nature microbiology

Article

https://doi.org/10.1038/s41564-023-01439-2

A predicted CRISPR-mediated symbiosis between uncultivated archaea

Received: 9 May 2023

Accepted: 23 June 2023

Published online: 27 July 2023

Check for updates

Sarah P. Esser[1,2,17], Janina Rahlff[2,14,17], Weishu Zhao[3,15], 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[2,16], 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] & Alexander J. Probst[1,2,12,13] ✉

CRISPR–Cas systems defend prokaryotic cells from invasive DNA of viruses, plasmids and other mobile genetic elements. Here, we show using metagenomics, metatranscriptomics and single-cell genomics that CRISPR systems of widespread, uncultivated archaea can also target chromosomal DNA of archaeal episymbionts of the DPANN superphylum. Using meta-omics datasets from Crystal Geyser and Horonobe Underground Research Laboratory, we find that CRISPR spacers of the hosts *Candidatus* Altiarchaeum crystalense and *Ca*. A. horonobense, respectively, match putative essential genes in their episymbionts' genomes of the genus *Ca*. Huberiarchaeum and that some of these spacers are expressed in situ. Metabolic interaction modelling also reveals complementation between host–episymbiont systems, on the basis of which we propose that episymbionts are either parasitic or mutualistic depending on the 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.

Clustered regularly interspaced short palindromic repeats associated systems (CRISPR–Cas) facilitate adaptive prokaryotic immunity via cleavage of mobile genetic elements (MGEs), for example, viruses and plasmids[1]. CRISPR loci consist of a series of direct repeat sequences interspaced by short variable fragments, that is, spacers, flanked by *cas* genes. Upon exposure to new 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 on the basis of matches to protospacers, corresponding nucleic acid regions in the MGE[2].

CRISPR systems exhibit remarkable diversity and functional plasticity including roles in non-defensive functions (reviewed by ref. 3). Six main types of CRISPR–Cas systems have been described, including different subtypes, for example, type I-A to I-F, depending on signature genes and their arrangements[4,5]. 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 protospacer. Rather than relying on a defined PAM for target recognition, other CRISPR systems (for example, type III) generally evaluate the extent of hybridization between the flanking portions of the crRNA (called protospacer-flanking sequence) and the target[6,7]. While CRISPR–Cas systems are widely distributed, they

A full list of affiliations appears at the end of the paper. ✉e-mail: alexander.probst@uni-due.de

are more common in archaea (in ~85% of genomes) than in bacteria (in ~40% of genomes; reviewed by ref. 5).

Branching from the archaeal tree of life, the DPANN superphylum including Diapherotrites, Parvarchaeota, Aenigmarchaeota, Nanoarchaeota and Nanohaloarchaeota and several other recently proposed phyla[8,9], comprises a vast collection of microorganisms remarkably small in size and enigmatic due to the scarcity of cultivated representatives[10,11]. Insights into the physiological characteristics of DPANN archaea arise primarily from detailed analyses of cocultivation with amenable microorganisms[12,13] and/or imaging of environmental samples[14]. 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[13,15–18] or even as intracellular symbionts[19]. The independent and autotrophic *Candidatus* Altiarchaeum sp. is host organism to another uncultivated DPANN archaeon, *Ca*. Huberiarchaeum crystalense[14,20].

Previous evidence suggested that certain DPANN archaea can fuse their cytoplasm with that of their hosts[14,15,21–23] and even exchange enzymes[21]. Cytoplasm fusion could in theory facilitate transfer of metabolites from the host to the symbiont consequently rendering such a symbiosis potentially parasitic. Hence, we investigated the symbiotic nature of the uncultivated DPANN *Ca*. Altiarchaeum and its uncultivated DPANN episymbiont *Ca*. Huberiarchaeum using meta-omics and metabolic modelling in two independent subsurface ecosystems (Fig. 1a). On the basis of the complex interaction of *Ca*. Altiarchaea and viruses in deep subsurface ecosystems, we examined the encoded CRISPR systems and analysed the targets of their respective spacer populations in two independent subsurface ecosystems (accessible through Crystal Geyser (CG, Utah, USA) and Horonobe Underground Research Laboratory (HURL, Hokkaido, Japan), where we identified both the host and the symbiont being associated on the basis of fluorescence in situ hybridization (FISH). Our findings demonstrate that a substantial portion of the host spacer population targets the genomes of the episymbionts having the same PAM sequence as the respectively targeted viruses in the ecosystems. In addition, the host CRISPR systems also target the host chromosome, on the basis of which genome-centric metabolic modelling predicted a mutualistic relationship between host and episymbiont as a function of metabolic complementation. Based on our results, we suggest that CRISPR–Cas systems play an integral role in mediating archaeal host–DPANN interactions.

Results

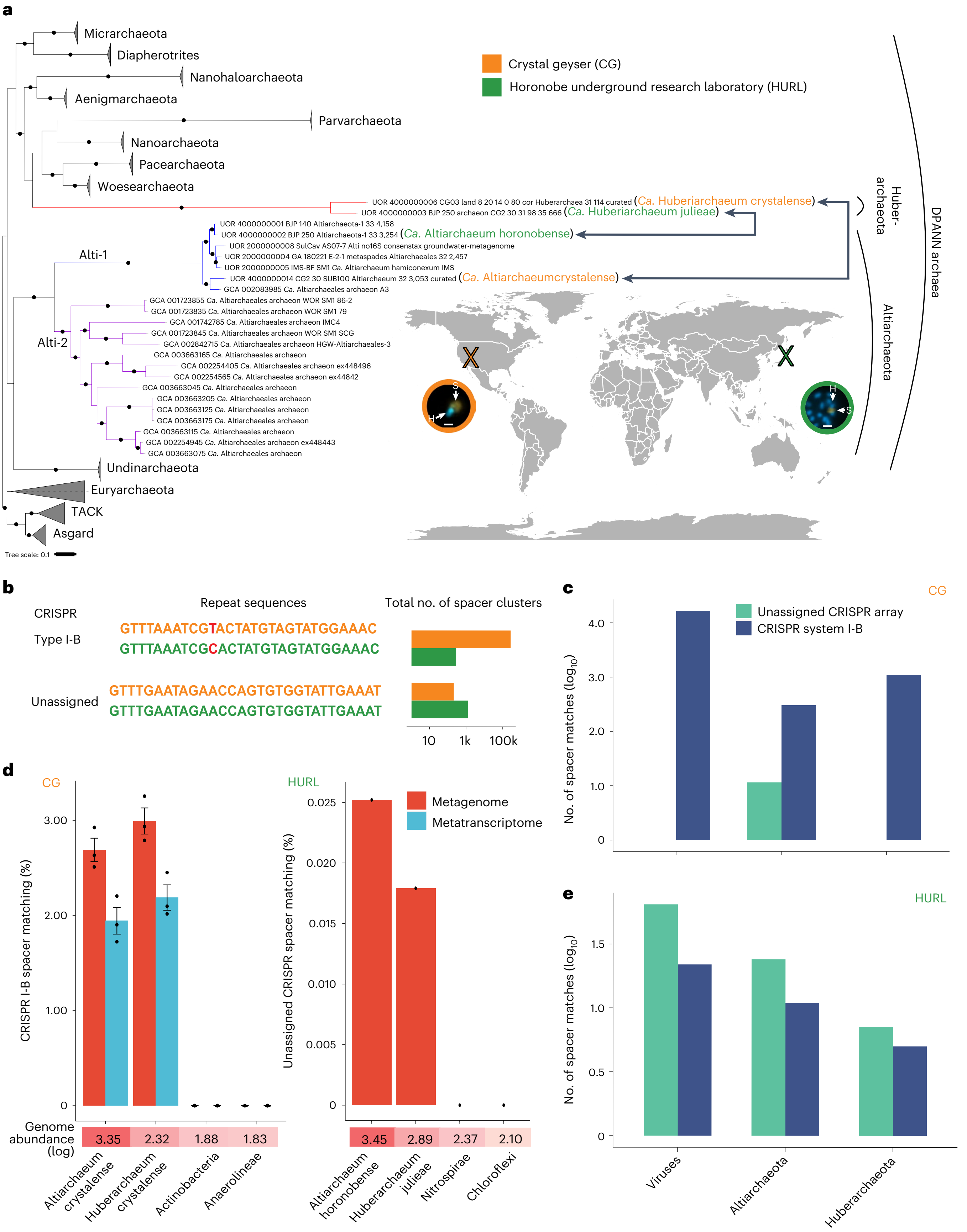
A CRISPR–Cas system targets archaeal episymbionts' genomes

Two subsurface ecosystems separated by 8,255 km (Fig. 1a) and derived from different geological formations[20,24], that is, a Wingate Sandstone-hosted aquifer of the Colorado Plateau at ~350 m depth (the CG ecosystem)[20,25–27] and a diatomaceous/siliceous mudstone-hosted aquifer of the HURL[24] at ~250 m depth, were dominated by two species of *Ca*. Altiarchaea (up to 24.5% and 51.6% of the community, respectively). We show their association with cells of *Ca*. Huberiarchaea, their DPANN episymbiont, using species-specific FISH (Fig. 1a). The *Ca*. Altiarchaea genomes retrieved from CG and HURL were shown to encode a CRISPR system type I-B and an abundant CRISPR array, which could not be assigned to a specific *cas* gene casette, but has been reported for other *Ca*. Altiarchaea species[25,28]. Confidence in assigning the CRISPR–Cas system to its correct metagenome-assembled genome (MAG; Altiarchaea genomes $n = 1$; Supplementary Tables 1 and 2) derives from the exceedingly high abundance of *Ca*. Altiarchaea genome fragments in the CG samples (Supplementary Fig. 1). In addition, within 219 single-cell amplified genomes (SAGs; Altiarchaea SAGs $n = 7$; Supplementary Table 3) from CG, only *Ca*. Altiarchaea bore the corresponding consensus direct repeat sequence (see Extended Data Fig. 1 for additional correlation-based evidence), which were remarkably well-conserved across ecosystems[28] (Fig. 1b).

Analyses of spacers from *Ca*. Altiarchaeum crystalense detected in 66 CG metagenomes over 6 years of surveillance (1.07 terabases (Tb) of sequencing data; Supplementary Table 1) revealed 297,531 distinct spacer clusters (Fig. 1b), indicative of a complex CRISPR spacer repertoire system for this organism (Supplementary Figs. 2 and 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. 1c, Supplementary Table 6, Extended Data Fig. 2 and Supplementary Figs. 4–6; details in Supplementary Results). The PAM sequence 5′-TTN-3′ was identified on viral targets matched by type I-B spacers (Supplementary Fig. 7). However, we were unable to experimentally confirm this PAM using an established PAM assay[29] or to assess GFP repression[30] using the 5′-TTN-3′ PAM in a cell-free transcription–translation (TXTL) system[29], probably due to the divergent settings (including temperature) of the host environment compared to those used in the established assay[29].

The finding that all virus-matched spacers detected in the exhaustive CG survey derived from the CRISPR system type I-B and the ubiquitous nature of this system in *Ca*. Altiarchaea worldwide[28] suggests that the type I-B system serves as a primary line of defence against viruses infecting these archaea. A substantial fraction of the spacers matched microbial genomes, including those of *Ca*. A. crystalense, that is, 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. 1c,d and Fig. 2a–d). The relative proportion of spacers matching the episymbiont was greater than that matching the host genome (Fig. 1c,d), indicative of biased acquisition, negative selection of self-targeting spacers, or a positive selection for spacers

Fig. 1 | Phylogenetic positioning of *Ca*. Altiarchaea and *Ca*. Huberarchaea, sampling locations, FISH analysis and CRISPR–Cas targets. a, Phylogenetic tree of archaea highlighting *Ca*. Altiarchaeum and *Ca*. Huberiarchaeum of the sampling locations CG (orange) and HURL (green). Fluorescence pictures 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*. Altiarchaea CRISPR systems, their associated conserved direct repeat sequences (with exception of a point mutation marked in red) and the number of spacer clusters (97% nucleotide identity) arising from the two sampling sites. k, 1,000. **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 16,561 of CRISPR type I-B system), 17 binned genomes of *Ca*. Altiarchaeum crystalense (total number of spacer matches: 115 of unassigned CRISPR system and 1,311 of CRISPR system type I-B) 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 (Supplementary Table 2). Spacers were derived from the complete 66-sample metagenomic dataset. **d**, Percentage of CRISPR system type 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 were available. The dataset of HURL is referring to one metagenome, as no other data were available. (Means and standard deviations: 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 type I-B), two binned genomes of *Ca*. Altiarchaeum horonobense (total number of spacer matches: 19 of unassigned CRISPR system and two of CRISPR system type I-B) and one binned genome of *Ca*. Huberiarchaeum julieae (total number of spacer matches: seven of unassigned CRISPR system and zero of CRISPR system type I-B) originating from the HURL site. Spacers were derived from one metagenomic dataset.


a
Micrarchaeota
Diapherotrites
Nanohaloarchaeota
Aenigmarchaeota
Parvarchaeota
Nanoarchaeota
Pacearchaeota
Woesearchaeota
UOR 4000000006 CG03 land 8 20 14 0 80 cor Huberarchaea 31 114 curated (Ca. Huberiarchaeum crystalense)
UOR 4000000003 BJP 250 archaeon CG2 30 31 98 35 666 (Ca. Huberiarchaeum julieae)
UOR 4000000001 BJP 140 Altiarchaeota-1 33 4,158
UOR 4000000002 BJP 250 Altiarchaeota-1 33 3,254 (Ca. Altiarchaeum horonobense)
UOR 2000000008 SulCav AS07-7 Alti no16S consenstax groundwater-metagenome
UOR 2000000004 GA 180221 E-2-1 metaspades Altiarchaeales 32 2,457
UOR 2000000005 IMS-BF SM1 Ca. Altiarchaeum hamiconexum IMS
UOR 4000000014 CG2 30 SUB100 Altiarchaeum 32 3,053 curated (Ca. Altiarchaeumcrystalense)
GCA 002083985 Ca. Altiarchaeales archaeon A3
Alti-1
GCA 001723855 Ca. Altiarchaeales archaeon WOR SM1 86-2
GCA 001723835 Ca. Altiarchaeales archaeon WOR SM1 79
GCA 001742785 Ca. Altiarchaeales archaeon IMC4
GCA 001723845 Ca. Altiarchaeales archaeon WOR SM1 SCG
GCA 002842715 Ca. Altiarchaeales archaeon HGW-Altiarchaeales-3
GCA 003663165 Ca. Altiarchaeales archaeon
GCA 002254405 Ca. Altiarchaeales archaeon ex448496
GCA 002254565 Ca. Altiarchaeales archaeon ex44842
GCA 003663045 Ca. Altiarchaeales archaeon
GCA 003663205 Ca. Altiarchaeales archaeon
GCA 003663125 Ca. Altiarchaeales archaeon
GCA 003663175 Ca. Altiarchaeales archaeon
GCA 003663115 Ca. Altiarchaeales archaeon
GCA 002254945 Ca. Altiarchaeales archaeon ex448443
GCA 003663075 Ca. Altiarchaeales archaeon
Alti-2
Undinarchaeota
Euryarchaeota
TACK
Asgard
Tree scale: 0.1
🟧 Crystal geyser (CG)
🟩 Horonobe underground research laboratory (HURL)
Huber-archaeota
Altiarchaeota
DPANN archaea
X
X
S
H
H
S
b
CRISPR
Type I-B
Unassigned
Repeat sequences
GTTTAAATCGTACTATGTAGTATGGAAAC
GTTTAAATCGCACTATGTAGTATGGAAAC
GTTTGAATAGAACCAGTGTGGTATTGAAAT
GTTTGAATAGAACCAGTGTGGTATTGAAAT
Total no. of spacer clusters
10
1k
100k
c
CG
Unassigned CRISPR array
CRISPR system I-B
No. of spacer matches (log10)
4.0
3.0
2.0
1.0
0
d
CG
CRISPR I-B spacer matching (%)
3.00
2.00
1.00
0
Genome abundance (log)
3.35
2.32
1.88
1.83
Altiarchaeum crystalense
Huberarchaeum crystalense
Actinobacteria
Anaerolineae
HURL
Unassigned CRISPR spacer matching (%)
0.025
0.020
0.015
0.010
0.005
0
Metagenome
Metatranscriptome
3.45
2.89
2.37
2.10
Altiarchaeum horonobense
Huberarchaeum julieae
Nitrospirae
Chloroflexi
e
HURL
No. of spacer matches (log10)
1.5
1.0
0.5
0
Viruses
Altiarchaeota
Huberarchaeota

from the episymbiont's genome. The positions of these spacers in the CRISPR type I-B array encoded in an altiarchaeal SAG suggest that these spacers prevailed in the system for extended periods (Fig. 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 one-sided signed rank test, target sites = 196, FDR-corrected $P < 0.05$), 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 paired one-sided signed rank test, target sites = 73, FDR-corrected $P < 0.05$; Supplementary Results and Extended Data Fig. 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 $P < 0.05$; details in Supplementary Results and Supplementary Table 5). 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 (Supplementary Fig. 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 (Supplementary Fig. 7). As shown for other microbial communities, self-targeting can result in cell suicide (reviewed in ref. 31) or transcriptional regulation[32] of genes influencing the fitness of the microbial population and can thus reduce the strain variation within an ecosystem[33]. However, lack of the PAM, the essential motif for successful targeting of DNA by CRISPR system type I (reviewed in ref. 34) 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. On the basis of 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[14,15,21–23]. This direct interaction of the cytoplasms of the host and the symbiont and a potentially predatory nature of the symbiont[14,20], probably underlie the evolution of a direct assault on the episymbiont's genome by the Altiarchaea CRISPR system (Fig. 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. 2c and Supplementary Table 5). Methyltransferases protect DNA against cleavage by restriction enzymes[35]. Inactivation of such a methyltransferase might increase vulnerability of the episymbiont towards enzymatic cleavage by the host.

CRISPR targeting in two independent host–episymbiont systems

Targeting of episymbiont's genomes by altiarchaeal 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 use CRISPR spacers of an unassigned array (that is, no *cas* genes in direct vicinity could be detected due to genome fragmentation but the direct repeat sequence is identical to CRISPR systems type III of other Altiarchaea[28]) to potentially ward off *Ca.* Huberiarchaeum julieae episymbionts and viral invaders (Fig. 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. 1d), their relative abundance in the metagenome was nearly twofold lower (Fig. 1e). The ecosystem-specific involvement of CRISPR–Cas I-B along with the unassigned array targeting of the episymbiont's genomes in two distinct subsurface ecosystems seemingly indicates an independent evolution of defence against intruding DNA, which aligns with previous investigations that demonstrated a strict biogeography of *Ca.* Altiarchaea core genomes and site-specific evolution[36]. 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[37].

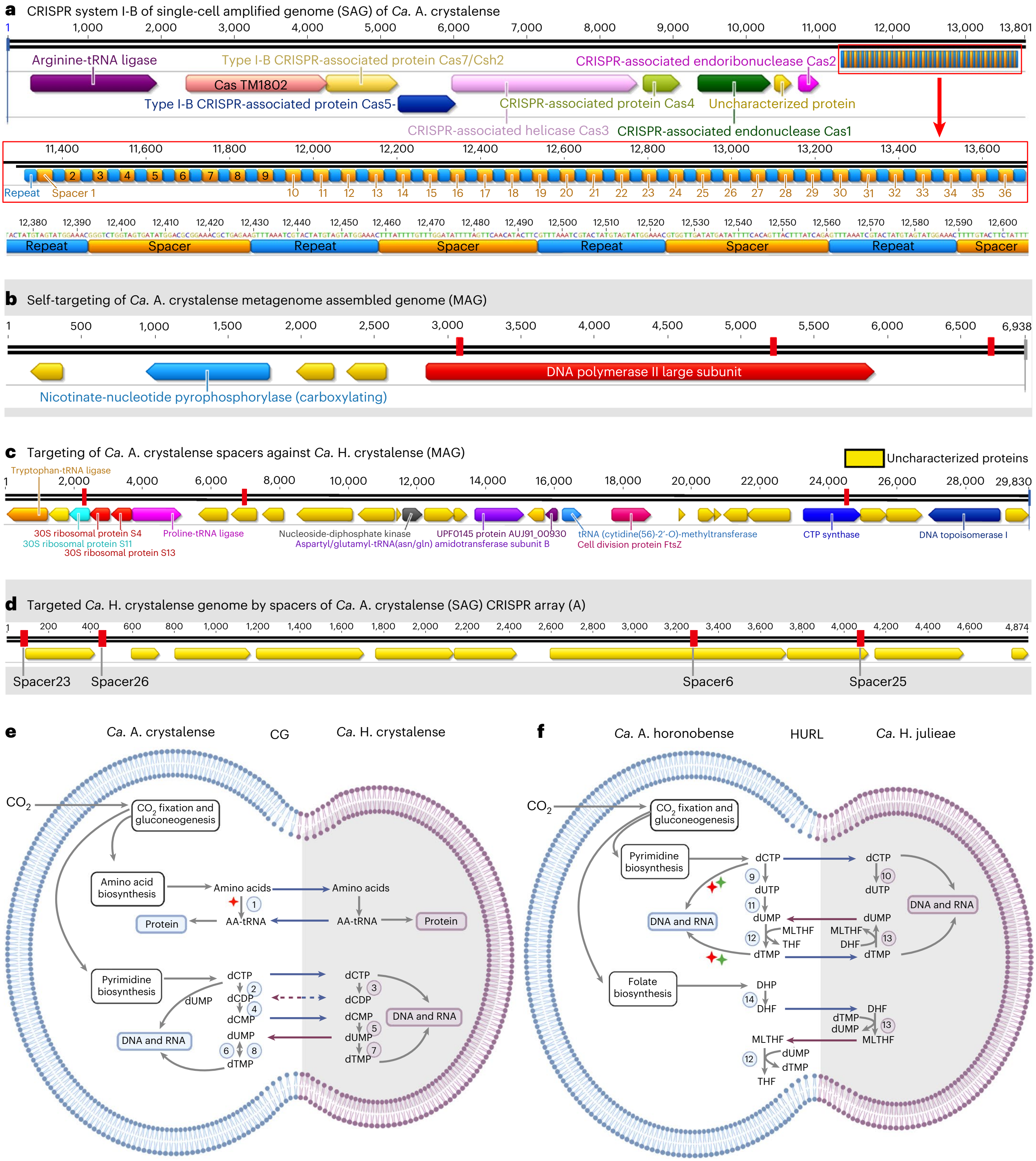
Spacers of the unassigned CRISPR array were detected in much greater diversity than those of CRISPR I-B systems at the HURL site (Fig. 1d,e). While spacers of the unassigned array of CG-derived *Ca.* A. crystalense self-target chromosomal gene sequences, the spacers of the unassigned array of *Ca.* A. horonobense's self-target intergenic regions (Fig. 1c and Supplementary Table 7). Notably, it has been demonstrated in haloarchaea that self-targeting does not necessarily lead to cell suicide[38]. Assuming that the CRISPR–Cas interference is associated with a defence against the symbiont, a plethora of spacers present at CG might effectively repress the symbiont (host:symbiont 11:1 based

Fig. 2 | Example of *Ca.* Altiarchaea CRISPR–Cas type I-B loci, gene targets on host and episymbionts' genomes and metabolic interactions between *Ca.* Altiarchaea and *Ca.* Huberiarchaea as inferred from genome-scaled metabolic modelling. **a**, Example of CRISPR system type I-B locus of *Ca.* A. crystalense with assembled CRISPR array from a SAG (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 altiarchaeal 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**,**f**, Metabolic interactions between *Ca.* Altiarchaeum and *Ca.* Huberiarchaeum in CG (17 genomes of *Ca.* A. crystalense and 11 of *Ca.* H. crystalense and spacers extracted from transcriptomes) (**e**) and in HURL (one genome of A. horonobense and one *Ca.* H. julieae) (**f**). 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 labelled with red diamonds and the unassigned type labelled with green diamonds). While most compounds were produced by *Ca.* Altiarchaea, 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).

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 metabolically complement self-targeted hosts

We applied genome-scale metabolic modelling 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 analysed. MAGs (ten genomes of *Ca*. A. crystalense, ten of *Ca*. H. crystalense, one of *Ca*. A. horonobense and 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 ecosystems were used to render genome-scale metabolic reconstructions. Although we applied thorough manual data curation[20,27], the genomes were fairly fragmented (average $N50_{host/CG} = 8{,}067.24$, average $N50_{symbtiont/CG} = 14{,}983.73$, average


a CRISPR system I-B of single-cell amplified genome (SAG) of Ca. A. crystalense
1 1,000 2,000 3,000 4,000 5,000 6,000 7,000 8,000 9,000 10,000 11,000 12,000 13,000 13,801
Arginine-tRNA ligase
Type I-B CRISPR-associated protein Cas7/Csh2
CRISPR-associated endoribonuclease Cas2
Cas TM1802
Type I-B CRISPR-associated protein Cas5-
CRISPR-associated protein Cas4
Uncharacterized protein
CRISPR-associated helicase Cas3
CRISPR-associated endonuclease Cas1
11,400 11,600 11,800 12,000 12,200 12,400 12,600 12,800 13,000 13,200 13,400 13,600
2 3 4 5 6 7 8 9
Repeat
Spacer 1
10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31 32 33 34 35 36
12,380 12,390 12,400 12,410 12,420 12,430 12,440 12,450 12,460 12,470 12,480 12,490 12,500 12,510 12,520 12,530 12,540 12,550 12,560 12,570 12,580 12,590 12,600
Repeat Spacer Repeat Spacer Repeat Spacer Repeat Spacer
b Self-targeting of Ca. A. crystalense metagenome assembled genome (MAG)
1 500 1,000 1,500 2,000 2,500 3,000 3,500 4,000 4,500 5,000 5,500 6,000 6,500 6,938
DNA polymerase II large subunit
Nicotinate-nucleotide pyrophosphorylase (carboxylating)
c Targeting of Ca. A. crystalense spacers against Ca. H. crystalense (MAG)
Uncharacterized proteins
Tryptophan-tRNA ligase
1 2,000 4,000 6,000 8,000 10,000 12,000 14,000 16,000 18,000 20,000 22,000 24,000 26,000 28,000 29,830
30S ribosomal protein S4
30S ribosomal protein S11
30S ribosomal protein S13
Proline-tRNA ligase
Nucleoside-diphosphate kinase
UPF0145 protein AUJ91_00930
Aspartyl/glutamyl-tRNA(asn/gln) amidotransferase subunit B
tRNA (cytidine(56)-2′-O)-methyltransferase
Cell division protein FtsZ
CTP synthase
DNA topoisomerase I
d Targeted Ca. H. crystalense genome by spacers of Ca. A. crystalense (SAG) CRISPR array (A)
1 200 400 600 800 1,000 1,200 1,400 1,600 1,800 2,000 2,200 2,400 2,600 2,800 3,000 3,200 3,400 3,600 3,800 4,000 4,200 4,400 4,600 4,874
Spacer23 Spacer26 Spacer6 Spacer25
e Ca. A. crystalense CG Ca. H. crystalense
CO2
CO2 fixation and gluconeogenesis
Amino acid biosynthesis
Amino acids
Amino acids
1
Protein
AA-tRNA
AA-tRNA
Protein
Pyrimidine biosynthesis
dCTP 2 dCDP 4 dCMP dUMP 6 8 dTMP
dUMP
DNA and RNA
dCTP 3 dCDP dCMP 5 dUMP 7 dTMP
DNA and RNA
f Ca. A. horonobense HURL Ca. H. julieae
CO2
CO2 fixation and gluconeogenesis
Pyrimidine biosynthesis
dCTP 9 dUTP 11 dUMP 12 dTMP
MLTHF THF
DNA and RNA
dCTP 10 dUTP dUMP 13 dTMP
MLTHF DHF
DNA and RNA
Folate biosynthesis
DHP 14 DHF
DHF
dTMP dUMP 13 MLTHF
MLTHF 12 dUMP dTMP THF

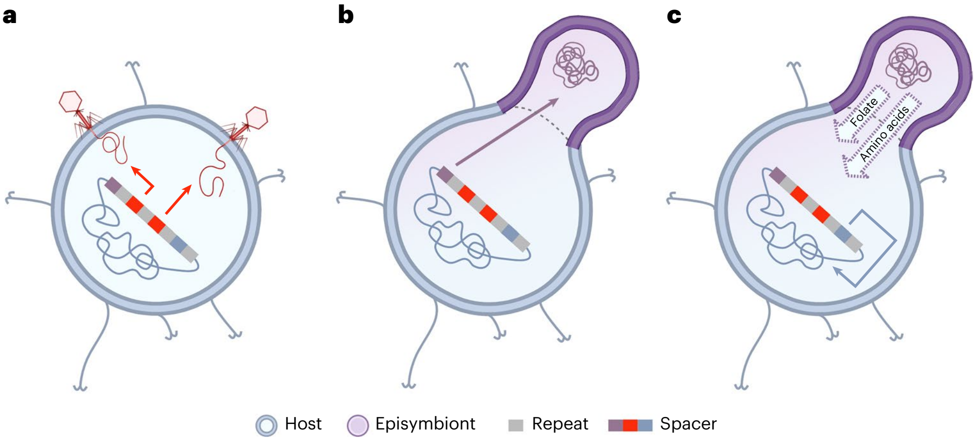


Fig. 3 | Illustration of the newly proposed functionality of CRISPR–Cas systems within *Ca.* Altiarchaea. a, Viral targeting: CRISPR–Cas system targets the genomes of MGEs that infect the cell (current state of knowledge). **b**, Targeting of episymbiont: CRISPR–Cas system targets the genome of the episymbiont *Ca.* Huberiarchaeum to defend against the parasite. **c**, Self-targeting and respective metabolic complementation: self-targeting of CRISPR–Cas in Altiarchaea 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. Figure created with BioRender.com.

$N50_{host/HURL}$ = 3,604, average $N50_{symbtiont/HURL}$ = 4,115, whereas N50 is defined as shortest contig/scaffold length which must be included for covering 50% of the genome), 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 modelling of these metabolic networks facilitated an assessment of host–symbiont metabolic complementarity (Fig. 2e,f and Supplementary Fig. 8). Models of the CG and HURL environment both revealed a substantial 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 (Supplementary Tables 8–11; details in Supplementary Results), supporting the notion that *Ca.* Altiarchaeum is a primary producer, while *Ca.* Huberiarchaeum relies on its host for carbon and energy sources[14].

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 (Supplementary Results), suggesting a collaborative effort of synthesizing pyrimidine (Supplementary Fig. 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 (Supplementary Fig. 8c and Fig. 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 (Supplementary Fig. 8 and Supplementary Results). 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 (Figs. 2d,e and Fig. 3 and Supplementary Fig. 8a,c–f). The respective protein sequences were not horizontally transferred between *Ca.* Altiarchaea and *Ca.* Huberiarchaea on the basis of 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 harboured a definitive PAM (5′-TTN-3′ associated with other I-B systems[39]), no such clear motif was detected in the host protospacers. Here, the second base of the putative PAM region, exhibited a fourfold greater than the average single nucleotide polymorphism (SNP) rate of genes (Supplementary Fig. 9; details in Supplementary Results). Mutations in the PAM region diverging from the 5′-TTN-3′ motif would prevent self-targeting at least for parts of the altiarchaeal population[40] 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[41], 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 ref. 31). 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 re-established. We thus suggest 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.

Interphylum interactions in other symbiotic archaea

To facilitate the overlay of our findings on other potential archaeal host–DPANN episymbiont relationships, we analysed CRISPR spacer matches between all archaeal genomes publicly available in NCBI GenBank (7,012 genomes: state May 2021; Supplementary Table 4). After having extracted 106,641 spacer sequences, 39,875 distinct spacer-to-protospacer matches across all genomes were detected. Few contigs carrying CRISPR arrays (for example, 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. 4). Network analyses showed the genomes of the DPANN *Ca.*

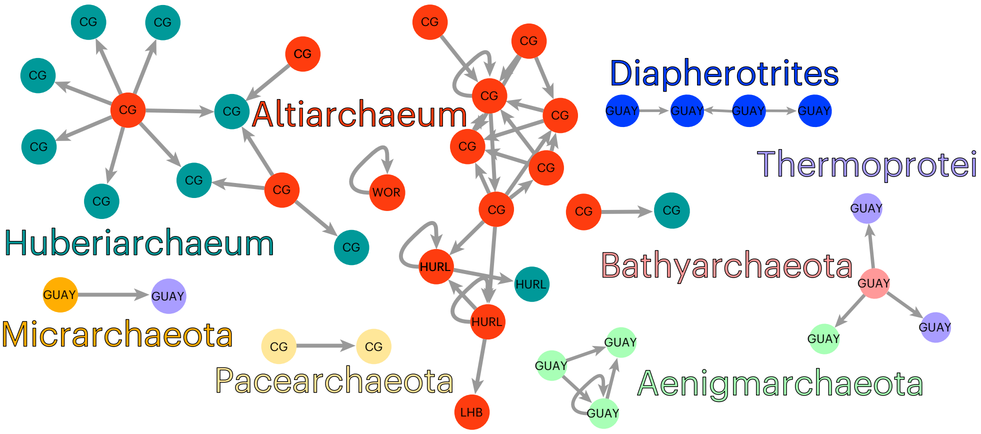


Fig. 4 | Directed spacer interaction of DPANN archaea derived from the analysis of 7,012 publicly available archaeal genomes. Nodes correspond to archaeal genomes. Boomerang and linear grey arrows indicate self-targeting and non-self (including interspecies) targeting spacers, respectively. With the exception of Thermoprotei and Bathyarchaeota, all of the archaea pictured belong to the DPANN superphylum. Colours represent the phylogenetic affiliation of genomes. Genomes of *Ca.* Altiarchaeum and *Ca.* Huberiarchaeum derive primarily from CG. Genomes coded according to their corresponding ecosystem: CG, Crystal Geyser[20,25,27]; LHB, Lake Huron Basin[115]; WOR, White Oak River[116]; GUAY, Guaymas Basin[44]; HURL, Horonobe Underground Research Laboratory[24].

Aenigmarchaeota and *Ca.* Altiarchaeota (Fig. 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. 4), which was also previously shown for other archaea[37,42]. Well-established DPANN–host cocultures, for example, *Ignicoccus hospitalis* and *Nanoarchaeum equitans*[43] 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[44], our approach enabled the a priori prediction of DPANN–host interactions on the basis of CRISPR–Cas genome targeting (Fig. 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, for example, encode for the LamGL domain-containing protein, which is inter alia responsible for the binding of sulfated glycolipids[45,46] and the ribonucleoside triphosphate reductase, amenable for catalysis of the conversion of ribonucleotides into deoxyribonucleotides[47].

Another 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 centimetres apart in depth[44]. 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 that spacer–protospacer matches are a useful tool for identifying in silico host–symbiont interactions of uncultivated archaea on the basis of metagenomic analyses.

Discussion

The findings discussed here demonstrate that archaeal CRISPR–Cas systems acquire resistance not only to genomes of foreign MGEs[1] and closely related species[32] 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 defence against the episymbiont. On the basis of our results from single-cell genomic data, metagenomes and metatranscriptomes, we suggest that a double-edged sword drives the evolution of microbial populations, that is, CRISPR–Cas-mediated defences probably render a major fraction of the DPANN episymbiont population truly parasitic, while the remainder seem to support the host in a mutualistic fashion.

Future studies should be set out with the aim of cultivating the host–symbiont system to validate the herein proposed CRISPR–Cas interference. However, cultivation attempts might selectively enrich for systems with mutualistic relationship and in silico screening of currently existing host–DPANN cocultures for spacer targeting of the episymbiont's genome were devoid of such an interaction, including the well-known archaeal system *Ignicoccus hospitalis* and *Nanoarchaeum equitans*[13]. Consequently, genetic engineering of the host and the symbiont will be necessary to eventually clarify the relationship between hosts and DPANN symbionts—may it be mutualistic, parasitic, or a mixed population model as suggested by our findings.

Methods

RNA extraction and metatranscriptomic sequencing

Samples for transcriptomics were collected along with DNA samples as previously published[20]. For the samples CG05, CG08 and CG16 we filtered about 189, 151 and 151 l of geyser-erupted water, respectively. The MoBio PowerMax Soil DNA kit, now rebranded as the Qiagen DNeasy PowerMax Soil kit (Qiagen), 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 min 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 min. 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). Sequencing libraries were created using the Illumina TruSeq Stranded mRNA Library Prep Kit, following manufacturer's protocol (Illumina). Libraries were sequenced on the Illumina HiSeq 2500 platform (Illumina).

Sample preparation for FISH

Groundwater for FISH analysis was sampled to visualize the Altiarchaea–Huberarchaea relationship within the CG and HURL ecosystem. 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) through the filter to exchange the sample water with fixative. Fixation was performed for 1 h in the dark. Within the filter holder, a washing step with 3 × 20 ml of phosphate buffered saline (PBS) (concentration 1 v/v%) was done, followed by alternating washing and incubation with ethanol with 50, 70 and 100% (v/v)% for 10 min 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 min. Filter samples for FISH from CG of the sampling campaign in 2021 were covered and stored in RNAlater (Invitrogen by Thermo Fisher Scientific; ref. AM7021).

Imaging of FISH samples

FISH was performed according to ref. 14 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 (v. 3.4.91.00000) (Carl Zeiss Microscopy GmbH). Imaging was carried out by using the 110×/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 in figshare (https://doi.org/10.6084/m9.figshare.22739849).

Metagenome assembly and genome reconstruction

Omic datasets generated from sampling campaigns for CG[20,27] and HURL[24] were downloaded from the NCBI Sequence Read Archive (SRA) in April 2019. Please refer to Supplementary Table 1 (metagenomic and metatranscriptomic datasets), Supplementary Table 2 (genome accessions of *Ca*. Aliarchaeum and *Ca*. Huberiarchaeum) and Supplementary Table 3 (single-cell genomic dataset) for all accession numbers of publicly available datasets and generated genomes used in this study. 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, v.37.09) and Sickle[48] (v.1.33). The MetaSPAdes[49] (v.3.10) and Bowtie2[50] utilities (--sensitive, v.2.3.5.1) were applied to assemble reads and estimate coverage, respectively. Scaffolds <1 kilobases were excluded from further analysis. The interactive uBin[51] software (v.0.9.14) was used to segregate the genomes of *Ca*. Altiarchaeum horonobense and *Ca*. Huberiarchaeum julieae on the basis of %GC content, taxonomy and coverage information. To determine genome completeness and contamination we used checkM[52] (v.1.2.2) (Supplementary Table 2). Previously published *Ca*. Huberiarchaeum genomes generated from each of the CG and HURL ecosystems 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*. Huberiarchaea and *Ca*. Altiarchaea phylogenetically. The accession numbers of all genomes within the reference datasets can be found in the Supplementary Data within the phylogenetic tree (with the suffix 'GCA_'). To avoid redundancy, all genomes annotated as *Ca*. Altiarchaea on NCBI (June 2019), previously published *Ca*. Altiarchaea genomes[36] 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 (ref. 53) with the Phylosift[54] marker HMM profiles and an e-value cutoff of 1×10^{-5}. All DNA sequences were aligned with MUSCLE v.3.8.31 (ref. 55) (default parameters) and manually curated to fuse fragmented genes and remove distant homologues and paralogous copies. One *Ca*. Altiarchaea genome (GCA_003663105) was probably 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) (ref. 56) and concatenated into one supermatrix (200 taxa; 5,974 amino acid positions). Phylogenies were reconstructed with IQ-TREE 2 (ref. 57) (v.2.0.5), first using ModelFinder[58], then using that phylogeny as a guide, with the PMSF model[59] (LG + C60 + F + G). Branch supports were calculated using 1,000 ultrafast bootstrap[60] and 1,000 SH-aLRT (ref. 61) replicates and the aBayes[62] test and trees were visualized in iTOL[63] (v.5).

Naming of archaeal species

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 and average amino acid identity cutoffs along with phylogenetic analyses (Fig. 1 and Supplementary Data), 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).

Phylogenetic reconstruction of individual metabolic genes

For the phylogenies of lysine and phenylalanine (subunit B) transfer RNA 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 1,808 archaeal and 25,118 bacterial genomes (all genomes of the respective domain on NCBI as of 1 June 2019 dereplicated at species level) with DIAMOND v.2.0.15.153 (ref. 64). The maximum number of target sequences (-k 400) was determined by trying different numbers (100, 200, 400, 800, 1,000, 0/all), aligning with MAFFT FFT-NS-2 (v.7.505) and running a preliminary phylogeny (BioNJ or PhyML without tree topology optimization) in Seaview v.5.0.4 (ref. 65). We picked the number that we deemed to give a reasonable view of the origin of each sequence, without including too many divergent homologues 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 semimanually (https://github.com/ProbstLab/Adam_Kolyfetis_2021_methanogenesis/blob/master/fuse_sequences.py) to fuse fragmented sequences, realigned as before and trimmed with BMGE[56] (BLOSUM30). Phylogenies were reconstructed with IQ-TREE 2 (ref. 57) using ModelFinder[58] for the model selection and branch supports calculated using 1,000 ultrafast bootstrap[60] and 1,000 SH-aLRT replicates.

CRISPR system extraction and viral sequence determination

The CRISPR systems of 18 distinct *Ca*. Altiarchaeum crystalense genomes[20,26] (Supplementary Table 2) and one *Ca*. Altiarchaeum horonobense genome[24] (Supplementary Table 2) were extracted with CRISPRCasFinder[66] (v.1.2) and annotated *cas* genes were used to identify CRISPR–Cas cassettes. Two resulting consensus direct repeat sequences were used as input for MetaCRAST[67] (-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 base pairs (bp), longer than 57 bp or harbouring homopolymers of six or more identical bases in a row were excluded. Spacers were clustered to 97% nucleotide identity using CD-hit[68] (v.4.8.1) 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[69]. To avoid false positive predictions of self-targeting and episymbiont targeting, we masked prophage region, predicted by VirSorter[70] (category 1–3, 4–6) and transposon regions, predicted by ISEScan[71]. 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 7 SAGs), 11 *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 (https://doi.org/10.6084/m9.figshare.22738568).

Detection, dereplication and analysis of DNA viral scaffolds

Assembled metagenomes were used to extract and predict viral and putatively viral sequences as previously performed[28]. In brief, predicted viral operational taxonomic units (vOTUs) >3 kb were dereplicated via USEARCH[72] at 95% nucleotide identity resulting in centroid sequences for downstream analysis. The vOTUs were identified via blastn[73] (--short, filtering for 80% similarity, v.2.9.0+) of CRISPR-derived spacers against centroid vOTUs. Completeness and origin (host, viral and unclassified) of vOTUs was assessed using CheckV v.0.4.0 (ref. 74). Clustering of viral sequences with a recent viral Refseq database[75] (release July 2022). and previously detected Altiarchaea-targeting viruses[28] was performed using vConTACT2 (refs. 76,77) v.0.11.3, VICTOR[78] (using nucleic acid sequences) and VIRIDIC[79] 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[80] method under settings recommended for prokaryotic viruses[78]. 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[81] for the distance formula D0. Branch support was inferred from 100 pseudobootstrap replicates each. Trees were rooted at the midpoint[82]. Visualization of viral clusters identified with vConTACT2 in conjunction with the viral RefSeq database was performed using Cytoscape v.3.9.02 (ref. 83). In addition, a circular proteomic tree with viral genomes using the Virus-Host DB: RefSeq release 217 was built using ViPTree v.3.5 (ref. 84). Within ViPTree, double-stranded DNA was selected as nucleic acid type and 'any host' chosen as host category.

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 (ref. 50) with default settings. Mappings were filtered to remove hits with more than three mismatches using SAMtools[85] (v.1.10). Genomecov from BEDtools (v.2.27.1) was used to calculate coverage per position[86]. The first and last 150 bp of each scaffold and possible transposons and viruses were masked by setting the breadth to zero. Mean breadths from sliding windows of 35 bp were calculated. In addition, all positions with a coverage lower than ten were excluded. The median coverage of each scaffold (δ) serves to differentiate high and low breadth. Wilcoxon signed rank tests (standard function R; ref. 87) 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 1,000 times for each scaffold.

Models for *Ca*. Altiarchaea and *Ca*. Huberarchaea 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 Supplementary Table 2) were based on MAGs and SAGs identified from CG (AltiCG-HuberCG model) and HURL (AltiHURL-HuberHURL model). 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[88,89]. 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 (Supplementary Table 8). The AltiHURL-HuberHURL model included 388 *Ca*. A. horonobense and 78 *Ca*. H. julieae genes, associated with 495 and 128 reactions, respectively (Supplementary Table 9). 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.

Details of the model are represented in Supplementary Tables 8–11. The CG model was based on the prediction of metabolic pathways using combined annotation of all MAGs and SAGs identified from this and a previous study[14]. Protein sequences annotated from the individual MAGs and SAGs were clustered at 100% amino acid identity using CD-HIT[68,90], followed by a pangenome analysis to capture metabolic capacities represented by the entire population. Automated metabolic reconstruction was performed on the basis of orthologue mapping to (1) existing models of other archaeal strains, that is *Pyrococcus furiosus*, *Thermococcus eurythermalis*, *Methanosarcina barkeri* and *Methanococcus maripaludis*[91,92] and (2) public databases, such as the Kyoto Encyclopedia of Genes and Genomes[93] (KEGG v. > 94.0), EggNOG (v. 5.0) (ref. 94) and Transporter Classification Database (2016) (ref. 95). Extensive manual curations were carried out following the automated reconstruction to integrate prior annotations of the metabolisms of *Ca*. Altiarchaeum and *Ca*. Huberiarchaeum[14,20], as well as latest biochemical evidence of enzymatic functions in archaeal organisms (Supplementary Tables 8 and 9). 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 (Supplementary Tables 8 and 9). The biosynthesis of macromolecules (for example, 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 on the basis of 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[96]. The stoichiometry of protein biosynthesis was calculated on the basis of the average composition of amino acids in the corresponding proteome and the associated energy cost was estimated on the basis of the mechanism of protein synthesis[97], 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 on the basis of experimental measurements of the weight compositions of core lipids and header groups of *Ca*. Altiarchaeum or *Ca*. Huberiarchaeum of the respective system[20]. 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. Relative abundance (based on coverage) of the respective genomes was calculated via metagenomic read mapping with Bowtie2 (ref. 50) (--sensitive mode). The CG- and HURL-specific *Ca*. Altiarchaeum and *Ca*. Huberiarchaeum biomass were then combined on the basis of 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 Supplementary Results).

Metabolic modelling and reconstruction

Consensus models of *Ca*. Altiarchaeum to *Ca*. Huberiarchaeum were constructed on the basis of collections of MAGs and SAGs from CG (20 MAGs and 8 SAGs) and HURL (2 MAGs) (Supplementary Table 2) to capture the metabolic potential of each population. Candidate genes were first identified on the basis of a pangenome analysis, which was performed following orthologue identification using a bidirectional best-hit approach[98]. 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 (v.1.0) package[88,89] using the cplex solver (v.12.7.1.0). Simulations targeted the growth optimization of *Ca*. Altiarchaeum while applying the metabolic reactions 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 (Supplementary Tables 9 and 10). Comparative analyses based on computational simulations were carried out in the presence or absence of CRISPR-targeted genes. This enabled the identification of changes in metabolite transfer and/or metabolic collaboration between *Ca*. Altiarchaeum and *Ca*. Huberiarchaeum (Supplementary Fig. 8) upon targeting specific genes with spacers. Metabolic gaps in the production of biomass components by *Ca*. Altiarchaea were identified using the PSAMM fluxcheck and gapcheck functions[88,89] Candidate gap-filling reactions for unblocking each biomass component were identified using the PSAMM fastgapfill implementation with the KEGG reaction database[93] 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[88,89]. 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.

Metabolic simulations were performed with PSAMM v.1.0 using the IBM ILOG CPLEX Optimizer v.12.7.1.0 (https://www.ibm.com/products/ilog-cplex-optimization-studio). Simulation of the *Ca*. Altiarchaeum–*Ca*. Huberiarchaeum metabolism was formulated with exchange constraints that represent the corresponding in situ geochemical measurements in the CG[20] and HURL[24]. These geochemical measurements included the ion concentrations in porewater and the compositions of headspace gas (Supplementary Tables 8–11). Some measurements, for example, 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.

To identify complementary metabolic processes between *Ca*. Altiarchaea and *Ca*. Huberiarchaea, the PSAMM fastgapfill implementation[88,89] was applied to optimize the *Ca*. Altiarchaea 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 metabolites were subsequently confirmed by showing that the removal of any metabolite exchange would lead to a non-viable *Ca*. Altiarchaea 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 (Supplementary Tables 8 and 9). 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 attacks targeted by the CRISPR system (Supplementary Tables 10 and 11). 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 v.12.7.1.0; a zero range is defined as any fluxes within 1×10^{-6} from zero.

PAM analysis of *Ca*. Altiarchaea, *Ca*. Huberarchaea and viruses

Applying CRISPRTarget[99] (accessed in June 2020) with default settings, PAMs were identified within the genomes of *Ca*. Altiarchaea, *Ca*. Huberarchaea and viruses using spacers bearing 80% sequence similarity. CRISPRTarget results were screened with WebLogo[100,101] (v.2.8.2) in batches of 10,000 8 bp 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 modelling) were aligned to nine different MAGs (Supplementary Table 2) and analysed individually by using BBMap (https://sourceforge.net/projects/bbmap/, v.38.92) (default parameters). SNPs were predicted using the VarScan[102] pileup2snp command (v.2.4.3; 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'.

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

The CRISPR–Cas gene cassette (Cas1, Cas2, Cas3, Cas4, Cas5 and Cas8b) of one SAG of *Ca*. Altiarchaeum was used in gene synthesis. The Cas6 gene was annotated in two other SAGs of *Ca*. Altiarchaea, 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 (ref. 103) and an *Escherichia coli* codon frequency table. The synthetic DNA fragments were obtained from Twist Bioscience, 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) 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 Supplementary Table 13.

CRISPR–Cas activity assay in TXTL

The activity of the *Ca*. Altiarchaea type I-B CRISPR–Cas system was tested in a cell-free (TXTL) system. Circular or linear DNA constructs that were added to a TXTL reaction were transcribed and translated and RNAs and proteins were produced[104]. The reaction conditions of the TXTL reactions performed here were adapted from ref. 29. A deGFP reporter plasmid was generated with site-directed mutagenesis using p70a_deGFP_PacI (ref. 29) 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 (ref. 104) encoding the T7 RNA polymerase and isopropyl β-D-1-thiogalactopyranoside (IPTG) 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 of TXTL reaction were prepared in Costar 3357 96-well V-bottom plates (Corning) with Costar 2080 cover mats (Corning) using the liquid handling machine Echo525 (Beckman Coulter) including the following components: 2.25 µl of myTXTL Sigma 70 MasterMix (Arbor Biosciences), 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 of 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) at 485/528 nm excitation/emission[105]. 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 $P > 0.05$ 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*. Altiarchaea or the motif TTTTC being a non-recognized PAM. All reactions were performed in triplicate.

PAM assay in TXTL (PAM-DETECT)

To reveal the PAM diversity recognized by the type I-B system of *Ca*. Altiarchaea, PAM-DETECT (PAM DETermination with Enrichment-based Cell-free TXTL) was performed. A detailed protocol can be found in ref. 29. 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 harbouring 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 (specified below). Separate 6 µl of TXTL reactions were prepared containing one or the other Cascade Master mix mentioned above. TXTL reactions contained: 4.5 µl of myTXTL Sigma 70 Master mix, 0.2 nM pET28a_T7RNAP (ref. 29), 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)[29]. 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} of PacI (NEB) in 1× 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} of Proteinase K (Cytiva) 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 adaptors with unique dual indices were added in two amplifications steps using KAPA HiFi HotStart Library Amplification Kit (KAPA Biosystems) 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 of purified amplicons generated with the first PCR in a 50 µl reaction and 18 cycles. Next-generation sequencing 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 refs. 106,107 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*. Altiarchaea. PAM-DETECT assays were performed in duplicate.

CRISPR–Cas interactions across archaeal diversity

All archaeal genomes housed in the publicly accessible NCBI database (May 2021; Supplementary Table 4) were screened for viral sequence contaminants using VirSorter[70] (default settings) and all respective hits, as well as annotated plasmids, were excluded from consideration. The CRISPRCasFinder[66] (v.2.0.3) utility was used to extract spacers, direct repeat 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[73] against all archaeal genomes to an 80% nucleotide similarity threshold and interactions between genomes based on CRISPR spacer matches were visualized in Cytoscape[83] (v.3.9.02). The taxonomy of each genome was pulled from the NCBI taxonomy database and in single cases validated using Genome Taxonomy Database[108–110] (GTDB-Tk classify, v.0.3.3, database r89). To avoid false positive predictions of self-targeting and episymbiont targeting, we masked prophage region, predicted by VirSorter[70] (category 1–6).

Reporting summary

Further information on research design is available in the Nature Portfolio Reporting Summary linked to this article.

Data availability

Metagenomic datasets generated from the CG[20,27] ecosystem in 2009, 2014 and 2015 (n = 66) and the HURL[24] environment (n = 2) were downloaded from the NCBI SRA in April 2019 (Supplementary Table 1). SAGs generated in a previous study[20] (n = 219) were retrieved from the Joint Genome Institute's Integrated Microbial Genomes and Microbiomes database[111] (Supplementary Table 3). The metagenome-derived genomes of *Ca*. A. crystalense and *Ca*. H. crystalense from CG are publicly accessible from NCBI (accession numbers in Supplementary Table 2). The genomes of *Ca*. A. horonobense and *Ca*. H. julieae from HURL were newly reconstructed in this investigation (Supplementary Table 2). All previously unpublished genomes used in this study are available in figshare https://doi.org/10.6084/m9.figshare.22339555 (ref. 112) and all viral genomes are available at https://doi.org/10.6084/m9.figshare.22738568(ref. 113). All raw FISH images are deposited here: https://doi.org/10.6084/m9.figshare.22739849(ref. 114).

Code availability

The code used in this publication is based on previously published code. Please refer to the Methods for information regarding the software and versions used.

References

1. Garneau, J. E. et al. The CRISPR/Cas bacterial immune system cleaves bacteriophage and plasmid DNA. *Nature* **468**, 67 (2010).
2. Andersson, A. F. & Banfield, J. F. Virus population dynamics and acquired virus resistance in natural microbial communities. *Science* **320**, 1047–1050 (2008).
3. Koonin, E. V. & Makarova, K. S. Evolutionary plasticity and functional versatility of CRISPR systems. *PLoS Biol*. **20**, e3001481 (2022).
4. Makarova, K. S. et al. An updated evolutionary classification of CRISPR–Cas systems. *Nat. Rev. Microbiol*. **13**, 722 (2015).
5. Makarova, K. S. et al. Evolutionary classification of CRISPR–Cas systems: a burst of class 2 and derived variants. *Nat. Rev. Microbiol*. **18**, 67–83 (2020).
6. Maniv, I., Jiang, W., Bikard, D. & Marraffini, L. A. Impact of different target sequences on type III CRISPR–Cas immunity. *J. Bacteriol*. **198**, 941 (2016).
7. Marraffini, L. A. & Sontheimer, E. J. Self versus non-self discrimination during CRISPR RNA-directed immunity. *Nature* **463**, 568–571 (2010).
8. Dombrowski, N., Lee, J.-H., Williams, T. A., Offre, P. & Spang, A. Genomic diversity, lifestyles and evolutionary origins of DPANN archaea. *FEMS Microbiol. Lett*. **366**, fnz008 (2019).
9. Rinke, C. et al. Insights into the phylogeny and coding potential of microbial dark matter. *Nature* **499**, 431–437 (2013).
10. Castelle, C. J. et al. Biosynthetic capacity, metabolic variety and unusual biology in the CPR and DPANN radiations. *Nat. Rev. Microbiol*. **16**, 629–645 (2018).
11. Sakai, H. D. et al. Insight into the symbiotic lifestyle of DPANN archaea revealed by cultivation and genome analyses. *Proc. Natl Acad. Sci. USA* **119**, e2115449119 (2022).
12. Jahn, U. et al. *Nanoarchaeum equitans* and *Ignicoccus hospitalis*: new insights into a unique, intimate association of two archaea. *J. Bacteriol*. **190**, 1743–1750 (2008).
13. Huber, H. et al. A new phylum of Archaea represented by a nanosized hyperthermophilic symbiont. *Nature* **417**, 63–67 (2002).
14. Schwank, K. et al. An archaeal symbiont–host association from the deep terrestrial subsurface. *ISME J*. **13**, 2135–2139 (2019).
15. Hamm, J. N. et al. Unexpected host dependency of Antarctic Nanohaloarchaeota. *Proc. Natl Acad. Sci. USA* **116**, 14661 (2019).
16. Munson-McGee, J. H. et al. Nanoarchaeota, their Sulfolobales host, and Nanoarchaeota virus distribution across Yellowstone National Park hot springs. *Appl. Environ. Microbiol*. **81**, 7860–7868 (2015).
17. Jarett, J. K. et al. Single-cell genomics of co-sorted Nanoarchaeota suggests novel putative host associations and diversification of proteins involved in symbiosis. *Microbiome* **6**, 161 (2018).
18. Wurch, L. et al. Genomics-informed isolation and characterization of a symbiotic Nanoarchaeota system from a terrestrial geothermal environment. *Nat. Commun*. **7**, 12115 (2016).
19. Hamm, J. N. et al. The parasitic lifestyle of an archaeal symbiont. Preprint at *bioarXiv* https://doi.org/10.1101/2023.02.24.5298342.24.529834v2 (2023).
20. Probst, A. J. et al. Differential depth distribution of microbial function and putative symbionts through sediment-hosted aquifers in the deep terrestrial subsurface. *Nat. Microbiol*. **3**, 328–336 (2018).
21. Heimerl, T. et al. A complex endomembrane system in the archaeon *Ignicoccus hospitalis* tapped by *Nanoarchaeum equitans*. *Front. Microbiol*. **8**, 1072 (2017).
22. Comolli, L. R. & Banfield, J. F. Inter-species interconnections in acid mine drainage microbial communities. *Front. Microbiol*. **5**, 367 (2014).
23. Baker, B. J. et al. Enigmatic, ultrasmall, uncultivated Archaea. *Proc. Natl Acad. Sci. USA* **107**, 8806–8811 (2010).
24. Hernsdorf, A. W. et al. Potential for microbial H_2 and metal transformations associated with novel bacteria and archaea in deep terrestrial subsurface sediments. *ISME J*. **11**, 1915–1929 (2017).
25. Probst, A. J. et al. Biology of a widespread uncultivated archaeon that contributes to carbon fixation in the subsurface. *Nat. Commun*. **5**, 5497 (2014).
26. Probst, A. J. et al. Genomic resolution of a cold subsurface aquifer community provides metabolic insights for novel microbes adapted to high CO_2 concentrations. *Environ. Microbiol*. **19**, 459–474 (2017).
27. Emerson, J. B., Thomas, B. C., Alvarez, W. & Banfield, J. F. Metagenomic analysis of a high carbon dioxide subsurface microbial community populated by chemolithoautotrophs and bacteria and archaea from candidate phyla. *Environ. Microbiol*. **18**, 1686–1703 (2016).
28. Rahlff, J. et al. Lytic archaeal viruses infect abundant primary producers in Earth's crust. *Nat. Commun*. **12**, 4642 (2021).
29. Wimmer, F., Mougiakos, I., Englert, F. & Beisel, C. L. Rapid cell-free characterization of multi-subunit CRISPR effectors and transposons. *Mol. Cell* **82**, 1210–1224.e6 (2022).
30. Marshall, R. et al. Rapid and scalable characterization of CRISPR technologies using an *E. coli* cell-free transcription-translation system. *Mol. Cell* **69**, 146–157.e3 (2018).
31. Heussler, G. E. & O'Toole, G. A. Friendly fire: biological functions and consequences of chromosomal targeting by CRISPR–Cas systems. *J. Bacteriol*. **198**, 1481–1486 (2016).
32. Stern, A., Keren, L., Wurtzel, O., Amitai, G. & Sorek, R. Self-targeting by CRISPR: gene regulation or autoimmunity? *Trends Genet*. **26**, 335–340 (2010).
33. Aklujkar, M. & Lovley, D. R. Interference with histidyl-tRNA synthetase by a CRISPR spacer sequence as a factor in the evolution of *Pelobacter carbinolicus*. *BMC Evol. Biol*. **10**, 230 (2010).
34. Bhaya, D., Davison, M. & Barrangou, R. CRISPR–Cas systems in bacteria and archaea: versatile small RNAs for adaptive defense and regulation. *Annu. Rev. Genet*. **45**, 273–297 (2011).
35. Wilson, G. G. Organization of restriction-modification systems. *Nucleic Acids Res*. **19**, 2539–2566 (1991).

36. Bornemann, T. L. V. et al. Genetic diversity in terrestrial subsurface ecosystems impacted by geological degassing. *Nat. Commun.* **13**, 284 (2022).
37. Turgeman-Grott, I. et al. Pervasive acquisition of CRISPR memory driven by inter-species mating of archaea can limit gene transfer and influence speciation. *Nat. Microbiol.* **4**, 177–186 (2019).
38. Stachler, A.-E. et al. High tolerance to self-targeting of the genome by the endogenous CRISPR–Cas system in an archaeon. *Nucleic Acids Res.* **45**, 5208–5216 (2017).
39. Vink, J. N. A., Baijens, J. H. L. & Brouns, S. J. J. PAM-repeat associations and spacer selection preferences in single and co-occurring CRISPR–Cas systems. *Genome Biol.* **22**, 281 (2021).
40. Pyenson, N. C., Gayvert, K., Varble, A., Elemento, O. & Marraffini, L. A. Broad targeting specificity during bacterial type III CRISPR–Cas immunity constrains viral escape. *Cell Host Microbe* **22**, 343–353 (2017).
41. Chabas, H., Müller, V., Bonhoeffer, S. & Regoes, R. R. Epidemiological and evolutionary consequences of different types of CRISPR-Cas systems. *PLoS Comput. Biol.* **18**, e1010329 (2022).
42. Brodt, A., Lurie-Weinberger, M. N. & Gophna, U. CRISPR loci reveal networks of gene exchange in archaea. *Biol. Direct* **6**, 65 (2011).
43. Paper, W. et al. *Ignicoccus hospitalis* sp. nov., the host of '*Nanoarchaeum equitans*'. *Int. J. Syst. Evol. Microbiol.* **57**, 803–808 (2007).
44. Dombrowski, N., Teske, A. P. & Baker, B. J. Expansive microbial metabolic versatility and biodiversity in dynamic Guaymas Basin hydrothermal sediments. *Nat. Commun.* **9**, 4999 (2018).
45. Hohenester, E. & Yurchenco, P. D. Laminins in basement membrane assembly. *Cell Adhes. Migr.* **7**, 56–63 (2013).
46. Hohenester, E. Laminin G-like domains: dystroglycan-specific lectins. *Curr. Opin. Struct. Biol.* **56**, 56–63 (2019).
47. Benner, S. A., Ellington, A. D. & Tauer, A. Modern metabolism as a palimpsest of the RNA world. *Proc. Natl Acad. Sci. USA* **86**, 7054–7058 (1989).
48. Joshi, N. A. & Fass, J. N. Sickle: a sliding-window, adaptive, quality-based trimming tool for FastQ files (v.1.33) *Github* https://github.com/najoshi/sickle (2011).
49. Nurk, S., Meleshko, D., Korobeynikov, A. & Pevzner, P. A. metaSPAdes: a new versatile metagenomic assembler. *Genome Res.* **27**, 824–834 (2017).
50. Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nat. Methods* **9**, 357–359 (2012).
51. Bornemann, T. L. V., Esser, S. P., Stach, T. L., Burg, T. & Probst, A. J. uBin—a manual refining tool for genomes from metagenomes. *Environ. Microbiol.* **25**, 1077–1083 (2023).
52. Parks, D. H., Imelfort, M., Skennerton, C. T., Hugenholtz, P. & Tyson, G. W. CheckM: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. *Genome Res.* **25**, 1043–1055 (2015).
53. Eddy, S. R. Accelerated profile HMM searches. *PLoS Comput. Biol.* **7**, e1002195 (2011).
54. Darling, A. E. et al. PhyloSift: phylogenetic analysis of genomes and metagenomes. *PeerJ* **2**, e243 (2014).
55. Edgar, R. C. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.* **32**, 1792–1797 (2004).
56. Criscuolo, A. & Gribaldo, S. BMGE (Block Mapping and Gathering with Entropy): a new software for selection of phylogenetic informative regions from multiple sequence alignments. *BMC Evol. Biol.* **10**, 210 (2010).
57. Minh, B. Q. et al. IQ-TREE 2: new models and efficient methods for phylogenetic inference in the genomic era. *Mol. Biol. Evol.* **37**, 1530–1534 (2020).
58. Kalyaanamoorthy, S., Minh, B. Q., Wong, T. K. F., von Haeseler, A. & Jermiin, L. S. ModelFinder: fast model selection for accurate phylogenetic estimates. *Nat. Methods* **14**, 587–589 (2017).
59. Wang, H.-C., Minh, B. Q., Susko, E. & Roger, A. J. Modeling site heterogeneity with posterior mean site frequency profiles accelerates accurate phylogenomic estimation. *Syst. Biol.* **67**, 216–235 (2017).
60. Hoang, D. T., Chernomor, O., von Haeseler, A., Minh, B. Q. & Vinh, L. S. UFBoot2: improving the ultrafast bootstrap approximation. *Mol. Biol. Evol.* **35**, 518–522 (2017).
61. Guindon, S. et al. New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. *Syst. Biol.* **59**, 307–321 (2010).
62. Anisimova, M., Gil, M., Dufayard, J.-F., Dessimoz, C. & Gascuel, O. Survey of branch support methods demonstrates accuracy, power, and robustness of fast likelihood-based approximation schemes. *Syst. Biol.* **60**, 685–699 (2011).
63. Letunic, I. & Bork, P. Interactive Tree Of Life (iTOL) v4: recent updates and new developments. *Nucleic Acids Res.* **47**, W256–W259 (2019).
64. Buchfink, B., Xie, C. & Huson, D. H. Fast and sensitive protein alignment using DIAMOND. *Nat. Methods* **12**, 59 (2014).
65. Gouy, M., Tannier, E., Comte, N. & Parsons, D. P. in *Multiple Sequence Alignment: Methods and Protocols* (ed. Katoh, K.) 241–260 (Springer, 2021).
66. Couvin, D. et al. CRISPRCasFinder, an update of CRISRFinder, includes a portable version, enhanced performance and integrates search for Cas proteins. *Nucleic Acids Res.* **46**, W246–W251 (2018).
67. Moller, A. G. & Liang, C. MetaCRAST: reference-guided extraction of CRISPR spacers from unassembled metagenomes. *PeerJ* **5**, e3788 (2017).
68. Fu, L., Niu, B., Zhu, Z., Wu, S. & Li, W. CD-HIT: accelerated for clustering the next-generation sequencing data. *Bioinformatics* **28**, 3150–3152 (2012).
69. Biswas, A., Fineran, P. C. & Brown, C. M. Accurate computational prediction of the transcribed strand of CRISPR non-coding RNAs. *Bioinformatics* **30**, 1805–1813 (2014).
70. Roux, S., Enault, F., Hurwitz, B. L. & Sullivan, M. B. VirSorter: mining viral signal from microbial genomic data. *PeerJ* **3**, e985 (2015).
71. Xie, Z. & Tang, H. ISEScan: automated identification of insertion sequence elements in prokaryotic genomes. *Bioinformatics* **33**, 3340–3347 (2017).
72. Edgar, R. C. Search and clustering orders of magnitude faster than BLAST. *Bioinformatics* **26**, 2460–2461 (2010).
73. Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. Basic local alignment search tool. *J. Mol. Biol.* **215**, 403–410 (1990).
74. Nayfach, S. et al. CheckV assesses the quality and completeness of metagenome-assembled viral genomes. *Nat. Biotechnol.* **39**, 578–585 (2021).
75. Cook, R. et al. INfrastructure for a PHAge REference. Database: identification of large-scale biases in the current collection of cultured phage genomes. *Phage* **2**, 214–223 (2021).
76. Bolduc, B. et al. vConTACT: an iVirus tool to classify double-stranded DNA viruses that infect archaea and bacteria. *PeerJ* **5**, e3243 (2017).
77. Bin Jang, H. et al. Taxonomic assignment of uncultivated prokaryotic virus genomes is enabled by gene-sharing networks. *Nat. Biotechnol.* **37**, 632–639 (2019).
78. Meier-Kolthoff, J. P. & Göker, M. VICTOR: genome-based phylogeny and classification of prokaryotic viruses. *Bioinformatics* **33**, 3396–3404 (2017).

79. Moraru, C., Varsani, A. & Kropinski, A. M. VIRIDIC—a novel tool to calculate the intergenomic similarities of prokaryote-infecting viruses. *Viruses* **12**, 1268 (2020).
80. Meier-Kolthoff, J. P., Auch, A. F., Klenk, H.-P. & Göker, M. Genome sequence-based species delimitation with confidence intervals and improved distance functions. *BMC Bioinform.* **14**, 60 (2013).
81. Lefort, V., Desper, R. & Gascuel, O. FastME 2.0: a comprehensive, accurate, and fast distance-based phylogeny inference program. *Mol. Biol. Evol.* **32**, 2798–2800 (2015).
82. Farris, J. S. Estimating phylogenetic trees from distance matrices. *Am. Nat.* **106**, 645–668 (1972).
83. Shannon, P. et al. Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res.* **13**, 2498–2504 (2003).
84. Nishimura, Y. et al. ViPTree: the viral proteomic tree server. *Bioinformatics* **33**, 2379–2380 (2017).
85. Li, H. et al. The sequence alignment/map format and SAMtools. *Bioinformatics* **25**, 2078–2079 (2009).
86. Quinlan, A. R. & Hall, I. M. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**, 841–842 (2010).
87. R Core Team. *R: A Language and Environment for Statistical Computing* (R Foundation for Statistical Computing, 2013).
88. Dufault-Thompson, K., Steffensen, J. L. & Zhang, Y. in *Metabolic Network Reconstruction and Modeling: Methods and Protocols* (ed. Fondi, M.) 131–150 (Springer, 2018).
89. Steffensen, J. L., Dufault-Thompson, K. & Zhang, Y. PSAMM: a portable system for the analysis of metabolic models. *PLoS Comput. Biol.* **12**, e1004732–e1004732 (2016).
90. Li, W. & Godzik, A. Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics* **22**, 1658–1659 (2006).
91. Gonnerman, M. C., Benedict, M. N., Feist, A. M., Metcalf, W. W. & Price, N. D. Genomically and biochemically accurate metabolic reconstruction of *Methanosarcina barkeri* Fusaro, iMG746. *Biotechnol. J.* **8**, 1070–1079 (2013).
92. Goyal, N., Widiastuti, H., Karimi, I. A. & Zhou, Z. A genome-scale metabolic model of *Methanococcus maripaludis* S_2 for CO_2 capture and conversion to methane. *Mol. Biosyst.* **10**, 1043–1054 (2014).
93. Kanehisa, M., Furumichi, M., Tanabe, M., Sato, Y. & Morishima, K. KEGG: new perspectives on genomes, pathways, diseases and drugs. *Nucleic Acids Res.* **45**, D353–D361 (2017).
94. Huerta-Cepas, J. et al. eggNOG 5.0: a hierarchical, functionally and phylogenetically annotated orthology resource based on 5090 organisms and 2502 viruses. *Nucleic Acids Res.* **47**, D309–D314 (2019).
95. Saier, M. H. Jr et al. The transporter classification database (TCDB): recent advances. *Nucleic Acids Res.* **44**, D372–D379 (2016).
96. Neidhardt, F. C., Neidhardt, F. C. N., Ingraham, J. L. & Schaechter, M. *Physiology of the Bacterial Cell: A Molecular Approach* (Sinauer Associates, 1990).
97. Nelson, D. L., Nelson, R. D. & Cox, M. M. *Lehninger Principles of Biochemistry* (W.H. Freeman, 2004).
98. Zhang, Y. & Sievert, S. Pan-genome analyses identify lineage- and niche-specific markers of evolution and adaptation in *Epsilonproteobacteria*. *Front. Microbiol.* **5**, 110 (2014).
99. Biswas, A., Gagnon, J. N., Brouns, S. J. J., Fineran, P. C. & Brown, C. M. CRISPRTarget: bioinformatic prediction and analysis of crRNA targets. *RNA Biol.* **10**, 817–827 (2013).
100. Crooks, G. E., Hon, G., Chandonia, J.-M. & Brenner, S. E. WebLogo: a sequence logo generator. *Genome Res.* **14**, 1188–1190 (2004).
101. Schneider, T. D. & Stephens, R. M. Sequence logos: a new way to display consensus sequences. *Nucleic Acids Res.* **18**, 6097–6100 (1990).
102. Koboldt, D. C. et al. VarScan 2: somatic mutation and copy number alteration discovery in cancer by exome sequencing. *Genome Res.* **22**, 568–576 (2012).
103. Oberortner, E., Cheng, J.-F., Hillson, N. J. & Deutsch, S. Streamlining the design-to-build transition with build-optimization software tools. *ACS Synth. Biol.* **6**, 485–496 (2017).
104. Garamella, J., Marshall, R., Rustad, M. & Noireaux, V. The All E. coli TX-TL Toolbox 2.0: a platform for cell-free synthetic biology. *ACS Synth. Biol.* **5**, 344–355 (2016).
105. Shin, J. & Noireaux, V. An *E. coli* cell-free expression toolbox: application to synthetic gene circuits and artificial cells. *ACS Synth. Biol.* **1**, 29–41 (2012).
106. Leenay, R. T. et al. Identifying and visualizing functional PAM diversity across CRISPR–Cas systems. *Mol. Cell* **62**, 137–147 (2016).
107. Ondov, B. D., Bergman, N. H. & Phillippy, A. M. Interactive metagenomic visualization in a web browser. *BMC Bioinform.* **12**, 385 (2011).
108. Chaumeil, P.-A., Mussig, A. J., Hugenholtz, P. & Parks, D. H. GTDB-Tk: a toolkit to classify genomes with the Genome Taxonomy Database. *Bioinformatics* **36**, 1925–1927 (2020).
109. Parks, D. H. et al. A standardized bacterial taxonomy based on genome phylogeny substantially revises the tree of life. *Nat. Biotechnol.* **36**, 996–1004 (2018).
110. Parks, D. H. et al. A complete domain-to-species taxonomy for bacteria and archaea. *Nat. Biotechnol.* **38**, 1079–1086 (2020).
111. Chen, I.-M. A. et al. IMG/M v.5.0: an integrated data management and comparative analysis system for microbial genomes and microbiomes. *Nucleic Acids Res.* **47**, D666–D677 (2019).
112. Esser, S. P. & Probst, A. J. Genomes of *Ca*. Altiarchaeum and *Ca*. Huberiarchaeum from Crystal Geyser and Horonobe Underground Research Laboratory. *figshare* https://doi.org/10.6084/m9.figshare.22339555 (2023).
113. Esser, S. P., Rahlff, J. & Probst, A. J. Viral operational taxonomic units (vOTUs) from Crystal Geyser. *figshare* https://doi.org/10.6084/m9.figshare.22738568.v1 (2023).
114. Turzynski, V., Esser, S. P. & Probst, A. J. Fluorescence in situ hybridization images of *Ca*. Altiarchaeum and *Ca*. Huberiarchaeu. *figshare* https://doi.org/10.6084/m9.figshare.22739849 (2023).
115. Sharrar, A. M. et al. Novel large sulfur bacteria in the metagenomes of groundwater-fed chemosynthetic microbial mats in the Lake Huron Basin. *Front. Microbiol.* **8**, 791 (2017).
116. Bird, J. T., Baker, B. J., Probst, A. J., Podar, M. & Lloyd, K. G. Culture independent genomic comparisons reveal environmental adaptations for Altiarchaeales. *Front. Microbiol.* **7**, 1221 (2016).
117. Posit team. Rstudio: Integrated development environment for R. https://posit.co/; version 2023.03.0+386 (2022).

Acknowledgements

This research 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 no. DFG PR1603/2-1) and through SPP 2141 (grant no. DFG BE6703/1-1). Genome-scale metabolic modelling 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 no. JPJ007597). The work (proposal https://doi.org/10.46936/10.25585/60000800) conducted by the US Department of Energy (DOE) Joint Genome Institute (https://ror.org/04xm1d337), a DOE Office of Science User Facility, is supported by the Office of Science of the US DOE operated under contract no. DE-AC02-05CH11231. M.P. is supported by the Austrian Science Funds (project MAINTAIN, DOC 69 doc.funds). P.S.A. was supported by a postdoctoral fellowship from the Alexander von Humboldt Foundation. J.P. was supported by Aker BP within the framework of the GeneOil Project given to A.J.P. J.R. was supported by the German Science Foundation (grant no. RA3432/1-1, project no. 446702140). Support by the German Federal Ministry of Education and Research within the project 'MultiKulti' (BMBF funding code: 161L0285E) is acknowledged. We thank K. Dreger for exemplary server administration and B. Siebers, I. Berg, J. F. Banfield and B. Meyer for insightful discussions.

Author contributions

S.P.E. and A.J.P. performed genome-resolved metagenomics, while S.P.E. and J.R. performed viromics. J.R. analysed viral genomes with input from M.K. CRISPR–Cas analyses were done by S.P.E., J.R. and A.J.P. SNP analysis was performed by M.P. and T.R. Genome-scale modelling was conducted by W.Z. and Y.Z. with input from S.P.E., P.A.F.G. and A.J.P. J.P. conducted the sliding window analysis with the input from S.P.E., J.R., and A.J.P. Phylogenomic analyses were carried out by P.S.A. T.L.V.B. provided bioinformatic assistance and K. Schwank, I.B. and V.T. performed microscopy and initial metabolic analyses. J.M. and W.B. resampled CG and J.L., T.W. and A.J.P. conducted RNA extraction and sequencing and S.P.E. analysed transcriptomes. J-F.C. synthesized the Cas genes with input from I.K.B., F.W. and C.B. performed binding, cleavage and PAM assays and J.L., J.J., Y.A., T.W. and A.J.P. generated/provided raw data. K. Sures and S.P.E. analysed the archaeal CRISPR–Cas interactions from published NCBI archaeal genomes. A.J.P. conceptualized the work. S.P.E., J.R., W.Z., Y.Z. and A.J.P. wrote the manuscript with input from all authors.

Competing interests

The authors declare no competing interests.

Additional information

Extended data is available for this paper at https://doi.org/10.1038/s41564-023-01439-2.

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41564-023-01439-2.

Correspondence and requests for materials should be addressed to Alexander J. Probst.

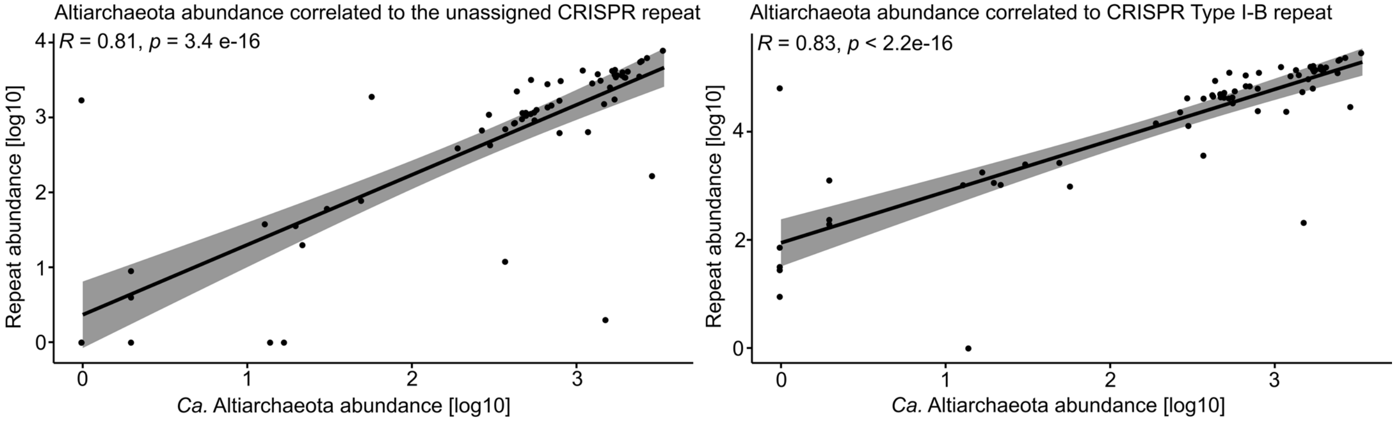
Peer review information *Nature Microbiology* thanks the anonymous reviewers for their contribution to the peer review of this work.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

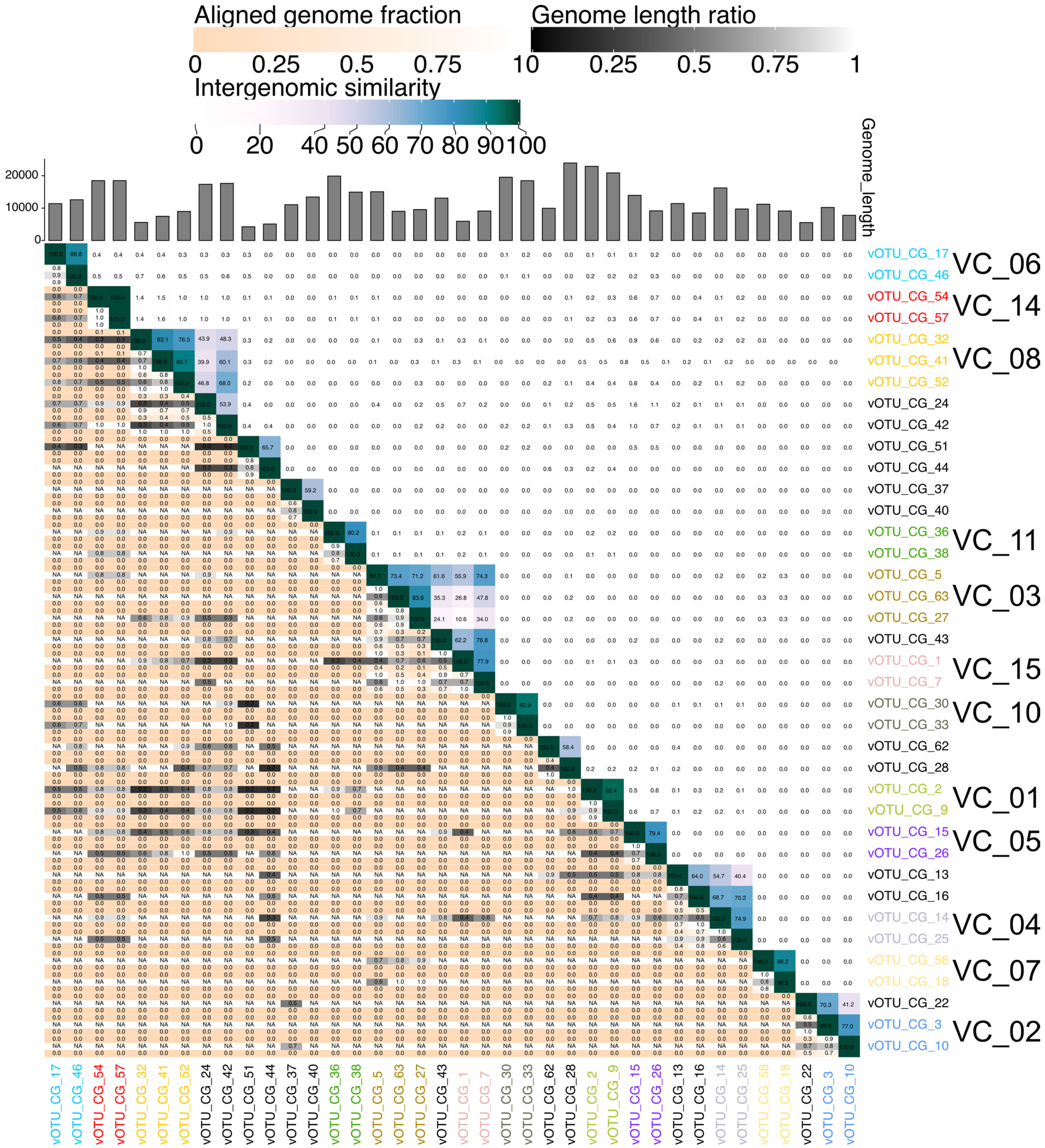
[1]Environmental Metagenomics, Research Center One Health Ruhr of the University Alliance Ruhr, Faculty of Chemistry, University of Duisburg-Essen, Essen, Germany. [2]Group for Aquatic Microbial Ecology, Environmental Microbiology and Biotechnology, University of Duisburg-Essen, Essen, Germany. [3]Department of Cell and Molecular Biology, College of the Environment and Life Sciences, University of Rhode Island, Kingston, RI, USA. [4]Computational Systems Biology, Centre for Microbiology and Environmental Systems Science, University of Vienna, Vienna, Austria. [5]Doctoral School in Microbiology and Environmental Science, University of Vienna, Vienna, Austria. [6]Helmholtz Institute for RNA-based Infection Research (HIRI), Helmholtz-Centre for Infection Research (HZI), Würzburg, Germany. [7]DOE Joint Genome Institute, Lawrence Berkeley National Laboratory, Berkeley, CA, USA. [8]School of Biological Sciences, University of Utah, Salt Lake City, UT, USA. [9]Institut Pasteur, Université Paris Cité, CNRS UMR6047, Archaeal Virology Unit, Paris, France. [10]Nuclear Fuel Cycle Engineering Laboratories, Japan Atomic Energy Agency, Tokai, Japan. [11]Medical faculty, University of Würzburg, Würzburg, Germany. [12]Centre of Water and Environmental Research (ZWU), University of Duisburg-Essen, Essen, Germany. [13]Centre of Medical Biotechnology (ZMB), University of Duisburg-Essen, Essen, Germany. [14]Present address: Centre for Ecology and Evolution in Microbial Model Systems (EEMiS), Department of Biology and Environmental Science, Linnaeus University, Kalmar, Sweden. [15]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, China. [16]Present address: University of Regensburg, Biochemistry III, Regensburg, Germany. [17]These authors contributed equally: Sarah P. Esser, Janina Rahlff. ✉e-mail: alexander.probst@uni-due.de

https://doi.org/10.1038/s41564-023-01439-2

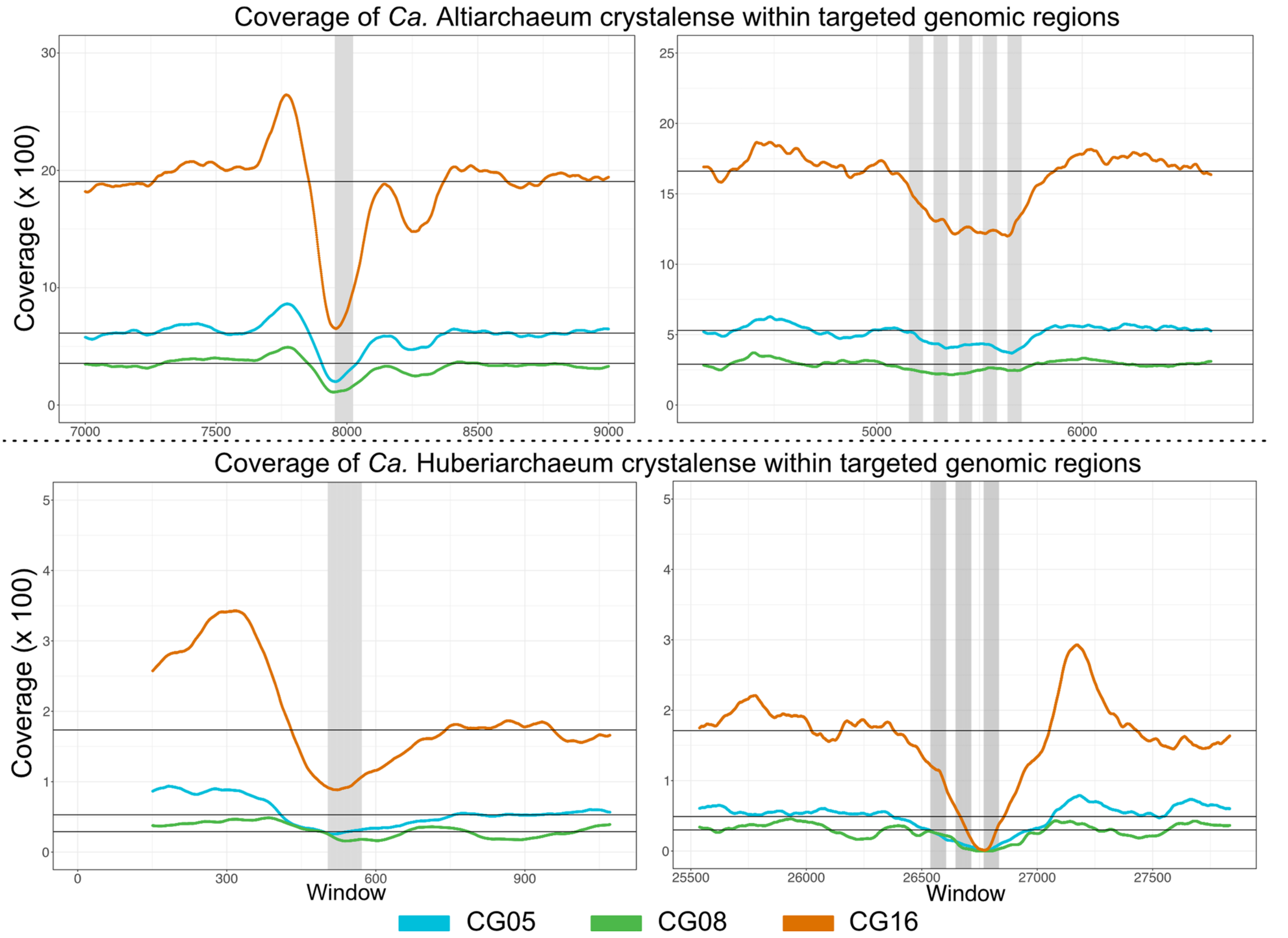


Extended Data Fig. 1 | Correlation of repeat abundance and abundance of Ca. Altiarchaea genomes. Spearman rank correlation (two-tailed) of logarithmic abundances of *Ca.* A. crystalense and logarithmic abundances of repeat sequences of the unassigned CRISPR array (p-value < 3.4 e^{-16}) and the CRISPR system type I-B (p < 2.2 e^{-16}) in metagenomes from CG (n = 66). The grey area depicts the confidence interval of 0.95. The line indicates that the correlation of the genome abundance and repeat abundance is linear. Visualization was performed with R[87,117] (version 3.6.1).

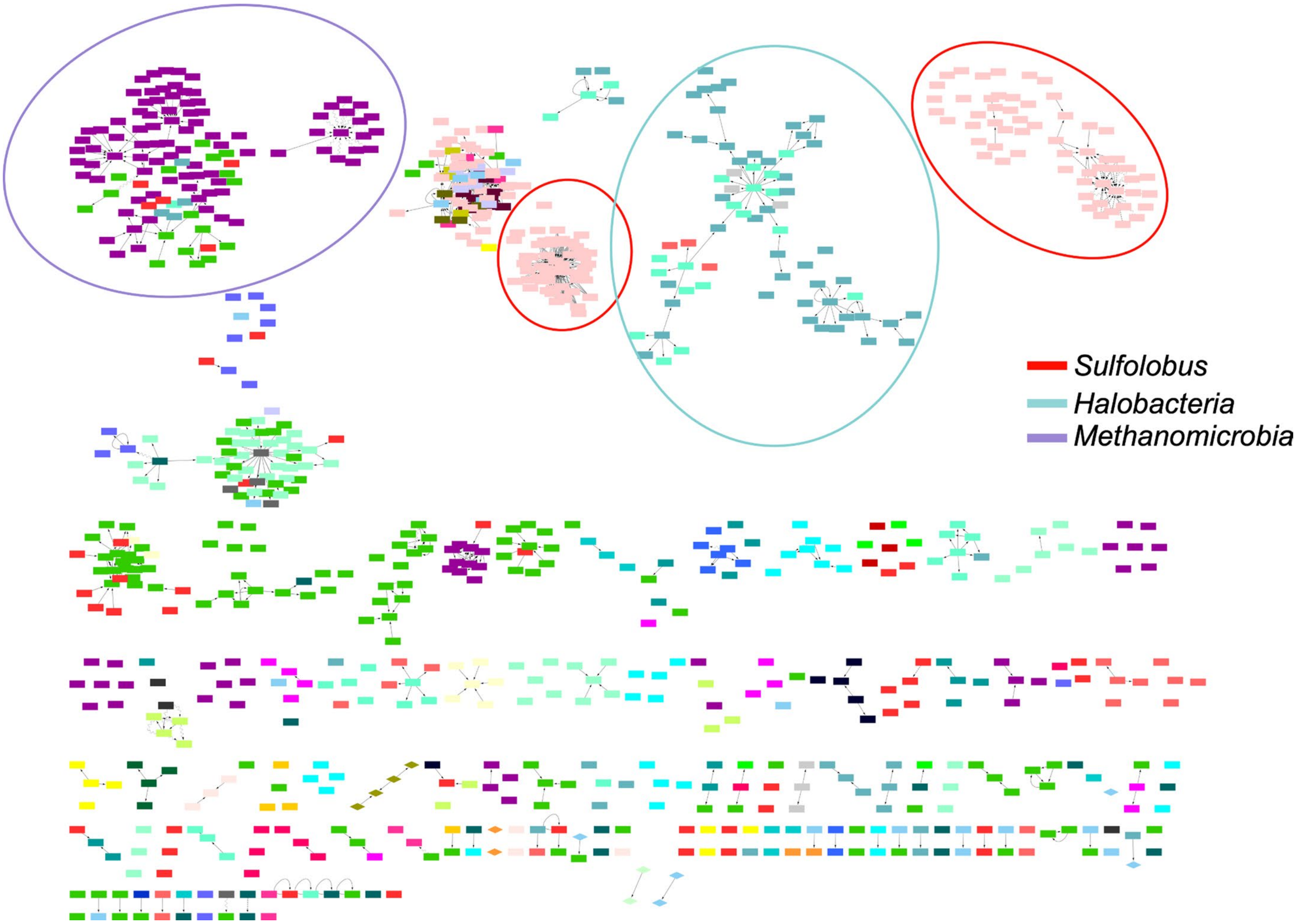
https://doi.org/10.1038/s41564-023-01439-2



Extended Data Fig. 2 | Viral clusters predicted by VIRIDIC[79]. Heatmap showing intergenomic similarity for viral scaffolds of viral clusters (VC_XY) and some singletons (black). Colouring of viral OTUs (vOTUs) according to Supplementary Table 6. 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 | Coverage analyses of scaffolds targeted by spacers from *Ca*. Altiarchaea. Coverage changes within targeted regions by CRISPR system type I-B 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 coloured 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 (Supplementary Fig. 1).



Extended Data Fig. 4 | Spacer targeting analyses of publicly available archaeal genomes. Directed spacer analysis of 7,012 publicly available archaeal genomes (Supplementary Table 4) 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. Edges are colored according to their relationship at least familiy level or lower. The clustering was illustrated with Cytoscape[83] (version 3.9.1). Please note that targeting within the same genus might limit the interspecies recombination, as demonstrated in haloarchaea[37], or reflect the presence of multiple conserved genomic regions between the genomes.

nature portfolio

Corresponding author(s): Alexander J. Probst

Last updated by author(s): Jun 19, 2023

Reporting Summary

Nature Portfolio wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Portfolio policies, see our Editorial Policies and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed	
☐	☒	The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☐	☒	A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☐	☒	The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☒	☐	A description of all covariates tested
☒	☐	A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☐	☒	A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☐	☒	For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☒	☐	For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒	☐	For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☐	☒	Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	Datasets were created based on DNA and RNA sequencing as well as based on fluorescence in situ hybridization (FISH) as stated in the manuscript.
Data analysis	Software code and web links are available from online sources and appropriately referenced to in the method section. A description of how we performed the analysis and which software version we used can be found in this respective methods section in the manuscript. A list of software and versions used is provided below. metaSPAdes (version 3.10), bowtie2 (version 2.3.5.1), uBin (version 0.9.14), bbduk (version 37.09), sickle (version 1.33), HMMER 3.2.1, Phylosift, MUSCLE v3.8.31, BMGE (BLOSUM30), IQ-TREE v2.0.5, ModelFinder, PMSF model (LG+C60+F+G), Ultrafast bootstrap, SH-aLRT, aBayes, iTOL (version 5), MetaCRAST, CD-hit (cersion 4.8.1), CRISPRDirection 2.0, VirSorter (version 1), ISEScan, PSAMM (v1.0), cplex solver (v12.7.1.0) CRISPRTarget (June 2020), WebLogo (v2.8.2), BBmap (v38.92), VarScan (v2.4.3), CRISPRCasFinder (version 2.0.3), GTDB-Tk classify (v0.3.3 r89), Cytoscape (version 3.9.02), blastn (v 2.9.0+), CheckV v0.11.3, Refseq Database (release July 2022), vContact2 (v. 0.11.3), INPHARED (v1.2), VICTOR, VIRIDIC (v1.0 r3.6), Virus-Host DB:RefSeq release 217, ViPTree version 3.5, Zen 3.4 Pro (version 3.4.91.00000), BOOST design software (v1.3.9), DIAMOND (version 2.0.15.153), MAFFT FFT-NS-2 (v7.505), Seaview (v 5.0.4), SAMtools (version 1.10), BEDtools (v2.27.1), R basic, R studio (v.2023.03.0+386), checkM (v1.2.2), EggNOG (v5.0), Biorender, Prodigal (v2.6.3), Microsoft Office (v16.74), KEGG (v >94.0), Transporter Classification Database (2016)

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Portfolio guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A description of any restrictions on data availability
- For clinical datasets or third party data, please ensure that the statement adheres to our policy

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) (Hokkaido, Japan) (Hernsdorf et al. 2017) environment (n = 1) 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 previously unpublished genomes used in this study are available in a Figshare folder 10.6084/m9.figshare.22339555 and all viral genomes are available here: 10.6084/m9.figshare.22738568. All raw FISH images are deposited here: https://doi.org/10.6084/m9.figshare.22739849.

Human research participants

Policy information about studies involving human research participants and Sex and Gender in Research.

Reporting on sex and gender	This does not apply to the present study.
Population characteristics	This does not apply to the present study.
Recruitment	This does not apply to the present study.
Ethics oversight	This does not apply to the present study.

Note that full information on the approval of the study protocol must also be provided in the manuscript.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

□ Life sciences □ Behavioural & social sciences ☒ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Ecological, evolutionary & environmental sciences study design

All studies must disclose on these points even when the disclosure is negative.

Study description	Our study deals with Altiarchaeota, a dominant archaeal primary producer of the deep subsurface and its interaction with viruses and the episymbiont of the uncultivated phylum Huberarchaeota. We investigated the host's CRISPR-Cas systems and associated spacer matches to the episymbiont genomes to investigate the nature of the relationship. We also transferred these findings to other archaeal symbiont-host associations.
Research sample	We used publicly available metagenomic/metatranscriptomic datasets of filtered groundwater samples from Crystal Geyser and Horonobe Underground Research Laboratory, in 350m and 250m depth, respectively. For FISH analyses we retrieved filter sets with microbes from the two sites as well.
Sampling strategy	The sampling strategies/procedures can be found in the method section in the manuscript. We analyzed 66 metagenomes from one ecosystem and compared the results to a distinct ecosystem with one publicly available metagenome, were the host-symbiont relationship was also found. Additionally, we also analyzed 3 metatranscriptomes of the respective site. Samples for FISH were additionally collected specifically in eruption periods of the geyser, which showed the respective symbiont-host association, and from the respective aquifer in HURL.
Data collection	Metagenomic data were downloaded from public datasets in April 2019 and June 2021.
Timing and spatial scale	Sampling of Crystal Geyser for metagenomes and metatranscriptomes was performed in 2009, 2014 and 2015. Sampling of the Horonobe Underground Research Laboratory was conducted between June 2012 and June 2013. Subsurface ecosystems rely on geological settings and geological time scales that do not affect the study. FISH samples of Crystal Geyser were collected in 2021 and of HURL in 2019. Dates of samplings are listed in the section "Field work, collection and transport". All further details are specified in the method section of our manuscript.

Data exclusions

We did not exclude any data.

Reproducibility

Detailed information can be found in the method section of the manuscript. Metagenomes had 66 replicates, transcriptomes had three biological replicates. We did not test for reproducibility of metagenomic sequencing with technical replicates but relied on biological replicates instead. Environmental parameters of the groundwater acquifer/sulfidic spring were extensively investigated previously and remained constant over years.

Randomization

This does not apply to the present study.

Blinding

This does not apply to the present study.

Did the study involve field work? ☒ Yes □ No

Field work, collection and transport

Field conditions

Metatranscriptomics: Mai 25/26 2015 at Crystal Geyser (location, see below). Samples were retrieved from the subsurface and thus the weather conditions are not relevant.
FISH: The sampling of Crystal Geyser was performed at the 13th of August 2021 within the minor eruption phase. Samples were retrieved from the subsurface and thus the weather conditions are not relevant.
HURL/FISH: The sampling of HURL was performed at the 9th of July 2019. Samples were retrieved from the subsurface and thus the weather conditions are not relevant.

Location

Crystal Geyser (Utah, USA) longitude: -110° 08' 7.58" W; latitude: 38° 56' 17.71" N; Honorobe Underground Research Laboratory (Japan) longitude: 141° 51' 34.55" E; latitude: 45° 02' 43.41" N

Access & import/export

As the sampling site is in the middle of a desert, the sampling campaign was performed by car. The samples were directly frozen at site at -80 degrees celsius and transported frozen.

Disturbance

There was no disturbance of the ecosystem caused by this study.

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study	
☒	□	Antibodies
☒	□	Eukaryotic cell lines
☒	□	Palaeontology and archaeology
☒	□	Animals and other organisms
☒	□	Clinical data
☒	□	Dual use research of concern

Methods

n/a	Involved in the study	
☒	□	ChIP-seq
☒	□	Flow cytometry
☒	□	MRI-based neuroimaging