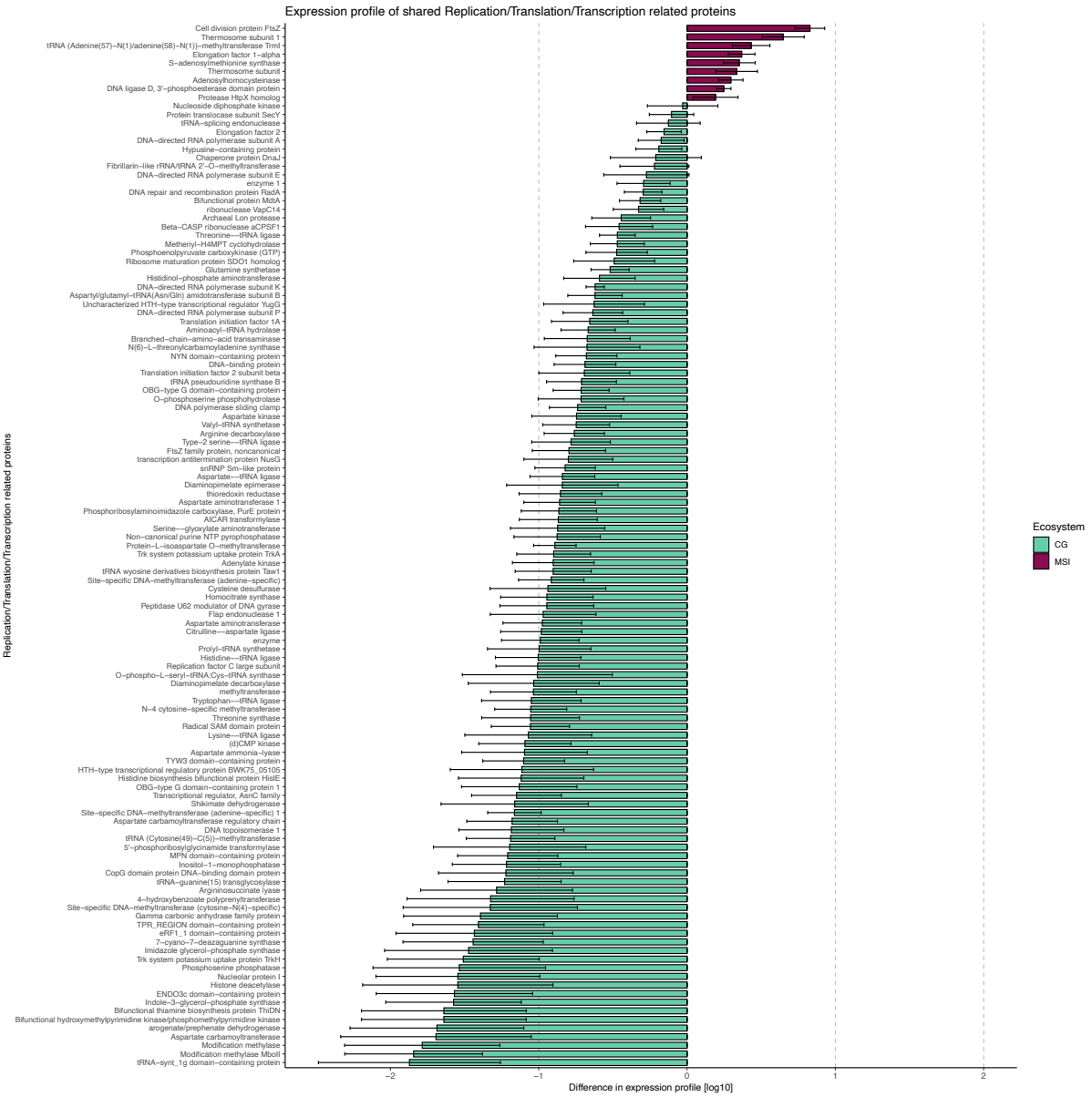


Figure S3.3.1.5| Difference in expression profile of replication, translation and transcription related proteins derived from metatranscriptomics data of CG (blue) and MSI (magenta). Visualization was performed with ggplot (Wickham, 2016) in R studio (Posit team, 2022; Team, 2013).

Publication 2: Now available as: Esser, S. P. *et al.* Differential expression of core metabolic functions in *Candidatus* Altiarchaeum inhabiting distinct subsurface ecosystems. *bioRxiv* 2023.11.20.567779 (2023).

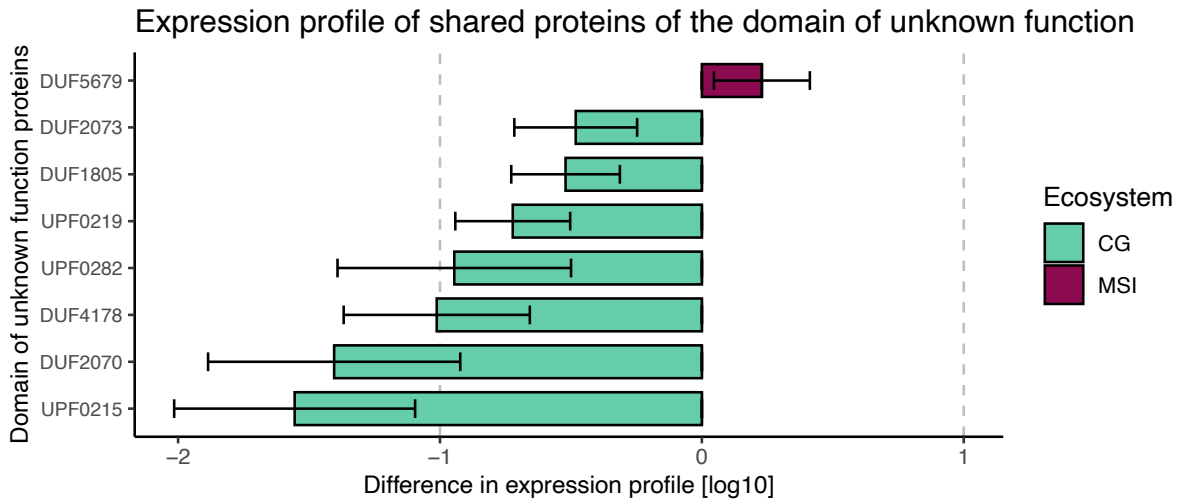


Figure S3.3.1.6| Difference in expression profile of proteins classified as "domain of unknown function" and derived from metatranscriptomics data of CG (blue) and MSI (magenta). Visualization was performed with ggplot (Wickham, 2016) in R studio (Posit team, 2022; Team, 2013).

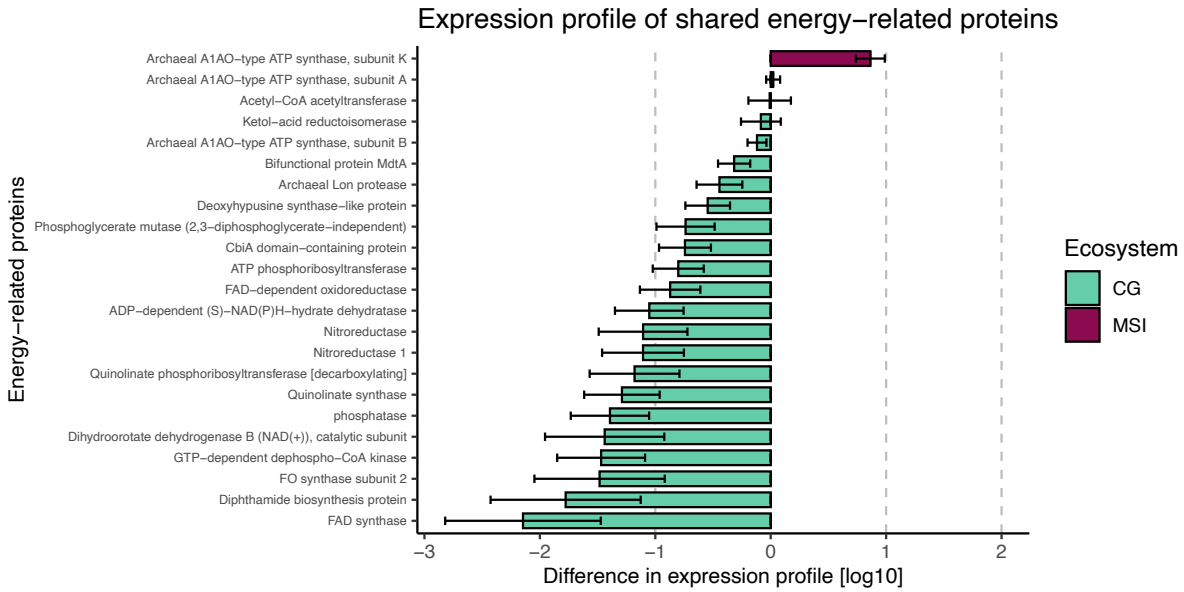


Figure S3.3.1.7| Difference in expression profile of energy related proteins derived from metatranscriptomics data of CG (blue) and MSI (magenta). Visualization was performed with ggplot (Wickham, 2016) in R studio (Posit team, 2022; Team, 2013).

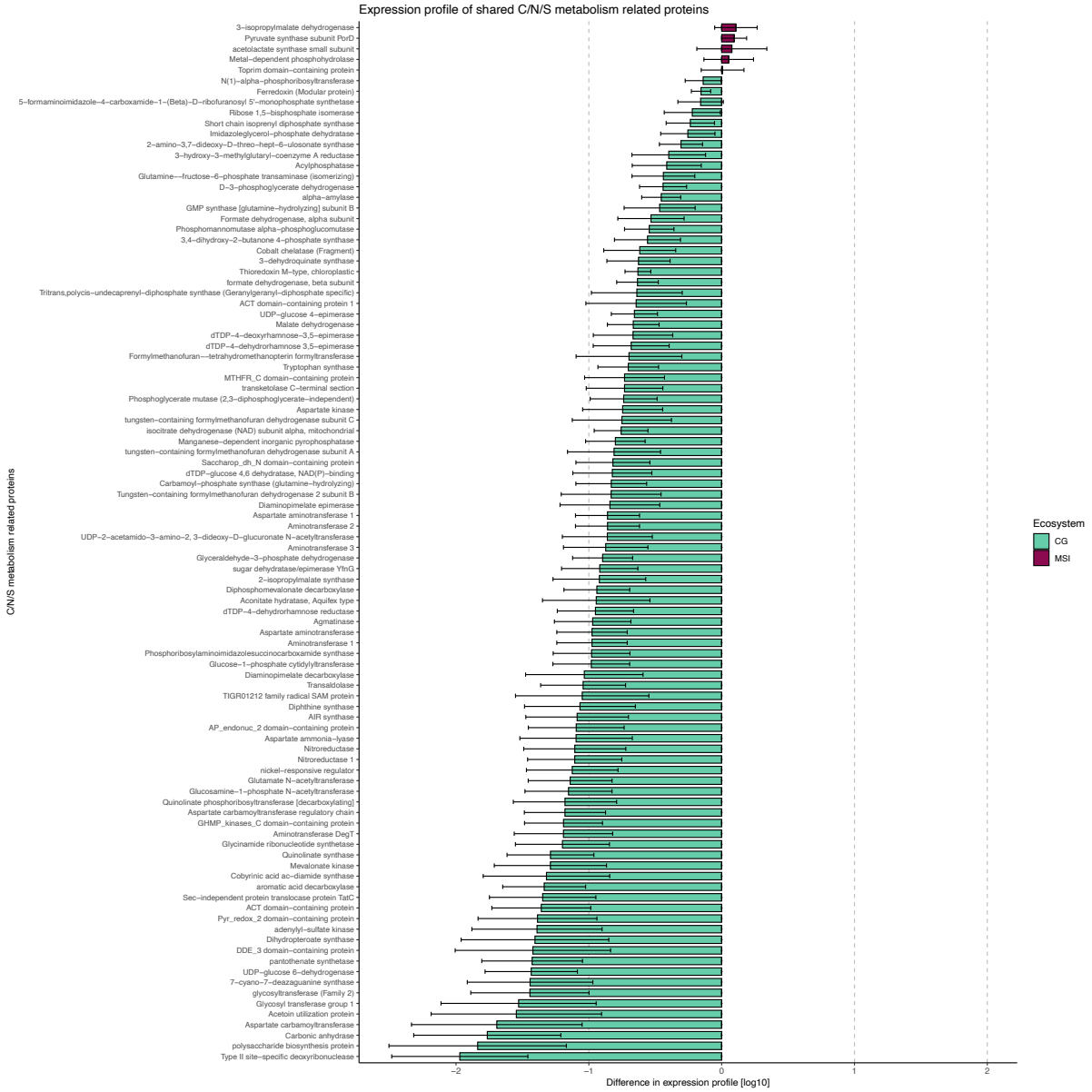


Figure S3.3.1.8| Difference in expression profile of carbon, nitrogen and sulfur metabolism related proteins derived from metatranscriptomics data of CG (blue) and MSI (magenta). Visualization was performed with ggplot (Wickham, 2016) in R studio (Posit team, 2022; Team, 2013).

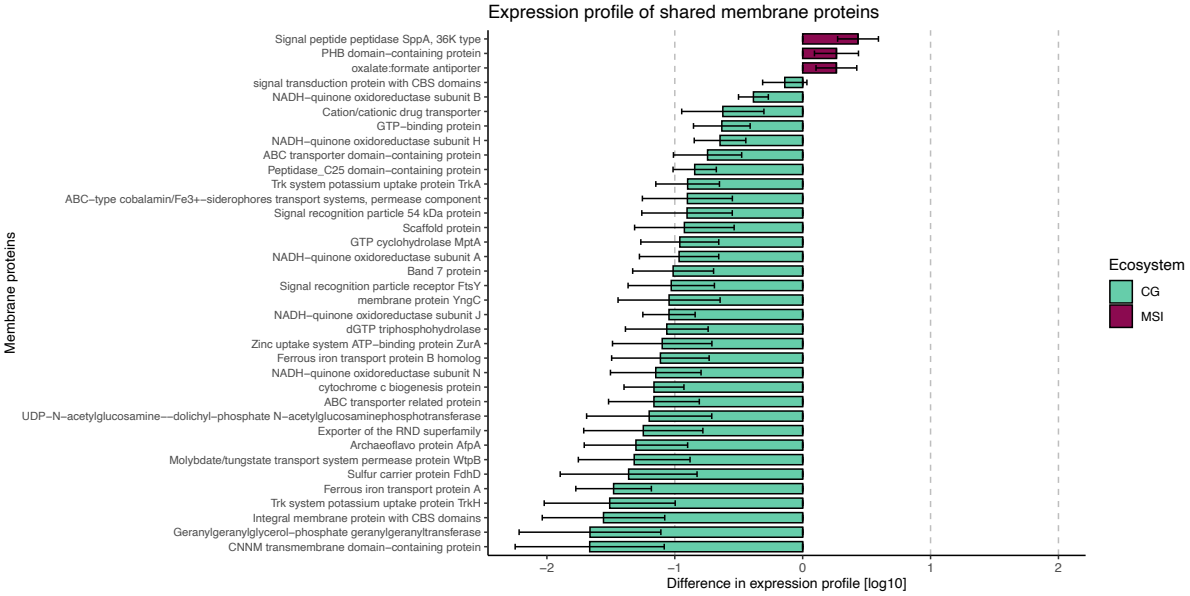


Figure S3.3.1.9| Difference in expression profile of membrane proteins derived from metatranscriptomics data of CG (blue) and MSI (magenta). Visualization was performed with ggplot (Wickham, 2016) in R studio (Posit team, 2022; Team, 2013).

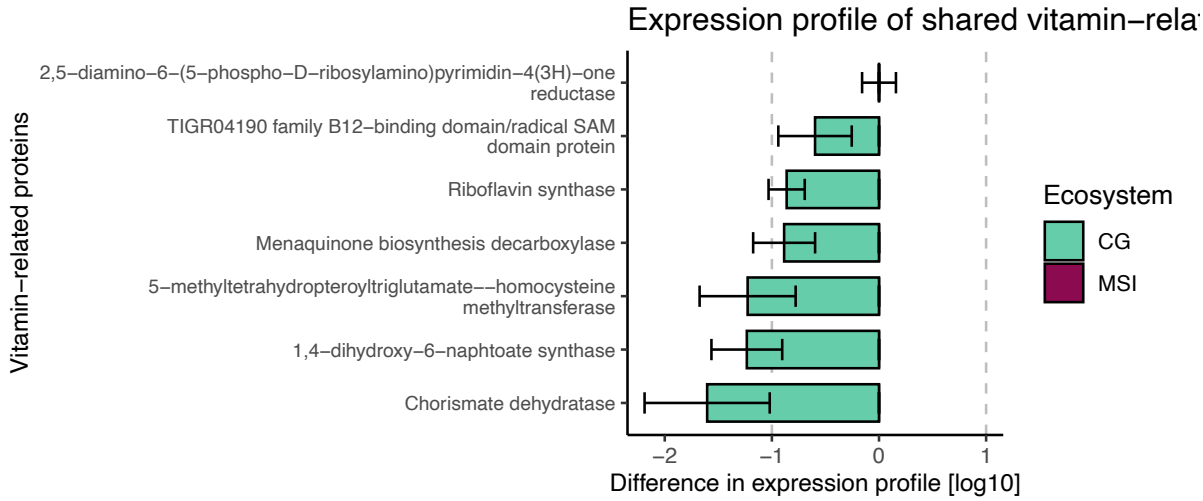


Figure S3.3.1.10| Difference in expression profile of vitamin related proteins derived from metatranscriptomics data of CG (blue) and MSI (magenta). Visualization was performed with ggplot (Wickham, 2016) in R studio (Posit team, 2022; Team, 2013).

Publication 2: Now available as: Esser, S. P. *et al.* Differential expression of core metabolic functions in *Candidatus* Altiarchaeum inhabiting distinct subsurface ecosystems. *bioRxiv* 2023.11.20.567779 (2023).

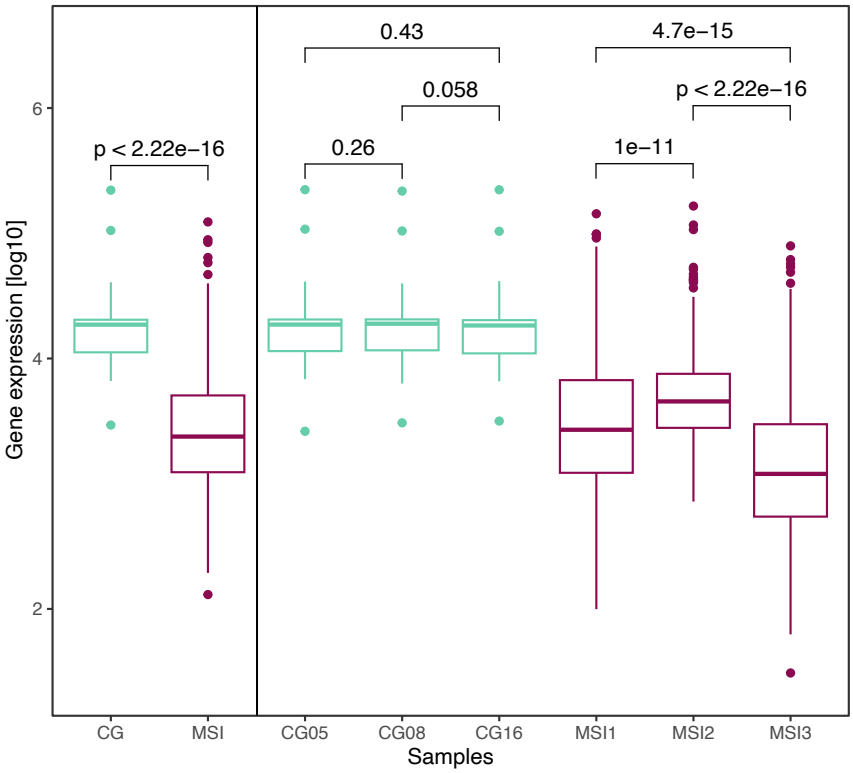


Figure S3.3.1.11| Inter and intra ecosystem comparison of the mean normalized gene expression of CG's and MSI's shared gene clusters (Kruskal-Wallis Anova, p-values within figure). CG showed a significantly higher expression of the core metabolic pathways then MSI. Data visualization was performed with ggplot (Wickham, 2016) in R studio (Posit team, 2022; Team, 2013).

Publication 2: Now available as: Esser, S. P. *et al.* Differential expression of core metabolic functions in *Candidatus* Altiarchaeum inhabiting distinct subsurface ecosystems. *bioRxiv* 2023.11.20.567779 (2023).

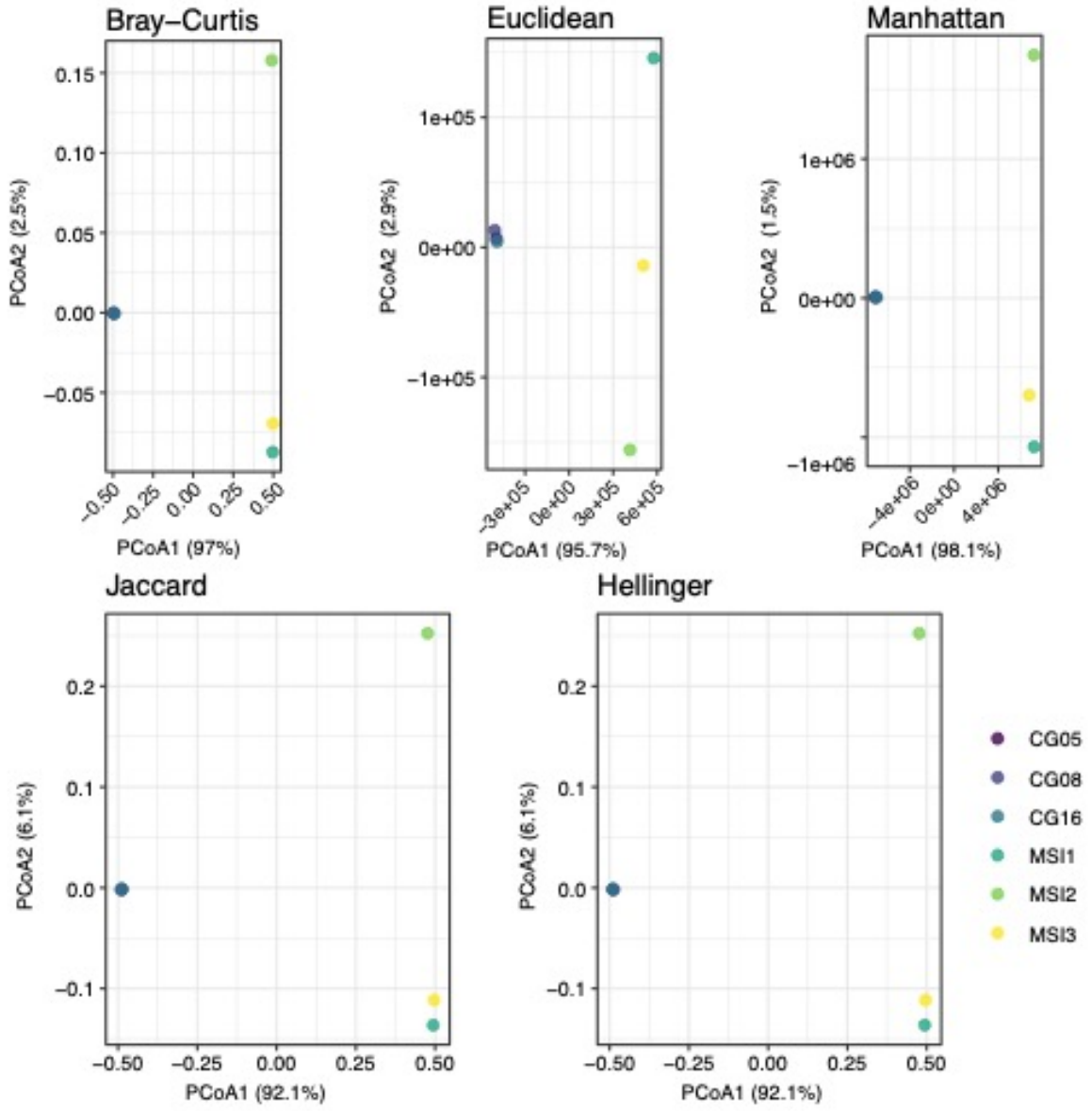


Figure S3.3.1.12| Principal Coordinate Analyses (PCoA) of normalized gene expression within the metatranscriptomic dataset. Compared are different calculations of the dissimilarity matrix. All variants of the PCoAs show that the datasets of Crystal Geyser and Muehlbacher sulfidic spring cluster separated from each other. Data visualization was performed with ggplot (Wickham, 2016) in R studio (Posit team, 2022; Team, 2013).

3. Supplementary Tables

Table S3.3.1.1| Accession Numbers and metadata of metatranscriptomics reads and *Ca.* Altiarchaea genomes used in this study

	Type of data	Ecosystem	Abbr./Genome name	Accession number	Year
1	Reads	CG	CG05	SAMN14515498 [NCBI]	2015
2	Reads	CG	CG08	SAMN14515403	2015
3	Reads	CG	CG16	SAMN14515402	2015
4	Reads	MSI	MSI1	/data4/IMS_Transcriptome/raw.d	2021
5	Reads	MSI	MSI2	/data4/IMS_Transcriptome/raw.d	2021
6	Reads	MSI	MSI3	/data4/IMS_Transcriptome/raw.d	2021
7	Genome	CG	Candidatus Altiarchaeum sp. CG_4_8_14_3_um_filter_33_2054	2786546689 [JGI IMG]	2014
8	Genome	CG	Candidatus Altiarchaeum sp. CG_4_9_14_0_8_um_filter_32_206	2786546690 [JGI IMG]	2014
9	Genome	CG	Candidatus Altiarchaeum sp. CG_4_10_14_0_8_um_filter_32_851	2786546691 [[JGI IMG]	2014
10	Genome	CG	Candidatus Altiarchaeum sp. CG2_30_32_3053	2786546688 [JGI IMG]	2009
11	Genome	CG	Candidatus Altiarchaeum sp. CG03_land_8_20_14_0_80_32_618	2786546693 [JGI IMG]	2014
12	Genome	CG	Candidatus Altiarchaeum sp. CG12_big_fil_rev_8_21_14_0_65_33_22	2786546692 [JGI IMG]	2014
13	Genome	CG	Candidatus Altiarchaeum sp. CG1_02_FULL	10.6084/m9.figshare.22339555	2009
14	Genome	CG	Candidatus Altiarchaeum sp. CG1_02_FULL_1	10.6084/m9.figshare.22339555	2009
15	Genome	CG	Candidatus Altiarchaeum sp. CG2_30_FULL	10.6084/m9.figshare.22339555	2009
16	Genome	CG	Candidatus Altiarchaeum sp. CG2_30_SUB10	10.6084/m9.figshare.22339555	2009
17	Genome	CG	Euryarchaeota archaeon JGI CrystG Aug3-3-F14	2693430074 [JGI IMG]	2014
18	Genome	CG	Euryarchaeota archaeon JGI CrystG Aug3-1-I17	2693430078 [JGI IMG]	2014
19	Genome	CG	Euryarchaeota archaeon JGI CrystG Aug3-3-C16	2693430067 [JGI IMG]	2014
20	Genome	CG	Euryarchaeota archaeon JGI CrystG Aug3-1-C22	2693430068 [JGI IMG]	2014
21	Genome	CG	Euryarchaeota archaeon JGI CrystG Aug3-3-I21	2693430072 [JGI IMG]	2014
22	Genome	CG	Euryarchaeota archaeon JGI CrystG Aug3-3-I7	2693430046 [JGI IMG]	2014
23	Genome	CG	Euryarchaeota archaeon JGI CrystG Aug3-1-F4	2693430045 [JGI IMG]	2018
24	Genome	MSI	Candidatus Altiarchaeum hamiconexum	SAMN18220766 [NCBI]	2014

4. List of supplementary tables

The tables of this publication can be found at the supporting CD within the folder Publication_2_Section_3.3.

Table S2 | Normalization factors calculated based on ten ribosomal proteins that are shared at 80% amino acid similarity in CG and MSI.

Table S3 | Raw and normalized coverages of shared gene clusters in CG and MSI. Normalization was done based in the in Table S2 calculated normalization factors.

Table S4 | Gene annotations of shared gene clusters in CG and MSI based the FunTaxDB 1.2. [accessed September 2022]

4. Discussion

As archaea are globally widespread and contribute to the cycling of carbon, nutrients, and bio-processable molecules, they are interesting to study. *Ca.* Altiarchaeum, with its global distribution in various ecosystems (Bornemann et al., 2022), is a good model organism to identify the role of archaea in the deep subsurface.

4.1. *Ca.* Altiarchaeum and its complex CRISPR systems

4.1.1. Self-targeting against directed targeting

In general, CRISPR systems are adapted to hinder the viral lysis of microbial communities and are therefore protecting these communities from extinction (reviewed in Terns and Terns, 2011). The manuscript (Publication 1, Section 3.2) collects insight into the construction and targeting aims of the CRISPR system type I-B and an unassigned CRISPR system of *Ca.* Altiarchaeum in CG and MSI. Next to viral and MGE targeting, which is the *status quo* of CRISPR interactions and described as the immune system of microorganisms, *Ca.* Altiarchaea acquire multiple spacers per array targeting the episymbiont or the genome of *Ca.* Altiarchaea itself. Later was addressed over time as self-targeting (*i.e.*, Levy et al., 2015; Wimmer and Beisel, 2020).

Studies on self-targeting by CRISPR systems revealed that 18% of all known CRISPR systems have self-targeting spacer encoded in the CRISPR array, with approx. 0.4% of all spacers fall into this category (Stern et al., 2010). Compared to the self-targeting spacer abundance of 0.4% in CRISPR arrays from public databases (Stern et al., 2010), the abundance of these self-targeting spacers in *Ca.* Altiarchaeum crystalense is increased to 0.9% (5,382 of 595,081 unclustered spacers). In addition to the high complexity of the CRISPR arrays of *Ca.* Altiarchaeum crystalense (Publication 1, Section 3.2-3.3) this increase in self-targeting spacer acquisition supports the assumption that the CRISPR systems of *Ca.* Altiarchaeum crystalense play an essential role in deep subsurface ecosystems. Why the CRISPR system of *Ca.* Altiarchaeum crystalense is acquiring an increased amount of self-targeting spacers is, until today, highly speculative, as the role of the self-targeting can either be regulation of gene expression or potential auto-immunity resulting in cell death (reviewed in Heussler and O'Toole, 2016). As members of the genus Altiarchaeota are not yet cultured, the nature of

the self-targeting stays speculative. As the protospacers within the *Ca.* Altiarchaeum crystalense genomes do not encode a defined protospacer adjacent motif (PAM; Publication 1), in contrast to the episymbiont and the viral genomes, the self-targeting of the CRISPR system type I-B might be either omitted or only a part of the population is actively affected by self-targeting. On the one hand, the absence of PAMs in subpopulations of *Ca.* Altiarchaea might prevent cell suicide but can on the other hand, in the subpopulations carrying the PAM sequence result in strain reduction or variation. The latter would reduce intra-species competition about nutrients and bio-processable products. The strain variation, dependent on CRISPR efficiency was previously shown by reduced inter-species mating in *Haloferax* species (Turgeman-Grott et al., 2019).

Focusing on the complexity of gene coding regions self-targeted by the CRISPR system IB of *Ca.* Altiarchaeum, it can be assumed that only subpopulations are affected. A possible reason for self-targeting within tRNA ligases, which in the case of *Ca.* Altiarchaea occurs within the gene coding region of the, *e.g.*, Phenylalanine—tRNA ligase and Valine–tRNA ligase, is the inhibition of growth resulting in a reduction of strain diversity within an environment. Self-targeting was shown previously for the histidyl—tRNA synthase in *Pelobacter fcarbinolicus* (Aklujkar and Lovley, 2010). Therefore, the targeting within tRNA ligases seems to be a common targeting site of self-targeting CRISPR system. This suggest that the CRISPR system functionality of *Ca.* Altiarchaea might be comparable to already known systems in bacteria. Aklujkar and Lovley hypothesize that the genome of *Pelobacter carbinolicus* is under constant evolutionary pressure by self-targeting spacers (Aklujkar and Lovley, 2010), which might be a transferable hypothesis for the genomes of *Ca.* Altiarchaea, as within the position of the PAM of CRISPR system IB is carrying a single nucleotide polymorphism, resulting in mutation of the PAM system (Publication 1; Section 3.2 and Section 3.3 Fig. S3.2.1.9).

Taking the unassigned CRISPR system into account, which might encode a CRISPR system type III based on the CRISPR cas genes found on fragmented scaffolds of *Ca.* Altiarchaeum (Rahlff et al., 2021), which does not necessitate a PAM for CRISPR system functionality (reviewed in Gleditzsch et al., 2019), the organism can not only target DNA but also might be able to target RNA, and therefore, repress transcriptional products (reviewed in Kolesnik et al., 2021). The co-occurrence of CRISPR system type I and type III in *Ca.*

Altiarchaea genomes can be an evolutionary loophole to counteract viral escape of CRISPR systems of type I, as previously shown for type I-F and III-B systems in *Marinomonas mediterranea*, a bacterium isolated from seawater (Silas et al., 2017). This co-occurrence would, in conclusion, mean that *Ca.* Altiarchaea can defend the cell from viral lysis, influence the amount of nutrients released to the environment and, therefore, also affect biogeochemical cycles within the subsurface.

4.1.2. Transposons as evolutionary driver of the symbiotic relationship?

A possible explanation for why *Ca.* Altiarchaea acquired spacers from the episymbiont can be a transposon that inserts genes or gene fragments of host and symbiont into the respective other genome and, therefore, shares genetic elements and drives coevolution, as previously described for symbiotic plant-microbe relationships (reviewed by Seidl and Thomma, 2017). Supportive of this hypothesis is that all publicly available genomes of *Ca.* Altiarchaeum crystalense (MAGs and SAGs) carry approximately seven transposon regions, detected and masked for spacer-protospacer matches in Publication 1, Section 3.2. In general, transposons can transfer DNA fragments, which can also be genomic locations to encode *cas* genes (Krupovic et al., 2017), intragenomically or as a mobile genetic element into other genomes and are therefore involved in horizontal gene transfer (Muñoz-López and García-Pérez, 2010). In addition, some of the transposons are conserved across samples of different time points and are targeted by self-targeting spacers of the CRISPR system type IB. However, the findings described in Publication 1 suggest that at least the tRNA ligases, that are self-targeted by *Ca.* Altiarchaea and complimented by *Ca.* Huberiarchaeum, do not originate within the same phylogenetic lineage, or are shared between host and symbiont. To conclude, if the CRISPR-mediated symbiosis of *Ca.* Altiarchaeum and *Ca.* Huberiarchaeum drives the evolution and increases the evolutionary development of the strain complexity, further analyses, ideally with a stable co-culture of the DPANN host-episymbiont system, would need to be done.

If self-targeting or directed targeting against genomes of other microorganisms is also prevalent not only in/between archaeal genomes but also in/between bacterial genomes or even inter-phyla targeting needs to be further quantified and analyzed in the future.

4.2. Cytoplasmic contact might facilitate symbiotic relationship

As *Ca.* Huberiarchaeum is a proposed symbiont of *Ca.* Altiarchaeum, the metabolite exchange or the interference of the CRISPR systems need to be facilitated by either direct cytoplasmic contact, transporting channels or traveling vesicles (Schwank et al., 2019). Thereby direct cytoplasmic contact was already previously shown for other host-symbiont systems (Baker et al., 2010; Comolli and Banfield, 2014; Hamm et al., 2019; Heimerl et al., 2017). Like the fusion of *Nanohaloarchaea* with their respective haloarchaea host (La Cono et al., 2020), it can be assumed that *Ca.* Altiarchaeum and *Ca.* Huberiarchaeum fuse their membrane and have cytoplasmic contact. This assumption is also supported by fluorescence *in-situ* hybridization images, where the fluorescence signals of host and episymbiont cells are in close proximity to each other (Schwank et al., 2019). Specifically for the activity of the CRISPR system of *Ca.* Altiarchaeum crystalense, direct cytoplasmic contact would be essential because, for CRISPR interference the CRISPR complex of the CRISPR system IB and the *cas3* helicase need access to the genome of the episymbiont.

Direct cytoplasmic contact was previously shown between an ARMAN archaeon and a *Thermoplasmatales* with cryo-electron tomography (Baker et al., 2010). In 2010 Baker and collaborators showed that *Thermoplasmatales* can penetrate the membrane of the ARMAN cell in the presence of vacuoles, contemporaneous with viral infections of the ARMAN cell (Baker et al., 2010). Transferring this finding to the symbiotic nature of *Ca.* Altiarchaea and *Ca.* Huberiarchaea, the acquisition of the genomic DNA of the episymbiont within the CRISPR system of *Ca.* Altiarchaea might be triggered by the cytoplasmic contact, while *Ca.* Huberiarchaea is infected by viruses. Problematic to proof this transfer of findings is that *Ca.* Huberiarchaeum does not encode a CRISPR system, which could be used to provide insight into the infection histories of invading MGEs (Barrangou et al., 2007a; Garneau et al., 2010; Horvath and Barrangou, 2010). If a virus might infect either both, the host and the symbiont or if a virus infects only the symbiont and through cytoplasmic contact *Ca.* Altiarchaea's CRISPR system adapt to the invading DNA would be interesting to study. To proof this hypothesis a stable co-culture alongside viruses infecting *Ca.* Huberiarchaea but not *Ca.* Altiarchaea would be necessary.

Within eukaryote-bacteria symbiotic partners, multiple symbionts can coexist with the same host, which results in a complex coadaptation of the genomes and transfer of

genetic information between the cells (Moran, 2007). Transfers of genetic information in Host-symbiont-viral associations are, to the best of my knowledge, only known for mosquitos in symbiosis with *Wolbachia*, a Proteobacteria harboring the West Nil virus (Alomar et al., 2023). This finding shows that in nature, it is possible to transfect microbial cells with genetic material not only from symbiotic partners or by horizontal gene transfer, but also genetic information of harbored viruses can get conducted. Transferring the transfection theory to the symbiosis of Ca. Altiarchaea and Ca. Huberiarchaea, it might be possible that the CRISPR system of *Ca.* Altiarchaea gets activated by the cytoplasmic contact with an actively infected symbiotic cell and, therefore, not only acquires spacers from the infecting virus but also from the symbiont. This concept would lead to the assumption that the symbiotic concept, to which *Ca.* Altiarchaea and *Ca.* Huberiarchaea underly is more likely mutualism than parasitism, as DNA pick-up of the symbiont, is rather random than directed.

4.3. Transfer of the proposed concepts to other archaeal symbiotic relationships

Comparing the presented results collected in Publications 1 and 2 (Sections 3.2 and 3.3) about the relationship of *Ca.* Altiarchaea and *Ca.* Huberiarchaea to other archaeal host-symbiont systems provides insights into the role of this symbiosis in nature.

The most prominent representative of a symbiosis between archaea and DPANN archaea is *Ignicoccus hospitalis* with its symbiont *Nanoarchaeum equitans* (Huber et al., 2002). Compared to the *Ca.* Altiarchaeum and *Ca.* Huberiarchaeum, *Ignicoccus hospitalis* did not acquire spacers from the symbiont *Nanoarchaeum equitans*, and the CRISPR system does not show differential expression triggered by the presence of the symbiont (Giannone et al., 2011), indicating a mutualistic relationship between the two archaea. This behavior of the CRISPR system contrasts with the found spacer acquisition in *Ca.* Altiarchaea, that describes a possible shift from a parasitic to a mutualistic lifestyle (Publication 1; Section 3.2). This assumption is based on the acquisition of spacers targeting the symbiont and of self-targeting spacers leading *Ca.* Altiarchaea cells to necessitate metabolite exchange for complete metabolic pathways (Publication 1, Section 3.2 Fig. 3.2.2). Interestingly, in agreement with the genomes of *Ca.* Altiarchaea (Bornemann et al., 2022; Probst et al., 2014; Probst and Moissl-Eichinger, 2015), the genome of *Ignicoccus hospitalis* seem to be reduced and focused on anaerobic metabolism (Podar et al., 2008). However, the genome of *Ignicoccus hospitalis*

does not encode any transposable elements (Forterre et al., 2009) in contrast to the population genomes of *Ca.* Altiarchaea (Publication 1, Section 3.2). This indicated that *Ca.* Altiarchaeum is affected by horizontal gene transfer while other archaeal hosts are likely not distressed by this kind of evolutionary pressure.

Moreover, *Ignicoccus* species forms cell appendages, named fibers, likely used for the adhesion of the cells on surfaces (Huber et al., 2012, 2002; Paper et al., 2007), which contrasts with *Ca.* Altiarchaeum hamiconexum *hami* connecting cells to biofilms (Moissl et al., 2005, 2003; Rudolph et al., 2004). However, within Crystal Geyser, the *hami* are expressed and visible, but the cells are not found in stable biofilms (Probst et al., 2018) what might be reasoned by the constant disturbance of the system by the eruption cycles of the geyser. An interesting study would be if the cell appendages facilitate the attachment of the episymbiont in archaea or if the biofilm formation is a major factor within the host-symbiont relationships of DPANN archaea and archaea. Interestingly, for successful co-cultivation of *Ca.* Micrarchaea, a member of the DPANN archaea, with its host *Ca.* Scheffleriplasma hospitalis, a member of the Thermoplasmatales lineage, the formation of a biofilm facilitates the stability of these co-cultures (Krause et al., 2022). Transferred this implements that also *Ca.* Altiarchaea and *Ca.* Huberiarchaea host-symbiont relationships might profit from the ability of biofilm formation of *Ca.* Altiarchaea.

Next to the well-established host-symbiont association *Ignicoccus hospitalis* and *Nanoarchaeum equitans*, *Nanoarchaeum* Nst1 was metagenomically linked to *Sulfolobales* Acd1 as host (Munson-McGee et al., 2015; Podar et al., 2013). As part of the cellular carbon metabolism, *Sulfolobales* encode a complete Hydroxybutyrate cycle, a near-complete citric acid cycle (TCA cycle), and a near-complete gluconeogenesis pathway. By contrast, the associated symbiont only encodes pathways from acetate to glucose as the main carbon metabolism (Podar et al., 2013). Noteworthy, contrasting to *Ca.* Huberiarchaeum, the symbiont *Nanoarchaeum* Nst1, and *Nanoarchaeum equitans* do not encode lipid synthesis pathways (Podar et al., 2013; Schwank et al., 2019). Not only the metabolic capacity of the episymbiont *Ca.* Huberiarchaeum is different to other known archaeal symbionts, but also the host's cell (*Ca.* Altiarchaea) structure and metabolism contrast with *Sulfolobales*. While *Ca.* Altiarchaea and *Ignicoccus hospitalis* have double layer membranes (Huber et al., 2000; Perras et al., 2014), *Sulfolobales* Acd1 has the characteristic S-layer membrane of the

Sulfolobales family (Veith et al., 2009), determined based on close phylogenetic relationship to other *Sulfolobales* and according to the encoded genes (Podar et al., 2013).

Within co-cultures of the DPANN symbiont *Ca.* Micrarchaeum and *Metallosphera* sp. AS-7, the dependency of active episymbiont replication seems to be reliant on the active replication cycles of the host (Sakai et al., 2022), which was also shown for *Ignicoccus hospitalis* and *Nanoarchaeum equitans* (Giannone et al., 2015; Jahn et al., 2008). Moreover, *Ca.* Micrarchaea as a symbiont can be found associated to multiple host (Sakai et al., 2022). This is in contrast to *Nanoarchaeum equitans*, which was only found in association with *Ignicoccus hospitalis* determined based on culture experiments (Jahn et al., 2008). Based on metagenomic analyses, *Ca.* Huberiarchaea was only found in association with *Ca.* Altiarchaea, what indicates that the nature of the *Ca.* Altiarchaea – *Ca.* Huberiarchaea host-symbiont association is rather closer related to *Nanoarchaeum equitans* and *Ignicoccus hospitalis* than to *Ca.* Micrarchaea and its divergent hosts.

The comparison of different host-symbiont systems within archaea shows that the relationship of *Ca.* Altiarchaea and *Ca.* Huberiarchaea has parts in common with intradomain symbiosis but, due to the CRISPR adaptation of the host, might originally evolved independently from other symbiosis relationships.

4.3.1. Comparison of the host-symbiont association to inter-domain symbiosis

Intradomain symbiosis is not only known at this point of research, but also interdomain symbiosis is a known concept, namely in *Ca.* Patescibacteria as bacterial symbiont and *Methanothrix* as archaeal host (Kuroda et al., 2022). Based on the genomically determined little metabolic capacity of members of the Patescibacteria group (Castelle et al., 2018), on signal peptides which were also found in pathogenic bacteria (Kuroda et al., 2022; McLean et al., 2020), and on growth experiments in culture, a pure parasitic lifestyle of members of the Patescibacteria was determined within wastewater treatment bioreactors (Kuroda et al., 2022). This contrasts the predicted shift from parasitism to mutualism in the *Ca.* Altiarchaeum-*Ca.* Huberiarchaeum host-symbiont association (Publication 1, Section 3.2). In *Nanoarchaeum equitans* and *Ignicoccus hospitalis* the nature of the symbiosis could not be definitely assigned (Jahn et al., 2008), which is also in contrast to the definite parasitic symbiosis of member of the Patescibacteria and *Methanothrix* (Kuroda et al., 2022). It would

be interesting to analyze if the CRISPR interference, predicted by spacer to protospacer matches, presented in Publication 1 (Section 3.2) can be used to further predict inter-domain symbiotic relationships.

Further analyses of symbiotic association, independent of phylogenetic assignments of the symbiotic partners, would provide insight into the possible interactions between microorganisms and can indicate nutrient and carbon fluxes within different ecosystems.

4.4. Shared core metabolism of *Ca.* Altiarchaea between distant ecosystems

While *Ca.* Altiarchaea is widespread across ecosystems, the analysis of genomes of *Ca.* Altiarchaeum hamiconexum (Regensburg, Germany) and *Ca.* Altiarchaeum crystalense (Utah, USA) shows that *Ca.* Altiarchaea has a conserved core genome with 527 genes shared at 80% amino acid similarity (Publication 2, Section 3.3). In addition to the in Publication 2 described and discussed genes (20 genes in Section 3.3 Fig. 3.2.2) several other genes were identified that belong to genetic collections such as energy-related proteins, transcription, translation, and replication-related, and genes involved in carbon, nitrate, and sulfur metabolism.

The differences in the expression profiles of the shared membrane proteins are interesting, as membrane associated proteins have an influence on the nutrient supply for the cells. Only three out of 36 membrane proteins are higher expressed in MSI than in CG (Publication 2, Section 3.3.1), which might indicate that in *Ca.* Altiarchaeum crystalense the metabolite exchange or the uptake of molecules and ions from the environment is more relevant than in *Ca.* Altiarchaeum hamiconexum. The increase in transmembrane protein expressions might be influenced by the attachment of the episymbiont in CG or is based on the environmental conditions of the ecosystem. Especially, the CNNM transmembrane proteins are 1.5 magnitudes higher expressed in CG than in MSI (Publication 2, Section 3.3 Fig. 3.3.2). These proteins either influence the manganese (2+) content within the cell, acting as direct transmembrane transporter (Funato et al., 2014) or as manganese sensor to control cellular manganese homeostasis (Bai et al., 2021). Considering that manganese (2+) was found in higher concentration in CG than in MSI, this expression seems to be somewhat affected by the environmental factors than the presence of the symbiont in the case of membrane proteins. As the direct cytoplasmic contact was previously proposed for *Ca.*

Altiarchaea and *Ca.* Huberiarchaeum (Schwank et al., 2019), which was already shown for other archaeal host-symbiont associations (Baker et al., 2010) this is not surprising.

Interestingly, although the *hamus* of *Ca.* Altiarchaea is a unique cell appendage for DPANN archaea, the gene encoding the protein is not part of the core gene set that is shared across the genomes of *Ca.* Altiarchaea in CG and MSI. The putative *hamus* subunit (UniRef100_A0A098ED57) from CG from a separate cluster, while the genes annotated identical in MSI clusters with S_layer_N domain-containing protein. As this S_layer_N domain protein is part of the ubiquitous S-layer membrane of organisms (Sleytr et al., 1993), the clustering with these is not surprising as *hami* extent the organism cell surface (reviewed in Probst and Moissl-Eichinger, 2015). This suggests that the *hamus* protein is under constant evolutionary pressure and divergent across ecosystems. However, most striking in this analysis is that the expression of the *hamus* protein in Crystal Geyser is extensively higher than in MSI, although *Ca.* Altiarchaeum hamiconexum form a visible biofilm within the latter ecosystem. This finding might be reasoned by the constant disturbance of the geyser due to the eruption phases in general (Probst et al., 2018) and, therefore, *Ca.* Altiarchaea compensate to counteracting and expressing more *hami*. This would furthermore underline the observations described within Publication 2, Section 3.3.

4.5. Future perspectives on investigating the host-symbiont association of *Ca.* Altiarchaea and *Ca.* Huberiarchaea

To determine if the host-symbiont association of *Ca.* Altiarchaea and *Ca.* Huberiarchaea is of a mutualistic or parasitic nature, different analytical and microbiological ways might be possible. Although a stable co-culture of the two DPANN archaea is one of the most promising ways to determine if the cells have cytoplasmic contact, further analyses with model systems might be possible. Investigations of CRISPR-Cas interference between *Ca.* Altiarchaea and *Ca.* Huberiarchaea might be possible with synthesized *cas* genes of *Ca.* Altiarchaea in further cell-free TXTL assays with modifications to the protocol used in Publication 1. Possible adaptations can be made with temperature, and a pressurized system might facilitate the protein folding of the cas genes and might support the CRISPR-Cas interference.

In addition, sequencing attempts to generate metatranscriptomes that cover the population genome of *Ca.* Huberiarchaeum in the ecosystems, where the symbiont is present,

would increase the knowledge about the dependency of the symbiont on the host's metabolism. This analysis would also leave the opportunity to investigate which proteins are upregulated and if the in Publication 1 (Section 3.2) predicted CRISPR-Cas interference is not only visible in the metagenomic coverage (DNA splicing) but might also have effects on transcriptome expressions.

5. Zusammenfassung

Die aquatische und terrestrische Tiefenbiosphäre zeichnet sich durch diverse Ökosysteme aus, in denen komplexe und divergente mikrobielle Gemeinschaften vorkommen. Durch Analysen, welche nicht auf Kultivierungsmethoden basieren, wurden mehrere Zweige des „Baum des Lebens“ phylogenetisch erweitert und dementsprechend wurde die „mikrobielle dunkele Materie“ tiefergehend beleuchtet. Die nähere Betrachtung von aquatischen Ökosystemen, wie Grundwasser und marine Ökosysteme, brachte im archaeellen Zweig neue Vertreter zum Vorschein, vor allem im DPANN Superphylum. Zu diesem Superphylum gehört auch das Archaeum *Candidatus* Altiarchaeum (*Ca.*), welches weltweit verbreitetet und in verschiedenen, chemisch komplexen Ökosystem, wie Geysiren und terrestrischen Aquiferen, gefunden wurde. Metagenomische Analysen von Populationsgenomen des *Ca.* Altiarchaeums zeigten sowohl die genetischen Elemente von CRISPR Systemen, welche als Immunsystem der Mikroorganismen bekannt ist, als auch deuteten diese Analysen auf einen symbiotischen Partner namens *Ca.* Huberiarchaeum hin. Bis zu diesem Punkt ist die Komplexität des CRISPR Systems und dessen Ziele unbekannt. Die nachfolgenden Studien illustrieren welche Ziele durch das CRISPR-Cas System attackiert werden und vergleicht die Expressionsraten von genomisch konservierten metabolischen Basisfunktionen in verschiedenen Ökosystemen, welche um den Globus verteilt sind. Die Analysen der CRISPR-Cas Interferenz, welche mit kurzen Sequenzen, die aus dem CRISPR System isoliert werden, durchgeführt werden, zeigten, dass sowohl Viren als auch genomische Regionen des Wirtes und des Symbionten Ziele des Immunsystems sind. Abundanzveränderungen in den attackierten Regionen des Genoms des Altiarchaeums selbst und des Symbionten deuten darauf hin, dass die genomische DNA beider Mikroorganismen vom CRISPR-Cas System, genauer den *cas* Proteinen, geschnitten wird. Werden die dargestellen Analysen auf publizierte archaeelle Genome übertragen, wird ersichtlich, dass das Attackieren von genomischer DNA anderer Archaeen phylumübergreifend und vor allem in extremen Ökosystemen der Tiefenbiosphäre prävalent ist. Die dargestellten Ergebnisse verdeutlichen, dass das Attackieren des Symbionten Einfluss auf die Art der Symbiose haben kann und impliziert, basierend auf der Analyse der metabolischen Abhängigkeiten, eine Verschiebung von Parasitismus zu Mutualismus. Zudem wurde das Expressionsprofil der metabolischen Kernfunktionen von *Ca.* Altiarchaeum in Assoziation und in Abwesenheit des Symbionten bestimmt. Obwohl die

untersuchten Ökosysteme unterschiedliche chemische und biochemische Zusammensetzungen haben, implizieren die Ergebnisse, dass *Ca.* Altiarchaeum in dem einen Ökosystem hauptsächlich von attackierenden Viren und in einem anderen vorwiegend durch ökologische Stressfaktoren beeinflusst wird. Verglichen mit anderen bekannten archaeellen Wirt-Symbiont Systemen, die auch DPANN Archaeen als Symbionten beinhaltet, werden durch die Analyse der CRISPR-Cas Interferenz neue Erkenntnisse über die Interaktionen der Archaeen in der terrestrischen Tiefenbiosphäre gewonnen. Im Allgemeinen beleuchtet diese Dissertation die Komplexität des Immunsystems eines Archaeums des Superphylums DPANN, welches im tiefen terrestrischen aquatischen Untergrunds zu finden ist, und erweitert das Wissen über, bis zum heutigen Zeitpunkt, unkultivierte archaeelle symbiotische Partner. Zudem vertieft der Vergleich der Expressionsraten der konservierten metabolischen Basisfunktionen des Altiarchaeums die Erkenntnis über den Gebrauch von vielfältigen metabolischen Funktionen in verschiedenen Ökosystemen.

6. Bibliography

A

Abdurakhmonov, I.Y., 2016. Bioinformatics: Updated Features and Applications. IntechOpen.

Agrawal, K., Verma, P., 2021. Chapter 23. "Omics"—A Step Toward Understanding of Complex Diversity of the Microbial Community | Elsevier Enhanced Reader [WWW Document]. https://doi.org/10.1016/B978-0-12-821881-5.00023-4

Aguiar-Pulido, V., Huang, W., Suarez-Ulloa, V., Cickovski, T., Mathee, K., Narasimhan, G., 2016. Metagenomics, Metatranscriptomics, and Metabolomics Approaches for Microbiome Analysis. Evolutionary Bioinformatics Online 12, 5. https://doi.org/10.4137/EBO.S36436

Aklujkar, M., Lovley, D.R., 2010. Interference with histidyl-tRNA synthetase by a CRISPR spacer sequence as a factor in the evolution of Pelobacter carbinolicus. BMC Evolutionary Biology 10, 230. https://doi.org/10.1186/1471-2148-10-230

Alkhnbashi, O.S., Shah, S.A., Garrett, R.A., Saunders, S.J., Costa, F., Backofen, R., 2016. Characterizing leader sequences of CRISPR loci. Bioinformatics 32, i576–i585. https://doi.org/10.1093/bioinformatics/btw454

Alomar, A.A., Pérez-Ramos, D.W., Kim, D., Kendziorski, N.L., Eastmond, B.H., Alto, B.W., Caragata, E.P., 2023. Native Wolbachia infection and larval competition stress shape fitness and West Nile virus infection in Culex quinquefasciatus mosquitoes. Frontiers in Microbiology 14.

Altschul, S.F., Gish, W., Miller, W., Myers, E.W., Lipman, D.J., 1990. Basic local alignment search tool. Journal of Molecular Biology 215, 403–410. https://doi.org/10.1016/S0022-2836(05)80360-2

Anderson, R., Brazelton, W., Baross, J., 2011. Is the Genetic Landscape of the Deep Subsurface Biosphere Affected by Viruses? Frontiers in Microbiology 2. https://doi.org/10.3389/fmicb.2011.00219

Andersson, A.F., Banfield, J.F., 2008. Virus population dynamics and acquired virus resistance in natural microbial communities. Science 320, 1047–50. https://doi.org/10.1126/science.1157358

Anisimova, M., Gil, M., Dufayard, J.-F., Dessimoz, C., Gascuel, O., 2011. Survey of Branch Support Methods Demonstrates Accuracy, Power, and Robustness of Fast Likelihood-based Approximation Schemes. Systematic Biology 60, 685–699. https://doi.org/10.1093/sysbio/syr041

Ari, Ş., Arikan, M., 2016. Next-Generation Sequencing: Advantages, Disadvantages, and Future, in: Hakeem, K.R., Tombuloğlu, H., Tombuloğlu, G. (Eds.), Plant Omics: Trends and Applications. Springer International Publishing, Cham, pp. 109–135. https://doi.org/10.1007/978-3-319-31703-8_5

B

Bai, Z., Feng, J., Franken, G.A.C., Al'Saadi, N., Cai, N., Yu, A.S., Lou, L., Komiya, Y., Hoenderop, J.G.J., Baaij, J.H.F. de, Yue, L., Runnels, L.W., 2021. CNNM proteins selectively bind to the TRPM7 channel to stimulate divalent cation entry into cells. PLOS Biology 19, e3001496. https://doi.org/10.1371/journal.pbio.3001496

Baker, B.J., Comolli, L.R., Dick, G.J., Hauser, L.J., Hyatt, D., Dill, B.D., Land, M.L., VerBerkmoes, N.C., Hettich, R.L., Banfield, J.F., 2010. Enigmatic, ultrasmall, uncultivated Archaea. Proc. Natl. Acad. Sci. U.S.A. 107, 8806–8811. https://doi.org/10.1073/pnas.0914470107

Baker, B.J., Tyson, G.W., Webb, R.I., Flanagan, J., Hugenholtz, P., Allen, E.E., Banfield, J.F., 2006. Lineages of Acidophilic Archaea Revealed by Community Genomic Analysis. Science 314, 1933–1935. https://doi.org/10.1126/science.1132690

Baker, E.T., German, C.R., 2004. On the global distribution of hydrothermal vent fields. Mid-Ocean Ridges: Hydrothermal Interactions Between the Lithosphere and Oceans, Geophys. Monogr. Ser 148, 245–266.

Baross, J.A., Hoffman, S.E., 1985. Submarine hydrothermal vents and associated gradient environments as sites for the origin and evolution of life. Origins of life and evolution of the biosphere 15, 327–345. https://doi.org/10.1007/BF01808177

Barrangou, R., Allen, C.C.R., Deveau, H., Richards, M., Boyaval, P., Moineau, S., Romero, D.A., Horvath, P., 2007a. CRISPR provides acquired resistance against viruses in prokaryotes. Science 315, 1709–12. https://doi.org/10.1126/science.1138140

Barrangou, R., Fremaux, C., Deveau, H., Richards, M., Boyaval, P., Moineau, S., Romero, D.A., Horvath, P., 2007b. CRISPR Provides Acquired Resistance Against Viruses in Prokaryotes. Science 315, 1709–1712. https://doi.org/10.1126/science.1138140

Benner, S.A., Ellington, A.D., Tauer, A., 1989. Modern metabolism as a palimpsest of the RNA world. Proceedings of the National Academy of Sciences 86, 7054–7058. https://doi.org/10.1073/pnas.86.18.7054

Bentley, D.R., Balasubramanian, S., Swerdlow, H.P., Smith, G.P., Milton, J., Brown, C.G., Hall, K.P., Evers, D.J., Barnes, C.L., Bignell, H.R., Boutell, J.M., Bryant, J., Carter, R.J., Keira Cheetham, R., Cox, A.J., Ellis, D.J., Flatbush, M.R., Gormley, N.A., Humphray, S.J., Irving, L.J., Karbelashvili, M.S., Kirk, S.M., Li, H., Liu, X., Maisinger, K.S., Murray, L.J., Obradovic, B., Ost, T., Parkinson, M.L., Pratt, M.R., Rasolonjatovo, I.M.J., Reed, M.T., Rigatti, R., Rodighiero, C., Ross, M.T., Sabot, A., Sankar, S.V., Scally, A., Schroth, G.P., Smith, M.E., Smith, V.P., Spiridou, A., Torrance, P.E., Tzonev, S.S., Vermaas, E.H., Walter, K., Wu, X., Zhang, L., Alam, M.D., Anastasi, C., Aniebo, I.C., Bailey, D.M.D., Bancarz, I.R., Banerjee, S., Barbour, S.G., Baybayan, P.A., Benoit, V.A., Benson, K.F., Bevis, C., Black, P.J., Boodhun, A., Brennan, J.S., Bridgham, J.A., Brown, R.C., Brown, A.A., Buermann, D.H., Bundu, A.A., Burrows, J.C., Carter, N.P., Castillo, N., Chiara E. Catenazzi, M., Chang, S., Neil Cooley, R., Crake, N.R., Dada, O.O., Diakoumakos, K.D., Dominguez-Fernandez, B., Earnshaw, D.J., Egbujor, U.C., Elmore, D.W., Etchin, S.S., Ewan, M.R., Fedurco, M., Fraser, L.J., Fuentes Fajardo, K.V., Scott Furey, W., George, D., Gietzen, K.J., Goddard, C.P., Golda, G.S., Granieri, P.A., Green, D.E., Gustafson, D.L., Hansen, N.F., Harnish, K., Haudenschild, C.D., Heyer, N.I., Hims, M.M., Ho, J.T., Horgan, A.M., Hoschler, K., Hurwitz, S., Ivanov, D.V., Johnson, M.Q., James, T., Huw Jones, T.A., Kang, G.-D., Kerelska, T.H., Kersey, A.D., Khrebtukova, I., Kindwall, A.P., Kingsbury, Z., Kokko-Gonzales, P.I., Kumar, A., Laurent, M.A., Lawley, C.T., Lee, S.E., Lee, X., Liao, A.K., Loch, J.A., Lok, M., Luo, S., Mammen, R.M., Martin, J.W., McCauley, P.G., McNitt, P., Mehta, P., Moon, K.W., Mullens, J.W., Newington, T., Ning, Z., Ling Ng, B., Novo, S.M., O'Neill, M.J., Osborne, M.A., Osnowski, A., Ostadan, O., Paraschos, L.L., Pickering, L., Pike, Andrew C., Pike, Alger C., Chris Pinkard, D., Pliskin, D.P., Podhasky, J., Quijano, V.J., Raczy, C., Rae, V.H., Rawlings, S.R., Chiva Rodriguez, A., Roe, P.M., Rogers, John, Rogert Bacigalupo, M.C., Romanov, N., Romieu, A., Roth, R.K., Rourke,

N.J., Ruediger, S.T., Rusman, E., Sanches-Kuiper, R.M., Schenker, M.R., Seoane, J.M., Shaw, R.J., Shiver, M.K., Short, S.W., Sizto, N.L., Sluis, J.P., Smith, M.A., Ernest Sohna Sohna, J., Spence, E.J., Stevens, K., Sutton, N., Szajkowski, L., Tregidgo, C.L., Turcatti, G., vandeVondele, S., Verhovsky, Y., Virk, S.M., Wakelin, S., Walcott, G.C., Wang, J., Worsley, G.J., Yan, J., Yau, L., Zuerlein, M., Rogers, Jane, Mullikin, J.C., Hurles, M.E., McCooke, N.J., West, J.S., Oaks, F.L., Lundberg, P.L., Klenerman, D., Durbin, R., Smith, A.J., 2008. Accurate whole human genome sequencing using reversible terminator chemistry. Nature 456, 53–59. https://doi.org/10.1038/nature07517

Berg, I.A., Kockelkorn, D., Ramos-Vera, W.H., Say, R.F., Zarzycki, J., Hügler, M., Alber, B.E., Fuchs, G., 2010. Autotrophic carbon fixation in archaea. Nature Reviews Microbiology 8, 447–460. https://doi.org/10.1038/nrmicro2365

Bhaya, D., Davison, M., Barrangou, R., 2011. CRISPR-Cas Systems in Bacteria and Archaea: Versatile Small RNAs for Adaptive Defense and Regulation. Annu. Rev. Genet. 45, 273–297. https://doi.org/10.1146/annurev-genet-110410-132430

Bi, E., Lutkenhaus, J., 1991. FtsZ ring structure associated with division in Escherichia coli. Nature 354, 161–164. https://doi.org/10.1038/354161a0

Bin Jang, H., Bolduc, B., Zablocki, O., Kuhn, J.H., Roux, S., Adriaenssens, E.M., Brister, J.R., Kropinski, A.M., Krupovic, M., Lavigne, R., Turner, D., Sullivan, M.B., 2019. Taxonomic assignment of uncultivated prokaryotic virus genomes is enabled by gene-sharing networks. Nature Biotechnology 37, 632–639. https://doi.org/10.1038/s41587-019-0100-8

Bintrim, S.B., Donohue, T.J., Handelsman, J., Roberts, G.P., Goodman, R.M., 1997. Molecular phylogeny of Archaea from soil. Proceedings of the National Academy of Sciences 94, 277–282. https://doi.org/10.1073/pnas.94.1.277

Bird, J.T., Baker, B.J., Probst, A.J., Podar, M., Lloyd, K.G., 2016. Culture Independent Genomic Comparisons Reveal Environmental Adaptations for Altiarchaeales. Frontiers in Microbiology 7. https://doi.org/10.3389/fmicb.2016.01221

Biswas, A., Fineran, P.C., Brown, C.M., 2014. Accurate computational prediction of the transcribed strand of CRISPR non-coding RNAs. Bioinformatics 30, 1805–13. https://doi.org/10.1093/bioinformatics/btu114

Biswas, A., Gagnon, J.N., Brouns, S.J.J., Fineran, P.C., Brown, C.M., 2013. CRISPRTarget: bioinformatic prediction and analysis of crRNA targets. RNA Biol 10, 817–827. https://doi.org/10.4161/rna.24046

Bolduc, B., Jang, H.B., Doulcier, G., You, Z.-Q., Roux, S., Sullivan, M.B., 2017. vConTACT: an iVirus tool to classify double-stranded DNA viruses that infect Archaea and Bacteria. PeerJ 5, e3243. https://doi.org/10.7717/peerj.3243

Bolotin, A., Quinquis, B., Sorokin, A., Ehrlich, S.D., 2005. Clustered regularly interspaced short palindrome repeats (CRISPRs) have spacers of extrachromosomal origin. Microbiology 151, 2551–2561.

Bondy-Denomy, J., Garcia, B., Strum, S., Du, M., Rollins, M.F., Hidalgo-Reyes, Y., Wiedenheft, B., Maxwell, K.L., Davidson, A.R., 2015. Multiple mechanisms for CRISPR–Cas inhibition by anti-CRISPR proteins. Nature 526, 136–139. https://doi.org/10.1038/nature15254

Bondy-Denomy, J., Pawluk, A., Maxwell, K.L., Davidson, A.R., 2012. Bacteriophage genes that inactivate the CRISPR/Cas bacterial immune system. Nature 493, 429. https://doi.org/10.1038/nature11723 https://www.nature.com/articles/nature11723#supplementary-information

Bornemann, T.L.V., Adam, P.S., Turzynski, V., Schreiber, U., Figueroa-Gonzalez, P.A., Rahlff, J., Köster, D., Schmidt, T.C., Schunk, R., Krauthausen, B., Probst, A.J., 2022. Genetic diversity in terrestrial subsurface ecosystems impacted by geological degassing. Nature Communications 13, 284. https://doi.org/10.1038/s41467-021-27783-7

Bornemann, T.L.V., Esser, S.P., Stach, T.L., Burg, T., Probst, A.J., 2023. uBin – a manual refining tool for genomes from metagenomes. Environmental Microbiology. https://doi.org/10.1111/1462-2920.16351

Brissette, J.L., Russel, M., Weiner, L., Model, P., 1990. Phage shock protein, a stress protein of Escherichia coli. Proceedings of the National Academy of Sciences 87, 862–866. https://doi.org/10.1073/pnas.87.3.862

Brodt, A., Lurie-Weinberger, M.N., Gophna, U., 2011. CRISPR loci reveal networks of gene exchange in archaea. Biology Direct 6, 65. https://doi.org/10.1186/1745-6150-6-65

Brokowski, C., Adli, M., 2019. CRISPR Ethics: Moral Considerations for Applications of a Powerful Tool. Journal of Molecular Biology, CRISPR: from the basic biology to its technological applications 431, 88–101. https://doi.org/10.1016/j.jmb.2018.05.044

Brouns, S.J.J., Jore, M.M., Lundgren, M., Westra, E.R., Slijkhuis, R.J.H., Snijders, A.P.L., Dickman, M.J., Makarova, K.S., Koonin, E.V., van der Oost, J., 2008a. Small CRISPR RNAs Guide Antiviral Defense in Prokaryotes. Science 321, 960–964. https://doi.org/10.1126/science.1159689

Brouns, S.J.J., Jore, M.M., Lundgren, M., Westra, E.R., Slijkhuis, R.J.H., Snijders, A.P.L., Dickman, M.J., Makarova, K.S., Koonin, E.V., van der Oost, J., 2008b. Small CRISPR RNAs Guide Antiviral Defense in Prokaryotes. Science 321, 960–964. https://doi.org/10.1126/science.1159689

Brown, C.T., Hug, L.A., Thomas, B.C., Sharon, I., Castelle, C.J., Singh, A., Wilkins, M.J., Wrighton, K.C., Williams, K.H., Banfield, J.F., 2015. Unusual biology across a group comprising more than 15% of domain Bacteria. Nature 523, 208–211. https://doi.org/10.1038/nature14486

Buchfink, B., Xie, C., Huson, D.H., 2014. Fast and sensitive protein alignment using DIAMOND. Nature Methods 12, 59. https://doi.org/10.1038/nmeth.3176 https://www.nature.com/articles/nmeth.3176#supplementary-information

C

Castelle, C.J., Banfield, J.F., 2018. Major New Microbial Groups Expand Diversity and Alter our Understanding of the Tree of Life. Cell 172, 1181–1197. https://doi.org/10.1016/j.cell.2018.02.016

Castelle, C.J., Brown, C.T., Anantharaman, K., Probst, A.J., Huang, R.H., Banfield, J.F., 2018. Biosynthetic capacity, metabolic variety and unusual biology in the CPR and DPANN radiations. Nature Reviews Microbiology 16, 629–645. https://doi.org/10.1038/s41579-018-0076-2

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., Taylor, R.C., Williams, K.H., Banfield, J.F., 2015. Genomic Expansion of Domain Archaea Highlights Roles for Organisms from New Phyla in Anaerobic Carbon Cycling. Current Biology 25, 690–701. https://doi.org/10.1016/j.cub.2015.01.014

Chabas, H., Müller, V., Bonhoeffer, S., Regoes, R.R., 2022. Epidemiological and evolutionary consequences of different types of CRISPR-Cas systems. PLOS Computational Biology 18, e1010329. https://doi.org/10.1371/journal.pcbi.1010329

Chaumeil, P.-A., Mussig, A.J., Hugenholtz, P., Parks, D.H., 2020. GTDB-Tk: a toolkit to classify genomes with the Genome Taxonomy Database. Bioinformatics 36, 1925–1927. https://doi.org/10.1093/bioinformatics/btz848

Cheeseman, P., Toms-Wood, A., Wolfe, R.S., 1972. Isolation and Properties of a Fluorescent Compound, Factor420, from Methanobacterium Strain M.o.H. Journal of Bacteriology 112, 527–531. https://doi.org/10.1128/jb.112.1.527-531.1972

Chen, I.-M.A., Chu, K., Palaniappan, K., Pillay, M., Ratner, A., Huang, J., Huntemann, M., Varghese, N., White, J.R., Seshadri, R., Smirnova, T., Kirton, E., Jungbluth, S.P., Woyke, T., Eloe-Fadrosh, E.A., Ivanova, N.N., Kyrpides, N.C., 2019. IMG/M v.5.0: an integrated data management and comparative analysis system for microbial genomes and microbiomes. Nucleic Acids Res 47, D666–D677. https://doi.org/10.1093/nar/gky901

Chen, L.-X., Méndez-García, C., Dombrowski, N., Servín-Garcidueñas, L.E., Eloe-Fadrosh, E.A., Fang, B.-Z., Luo, Z.-H., Tan, S., Zhi, X.-Y., Hua, Z.-S., Martinez-Romero, E., Woyke, T., Huang, L.-N., Sánchez, J., Peláez, A.I., Ferrer, M., Baker, B.J., Shu, W.-S., 2018. Metabolic versatility of small archaea Micrarchaeota and Parvarchaeota. ISME J 12, 756–775. https://doi.org/10.1038/s41396-017-0002-z

Chénard, C., Lauro, F.M., 2017. Microbial ecology of extreme environments. Springer.

Comolli, L.R., Banfield, J.F., 2014. Inter-species interconnections in acid mine drainage microbial communities. Frontiers in Microbiology 5, 367. https://doi.org/10.3389/fmicb.2014.00367

Cook, R., Brown, N., Redgwell, T., Rihtman, B., Barnes, M., Clokie, M., Stekel, D.J., Hobman, J., Jones, M.A., Millard, A., 2021. INfrastructure for a PHAge REference Database: Identification of Large-Scale Biases in the Current Collection of Cultured Phage Genomes. PHAGE 2, 214–223. https://doi.org/10.1089/phage.2021.0007

Corliss, J.B., Dymond, J., Gordon, L.I., Edmond, J.M., von Herzen, R.P., Ballard, R.D., Green, K., Williams, D., Bainbridge, A., Crane, K., van Andel, T.H., 1979. Submarine Thermal Springs on the Galápagos Rift. Science 203, 1073–1083. https://doi.org/10.1126/science.203.4385.1073

Couvin, D., Bernheim, A., Toffano-Nioche, C., Touchon, M., Michalik, J., Neron, B., Rocha, E.P.C., Vergnaud, G., Gautheret, D., Pourcel, C., 2018. CRISPRCasFinder, an update of CRISRFinder, includes a portable version, enhanced performance and integrates search for Cas proteins. Nucleic Acids Res 46, W246–W251. https://doi.org/10.1093/nar/gky425

Criscuolo, A., Gribaldo, S., 2010. BMGE (Block Mapping and Gathering with Entropy): a new software for selection of phylogenetic informative regions from multiple sequence alignments. BMC Evolutionary Biology 10, 210. https://doi.org/10.1186/1471-2148-10-210

Crooks, G.E., Hon, G., Chandonia, J.-M., Brenner, S.E., 2004. WebLogo: A Sequence Logo Generator. Genome Research 14, 1188–1190. https://doi.org/10.1101/gr.849004

D

Darling, A.E., Jospin, G., Lowe, E., Matsen, F.A., IV, Bik, H.M., Eisen, J.A., 2014. PhyloSift: phylogenetic analysis of genomes and metagenomes. PeerJ 2, e243. https://doi.org/10.7717/peerj.243

De León, K., Gerlach, R., Peyton, B., Fields, M., 2013. Archaeal and bacterial communities in three alkaline hot springs in Heart Lake Geyser Basin, Yellowstone National Park. Frontiers in Microbiology 4.

DeLong, E.F., 1998. Everything in moderation: Archaea as 'non-extremophiles.' Current Opinion in Genetics & Development 8, 649–654. https://doi.org/10.1016/S0959-437X(98)80032-4

Delsuc, F., Brinkmann, H., Philippe, H., 2005. Phylogenomics and the reconstruction of the tree of life. Nat Rev Genet 6, 361–375. https://doi.org/10.1038/nrg1603

Deming, J.W., 2002. Psychrophiles and polar regions. Curr Opin Microbiol 5, 301–309. https://doi.org/10.1016/s1369-5274(02)00329-6

Den Blaauwen Tanneke, Buddelmeijer Nienke, Aarsman Mirjam E. G., Hameete Cor M., Nanninga Nanne, 1999. Timing of FtsZ Assembly in Escherichia coli. Journal of Bacteriology 181, 5167–5175. https://doi.org/10.1128/JB.181.17.5167-5175.1999

Deveau, H., Barrangou, R., Garneau, J.E., Labonté, J., Fremaux, C., Boyaval, P., Romero, D.A., Horvath, P., Moineau, S., 2008. Phage response to CRISPR-encoded resistance in Streptococcus thermophilus. J Bacteriol 190, 1390–1400. https://doi.org/10.1128/JB.01412-07

Dhillon, A., Teske, A., Dillon, J., Stahl, D.A., Sogin, M.L., 2003. Molecular Characterization of Sulfate-Reducing Bacteria in the Guaymas Basin. Applied and Environmental Microbiology 69, 2765–2772. https://doi.org/10.1128/AEM.69.5.2765-2772.2003

Dombrowski, N., Lee, J.-H., Williams, T.A., Offre, P., Spang, A., 2019. Genomic diversity, lifestyles and evolutionary origins of DPANN archaea. FEMS Microbiology Letters 366. https://doi.org/10.1093/femsle/fnz008

Dombrowski, N., Teske, A.P., Baker, B.J., 2018. Expansive microbial metabolic versatility and biodiversity in dynamic Guaymas Basin hydrothermal sediments. Nature Communications 9, 4999. https://doi.org/10.1038/s41467-018-07418-0

Doudna, J.A., Charpentier, E., 2014. The new frontier of genome engineering with CRISPR-Cas9. Science 346, 1258096. https://doi.org/10.1126/science.1258096

Dufault-Thompson, K., Steffensen, J.L., Zhang, Y., 2018. Using PSAMM for the Curation and Analysis of Genome-Scale Metabolic Models, in: Fondi, M. (Ed.), Metabolic Network Reconstruction and Modeling: Methods and Protocols. Springer New York, New York, NY, pp. 131–150. https://doi.org/10.1007/978-1-4939-7528-0_6

E

Eddy, S.R., 2011. Accelerated Profile HMM Searches. PLOS Computational Biology 7, e1002195. https://doi.org/10.1371/journal.pcbi.1002195

Edgar, R.C., 2010. Search and clustering orders of magnitude faster than BLAST. Bioinformatics 26, 2460–2461. https://doi.org/10.1093/bioinformatics/btq461 %J Bioinformatics

Edgar, R.C., 2004. MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Research 32, 1792–1797. https://doi.org/10.1093/nar/gkh340

Edgcomb, V.P., Kysela, D.T., Teske, A., de Vera Gomez, A., Sogin, M.L., 2002. Benthic eukaryotic diversity in the Guaymas Basin hydrothermal vent environment. Proceedings of the National Academy of Sciences 99, 7658–7662. https://doi.org/10.1073/pnas.062186399

Eirich, L.D., Vogels, G.D., Wolfe, R.S., 1979. Distribution of coenzyme F420 and properties of its hydrolytic fragments. Journal of Bacteriology 140, 20–27. https://doi.org/10.1128/jb.140.1.20-27.1979

Eirich, L.D., Vogels, G.D., Wolfe, R.S., 1978. Proposed structure for coenzyme F420 from Methanobacterium. Biochemistry 17, 4583–4593.

Emerson, J.B., Thomas, B.C., Alvarez, W., Banfield, J.F., 2016. Metagenomic analysis of a high carbon dioxide subsurface microbial community populated by chemolithoautotrophs and bacteria and archaea from candidate phyla. Environmental Microbiology 18, 1686–1703. https://doi.org/10.1111/1462-2920.12817

Erdmann, S., Garrett, R.A., 2012. Selective and hyperactive uptake of foreign DNA by adaptive immune systems of an archaeon via two distinct mechanisms. Molecular Microbiology 85, 1044–1056. https://doi.org/10.1111/j.1365-2958.2012.08171.x

Eren, A.M., Vineis, J.H., Morrison, H.G., Sogin, M.L., 2013. A Filtering Method to Generate High Quality Short Reads Using Illumina Paired-End Technology. PLOS ONE 8, e66643. https://doi.org/10.1371/journal.pone.0066643

F

Farris, J.S., 1972. Estimating Phylogenetic Trees from Distance Matrices. The American Naturalist 106, 645–668. https://doi.org/10.1086/282802

Fiehn, O., 2002. Metabolomics — the link between genotypes and phenotypes, in: Town, C. (Ed.), Functional Genomics. Springer Netherlands, Dordrecht, pp. 155–171. https://doi.org/10.1007/978-94-010-0448-0_11

Flemming, H.-C., Wuertz, S., 2019. Bacteria and archaea on Earth and their abundance in biofilms. Nature Reviews Microbiology 17, 247–260. https://doi.org/10.1038/s41579-019-0158-9

Forterre, P., 2013. The Common Ancestor of Archaea and Eukarya Was Not an Archaeon. Archaea 2013, 372396. https://doi.org/10.1155/2013/372396

Forterre, P., Gribaldo, S., Brochier-Armanet, C., 2009. Happy together: genomic insights into the unique Nanoarchaeum/Ignicoccus association. Journal of Biology 8, 7. https://doi.org/10.1186/jbiol110

Fox, G.E., Magrum, L.J., Balch, W.E., Wolfe, R.S., Woese, C.R., 1977. Classification of methanogenic bacteria by 16S ribosomal RNA characterization. Proceedings of the National Academy of Sciences 74, 4537–4541. https://doi.org/10.1073/pnas.74.10.4537

Fu, L., Niu, B., Zhu, Z., Wu, S., Li, W., 2012. CD-HIT: accelerated for clustering the next-generation sequencing data. Bioinformatics 28, 3150–3152. https://doi.org/10.1093/bioinformatics/bts565

Funato, Y., Yamazaki, D., Mizukami, S., Du, L., Kikuchi, K., Miki, H., 2014. Membrane protein CNNM4–dependent Mg^{2+} efflux suppresses tumor progression. J Clin Invest 124, 5398–5410. https://doi.org/10.1172/JCI76614

G

Gaci, N., Borrel, G., Tottey, W., O'Toole, P.W., Brugère, J.-F., 2014. Archaea and the human gut: New beginning of an old story. World J Gastroenterol 20, 16062–16078. https://doi.org/10.3748/wjg.v20.i43.16062

Garamella, J., Marshall, R., Rustad, M., Noireaux, V., 2016. The All E. coli TX-TL Toolbox 2.0: A Platform for Cell-Free Synthetic Biology. ACS Synth. Biol. 5, 344–355. https://doi.org/10.1021/acssynbio.5b00296

Garneau, J.E., Dupuis, M.-È., Villion, M., Romero, D.A., Barrangou, R., Boyaval, P., Fremaux, C., Horvath, P., Magadán, A.H., Moineau, S., 2010. The CRISPR/Cas bacterial immune system cleaves bacteriophage and plasmid DNA. Nature 468, 67. https://doi.org/10.1038/nature09523 https://www.nature.com/articles/nature09523#supplementary-information

Garrett, R.A., Vestergaard, G., Shah, S.A., 2011. Archaeal CRISPR-based immune systems: exchangeable functional modules. Trends in Microbiology 19, 549–556. https://doi.org/10.1016/j.tim.2011.08.002

Giannone, R.J., Huber, H., Karpinets, T., Heimerl, T., Küper, U., Rachel, R., Keller, M., Hettich, R.L., Podar, M., 2011. Proteomic Characterization of Cellular and Molecular Processes that Enable the Nanoarchaeum equitans-Ignicoccus hospitalis Relationship. PLoS ONE 6, e22942. https://doi.org/10.1371/journal.pone.0022942

Giannone, R.J., Wurch, L.L., Heimerl, T., Martin, S., Yang, Z., Huber, H., Rachel, R., Hettich, R.L., Podar, M., 2015. Life on the edge: functional genomic response of Ignicoccus hospitalis to the presence of Nanoarchaeum equitans. The Isme Journal 9, 101–114. https://doi.org/10.1038/ismej.2014.112

Gleditzsch, D., Pausch, P., Müller-Esparza, H., Özcan, A., Guo, X., Bange, G., Randau, L., 2019. PAM identification by CRISPR-Cas effector complexes: diversified mechanisms and structures. RNA Biology 16, 504–517. https://doi.org/10.1080/15476286.2018.1504546

Gómez, F., 2011. Extreme Environment, in: Gargaud, M., Amils, R., Quintanilla, J.C., Cleaves, H.J. (Jim), Irvine, W.M., Pinti, D.L., Viso, M. (Eds.), Encyclopedia of Astrobiology. Springer Berlin Heidelberg, Berlin, Heidelberg, pp. 570–572. https://doi.org/10.1007/978-3-642-11274-4_566

Gonnerman, M.C., Benedict, M.N., Feist, A.M., Metcalf, W.W., Price, N.D., 2013. Genomically and biochemically accurate metabolic reconstruction of Methanosarcina barkeri Fusaro, iMG746. Biotechnology Journal 8, 1070–1079. https://doi.org/10.1002/biot.201200266

Gouy, M., Tannier, E., Comte, N., Parsons, D.P., 2021. Seaview Version 5: A Multiplatform Software for Multiple Sequence Alignment, Molecular Phylogenetic Analyses, and Tree Reconciliation, in: Katoh, K. (Ed.), Multiple Sequence Alignment: Methods and Protocols. Springer US, New York, NY, pp. 241–260. https://doi.org/10.1007/978-1-0716-1036-7_15

Goyal, N., Widiastuti, H., Karimi, I.A., Zhou, Z., 2014. A genome-scale metabolic model of Methanococcus maripaludis S2 for CO2 capture and conversion to methane. Mol. BioSyst. 10, 1043–1054. https://doi.org/10.1039/C3MB70421A

Grissa, I., Vergnaud, G., Pourcel, C., 2007. CRISPRFinder: a web tool to identify clustered regularly interspaced short palindromic repeats. Nucleic Acids Res 35, W52-7. https://doi.org/10.1093/nar/gkm360

Guindon, S., Dufayard, J.-F., Lefort, V., Anisimova, M., Hordijk, W., Gascuel, O., 2010. New Algorithms and Methods to Estimate Maximum-Likelihood Phylogenies: Assessing the Performance of PhyML 3.0. Systematic Biology 59, 307–321. https://doi.org/10.1093/sysbio/syq010

H

Hamilton, T.L., Jones, D.S., Schaperdoth, I., Macalady, J.L., 2015. Metagenomic insights into S(0) precipitation in a terrestrial subsurface lithoautotrophic ecosystem. Frontiers in Microbiology 5.

Hamm, J.N., Erdmann, S., Eloe-Fadrosh, E.A., Angeloni, A., Zhong, L., Brownlee, C., Williams, T.J., Barton, K., Carswell, S., Smith, M.A., Brazendale, S., Hancock, A.M., Allen, M.A., Raftery, M.J., Cavicchioli, R., 2019. Unexpected host dependency of Antarctic Nanohaloarchaeota. Proc Natl Acad Sci USA 116, 14661. https://doi.org/10.1073/pnas.1905179116

Hamm, J.N., Liao, Y., Kügelgen, A. von, Dombrowski, N., Landers, E., Brownlee, C., Johansson, E.M.V., Whan, R.M., Baker, M.A.B., Baum, B., Bharat, T.A.M., Duggin, I.G., Spang, A., Cavicchioli, R., 2023. The intracellular lifestyle of an archaeal symbiont. bioRxiv 2023.02.24.529834. https://doi.org/10.1101/2023.02.24.529834

He, C., Keren, R., Whittaker, M.L., Farag, I.F., Doudna, J.A., Cate, J.H.D., Banfield, J.F., 2021. Genome-resolved metagenomics reveals site-specific diversity of episymbiotic CPR bacteria and DPANN archaea in groundwater ecosystems. Nature Microbiology 6, 354–365. https://doi.org/10.1038/s41564-020-00840-5

He, L., St. John James, M., Radovcic, M., Ivancic-Bace, I., Bolt, E.L., 2020. Cas3 Protein—A Review of a Multi-Tasking Machine. Genes 11. https://doi.org/10.3390/genes11020208

Head, I.M., Saunders, J.R., Pickup, R.W., 1998. Microbial Evolution, Diversity, and Ecology: A Decade of Ribosomal RNA Analysis of Uncultivated Microorganisms. Microbial Ecology 35, 1–21. https://doi.org/10.1007/s002489900056

Heidelberg, J.F., Nelson, W.C., Schoenfeld, T., Bhaya, D., 2009. Germ Warfare in a Microbial Mat Community: CRISPRs Provide Insights into the Co-Evolution of Host and Viral Genomes. PLOS ONE 4, e4169. https://doi.org/10.1371/journal.pone.0004169

Heimerl, T., Flechsler, J., Pickl, C., Heinz, V., Salecker, B., Zweck, J., Wanner, G., Geimer, S., Samson, R.Y., Bell, S.D., Huber, H., Wirth, R., Wurch, L., Podar, M., Rachel, R., 2017. A Complex Endomembrane System in the Archaeon Ignicoccus hospitalis Tapped by Nanoarchaeum equitans. Frontiers in Microbiology 8, 1072. https://doi.org/10.3389/fmicb.2017.01072

Hendrix, R.W., Smith, M.C.M., Burns, R.N., Ford, M.E., Hatfull, G.F., 1999. Evolutionary relationships among diverse bacteriophages and prophages: All the world's a phage. Proceedings of the National Academy of Sciences 96, 2192–2197. https://doi.org/10.1073/pnas.96.5.2192

Henneberger, R., Moissl, C., Amann, T., Rudolph, C., Huber, R., 2006. New Insights into the Lifestyle of the Cold-Loving SM1 Euryarchaeon: Natural Growth as a Monospecies Biofilm in the Subsurface. Applied and Environmental Microbiology 72, 192–199. https://doi.org/10.1128/AEM.72.1.192-199.2006

Hernsdorf, A.W., Amano, Y., Miyakawa, K., Ise, K., Suzuki, Y., Anantharaman, K., Probst, A., Burstein, D., Thomas, B.C., Banfield, J.F., 2017. Potential for microbial H2 and metal transformations associated with novel bacteria and archaea in deep terrestrial subsurface sediments. The ISME Journal 11, 1915–1929. https://doi.org/10.1038/ismej.2017.39

Herrera, A., Cockell, C.S., 2007. Exploring microbial diversity in volcanic environments: A review of methods in DNA extraction. Journal of Microbiological Methods 70, 1–12. https://doi.org/10.1016/j.mimet.2007.04.005

Heussler, G.E., O'Toole, G.A., 2016. Friendly Fire: Biological Functions and Consequences of Chromosomal Targeting by CRISPR-Cas Systems. Journal of Bacteriology 198, 1481–1486. https://doi.org/10.1128/JB.00086-16

Hille, F., Charpentier, E., 2016. CRISPR-Cas: biology, mechanisms and relevance. Philosophical Transactions of the Royal Society B: Biological Sciences 371, 20150496. https://doi.org/10.1098/rstb.2015.0496

Hinrichs, K.-U., Hayes, J.M., Sylva, S.P., Brewer, P.G., DeLong, E.F., 1999. Methane-consuming archaebacteria in marine sediments. Nature 398, 802–805. https://doi.org/10.1038/19751

Hoang, D.T., Chernomor, O., von Haeseler, A., Minh, B.Q., Vinh, L.S., 2017. UFBoot2: Improving the Ultrafast Bootstrap Approximation. Molecular Biology and Evolution 35, 518–522. https://doi.org/10.1093/molbev/msx281

Hohenester, E., 2019. Laminin G-like domains: dystroglycan-specific lectins. Current Opinion in Structural Biology 56, 56–63. https://doi.org/10.1016/j.sbi.2018.11.007

Hohenester, E., Yurchenco, P.D., 2013. Laminins in basement membrane assembly. null 7, 56–63. https://doi.org/10.4161/cam.21831

Horvath, P., Barrangou, R., 2010. CRISPR/Cas, the immune system of bacteria and archaea. Science 327, 167–70. https://doi.org/10.1126/science.1179555

Huber, H., Burggraf, S., Mayer, T., Wyschkony, I., Rachel, R., Stetter, K.O., 2000. Ignicoccus gen. nov., a novel genus of hyperthermophilic, chemolithoautotrophic Archaea, represented by two new species, Ignicoccus islandicus sp nov and Ignicoccus pacificus sp nov. and Ignicoccus pacificus sp. nov. International Journal of Systematic and Evolutionary Microbiology 50, 2093–2100. https://doi.org/10.1099/00207713-50-6-2093

Huber, H., Hohn, M.J., Rachel, R., Fuchs, T., Wimmer, V.C., Stetter, K.O., 2002. A new phylum of Archaea represented by a nanosized hyperthermophilic symbiont. Nature 417, 63–67. https://doi.org/10.1038/417063a

Huber, H., Hohn, M.J., Stetter, K.O., Rachel, R., 2003. The phylum Nanoarchaeota: Present knowledge and future perspectives of a unique form of life. Research in Microbiology 154, 165–171. https://doi.org/10.1016/S0923-2508(03)00035-4

Huber, H., Küper, U., Daxer, S., Rachel, R., 2012. The unusual cell biology of the hyperthermophilic Crenarchaeon Ignicoccus hospitalis. Antonie van Leeuwenhoek 102, 203–219. https://doi.org/10.1007/s10482-012-9748-5

Huerta-Cepas, J., Szklarczyk, D., Heller, D., Hernández-Plaza, A., Forslund, S.K., Cook, H., Mende, D.R., Letunic, I., Rattei, T., Jensen, L.J., von Mering, C., Bork, P., 2019. eggNOG

5.0: a hierarchical, functionally and phylogenetically annotated orthology resource based on 5090 organisms and 2502 viruses. Nucleic Acids Res 47, D309–D314. https://doi.org/10.1093/nar/gky1085

Hwang, Y., Roux, S., Coclet, C., Krause, S.J.E., Girguis, P.R., 2023. Viruses interact with hosts that span distantly related microbial domains in dense hydrothermal mats. Nature Microbiology. https://doi.org/10.1038/s41564-023-01347-5

Hyatt, D., Chen, G.L., Locascio, P.F., Land, M.L., Larimer, F.W., Hauser, L.J., 2010. Prodigal: prokaryotic gene recognition and translation initiation site identification. BMC Bioinformatics 11, 119. https://doi.org/10.1186/1471-2105-11-119

I

Ishino Y, Shinagawa H, Makino K, Amemura M, Nakata A, 1987. Nucleotide sequence of the iap gene, responsible for alkaline phosphatase isozyme conversion in Escherichia coli, and identification of the gene product. Journal of Bacteriology 169, 5429–5433. https://doi.org/10.1128/jb.169.12.5429-5433.1987

J

Jahn, U., Gallenberger, M., Junglas, B., Eisenreich, W., Stetter, K.O., Rachel, R., Huber, H. %J J. of bacteriology, 2008. Nanoarchaeum equitans and Ignicoccus hospitalis: new insights into a unique, intimate association of two archaea. Journal of Bacteriology 190, 1743–1750.

Jahn, U., Summons, R., Sturt, H., Grosjean, E., Huber, H., 2004. Composition of the lipids of Nanoarchaeum equitans and their origin from its host Ignicoccus sp. strain KIN4/I. Archives of Microbiology 182, 404–413. https://doi.org/10.1007/s00203-004-0725-x

Jannasch, H.W., Mottl, M.J., 1985. Geomicrobiology of Deep-Sea Hydrothermal Vents. Science 229, 717–725. https://doi.org/10.1126/science.229.4715.717

Jansen, R., Embden, J.D.A. van, Gaastra, W., Schouls, L.M., 2002. Identification of genes that are associated with DNA repeats in prokaryotes 43, 1565–1575. https://doi.org/10.1046/j.1365-2958.2002.02839.x

Jarett, J.K., Nayfach, S., Podar, M., Inskeep, W., Ivanova, N.N., Munson-McGee, J., Schulz, F., Young, M., Jay, Z.J., Beam, J.P., Kyrpides, N.C., Malmstrom, R.R., Stepanauskas, R., Woyke, T., 2018. Single-cell genomics of co-sorted Nanoarchaeota suggests novel putative host associations and diversification of proteins involved in symbiosis. Microbiome 6, 161. https://doi.org/10.1186/s40168-018-0539-8

Jore, M.M., Lundgren, M., van Duijn, E., Bultema, J.B., Westra, E.R., Waghmare, S.P., Wiedenheft, B., Pul, Ü., Wurm, R., Wagner, R., Beijer, M.R., Barendregt, A., Zhou, K., Snijders, A.P.L., Dickman, M.J., Doudna, J.A., Boekema, E.J., Heck, A.J.R., van der Oost, J., Brouns, S.J.J., 2011. Structural basis for CRISPR RNA-guided DNA recognition by Cascade. Nature Structural & Molecular Biology 18, 529–536. https://doi.org/10.1038/nsmb.2019

Joshi, N.A., Fass, J.N., 2011. Sickle: A sliding-window, adaptive, quality-based trimming tool for FastQ files (Version 1.33) [Software].

Jung, J., Kim, J.-S., Taffner, J., Berg, G., Ryu, C.-M., 2020. Archaea, tiny helpers of land plants. Computational and Structural Biotechnology Journal 18, 2494–2500. https://doi.org/10.1016/j.csbj.2020.09.005

K

Kalyaanamoorthy, S., Minh, B.Q., Wong, T.K.F., von Haeseler, A., Jermiin, L.S., 2017. ModelFinder: fast model selection for accurate phylogenetic estimates. Nature Methods 14, 587–589. https://doi.org/10.1038/nmeth.4285

Kanehisa, M., Furumichi, M., Tanabe, M., Sato, Y., Morishima, K., 2017. KEGG: new perspectives on genomes, pathways, diseases and drugs. Nucleic Acids Res 45, D353–D361. https://doi.org/10.1093/nar/gkw1092

Kchouk, M., Gibrat, J.-F., Elloumi, M., 2017. Generations of sequencing technologies: from first to next generation. Electromagnetic Biology and Medicine 9, 8-p.

Kiro, R., Shitrit, D., Qimron, U., 2014. Efficient engineering of a bacteriophage genome using the type I-E CRISPR-Cas system. RNA Biology 11, 42–44. https://doi.org/10.4161/rna.27766

Koboldt, D.C., Zhang, Q., Larson, D.E., Shen, D., McLellan, M.D., Lin, L., Miller, C.A., Mardis, E.R., Ding, L., Wilson, R.K., 2012. VarScan 2: Somatic mutation and copy number alteration discovery in cancer by exome sequencing. Genome Research 22, 568–576. https://doi.org/10.1101/gr.129684.111

Kolesnik, M.V., Fedorova, I., Karneyeva, K.A., Artamonova, D.N., Severinov, K.V., 2021. Type III CRISPR-Cas Systems: Deciphering the Most Complex Prokaryotic Immune System. Biochemistry Moscow 86, 1301–1314. https://doi.org/10.1134/S0006297921100114

Koonin, E.V., 2015. Origin of eukaryotes from within archaea, archaeal eukaryome and bursts of gene gain: eukaryogenesis just made easier? Philos Trans R Soc Lond B Biol Sci 370, 20140333. https://doi.org/10.1098/rstb.2014.0333

Koonin, E.V., Makarova, K.S., 2022. Evolutionary plasticity and functional versatility of CRISPR systems. PLOS Biology 20, e3001481. https://doi.org/10.1371/journal.pbio.3001481

Koonin, E.V., Makarova, K.S., Zhang, F., 2017. Diversity, classification and evolution of CRISPR-Cas systems. Current Opinion in Microbiology, Environmental microbiology * CRISPRcas9 37, 67–78. https://doi.org/10.1016/j.mib.2017.05.008

Krassowski, M., Das, V., Sahu, S.K., Misra, B.B., 2020. State of the Field in Multi-Omics Research: From Computational Needs to Data Mining and Sharing. Frontiers in Genetics 11.

Krause, S., Gfrerer, S., von Kügelgen, A., Reuse, C., Dombrowski, N., Villanueva, L., Bunk, B., Spröer, C., Neu, T.R., Kuhlicke, U., Schmidt-Hohagen, K., Hiller, K., Bharat, T.A.M., Rachel, R., Spang, A., Gescher, J., 2022. The importance of biofilm formation for cultivation of a Micrarchaeon and its interactions with its Thermoplasmatales host. Nat Commun 13, 1735. https://doi.org/10.1038/s41467-022-29263-y

Krupovic, M., Béguin, P., Koonin, E.V., 2017. Casposons: mobile genetic elements that gave rise to the CRISPR-Cas adaptation machinery. Curr Opin Microbiol 38, 36–43. https://doi.org/10.1016/j.mib.2017.04.004

Kunin, V., Engelbrektson, A., Ochman, H., Hugenholtz, P., 2010. Wrinkles in the rare biosphere: pyrosequencing errors can lead to artificial inflation of diversity estimates. Environmental Microbiology 12, 118–123. https://doi.org/10.1111/j.1462-2920.2009.02051.x

Kuroda, K., Yamamoto, K., Nakai, R., Hirakata, Y., Kubota, K., Nobu, M.K., Narihiro, T., 2022. Symbiosis between Candidatus Patescibacteria and Archaea Discovered in

Wastewater-Treating Bioreactors. mBio 13, e01711-22. https://doi.org/10.1128/mbio.01711-22

L

La Cono, V., Messina, E., Rohde, M., Arcadi, E., Ciordia, S., Crisafi, F., Denaro, R., Ferrer, M., Giuliano, L., Golyshin, P.N., Golyshina, O.V., Hallsworth, J.E., La Spada, G., Mena, M.C., Merkel, A.Y., Shevchenko, M.A., Smedile, F., Sorokin, D.Y., Toshchakov, S.V., Yakimov, M.M., 2020. Symbiosis between nanohaloarchaeon and haloarchaeon is based on utilization of different polysaccharides. Proceedings of the National Academy of Sciences 117, 20223–20234. https://doi.org/10.1073/pnas.2007232117

Langmead, B., Salzberg, S.L., 2012. Fast gapped-read alignment with Bowtie 2. Nature Methods 9, 357–359. https://doi.org/10.1038/nmeth.1923

Leenay, R.T., Maksimchuk, K.R., Slotkowski, R.A., Agrawal, R.N., Gomaa, A.A., Briner, A.E., Barrangou, R., Beisel, C.L., 2016. Identifying and Visualizing Functional PAM Diversity across CRISPR-Cas Systems. Molecular Cell 62, 137–147. https://doi.org/10.1016/j.molcel.2016.02.031

Lefort, V., Desper, R., Gascuel, O., 2015. FastME 2.0: A Comprehensive, Accurate, and Fast Distance-Based Phylogeny Inference Program. Molecular Biology and Evolution 32, 2798–2800. https://doi.org/10.1093/molbev/msv150

Letunic, I., Bork, P., 2019. Interactive Tree Of Life (iTOL) v4: recent updates and new developments. Nucleic Acids Research 47, W256–W259. https://doi.org/10.1093/nar/gkz239

Levy, A., Goren, M.G., Yosef, I., Auster, O., Manor, M., Amitai, G., Edgar, R., Qimron, U., Sorek, R., 2015. CRISPR adaptation biases explain preference for acquisition of foreign DNA. Nature 520, 505–510. https://doi.org/10.1038/nature14302

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., Durbin, R., 1000 Genome Project Data Processing Subgroup, 2009. The Sequence Alignment/Map format and SAMtools. Bioinformatics 25, 2078–2079. https://doi.org/10.1093/bioinformatics/btp352

Li, W., Godzik, A., 2006. Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. Bioinformatics 22, 1658–1659. https://doi.org/10.1093/bioinformatics/btl158

Liao, Y., Ithurbide, S., Evenhuis, C., Löwe, J., Duggin, I.G., 2021. Cell division in the archaeon Haloferax volcanii relies on two FtsZ proteins with distinct functions in division ring assembly and constriction. Nature Microbiology 6, 594–605. https://doi.org/10.1038/s41564-021-00894-z

Liao, Y., Shi, W., 2020. Read trimming is not required for mapping and quantification of RNA-seq reads at the gene level. NAR Genomics and Bioinformatics 2, lqaa068. https://doi.org/10.1093/nargab/lqaa068

Lillestøl, R., Redder, P., Garrett, R.A., Brügger, K., 2006. A putative viral defence mechanism in archaeal cells. Archaea 2, 59–72.

Liu, S., Wang, Z., Zhu, R., Wang, F., Cheng, Y., Liu, Y., 2021. Three Differential Expression Analysis Methods for RNA Sequencing: limma, EdgeR, DESeq2. JoVE e62528. https://doi.org/10.3791/62528

Liu, X., Li, M., Castelle, C.J., Probst, A.J., Zhou, Z., Pan, J., Liu, Y., Banfield, J.F., Gu, J.-D., 2018. Insights into the ecology, evolution, and metabolism of the widespread Woesearchaeotal lineages. Microbiome 6, 102. https://doi.org/10.1186/s40168-018-0488-2

Ljungdhal, L., 1986. The autotrophic pathway of acetate synthesis in acetogenic bacteria. Annual Reviews in Microbiology 40, 415–450.

Love, M.I., Huber, W., Anders, S., 2014. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biology 15, 550. https://doi.org/10.1186/s13059-014-0550-8

Luscombe, N.M., Greenbaum, D., Gerstein, M., 2001. What is bioinformatics? A proposed definition and overview of the field. Methods Inf Med 40, 346–58.

M

Magrum, L.J., Luehrsen, K.R., Woese, C.R., 1978. Are extreme halophiles actually “bacteria”? Journal of Molecular Evolution 11, 1–8. https://doi.org/10.1007/BF01768019

Maidak, B.L., Olsen, G.J., Larsen, N., Overbeek, R., McCaughey, M.J., Woese, C.R., 1997. The RDP (Ribosomal Database Project). Nucleic Acids Research 25, 109–110. https://doi.org/10.1093/nar/25.1.109

Makarova, K.S., Haft, D.H., Barrangou, R., Brouns, S.J.J., Charpentier, E., Horvath, P., Moineau, S., Mojica, F.J.M., Wolf, Y.I., Yakunin, A.F., van der Oost, J., Koonin, E.V., 2011. Evolution and classification of the CRISPR–Cas systems. Nature Reviews Microbiology 9, 467. https://doi.org/10.1038/nrmicro2577 https://www.nature.com/articles/nrmicro2577#supplementary-information

Makarova, K.S., Koonin, E.V., 2015. Annotation and Classification of CRISPR-Cas Systems. Methods Mol Biol 1311, 47–75. https://doi.org/10.1007/978-1-4939-2687-9_4

Makarova, K.S., Wolf, Y.I., Alkhnbashi, O.S., Costa, F., Shah, S.A., Saunders, S.J., Barrangou, R., Brouns, S.J.J., Charpentier, E., Haft, D.H., Horvath, P., Moineau, S., Mojica, F.J.M., Terns, R.M., Terns, M.P., White, M.F., Yakunin, A.F., Garrett, R.A., van der Oost, J., Backofen, R., Koonin, E.V., 2015. An updated evolutionary classification of CRISPR–Cas systems. Nature Reviews Microbiology 13, 722. https://doi.org/10.1038/nrmicro3569 https://www.nature.com/articles/nrmicro3569#supplementary-information

Makarova, K.S., Wolf, Y.I., Iranzo, J., Shmakov, S.A., Alkhnbashi, O.S., Brouns, S.J.J., Charpentier, E., Cheng, D., Haft, D.H., Horvath, P., Moineau, S., Mojica, F.J.M., Scott, D., Shah, S.A., Siksnys, V., Terns, M.P., Venclovas, Č., White, M.F., Yakunin, A.F., Yan, W., Zhang, F., Garrett, R.A., Backofen, R., van der Oost, J., Barrangou, R., Koonin, E.V., 2020. Evolutionary classification of CRISPR–Cas systems: a burst of class 2 and derived variants. Nature Reviews Microbiology 18, 67–83. https://doi.org/10.1038/s41579-019-0299-x

Makarova, K.S., Wolf, Y.I., Koonin, E.V., 2018. Classification and Nomenclature of CRISPR-Cas Systems: Where from Here? CRISPR J 1, 325–336. https://doi.org/10.1089/crispr.2018.0033

Maniv, I., Jiang, W., Bikard, D., Marraffini, L.A., 2016. Impact of Different Target Sequences on Type III CRISPR-Cas Immunity. J. Bacteriol. 198, 941. https://doi.org/10.1128/JB.00897-15

Marchesi, J.R., Ravel, J., 2015. The vocabulary of microbiome research: a proposal. Microbiome 3, 31. https://doi.org/10.1186/s40168-015-0094-5

Marcy, Y., Ouverney, C., Bik, E.M., Lösekann, T., Ivanova, N., Martin, H.G., Szeto, E., Platt, D., Hugenholtz, P., Relman, D.A., Quake, S.R., 2007. Dissecting biological "dark matter" with single-cell genetic analysis of rare and uncultivated TM7 microbes from the human mouth. Proceedings of the National Academy of Sciences 104, 11889–11894. https://doi.org/10.1073/pnas.0704662104

Maron, P.-A., Ranjard, L., Mougel, C., Lemanceau, P., 2007. Metaproteomics: A New Approach for Studying Functional Microbial Ecology. Microb Ecol 53, 486–493. https://doi.org/10.1007/s00248-006-9196-8

Marraffini, L.A., Sontheimer, E.J., 2010a. CRISPR interference: RNA-directed adaptive immunity in bacteria and archaea. Nature Reviews Genetics 11, 181. https://doi.org/10.1038/nrg2749

Marraffini, L.A., Sontheimer, E.J., 2010b. Self versus non-self discrimination during CRISPR RNA-directed immunity. Nature 463, 568–571. https://doi.org/10.1038/nature08703

Marshall, R., Maxwell, C.S., Collins, S.P., Jacobsen, T., Luo, M.L., Begemann, M.B., Gray, B.N., January, E., Singer, A., He, Y., Beisel, C.L., Noireaux, V., 2018. Rapid and Scalable Characterization of CRISPR Technologies Using an E. coli Cell-Free Transcription-Translation System. Molecular Cell 69, 146-157.e3. https://doi.org/10.1016/j.molcel.2017.12.007

Martin, W., Baross, J., Kelley, D., Russell, M.J., 2008. Hydrothermal vents and the origin of life. Nature Reviews Microbiology 6, 805–814. https://doi.org/10.1038/nrmicro1991

McGinn, J., Marraffini, L.A., 2016. CRISPR-Cas Systems Optimize Their Immune Response by Specifying the Site of Spacer Integration. Molecular Cell 64, 616–623. https://doi.org/10.1016/j.molcel.2016.08.038

McLean, J.S., Bor, B., Kerns, K.A., Liu, Q., To, T.T., Solden, L., Hendrickson, E.L., Wrighton, K., Shi, W., He, X., 2020. Acquisition and Adaptation of Ultra-small Parasitic Reduced Genome Bacteria to Mammalian Hosts. Cell Reports 32, 107939. https://doi.org/10.1016/j.celrep.2020.107939

Méheust, R., Burstein, D., Castelle, C.J., Banfield, J.F., 2019. The distinction of CPR bacteria from other bacteria based on protein family content. Nat Commun 10, 4173–4173. https://doi.org/10.1038/s41467-019-12171-z

Meier-Kolthoff, J.P., Auch, A.F., Klenk, H.-P., Göker, M., 2013. Genome sequence-based species delimitation with confidence intervals and improved distance functions. BMC Bioinformatics 14, 60. https://doi.org/10.1186/1471-2105-14-60

Meier-Kolthoff, J.P., Göker, M., 2017. VICTOR: genome-based phylogeny and classification of prokaryotic viruses. Bioinformatics 33, 3396–3404. https://doi.org/10.1093/bioinformatics/btx440

Men, A.E., Wilson, P., Siemering, K., Forrest, S., 2008. Sanger DNA sequencing. Next Generation Genome Sequencing: Towards Personalized Medicine 1–11.

Minh, B.Q., Schmidt, H.A., Chernomor, O., Schrempf, D., Woodhams, M.D., von Haeseler, A., Lanfear, R., 2020. IQ-TREE 2: New Models and Efficient Methods for Phylogenetic Inference in the Genomic Era. Molecular Biology and Evolution 37, 1530–1534. https://doi.org/10.1093/molbev/msaa015

Moissl, C., Rachel, R., Briegel, A., Engelhardt, H., Huber, R., 2005. The unique structure of archaeal 'hami', highly complex cell appendages with nano-grappling hooks. Molecular Microbiology 56, 361–370. https://doi.org/10.1111/j.1365-2958.2005.04294.x

Moissl, C., Rudolph, C., Huber, R., 2002. Natural Communities of Novel Archaea and Bacteria with a String-of-Pearls-Like Morphology: Molecular Analysis of the Bacterial Partners 68, 933–937. https://doi.org/10.1128/AEM.68.2.933-937.2002 %J Applied and Environmental Microbiology

Moissl, C., Rudolph, C., Rachel, R., Koch, M., Huber, R., 2003. In situ growth of the novel SM1 euryarchaeon from a string-of-pearls-like microbial community in its cold biotope, its physical separation and insights into its structure and physiology. Archives of Microbiology 180, 211–217. https://doi.org/10.1007/s00203-003-0580-1

Mojica, F.J., Díez-Villaseñor, C., Soria, E., Juez, G., 2000. Biological significance of a family of regularly spaced repeats in the genomes of Archaea, Bacteria and mitochondria. Molecular microbiology 36, 244–246.

Mojica, F.J.M., Díez-Villaseñor, C., García-Martínez, J., Soria, E., 2005. Intervening Sequences of Regularly Spaced Prokaryotic Repeats Derive from Foreign Genetic Elements. Journal of Molecular Evolution 60, 174–182. https://doi.org/10.1007/s00239-004-0046-3

Moller, A.G., Liang, C., 2017. MetaCRAST: reference-guided extraction of CRISPR spacers from unassembled metagenomes. PeerJ 5:e3788, e3788. https://doi.org/10.7717/peerj.3788

Momper, L., Jungbluth, S.P., Lee, M.D., Amend, J.P., 2017. Energy and carbon metabolisms in a deep terrestrial subsurface fluid microbial community. The ISME Journal 11, 2319–2333. https://doi.org/10.1038/ismej.2017.94

Moran, N.A., 2007. Symbiosis as an adaptive process and source of phenotypic complexity. Proceedings of the National Academy of Sciences 104, 8627–8633. https://doi.org/10.1073/pnas.0611659104

Moraru, C., Varsani, A., Kropinski, A.M., 2020. VIRIDIC—A Novel Tool to Calculate the Intergenomic Similarities of Prokaryote-Infecting Viruses. Viruses 12. https://doi.org/10.3390/v12111268

Mount, D.W., 2004. Bioinformatics: Sequence and Genome Analysis, Cold Spring Harbor Laboratory Series. Cold Spring Harbor Laboratory Press.

Muñoz-López, M., García-Pérez, J.L., 2010. DNA Transposons: Nature and Applications in Genomics. Curr Genomics 11, 115–128. https://doi.org/10.2174/138920210790886871

Munson-McGee, J.H., Field, E.K., Bateson, M., Rooney, C., Stepanauskas, R., Young, M.J. %J A.E.M., 2015. Nanoarchaeota, their Sulfolobales host, and Nanoarchaeota virus distribution across Yellowstone National Park hot springs. Applied and Environmental Microbiology 81, 7860–7868.

Mushegian, A.R., 2020. Are There 10(31) Virus Particles on Earth, or More, or Fewer? J Bacteriol 202, e00052-20. https://doi.org/10.1128/JB.00052-20

N

Navgire, G.S., Goel, N., Sawhney, G., Sharma, M., Kaushik, P., Mohanta, Y.K., Mohanta, T.K., Al-Harrasi, A., 2022. Analysis and Interpretation of metagenomics data: an approach. Biological Procedures Online 24, 18. https://doi.org/10.1186/s12575-022-00179-7

Navin, N., Kendall, J., Troge, J., Andrews, P., Rodgers, L., McIndoo, J., Cook, K., Stepansky, A., Levy, D., Esposito, D., Muthuswamy, L., Krasnitz, A., McCombie, W.R., Hicks, J., Wigler, M., 2011. Tumour evolution inferred by single-cell sequencing. Nature 472, 90–94. https://doi.org/10.1038/nature09807

Nawy, T., 2014. Single-cell sequencing. Nat Methods 11, 18–18. https://doi.org/10.1038/nmeth.2771

Nayfach, S., Camargo, A.P., Schulz, F., Eloe-Fadrosh, E., Roux, S., Kyrpides, N.C., 2021. CheckV assesses the quality and completeness of metagenome-assembled viral genomes. Nature Biotechnology 39, 578–585. https://doi.org/10.1038/s41587-020-00774-7

Neidhardt, F.C., Neidhardt, F.C.N., Ingraham, J.L., Schaechter, M., 1990. Physiology of the Bacterial Cell: A Molecular Approach. Sinauer Associates.

Nelson, D.L., Nelson, R.D., Cox, M.M., 2004. Lehninger Principles of Biochemistry, Fourth Edition + Lecture Notebook. W.H. Freeman.

Nicholson, J.K., 2006. Global systems biology, personalized medicine and molecular epidemiology. Molecular Systems Biology 2, 52. https://doi.org/10.1038/msb4100095

Nicholson, J.K., Lindon, J.C., Holmes, E., 1999. "Metabonomics": understanding the metabolic responses of living systems to pathophysiological stimuli via multivariate statistical analysis of biological NMR spectroscopic data. Xenobiotica 29, 1181–1189. https://doi.org/10.1080/004982599238047

Nishimura, Y., Yoshida, T., Kuronishi, M., Uehara, H., Ogata, H., Goto, S., 2017. ViPTree: the viral proteomic tree server. Bioinformatics 33, 2379–2380. https://doi.org/10.1093/bioinformatics/btx157

Nurk, S., Meleshko, D., Korobeynikov, A., Pevzner, P.A., 2017. metaSPAdes: a new versatile metagenomic assembler. Genome Research 27, 824–834. https://doi.org/10.1101/gr.213959.116

O

Oberortner, E., Cheng, J.-F., Hillson, N.J., Deutsch, S., 2017. Streamlining the Design-to-Build Transition with Build-Optimization Software Tools. ACS Synth. Biol. 6, 485–496. https://doi.org/10.1021/acssynbio.6b00200

Offre, P., Spang, A., Schleper, C., 2013. Archaea in Biogeochemical Cycles. Annual Review of Microbiology 67, 437–457. https://doi.org/10.1146/annurev-micro-092412-155614

Ondov, B.D., Bergman, N.H., Phillippy, A.M., 2011. Interactive metagenomic visualization in a Web browser. BMC Bioinformatics 12, 385. https://doi.org/10.1186/1471-2105-12-385

O'Neil, D., Glowatz, H., Schlumpberger, M., 2013. Ribosomal RNA Depletion for Efficient Use of RNA-Seq Capacity. Current Protocols in Molecular Biology 103, 4.19.1-4.19.8. https://doi.org/10.1002/0471142727.mb0419s103

Oren, A., 2021. Nomenclature of prokaryotic 'Candidatus' taxa: establishing order in the current chaos. New Microbes New Infect 44, 100932. https://doi.org/10.1016/j.nmni.2021.100932

Ortiz-Alvarez, R., Casamayor, E.O., 2016. High occurrence of Pacearchaeota and Woesearchaeota (Archaea superphylum DPANN) in the surface waters of oligotrophic high-altitude lakes. Environmental Microbiology Reports 8, 210–217. https://doi.org/10.1111/1758-2229.12370

P

Pace, N.R., 2009. Mapping the Tree of Life: Progress and Prospects. Microbiology and Molecular Biology Reviews 73, 565–576. https://doi.org/10.1128/mmbr.00033-09

Pace, N.R., 2006. Time for a change. Nature 441, 289–289. https://doi.org/10.1038/441289a

Paper, W., Jahn, U., Hohn, M.J., Kronner, M., Näther, D.J., Burghardt, T., Rachel, R., Stetter, K.O., Huber, H., 2007. Ignicoccus hospitalis sp. nov., the host of 'Nanoarchaeum equitans.' International Journal of Systematic and Evolutionary Microbiology 57, 803–808. https://doi.org/10.1099/ijs.0.64721-0

Parks, D.H., Chuvochina, M., Chaumeil, P.-A., Rinke, C., Mussig, A.J., Hugenholtz, P., 2020. A complete domain-to-species taxonomy for Bacteria and Archaea. Nature Biotechnology 38, 1079–1086. https://doi.org/10.1038/s41587-020-0501-8

Parks, D.H., Chuvochina, M., Waite, D.W., Rinke, C., Skarshewski, A., Chaumeil, P.-A., Hugenholtz, P., 2018. A standardized bacterial taxonomy based on genome phylogeny substantially revises the tree of life. Nature Biotechnology 36, 996–1004. https://doi.org/10.1038/nbt.4229

Pawluk, A., Davidson, A.R., Maxwell, K.L., 2018. Anti-CRISPR: discovery, mechanism and function. Nat Rev Microbiol 16, 12–17. https://doi.org/10.1038/nrmicro.2017.120

Pei, A.Y., Oberdorf, W.E., Nossa, C.W., Agarwal, A., Chokshi, P., Gerz, E.A., Jin, Z., Lee, P., Yang, L., Poles, M., Brown, S.M., Sotero, S., DeSantis, T., Brodie, E., Nelson, K., Pei, Z., 2010. Diversity of 16S rRNA Genes within Individual Prokaryotic Genomes. Applied and Environmental Microbiology 76, 3886–3897. https://doi.org/10.1128/AEM.02953-09

Perras, A.K., Wanner, G., Klingl, A., Mora, M., Auerbach, A.K., Heinz, V., Probst, A.J., Huber, H., Rachel, R., Meck, S., Moissl-Eichinger, C., 2014. Grappling archaea: ultrastructural analyses of an uncultivated, cold-loving archaeon, and its biofilm. Frontiers in Microbiology 5.

Philippe, H., Delsuc, F., Brinkmann, H., Lartillot, N., 2005. Phylogenomics. Annual Review of Ecology, Evolution, and Systematics 36, 541–562. https://doi.org/10.1146/annurev.ecolsys.35.112202.130205

Phipps, B.M., Hoffmann, A., Stetter, K.O., Baumeister, W., 1991. A novel ATPase complex selectively accumulated upon heat shock is a major cellular component of thermophilic archaebacteria. The EMBO Journal 10, 1711–1722. https://doi.org/10.1002/j.1460-2075.1991.tb07695.x

Phipps, B.M., Typke, D., Hegerl, R., Volker, S., Hoffmann, A., Stetter, K.O., Baumeister, W., 1993. Structure of a molecular chaperone from a thermophilic archaebacterium. Nature 361, 475–477. https://doi.org/10.1038/361475a0

Podar, M., Anderson, I., Makarova, K.S., Elkins, J.G., Ivanova, N., Wall, M.A., Lykidis, A., Mavromatis, K., Sun, H., Hudson, M.E., Chen, W., Deciu, C., Hutchison, D., Eads, J.R., Anderson, A., Fernandes, F., Szeto, E., Lapidus, A., Kyrpides, N.C., Saier, M.H., Richardson, P.M., Rachel, R., Huber, H., Eisen, J.A., Koonin, E.V., Keller, M., Stetter, K.O., 2008. A genomic analysis of the archaeal system Ignicoccus hospitalis-

Nanoarchaeum equitans. Genome Biol 9, R158. https://doi.org/10.1186/gb-2008-9-11-r158

Podar, M., Makarova, K.S., Graham, D.E., Wolf, Y.I., Koonin, E.V., Reysenbach, A.-L., 2013. Insights into archaeal evolution and symbiosis from the genomes of a nanoarchaeon and its inferred crenarchaeal host from Obsidian Pool, Yellowstone National Park. Biology Direct 8, 9. https://doi.org/10.1186/1745-6150-8-9

Posit team, 2022. Rstudio: Integrated Development Environment for R.

Probst, A.J., Castelle, C.J., Singh, A., Brown, C.T., Anantharaman, K., Sharon, I., Hug, L.A., Burstein, D., Emerson, J.B., Thomas, B.C., Banfield, J.F., 2017. Genomic resolution of a cold subsurface aquifer community provides metabolic insights for novel microbes adapted to high CO2 concentrations 19, 459–474. https://doi.org/doi:10.1111/1462-2920.13362

Probst, A.J., Holman, H.-Y.N., DeSantis, T.Z., Andersen, G.L., Birarda, G., Bechtel, H.A., Piceno, Y.M., Sonnleitner, M., Venkateswaran, K., Moissl-Eichinger, C., 2013. Tackling the minority: sulfate-reducing bacteria in an archaea-dominated subsurface biofilm. The ISME Journal 7, 635–651. https://doi.org/10.1038/ismej.2012.133

Probst, A.J., Ladd, B., Jarett, J.K., Geller-McGrath, D.E., Sieber, C.M.K., Emerson, J.B., Anantharaman, K., Thomas, B.C., Malmstrom, R.R., Stieglmeier, M., Klingl, A., Woyke, T., Ryan, M.C., Banfield, J.F., 2018. Differential depth distribution of microbial function and putative symbionts through sediment-hosted aquifers in the deep terrestrial subsurface. Nature Microbiology 3, 328–336. https://doi.org/10.1038/s41564-017-0098-y

Probst, A.J., Moissl-Eichinger, C., 2015. “Altiarchaeales”: uncultivated archaea from the subsurface. Life (Basel) 5, 1381–95. https://doi.org/10.3390/life5021381

Probst, A.J., Weinmaier, T., Raymann, K., Perras, A., Emerson, J.B., Rattei, T., Wanner, G., Klingl, A., Berg, I.A., Yoshinaga, M., Viehweger, B., Hinrichs, K.-U., Thomas, B.C., Meck, S., Auerbach, A.K., Heise, M., Schintlmeister, A., Schmid, M., Wagner, M., Gribaldo, S., Banfield, J.F., Moissl-Eichinger, C., 2014. Biology of a widespread uncultivated archaeon that contributes to carbon fixation in the subsurface. Nature Communications 5, 5497. https://doi.org/10.1038/ncomms6497 https://www.nature.com/articles/ncomms6497#supplementary-information

Pyenson, N.C., Gayvert, K., Varble, A., Elemento, O., Marraffini, L.A., 2017. Broad Targeting Specificity during Bacterial Type III CRISPR-Cas Immunity Constrains Viral Escape. Cell Host & Microbe 22, 343-353.e3. https://doi.org/10.1016/j.chom.2017.07.016

Q

Quackenbush, J., 2002. Microarray data normalization and transformation. Nature Genetics 32, 496–501. https://doi.org/10.1038/ng1032

Quinlan, A.R., Hall, I.M., 2010. BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics 26, 841–842. https://doi.org/10.1093/bioinformatics/btq033

R

Rahlff, J., Turzynski, V., Esser, S.P., Monsees, I., Bornemann, T.L.V., Figueroa-Gonzalez, P.A., Schulz, F., Woyke, T., Klingl, A., Moraru, C., Probst, A.J., 2021. Lytic archaeal viruses

infect abundant primary producers in Earth's crust. Nature Communications 12, 4642. https://doi.org/10.1038/s41467-021-24803-4

Ran, F.A., 2014. CRISPR/Cas9: tools and applications for eukaryotic genome editing. NABC.

Rath, D., Amlinger, L., Rath, A., Lundgren, M., 2015. The CRISPR-Cas immune system: Biology, mechanisms and applications. Biochimie 117, 119–128. https://doi.org/10.1016/j.biochi.2015.03.025

Reeks, J., Naismith, J.H., White, M.F., 2013. CRISPR interference: a structural perspective. Biochemical Journal 453, 155–166. https://doi.org/10.1042/BJ20130316

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., Dodsworth, J.A., Hedlund, B.P., Tsiamis, G., Sievert, S.M., Liu, W.-T., Eisen, J.A., Hallam, S.J., Kyrpides, N.C., Stepanauskas, R., Rubin, E.M., Hugenholtz, P., Woyke, T., 2013. Insights into the phylogeny and coding potential of microbial dark matter. Nature 499, 431–437. https://doi.org/10.1038/nature12352

Roux, S., Enault, F., Hurwitz, B.L., Sullivan, M.B., 2015. VirSorter: mining viral signal from microbial genomic data. PeerJ 3:e985, e985. https://doi.org/10.7717/peerj.985

Rudolph, C., Moissl, C., Henneberger, R., Huber, R., 2004. Ecology and microbial structures of archaeal/bacterial strings-of-pearls communities and archaeal relatives thriving in cold sulfidic springs. FEMS Microbiology Ecology 50, 1–11. https://doi.org/10.1016/j.femsec.2004.05.006

Rudolph, C., Wanner, G., Huber, R., 2001. Natural Communities of Novel Archaea and Bacteria Growing in Cold Sulfurous Springs with a String-of-Pearls-Like Morphology 67, 2336–2344. https://doi.org/10.1128/AEM.67.5.2336-2344.2001 %J Applied and Environmental Microbiology

S

Saier, M.H., Jr, Reddy, V.S., Tsu, B.V., Ahmed, M.S., Li, C., Moreno-Hagelsieb, G., 2016. The Transporter Classification Database (TCDB): recent advances. Nucleic Acids Res 44, D372–D379. https://doi.org/10.1093/nar/gkv1103

Sakai, H.D., Nur, N., Kato, S., Yuki, M., Shimizu, M., Itoh, T., Ohkuma, M., Suwanto, A., Kurosawa, N., 2022. Insight into the symbiotic lifestyle of DPANN archaea revealed by cultivation and genome analyses. Proceedings of the National Academy of Sciences 119, e2115449119. https://doi.org/10.1073/pnas.2115449119

Sanger, F., Air, G.M., Barrell, B.G., Brown, N.L., Coulson, A.R., Fiddes, J.C., Hutchison, C.A., Slocombe, P.M., Smith, M., 1977. Nucleotide sequence of bacteriophage ϕX174 DNA. Nature 265, 687–695. https://doi.org/10.1038/265687a0

Schneider, T.D., Stephens, R.M., 1990. Sequence logos: a new way to display consensus sequences. Nucleic Acids Res 18, 6097–6100. https://doi.org/10.1093/nar/18.20.6097

Schwank, K., Bornemann, T.L.V., Dombrowski, N., Spang, A., Banfield, J.F., Probst, A.J., 2019. An archaeal symbiont-host association from the deep terrestrial subsurface. The ISME Journal 13, 2135–2139. https://doi.org/10.1038/s41396-019-0421-0

Seidl, M.F., Thomma, B.P.H.J., 2017. Transposable Elements Direct The Coevolution between Plants and Microbes. Trends in Genetics, Transposable Elements 33, 842–851. https://doi.org/10.1016/j.tig.2017.07.003

Shakya, M., Lo, C.-C., Chain, P.S.G., 2019. Advances and Challenges in Metatranscriptomic Analysis. Frontiers in Genetics 10.

Shannon, P., Markiel, A., Ozier, O., Baliga, N.S., Wang, J.T., Ramage, D., Amin, N., Schwikowski, B., Ideker, T., 2003. Cytoscape: A Software Environment for Integrated Models of Biomolecular Interaction Networks. Genome Research 13, 2498–2504. https://doi.org/10.1101/gr.1239303

Shapiro, E., Biezuner, T., Linnarsson, S., 2013. Single-cell sequencing-based technologies will revolutionize whole-organism science. Nat Rev Genet 14, 618–630. https://doi.org/10.1038/nrg3542

Sharrar, A.M., Flood, B.E., Bailey, J.V., Jones, D.S., Biddanda, B.A., Ruberg, S.A., Marcus, D.N., Dick, G.J., 2017. Novel Large Sulfur Bacteria in the Metagenomes of Groundwater-Fed Chemosynthetic Microbial Mats in the Lake Huron Basin. Frontiers in Microbiology 8, 791. https://doi.org/10.3389/fmicb.2017.00791

Shin, J., Noireaux, V., 2012. An E. coli Cell-Free Expression Toolbox: Application to Synthetic Gene Circuits and Artificial Cells. ACS Synth. Biol. 1, 29–41. https://doi.org/10.1021/sb200016s

Silas, S., Lucas-Elio, P., Jackson, S.A., Aroca-Crevillén, A., Hansen, L.L., Fineran, P.C., Fire, A.Z., Sánchez-Amat, A., 2017. Type III CRISPR-Cas systems can provide redundancy to counteract viral escape from type I systems. eLife 6, e27601. https://doi.org/10.7554/eLife.27601

Sleytr, U.B., Messner, P., Pum, D., Sára, M., 1993. Crystalline bacterial cell surface layers. Molecular Microbiology 10, 911–916. https://doi.org/10.1111/j.1365-2958.1993.tb00962.x

Solden, L., Lloyd, K., Wrighton, K., 2016. The bright side of microbial dark matter: lessons learned from the uncultivated majority. Current Opinion in Microbiology 31, 217–226. https://doi.org/10.1016/j.mib.2016.04.020

Spang, A., Caceres, E.F., Ettema, T.J.G., 2017. Genomic exploration of the diversity, ecology, and evolution of the archaeal domain of life. Science 357, eaaf3883. https://doi.org/10.1126/science.aaf3883

Spiess, F.N., Macdonald, K.C., Atwater, T., Ballard, R., Carranza, A., Cordoba, D., Cox, C., Garcia, V.M.D., Francheteau, J., Guerrero, J., Hawkins, J., Haymon, R., Hessler, R., Juteau, T., Kastner, M., Larson, R., Luyendyk, B., Macdougall, J.D., Miller, S., Normark, W., Orcutt, J., Rangin, C., 1980. East Pacific Rise: Hot Springs and Geophysical Experiments. Science 207, 1421–1433. https://doi.org/10.1126/science.207.4438.1421

Stachler, A.-E., Turgeman-Grott, I., Shtifman-Segal, E., Allers, T., Marchfelder, A., Gophna, U., 2017. High tolerance to self-targeting of the genome by the endogenous CRISPR-Cas system in an archaeon. Nucleic Acids Research 45, 5208–5216. https://doi.org/10.1093/nar/gkx150

Steffensen, J.L., Dufault-Thompson, K., Zhang, Y., 2016. PSAMM: A Portable System for the Analysis of Metabolic Models. PLoS Comput Biol 12, e1004732–e1004732. https://doi.org/10.1371/journal.pcbi.1004732

Stern, A., Keren, L., Wurtzel, O., Amitai, G., Sorek, R., 2010. Self-targeting by CRISPR: gene regulation or autoimmunity? Trends in Genetics 26, 335–340. https://doi.org/10.1016/j.tig.2010.05.008

Swarts, D.C., Mosterd, C., van Passel, M.W.J., Brouns, S.J.J., 2012. CRISPR Interference Directs Strand Specific Spacer Acquisition. PLOS ONE 7, e35888. https://doi.org/10.1371/journal.pone.0035888

T

Team, R.C., 2013. R: A language and environment for statistical computing. Vienna, Austria.

Terns, M.P., Terns, R.M., 2011. CRISPR-based adaptive immune systems. Current Opinion in Microbiology, Ecology and industrial microbiology / Special section: Archaea 14, 321–327. https://doi.org/10.1016/j.mib.2011.03.005

Teske, A., Hinrichs, K.-U., Edgcomb, V., de Vera Gomez, A., Kysela, D., Sylva, S.P., Sogin, M.L., Jannasch, H.W., 2002. Microbial Diversity of Hydrothermal Sediments in the Guaymas Basin: Evidence for Anaerobic Methanotrophic Communities. Applied and Environmental Microbiology 68, 1994–2007. https://doi.org/10.1128/AEM.68.4.1994-2007.2002

Teske, A., Wegener, G., Chanton, J.P., White, D., MacGregor, B., Hoer, D., de Beer, D., Zhuang, G., Saxton, M.A., Joye, S.B., Lizarralde, D., Soule, S.A., Ruff, S.E., 2021. Microbial Communities Under Distinct Thermal and Geochemical Regimes in Axial and Off-Axis Sediments of Guaymas Basin. Frontiers in Microbiology 12.

Tivey, M., 2014. Black and White smokers. Encyclopedia of Marine Geosciences. Springer, Dordrecht.

Tringe, S.G., von Mering, C., Kobayashi, A., Salamov, A.A., Chen, K., Chang, H.W., Podar, M., Short, J.M., Mathur, E.J., Detter, J.C., Bork, P., Hugenholtz, P., Rubin, E.M., 2005. Comparative Metagenomics of Microbial Communities. Science 308, 554–557. https://doi.org/10.1126/science.1107851

Turgeman-Grott, I., Joseph, S., Marton, S., Eizenshtein, K., Naor, A., Soucy, S.M., Stachler, A.-E., Shalev, Y., Zarkor, M., Reshef, L., Altman-Price, N., Marchfelder, A., Gophna, U., 2019. Pervasive acquisition of CRISPR memory driven by inter-species mating of archaea can limit gene transfer and influence speciation. Nature Microbiology 4, 177–186. https://doi.org/10.1038/s41564-018-0302-8

Turzynski, V., Griesdorn, L., Moraru, C., Soares, A., Simon, S.A., Stach, T.L., Rahlff, J., Esser, S.P., Probst, A.J., 2023. Virus-host dynamics in archaeal groundwater biofilms and the associated bacterial community composition. bioRxiv 2023.02.03.526798. https://doi.org/10.1101/2023.02.03.526798

Tyson, G.W., Banfield, J.F., 2008. Rapidly evolving CRISPRs implicated in acquired resistance of microorganisms to viruses. Environmental Microbiology 10, 200–207. https://doi.org/10.1111/j.1462-2920.2007.01444.x

V

Veith, A., Klingl, A., Zolghadr, B., Lauber, K., Mentele, R., Lottspeich, F., Rachel, R., Albers, S.-V., Kletzin, A., 2009. Acidianus, Sulfolobus and Metallosphaera surface layers: structure, composition and gene expression. Molecular Microbiology 73, 58–72. https://doi.org/10.1111/j.1365-2958.2009.06746.x

Vink, J.N.A., Baijens, J.H.L., Brouns, S.J.J., 2021. PAM-repeat associations and spacer selection preferences in single and co-occurring CRISPR-Cas systems. Genome Biology 22, 281. https://doi.org/10.1186/s13059-021-02495-9

W

Wang, H.-C., Minh, B.Q., Susko, E., Roger, A.J., 2017. Modeling Site Heterogeneity with Posterior Mean Site Frequency Profiles Accelerates Accurate Phylogenomic Estimation. Systematic Biology 67, 216–235. https://doi.org/10.1093/sysbio/syx068

Waters, E., Hohn, M.J., Ahel, I., Graham, D.E., Adams, M.D., Barnstead, M., Beeson, K.Y., Bibbs, L., Bolanos, R., Keller, M., Kretz, K., Lin, X., Mathur, E., Ni, J., Podar, M., Richardson, T., Sutton, G.G., Simon, M., Söll, D., Stetter, K.O., Short, J.M., Noordewier, M., 2003. The genome of Nanoarchaeum equitans: Insights into early archaeal evolution and derived parasitism. Proc Natl Acad Sci USA 100, 12984. https://doi.org/10.1073/pnas.1735403100

Westra, E.R., Buckling, A., Fineran, P.C., 2014. CRISPR–Cas systems: beyond adaptive immunity. Nature Reviews Microbiology 12, 317. https://doi.org/10.1038/nrmicro3241 https://www.nature.com/articles/nrmicro3241#supplementary-information

Westra, E.R., Swarts, D.C., Staals, R.H.J., Jore, M.M., Brouns, S.J.J., van der Oost, J., 2012. The CRISPRs, They Are A-Changin': How Prokaryotes Generate Adaptive Immunity. Annu. Rev. Genet. 46, 311–339. https://doi.org/10.1146/annurev-genet-110711-155447

Whittaker, R.H., 1969. New Concepts of Kingdoms of Organisms. Science 163, 150–160. https://doi.org/10.1126/science.163.3863.150

Wickham, H., 2016. ggplot2: elegant graphics for data analysis. Springer.

Wilmes, P., Bond, P.L., 2006. Metaproteomics: studying functional gene expression in microbial ecosystems. Trends in Microbiology 14, 92–97. https://doi.org/10.1016/j.tim.2005.12.006

Wilmes, P., Bond, P.L., 2004. The application of two-dimensional polyacrylamide gel electrophoresis and downstream analyses to a mixed community of prokaryotic microorganisms. Environmental Microbiology 6, 911–920. https://doi.org/10.1111/j.1462-2920.2004.00687.x

Wilson, G.G., 1991. Organization of restriction-modification systems. Nucleic Acids Research 19, 2539–2566. https://doi.org/10.1093/nar/19.10.2539

Wimmer, F., Beisel, C.L., 2020. CRISPR-Cas Systems and the Paradox of Self-Targeting Spacers. Frontiers in Microbiology 10, 3078. https://doi.org/10.3389/fmicb.2019.03078

Wimmer, F., Mougiakos, I., Englert, F., Beisel, C.L., 2022. Rapid cell-free characterization of multi-subunit CRISPR effectors and transposons. Molecular Cell 82, 1210-1224.e6. https://doi.org/10.1016/j.molcel.2022.01.026

Woese, C.R., Fox, G.E., 1977. Phylogenetic structure of the prokaryotic domain: The primary kingdoms. Proc Natl Acad Sci USA 74, 5088. https://doi.org/10.1073/pnas.74.11.5088

Woese, C.R., Kandler, O., Wheelis, M.L., 1990. Towards a natural system of organisms: proposal for the domains Archaea, Bacteria, and Eucarya. Proceedings of the National Academy of Sciences 87, 4576–4579. https://doi.org/10.1073/pnas.87.12.4576

Woese, C.R., Magrum, L.J., Fox, G.E., 1978. Archaebacteria. Journal of Molecular Evolution 11, 245–252. https://doi.org/10.1007/BF01734485

Wood, H.G., 1991. Life with CO or CO2 and H2 as a source of carbon and energy. The FASEB Journal 5, 156–163. https://doi.org/10.1096/fasebj.5.2.1900793

Wurch, L., Giannone, R.J., Belisle, B.S., Swift, C., Utturkar, S., Hettich, R.L., Reysenbach, A.-L., Podar, M., 2016. Genomics-informed isolation and characterization of a symbiotic Nanoarchaeota system from a terrestrial geothermal environment. Nature Communications 7, 12115. https://doi.org/10.1038/ncomms12115

X

Xie, Z., Tang, H., 2017. ISEScan: automated identification of insertion sequence elements in prokaryotic genomes. Bioinformatics 33, 3340–3347. https://doi.org/10.1093/bioinformatics/btx433

Xu, B., Liu, L., Song, G., 2022. Functions and Regulation of Translation Elongation Factors. Frontiers in Molecular Biosciences 8.

Yan, W.X., Hunnewell, P., Alfonse, L.E., Carte, J.M., Keston-Smith, E., Sothiselvam, S., Garrity, A.J., Chong, S., Makarova, K.S., Koonin, E.V., Cheng, D.R., Scott, D.A., 2019. Functionally diverse type V CRISPR-Cas systems. Science 363, 88–91. https://doi.org/10.1126/science.aav7271

Y

Yang, C., Chowdhury, D., Zhang, Z., Cheung, W.K., Lu, A., Bian, Z., Zhang, L., 2021. A review of computational tools for generating metagenome-assembled genomes from metagenomic sequencing data. Computational and Structural Biotechnology Journal 19, 6301–6314. https://doi.org/10.1016/j.csbj.2021.11.028

Yosef, I., Goren, M.G., Qimron, U., 2012. Proteins and DNA elements essential for the CRISPR adaptation process in Escherichia coli. Nucleic Acids Research 40, 5569–5576. https://doi.org/10.1093/nar/gks216

Youssef, N.H., Rinke, C., Stepanauskas, R., Farag, I., Woyke, T., Elshahed, M.S., 2015. Insights into the metabolism, lifestyle and putative evolutionary history of the novel archaeal phylum 'Diapherotrites.' ISME J 9, 447–460. https://doi.org/10.1038/ismej.2014.141

Z

Zhang, Y., Sievert, S., 2014. Pan-genome analyses identify lineage- and niche-specific markers of evolution and adaptation in Epsilonproteobacteria. Frontiers in Microbiology 5, 110. https://doi.org/10.3389/fmicb.2014.00110

Zhou, J., He, Z., Yang, Y., Deng, Y., Tringe, S.G., Alvarez-Cohen, L., 2015. High-Throughput Metagenomic Technologies for Complex Microbial Community Analysis: Open and Closed Formats. mBio 6, e02288-14. https://doi.org/10.1128/mBio.02288-14

Zong, C., Lu, S., Chapman, A.R., Xie, X.S., 2012. Genome-Wide Detection of Single-Nucleotide and Copy-Number Variations of a Single Human Cell. Science 338, 1622–1626. https://doi.org/10.1126/science.1229164

I. Content of supporting CD

The supporting CD contains the supplementary data files for the presented publications in section 3.2 and 3.3. The folders are named according to the following scheme:

1. Publication_1_section_3.2
2. Publication_2_section_3.3

The folders are divided into different subfolders, which include the main figures of the publications in pdf format, the supplementary tables and for Publication one the Extended Data Figures. The not yet published genomes and fluorescence *in-situ* hybridization images used within Publication 1 will be, due to their size, provided upon request but are also accessible using the DOI's within the publication. The read files of Publication 2 are also provided on request and currently in submission at NCBI. Additionally for the scope of this thesis the data can also be found under the following server path (/data4/IMS_Transcriptome/raw.d).

II. Eidesstattliche Erklärung

Ich bestätige hiermit an Eides statt, dass die vorliegende Arbeit ohne unzulässige Hilfe Dritter und ohne Benutzung anderer als der angegebenen Hilfsmittel angefertigt habe. Die aus anderen Quellen direkt oder indirekt übernommenen Daten und Konzepte sind unter Angabe des Literaturzitats gekennzeichnet.

Essen, den 06.06.2023 _____

Sarah Eßer