Functional assignment of gut-specific archaeal proteins in the human gut microbiome

Polina V. Novikova[1], Susheel Bhanu Busi[1], Alexander J. Probst[2], Patrick May[3], Paul Wilmes[1,4]

[1]Systems Ecology, Luxembourg Centre for Systems Biomedicine, University of Luxembourg, 7, avenue des Hauts-Fourneaux, L-4362 Esch-sur-Alzette, Luxembourg

[2]Environmental Metagenomics, Research Center One Health Ruhr of the University Alliance Ruhr, for Environmental Microbiology and Biotechnology, Department of Chemistry, University Duisburg-Essen, 2, Forsthausweg, 47057 Duisburg, Germany

[3]Bioinformatics Core, Luxembourg Centre for Systems Biomedicine, University of Luxembourg, 7, avenue des Hauts-Fourneaux, L-4362 Esch-sur-Alzette, Luxembourg

[4]Department of Life Sciences and Medicine, Faculty of Science, Technology and Medicine, University of Luxembourg, 7, avenue des Hauts-Fourneaux, L-4362 Esch-sur-Alzette, Luxembourg

Keywords: protein structure, archaea, methanogens, gut microbiome

Abstract

The human gastrointestinal tract is a habitat for diverse microbial taxa including archaea. Here, by establishing a unique *in silico* workflow, we resolve the identities of gut-specific archaeal proteins. We developed a strategy for protein functional annotation employing the most recent protein sequence- and structure-based approaches such as AlphaFold2, trRosetta, ProFunc and DeepFri. Where sequence-based approaches failed, our strategy led to the identification of archaeal proteins. Here, we identify a subset of archaeal gut-specific proteins involved in various defense strategies and virulence. For instance, *BglII restriction endonuclease* is the most commonly occurring unique protein involved in host genome defense against invading DNA. Furthermore, multiple adhesin-like proteins which likely facilitate archaeal cell adhesion to surfaces. We also resolve functions assisting degradation of toxic substances via propanediol utilization protein *pduA*. We identify a variety of glycosyltransferases involved in both N-linked glycosylation and O-glycosylation. In addition, we highlight several types of glycosyltransferases encoded by *Methanobrevibacter smithii* which may possibly be involved in the construction of O-glycosylated surfaces, beneficial for proliferation within the human intestine. Finally, we demonstrate the presence of proteins not only modulating interdomain horizontal gene transfer but also those possibly being the result of such genetic material exchange. In this context, we present evidence on horizontally transferred sporulation stage V proteins AE and AD from *Clostridium* species to *M. smithii.* In summary, our tailored *in silico* workflow allows the characterization of previously undescribed archaeal proteins which are specific to the gut environment and, thus, human microbiome function.

Importance

Our study characterizes gut-specific archaeal proteins in the human gastrointestinal tract. Our innovative *in silico* workflow leverages the latest protein structure-based tools which allows to bypass limitations introduced by traditional sequence-based methods. Here, we found archaeal proteins involved in host defense, surface adhesion, and those assisting in forming symbiotic relationships. Also, we discovered multiple archaeal glycosyltransferases involved in N-linked and less well understood O-glycosylation. Evidence of proteins regulating gene transfer and originating from horizontal gene transfer between bacteria and archaea was also found. In particular, these are sporulation stage V proteins possibly transferred to *Methanobrevibacter smithii* from Clostridium species. We believe that our study uncovers new roles for the common archaeon *Methanobrevibacter smithii* and contributes to the understanding of functional potential of gut-borne archaea and the growing recognition of their importance in the context of the human microbiome.

Introduction

In 1977, Carl Woese, George Fox and colleagues presented their pioneering research establishing a separate and previously unknown branch of the tree of life - the kingdom of Archaebacteria, later renamed the domain of Archaea [1]–[4]. The unravelling of the Asgardarchaea, which are truly different in their biochemical characteristics, led to the common understanding that the Eukarya originated within the tree of archaea [5]. Historically, archaea were only known to live in extreme environments. Over time, however, our knowledge of archaea has broadened to include an understanding of their broader importance, *cf.* [6], [7]. The ability of some archaea to thrive in extreme habitats and resist chemicals is thought to be due in part to their unique cell envelope structures. On the other hand, archaea living under moderate conditions possess unique metabolic capabilities and play important biogeochemical roles such as methanogenesis, anaerobic oxidation of methane and oxidation of ammonia [8], [9]. By pursuing diverse ecological strategies by metabolizing different organic and inorganic compounds for energy production, archaea have the characteristic of inhabiting a wide variety of environments [10]. Interestingly, archaea are also found in other

environments such as plants and the gastrointestinal tract of both humans and animals [11], [12], making them a small but important part of the gut microbiota. The proportion of archaeal organisms in the human gastrointestinal tract can be as high as 21%, with methanogenic archaea being the dominant group, accounting for 10% of all gut anaerobes [13], [14].

Archaeal proteins represent a means of adaptation to a wide spectrum of environments and the evidence of their unique biology. In recent years, the association of archaea especially to the human gut microbiome has expanded to several novel lineages [15]. Importantly, these host-associated taxa, such as the Methanomassilicoccales, have been reported to have beneficial impacts on human health [16]. Archaea have also been linked to the physiology of the skin [14], while *Methanosphaera stadtmanae* and *Methanobrevibacter smithii* have also being implicated in pro-inflammatory immune processes, thus affecting human physiology [17]. Given the current interest in the role of archaea in human health and disease, it is therefore crucial to elucidate the archaeal proteome, which is essential for understanding their functional potential. However, despite their importance, archaeal proteins are extremely difficult to study both experimentally and computationally. Lyu *et al*. have previously reported the potential for biotechnological applications in several archaeal genera [18]. These genera include *Methanobacteriales*, *Methanococcales*, *Methanopyrales, Methanomassiliicoccales, Methanocellales, Methanomicrobiales, Methanonatronarchaeales, and Methanosarcinales.* However, it should be noted that the current genetic toolboxes for any targeted genomic modifications are limited to mesophilic *Methanococcus* and *Methanosarcina*. A further complication arises from the fact that protein denaturation of archaeal proteins is challenging using standard proteomic methods [19]. Additionally, MS-based searches can be hampered by incorrect prediction of protein coding sequences, which is challenged by limited knowledge of ribosomal binding sites and promoter consensus sequences [20]. The biggest challenge remains unresolved - archaea are extremely difficult to isolate and culture under laboratory conditions, so metagenomic sequencing of environmental samples is becoming a promising solution for studying archaea and their relationships with other species and the habitats in which they live. Sequencing methods have allowed the tree of life to be extended, particularly in relation to archaea. Recently, Castelle *et al.,* Spang *et al.* and Zaremba-Niedzwiedzka *et al.* have further expanded the branches of the tree of life to include archaea [5], [21], [22] .

Nevertheless, uncultured archaea can mostly only be characterized via computational methods including protein prediction, annotation, and comparison with other proteins. Environmental genomics allows prediction of protein coding sequences (CDS), with subsequent derivation of functional information about them, thus characterizing functionalities of microorganisms in metagenomic samples. In metagenomics, metagenome-assembled genomes (MAGs) are reconstructed and characterized through several steps: a metagenomic sample is collected and its DNA is extracted and fragmented; following sequencing, short reads of fragmented DNA are assembled into longer genomic sequences – contigs, which are subsequently used to predict and functionally annotate encoded proteins. However, almost each step in this workflow hinders functional assignment due to a number of reasons. In the first place, metagenomic sequencing leaves many incomplete sequences, which compromises ORF prediction due to poor recognition by functional annotation tools or simply incorrect prediction [23], [24]. When ORFs are predicted, they must be annotated using functional annotation tools. Protein function annotation based on sequence comparison was historically used first and remains the most commonly used strategy. However, today this solution proves to be not particularly effective, especially in cases when analyzed proteins share distant homologies with existing proteins in databases [25]. Databases with archaeal proteins and functions are not entirely updated and the fall behind creates a two-fold challenge in sequence-based annotation of archaeal proteins. On one hand, Makarova et al. [26] report that archaeal ribosomal proteins L45 and L47, experimentally identified in 2011 [27] and pre-rRNA processing and ribosome biogenesis proteins of the NOL1/NOP2/fmu family characterized in 1998 [28] were not added to annotation pipelines by 2019 and were labelled as ‘hypothetical’. On the other hand, sequence similarity-based approaches are not able to capture relationships between proteins whose sequences diverged too far to point at their common ancestry when aligned with a known database protein. In this relation, unfortunately, archaea remain the least described domain of life and the lack of experimental data and well-curated and updated databases also result in spurious and incorrect protein annotations. In such situations, due to low sequence similarity, proteins remain unannotated as annotation algorithms aim at reducing

the number of false-positive assignments. Finally, the amount of such poorly annotated proteins results in reduced knowledge about functionality of a specific microorganism and their role in specific processes.

Over the last years, protein structure prediction for functional inference attracted much attention due to its potential for overcoming the sequence-function annotation gap [29]. As tertiary structure of a protein often remains more conserved than its sequence, similar structures of remote homologs can be used to circumvent limitations in functional assignment arising from sequence-based methods [30]. When methods based on sequence comparison report vague or missing annotations, examination of protein tertiary structure can reveal remote phylogenetic relationships and suggest functional roles [31]. In cases when a single protein has different and non-consistent annotations according to different databases (e.g., Pfam, KEGG, EGGNOG etc.), consideration of the respective protein structure could guide towards the actual function or at least narrow down the options. Up-to-date methods provide accurate computational approaches to address the bottlenecks in protein structure prediction and annotation. For instance, upon the input of a protein sequence AlphaFold2 regularly predicts its three-dimensional structure with nearly experimental accuracy thanks to the implementation of a neural network and training algorithms based on evolutionary, physical and geometric constraints of proteins [32]. trRosetta is a two-step approach powered by a deep learning network to predict inter-residue geometries and the Rosetta framework to guide structure prediction via applying restraints and running direct energy minimization [33].

Given all the pros and cons of sequence- and structure-based methods for protein function assignment, it seems reasonable to take advantage of both strategies. Accordingly, using as many strategies and sources as possible may benefit a more comprehensive insight into functions of uncharacterized proteins. Therefore, in this work we propose an integrated sequence- and structure-informed strategy for the annotation of “dark matter” archaeal gut proteins. We showcase the application of most recent protein structure prediction and annotation tools such as AlphaFold2, trRosetta, ProFunc and DeepFri. We find that structure-based approaches improve the annotation of archaeal gut-specific proteins where sequence-based consensus annotation methods struggle to provide confident results. Our results demonstrate the benefits of structure-based annotation of archaeal proteins within the human gut microbiome. We further characterize the functional domains and structures of 74 gut-specific archaeal protein clusters. These proteins facilitate the success of *M. smithii* within the human gut microbiome via factors associated with adaptation, existential and nutritional capacities. Collectively, our findings suggest that structure-based approaches are relevant and untapped methodology for annotating unique archaeal proteins.

Materials and Methods

Data selection

We used a total of 1190 archaeal MAGs derived from the Genomes from Earth’s Microbiomes (GEM) catalog [34] and the Unified Human Gastrointestinal Genome (UHGG) collection [35], and 285,835 bacterial MAGs from the UHGG collection, both accessed in November 2020. Genomes were extracted according to the available metadata of both datasets being filtered by taxonomy for archaea.

Gene prediction

For gene prediction we used Prodigal (V2.6.3: February, 2016) [36] on archaeal and bacterial MAGs from UHGG [35], while coding sequences (CDS) from the GEM catalog were downloaded from the provided source (https://portal.nersc.gov/GEM).

Protein clustering

Archaeal and bacterial proteins were each separately clustered with MMseqs2 (MM2) [37], [38] with the following parameters: --cov-mode 0 --min-seq-id 0.9 -c 0.9. In total, 1,873,481 archaeal and 87,282,994 bacterial proteins were clustered. Gut-specific archaeal proteins were used in the downstream analysis, representing 37% (707,754 proteins) of all predicted archaeal proteins at the time of metagenome collection (Dec 2021). A total of 61,123 MM2 clusters (out of 233,135) with ≥ 2 proteins and 1,967,480 MM2 clusters (out of 13,992,739) with ≥ 10 proteins were retained for the archaeal and bacterial datasets, respectively.

As mentioned above, archaeal gut-specific proteins were used, *i.e.,* proteins uniquely present in the human gut and not present in any other environment were selected. To solely resolve archaeal functions in the human gut microbiome, in our analysis we aimed to target only proteins occurring in the human gut and not having archaeal homologs in not human gut environments. This approach was chosen to identify unique functionalities of archaea in the gut environment. For this, all known proteins (from human gut and other environments) were clustered together and the following extraction criterium was applied: if a gut protein (extracted from UHGG and GEM) is clustered together with non-gut proteins (from GEM) by MMseqs2, it is discarded from the initial list of proteins. Sourmash [39], [40] was used on MM2 clusters to infer close and remote homologous protein groups. Only MM2 clusters where all proteins were annotated with KEGG were analyzed, which comprised 55,117 and 1,481,580 MM2 clusters for the archaeal and bacterial datasets, respectively. Sourmash allows the comparison of several protein fasta files using an algorithm-based on MinHash sketching. The distance matrix was produced using the following commands: *sourmash compute* (*--protein --input-is-protein -k 51), sourmash index* and *sourmash compare.* The distance matrix between MM2 clusters was processed, and hierarchical clustering was run with the Python clustering package scipy.cluster.hierarchy (https://docs.scipy.org/doc/scipy/reference/cluster.hierarchy.html). The threshold of 0.3**max_distance* was applied.

As a result, archaeal proteins were divided in two groups: those which shared KEGG KOs with bacterial proteins and archaeal proteins with unique KEGG KOs. The idea behind the division into such groups was the following: the first group was constituted to resolve archaeal functions in the gut in comparison to those fulfilled by bacteria; the second group was formed to extract unique functions of archaeal proteins in the human gut. Within the first group of archaeal and bacterial proteins with common KEGG KOs, annotation with the same KEGG KOs indicated sequence similarity of these proteins. Based on sequence similarity, two groups of archaeal proteins specific to the human gut environment were distinguished: archaeal-bacterial remote and close homologous proteins.

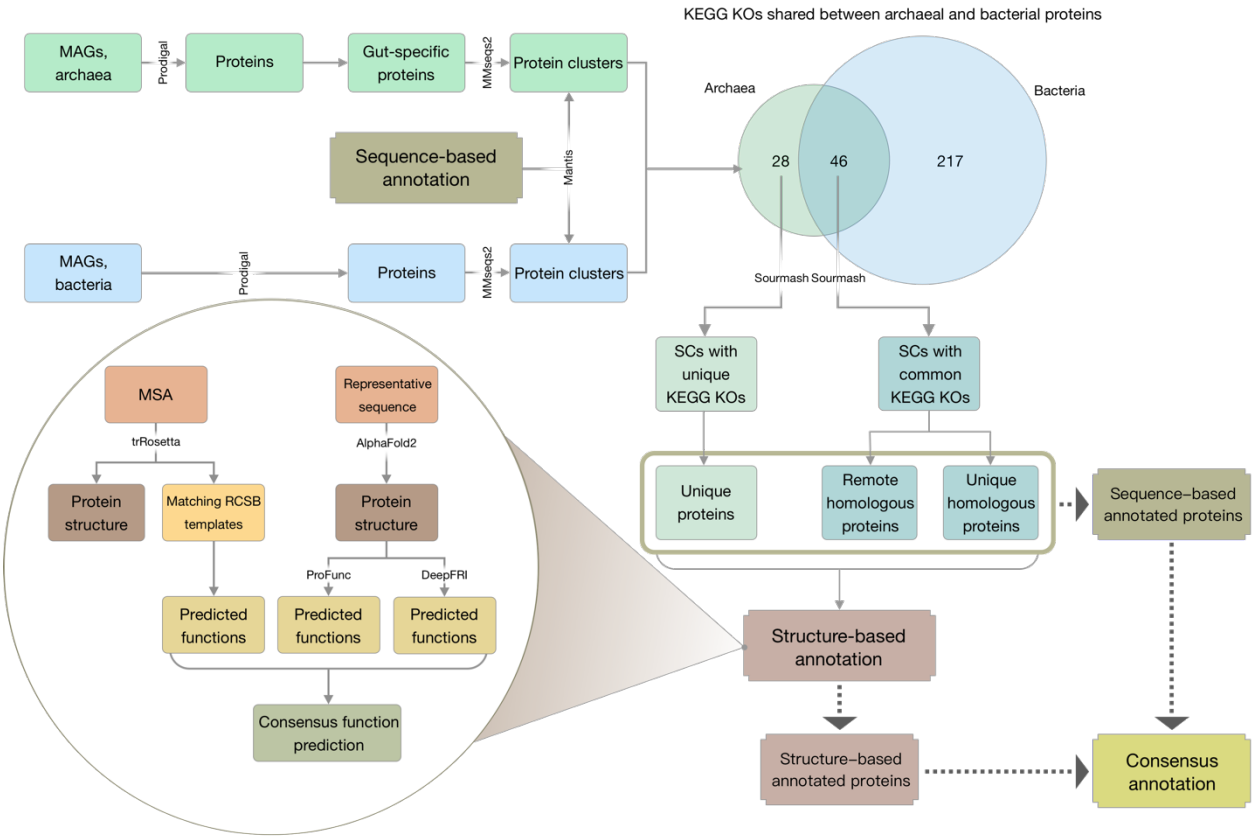

Figure 1. *Flowchart demonstrating major steps of the analysis. The Venn diagram demonstrates the amount of shared KEGG KOs assigned to archaeal and bacterial smash clusters.*

Protein functional assignment

Sequence-based annotation

Archaeal and bacterial proteins were annotated with KEGG orthologs, EGGNOG and Pfam using Mantis (release 1.5.4) [41]. KEGG orthologs were used as the main annotation resource, as they provide the most comprehensive functional resolution in comparison to EGGNOG and Pfam which were considered as auxiliary sources (see Supplementary Tables 1-2).

Structure-based annotation

For the functional assignment of each protein in clusters described above we used a combination of structure prediction and annotation tools. TrRosetta (TR) [33] and AlphaFold2 (AF) [32] were used as protein structure prediction tools. AF has demonstrated accuracy in prediction of three-dimensional structures of proteins outperforming other methods [42]. TR is outstanding in terms of rapid and accurate de-novo protein prediction [33]. For each tool, the predicted protein structure was then annotated separately. For targets with detectable homologous matches in Protein Data Bank (PDB), TR provides those hits, and the respective molecules reported in templates with the highest *identity* and *coverage*

features were used for protein structure annotation. By default, TR uses a template for prediction and provides a hit only if it satisfies the condition: confidence > 0.6, e-value < 0.001 and coverage > 0.3.

For instance, upon the input of a protein sequence AF regularly predicts its three-dimensional structure with nearly experimental accuracy thanks to the implementation of a neural network and training algorithms based on evolutionary, physical and geometric constraints of proteins [32]. TR is a two-step approach powered by a deep learning network to predict inter-residue geometries and the Rosetta framework to guide structure prediction via applying restraints and running direct energy minimization [33].

The model produced by AF was submitted to the ProFunc (PF) [43] web server. *Sequence search vs existing PDB entries* and 3D functional template searches sections from the ProFunc report were used for structure-based protein annotation. Sequence-based annotation was acquired from the section *BLAST search vs Uniprot*. The search is performed via running Blast of the target sequence against (1) protein sequences with known structures in the PDB to first extract obvious matches and (2) the UniProt database of protein sequences to access gene neighbourhoods as it provides clues to the target protein's role. Sequence and structure matches were selected according to the reported highest possible likelihood of being correct as follows: *certain matches (E-value <10-6)*, *probable matches (10-6 < E-value <0.01)*, *possible matches (0.01 < E-value < 0.1)*, and *long shots (0.1 < E-value < 10.0)*. Only certain matches were used for the functional assignment. DeepFri [44] was used as a 'hint' tool as it provides rather broad and general descriptions, which are however useful to either verify or refute the suggestions reported by AF and/or TR. DeepFri predictions with a certainty score >0.7 were taken into consideration. Our combined approach is a systematic agglomeration of methods allowing for better resolution of functions that pose challenges when applying traditional sequence-based methods.

Annotation selection

If TR- and AF-based annotations provided consistent results, the consensus was used as the final annotation of the protein function. However, when the reports gave differing results, we prioritized the result with higher certainty. For instance, when the confidence of the predicted model by TR was *very high* and template matches were provided, and AF-based ProFunc reported a match with a lower confidence (anything but *certain match*), the template hit by TR was be used as the primary source for the annotation. For convenience, the relationship between PF likelihood and TR TM-scores is shown in the Table 1. Similarly, any protein with a TR template match was considered as more reliable than an annotation with the *long shot* likelihood. In cases when there were no 3D functional hits, the annotation by TR was given priority. In cases when PF and TR provided annotations with the same level of significance/likelihood, the protein structure with highest *coverage* and *identity* were chosen. In this case, we mean coverage as the *coverage* feature as in TR and $\frac{longest\ fitted\ segment}{query\ sequence\ length}$ as in PF, and for identity we take *identity* as in TR and *percentage sequence identity* as in PF.

PF likelihood	PF e-value	TR significance score	TR TM-score
certain match	$<10^{-6}$	very high	>0.7
probable match	<0.01	high	>0.5
possible match	<0.1	medium	>0.4
long shot	<10	low	>0.3

Table 1. *Relationships between ProFunc (PF) likelihood and trRosetta (TR) TM-scores.*

Protein relative annotation abundance calculation

Relative occurrence or frequency of protein functions in the unique and the homologs groups was calculated. The measure was calculated as a relation between a number of proteins with a specific KO to the overall number of proteins of bacterial or archaeal proteins. For example, the relative occurrence of unique archaeal proteins annotated as *K20411* (smash cluster 1) is: $\frac{\#\ of\ proteins\ with\ K20411}{total\ number\ of\ archaeal\ proteins} * 10^6$.

Gene expression analysis

Gene expression was verified by mapping metatranscriptomic reads of faecal samples of healthy individuals and patients with T1DM [45]. The reads were mapped to fasta sequences of genes of interest using bwa mem [46] and mapping files were processed with SAMtools [47]. Mosdepth [48] was used to calculate mean read coverage per gene of interest.

Horizontal gene transfer analysis

Horizontal gene transfer (HGT) analysis was conducted using metaCHIP [49] on all *Methanobrevibacter smithii* MAGs available in included datasets using. Several *Methanobrevibacter_A oralis* genomes were also included for the comparison of the number of HGT events.

Gene synteny analysis

pyGenomeViz [50] was used to build gene synteny for all archaeal genes of interest. Gene coordinates predicted with Prodigal were used as an input. An interval of 10Kbp up- and downstream of the gene of interest was selected from the protein predictions. KEGG KOs were allocated based on the sequence-based annotations generated using Mantis [41]. The commercial *M. smithii* strain DSM 861 was used to computationally study the presence of genes from flanking regions of selective genes in a real archaeal culture.

Accuracy estimation

Any annotation was considered to be correct whenever the protein was assigned with a function clearly related to archaea and described in related literature. The rest of proteins were considered as incorrectly annotated. Following

this initial step, *sensitivity* was calculated as $\frac{N_{str}}{N_{str}+N_{seq}}$, *specificity* was calculated as $\frac{N_{seq}}{N_{seq}+N_{str}}$, *positive likelihood ratio (PLR)* - $\frac{sensitivity}{1-specificity}$, negative likelihood ratio (NLR) - $\frac{1-sensitivity}{specificity}$, where N_{seq} and N_{str} are the numbers of correct sequence- and structure-based annotations, respectively.

Phylogenetic analysis

In order to build phylogenetic trees for selective smash clusters (SCs), additional similar sequences were added from Uniprot [51] via running BLAST [52] on the consensus sequences representing SCs of interest. Further, Uniprot sequences and SC sequences were used to build trees. Multiple sequence alignments (MSAs) were built using MAFFT [53] and trimmed with BMGE [54] using BLOSUM95 similarity matrix and the default cut-off 0.5. Maximum Likelihood phylogenetic trees were built with IQ-TREE [55] and visualized with R library *ggtree* [56].

Data and code availability

Microbial MAGs from the UHGG collection are available from the MGnify FTP site at http://ftp.ebi.ac.uk/pub/databases/metagenomics/mgnify_genomes/, MAGs from GEM catalog are accessible at https://portal.nersc.gov/GEM/.

A description of the analyses including pre-processing steps along with the scripts for the main analysis, archaeal gut-specific unique and homologous smash clusters and synteny plots can be found at GitLab: https://gitlab.lcsb.uni.lu/polina.novikova/archaea-in-gut.

Results

Characterizing unique archaeal and bacterial proteins specific to the gut

In order to comprehensively characterize protein functions of archaea specific to the human gut, not only archaeal but also bacterial sequence data was used. The latter was used as a reference to distinguish unique proteins encoded and transcribed by archaea as well as archaeal proteins with homologs to bacterial ORFs. Proteins were assigned functions with Mantis using Pfam and EGGNOG [57].

In this study, we used KEGG KOs for protein sequence-based annotation as they offer the most comprehensive functional resolution. KEGG annotated protein clusters were subjected to Sourmash and subsequent hierarchical clustering, which merged them into larger groups designated as smash clusters (SCs) (Fig. 1). As a result, there were three types of SCs of archaeal and bacterial clusters which shared KEGG KOs: 10 close homologous SCs each containing archaeal and bacterial protein clusters, 36 remote homologous SCs with exclusively archaeal protein clusters, and 217 clusters with exclusively bacterial protein clusters. The term homology was used here with respect to bacterial proteins with the same KEGG KO, and proteins were considered to be close homologs when in addition to shared KEGG KOs archaeal and bacterial proteins were clustered by Sourmash. Remote homologs were those which shared KEGG KOs with the bacterial counterparts but were not clustered together. The second group of archaeal proteins with unique KEGG KOs contained 28 protein clusters. Final functional annotation for each SC was concluded using the information provided by the structure- and sequence-based annotation methods, and the analysis of protein structures revealed a handful of functions not available via annotation with sequence-based methods. The summary of

annotations provided by trRosetta, AlphaFold2.0-based ProFunc report, and DeepFRI were provided in the Supplementary materials.

Accuracy estimation

We estimated the accuracy of our structure-based strategy compared to sequence-based methods (Tab. 2). Accuracy metrics were calculated for the annotation of unique and homologous protein clusters separately.

	Unique proteins	Homologous proteins
sensitivity	73.3%	66.7%
specificity	58.5%	56.1%
positive likelihood ratio (PLR)	1.77	1.51
negative likelihood ratio (NLR)	0.46	0.6

Table 2. *Accuracy estimation for structure-based functional assignment over sequence-based methods for unique and homologous protein clusters.*

Occurrence and expression of archaeal proteins

The relative abundance of the identified proteins and the related processes in both groups was determined (Fig. 2). Interestingly, a few types of processes in which archaeal proteins were involved were apparent amongst the unique and homologous proteins. Among the unique archaeal protein clusters, the most commonly identified functions relate to adaptation to changing environments (smash clusters *u2, u3, u8, u23, u24*), protection mechanisms (smash clusters u1, u15, u22), RNA and DNA regulation (u4, u12), energy metabolism (*u7, u14, u25*) and cell wall integrity and maintenance (smash clusters u5, u6) (Tab. 2). In the homologous archaeal protein clusters, there were three types of abundant processes: those more common in unique archaea clusters (smash clusters *h1-6, h7, h8, h10, h13, h14, h15-18, h19, h21, h22, h25, h26, h29, h37, h33, h35, h38, h39, h41, h42, h43, h44, h45*), those more common in purely bacterial clusters (smash clusters *h9, h20, h32, h34, h40*) and clusters with almost equal distribution of archaeal and bacterial proteins (smash clusters *h11, h27, h31*) (Tab.3). Overall, the most frequently identified functions relate to adaptation to changing environments, various protection mechanisms (defense against foreign DNA, toxic compounds, oxidative stress; DNA repair), several energy metabolism pathways, processes providing structural integrity to the cell and others. In addition, analysis of faecal metatranscriptomic data demonstrated that the majority of encoded genes were indeed transcribed (Fig. 2). It was found that not all identified genes were highly expressed in relation to their gene abundance. However, two unique and 19 homologous genes which did show higher levels of expression were identified. Those with a known function included genes associated with adaptation (*h39*), defense (*h25, h35, h43*), DNA/RNA regulation (*h29, h30, h38*), energy metabolism (*h27, h28, h40*), the rest were unannotated (*u10, u18, h9, h13, h33, h34, h36, h37, h42, h44, h45*). The reason for such a phenomenon might have been that the earlier mentioned genes and the associated functions were more significant for the success of cells in the gut.

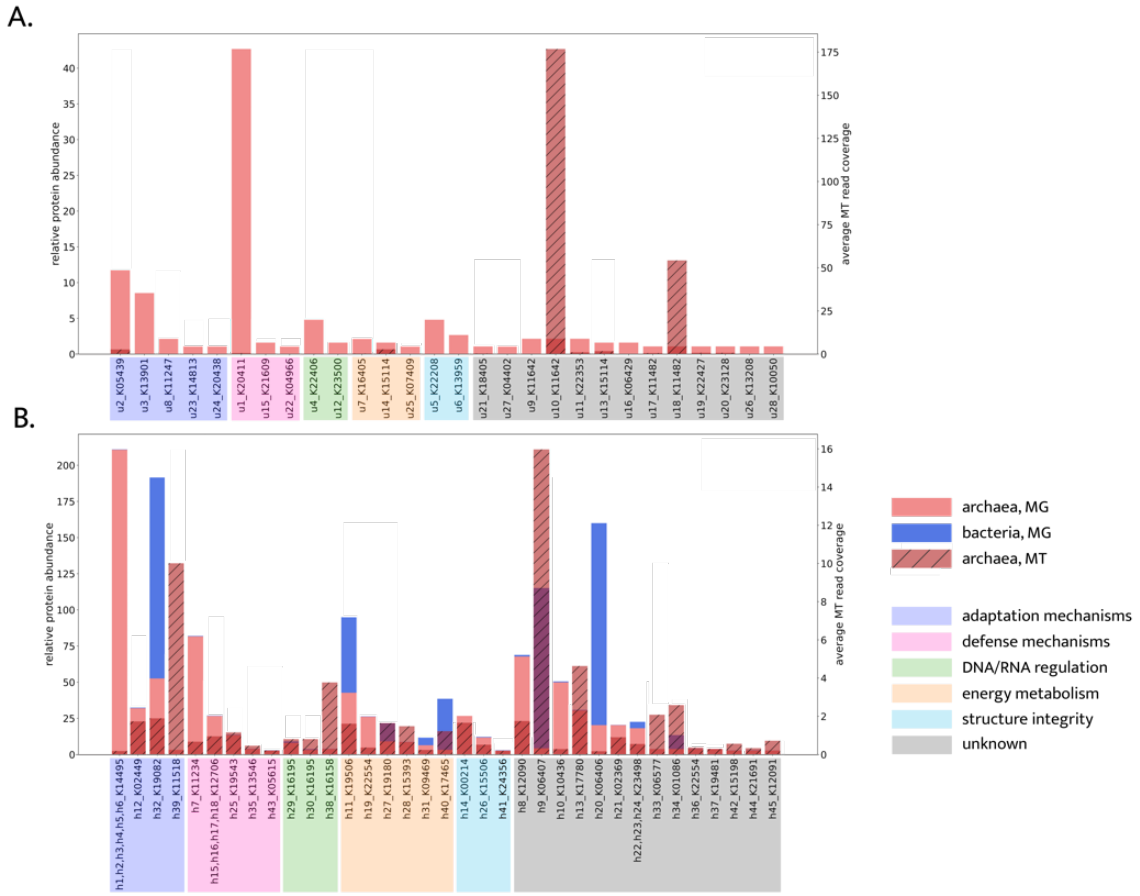

Figure 2. *Relative occurrence and average metatranscriptomic read coverage of proteins in the A. unique and B. homologous groups of clusters with archaeal proteins.*

We supplemented each protein functional assignment with an overview of the functions and reviewed its importance in the context of archaeal biology where possible (see Supplementary Materials). We only considered proteins that were annotated by both AF, PF, and TR. The resulting consensus structure annotations were shown in Supplementary Tables 1-2 for unique and homologous archaeal protein clusters, respectively. We assessed the disparity between sequence-based and structure-based consensus annotations by comparing the reported functions. For example, cluster *u1* was identified as proline-rich protein 5 with KEGG and double-stranded D endodeoxyribonuclease activity with EGGNOG. However, the consensus structure-based annotation identified Type II restriction endonuclease BglII.

Stability of the gene structure in *Methanobrevibacter smithii* genomes

The human gut archaeome is dominated by the presence of methanogenic archaea, and the most abundant of these microorganisms is *M. smithii* [54]. As reported earlier, *M. smithii* genomes have been implicated in both the health and disease of humans. It was yet unclear how many of the identified genes recovered in our analyses are conserved archaeal genes. Therefore, to elucidate this further, we performed an analysis of possible horizontal gene transfer

(HGT) to understand the stability of the gene structure in *M. smithii* genomes (Fig. 3). As all the identified and analyzed above proteins belonged to *Methanobrevibacter smithii*, we used all available MAGs in this analysis. Additionally, we used genomes of the *Methanobrevibacter_A oralis* species in order to provide a reference on how HGT occurred between genomes of species of the same genus. As a result, we observed that HGT events were elevated between genomes of two *M. smithii* species: *Methanobrevibacter_A smithii* and *GCF_000016525.1* (based on GTDB classification). 2.6% of the MAGs (n=27) were found to have exchanged up to10±3 genes with other MAGs.

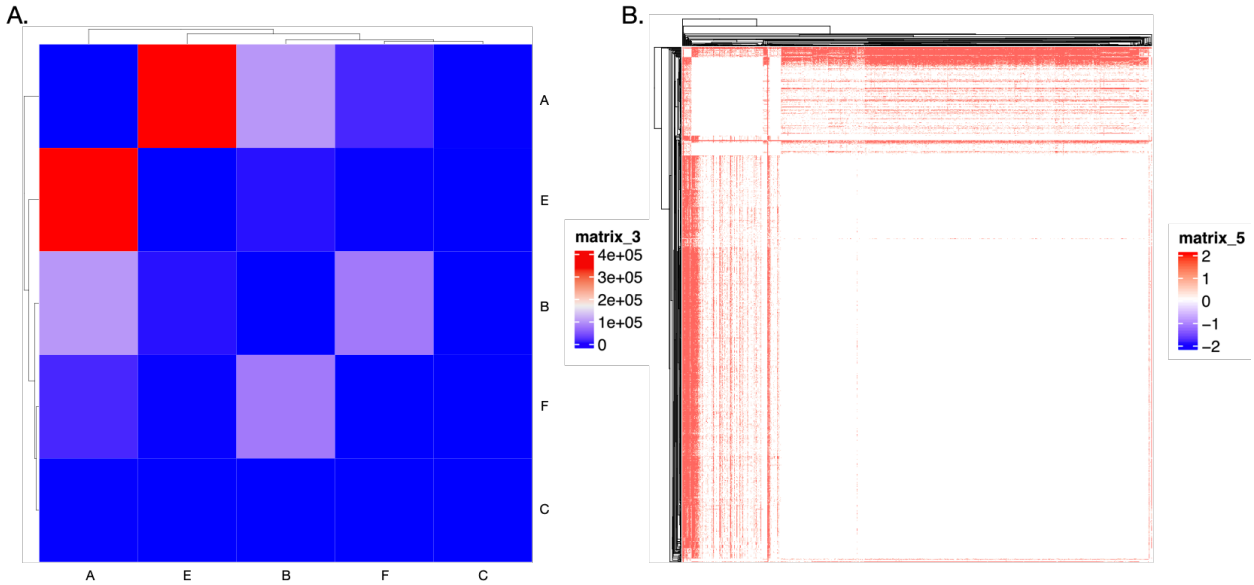

Figure 3. *Heatmaps demonstrating the intensity of HGT events between M. smithii genomes.* ***A****, HGT between taxonomic groups named as follows: A - Methanobrevibacter_A smithii, B - Methanobrevibacter_A smithii_A, C - Methanobrevibacter_A oralis, E - GCF_000016525.1, F - GCF_002252585.1;* ***B****, HGT events between individual genomes of same groups.*

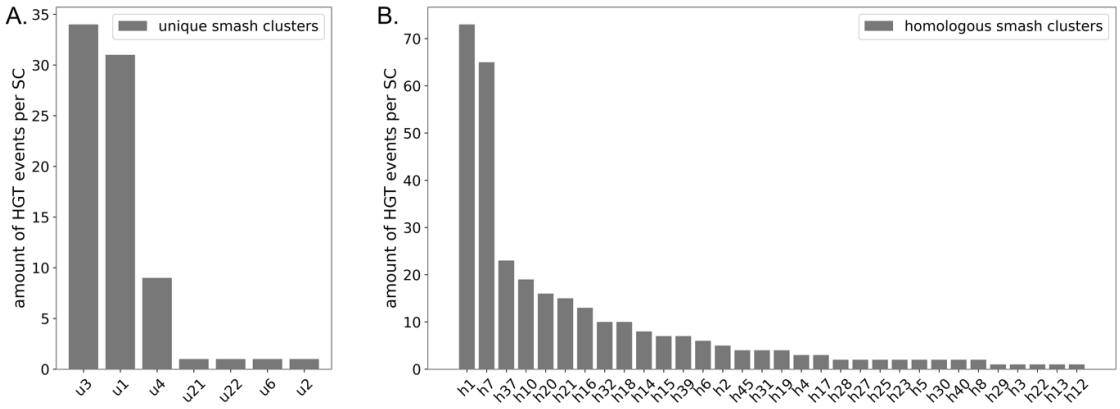

Figure 4. *Rank abundance plots for the number of horizontal gene transfer events in* ***A****, unique and* ***B****, homologous smash clusters.*

Importantly, substantiating the notion of genomic stability, we found that the flanking regions of the genes of interest were also fairly conserved among *M. smithii* genomes. Using analyses of synteny, we demonstrated representative examples of conserved synteny for genes encoded on different *M. smithii* genomes. Figure 5 demonstrates gene synteny for several examples of archaeal proteins, homologous and unique archaeal proteins with no known function (*h24* and *u18*, respectively).

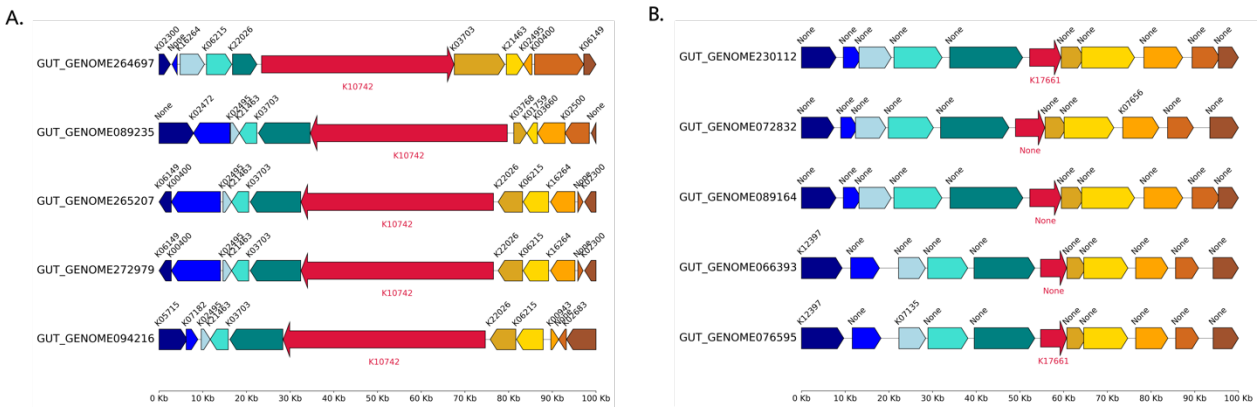

Figure 5. *Representative examples of archaeal gene synteny for specific genes:* ***A****, a homologous archaeal protein with no known function (h24), and* ***B****, a unique archaeal protein with no known function (u18).*

Analysis of stage V sporulation proteins AE and AD

Among the identified archaeal gut-specific proteins, we identified stage V sporulation proteins AE (spoVAE) and AD (spoVAD) (*h9* and *h20*). Incidentally, the MT data suggested that *h9* is expressed at a higher level than *h20*. To investigate the presence and genomic context of sporulation genes in archaeal genomes, we analyzed spoVAE and spoVAD. 250 SpoVAE and 250 spoVAD bacterial protein sequences were extracted from Uniprot. 12 and 38 spoVAE and spoVAD proteins, respectively, were reported to be found in environmental samples, while the rest of sequences

were encoded in isolate genomes. The corresponding isolate genomes belonged to different classes of Firmicutes phylum.

Phylogenetic trees built based on proteins from *h9* and *h20* together with similar proteins available in Uniprot (Supp. Fig. 1, 2) demonstrated that proteins from *h9* and *h20* are phylogenetically remote from other sequences and cluster together as separate branches in both cases. Proteins from *h9* and *h20* are compositionally distinct from other sequences which can be resolved through their MSAs with bacterial sequences from Uniprot.

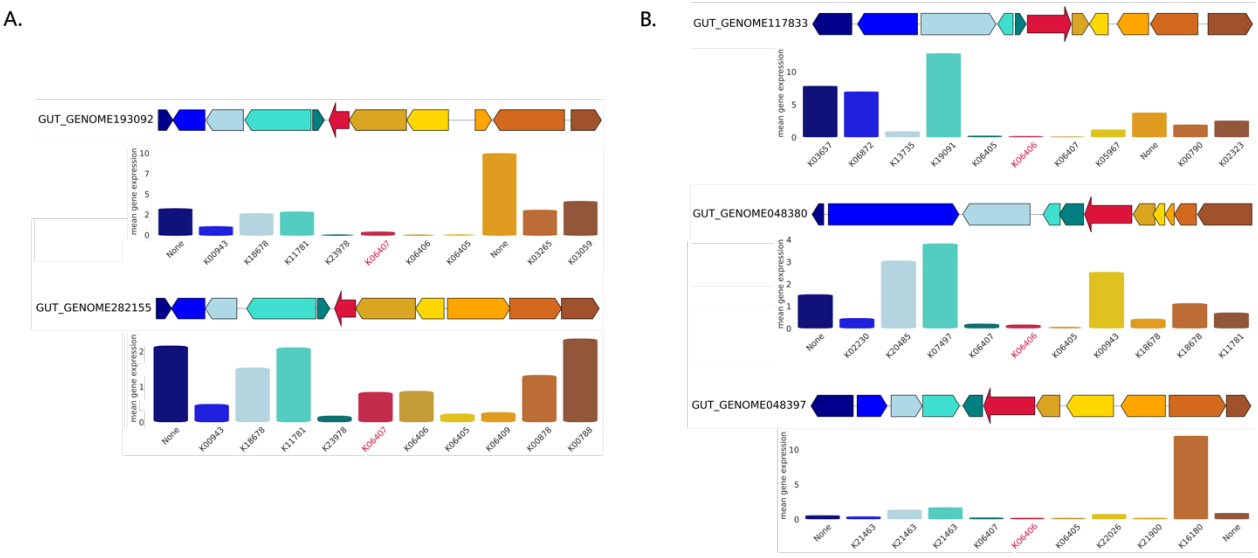

Figure 6. *Gene syntenies for sporulation stage V genes AE and AD from their respective smash clusters* ***A****, h9, and* ***B****, h20. Gene expression of target genes (spoVAE and spoVAD, in red) as well as genes from flanking regions are demonstrated below each sequence and are colored correspondingly. Genes with key archaeal functions:* ***A****, pyrimidine metabolism (K18678, phytol kinase), methane metabolism (K11781, 5-amino-6-(D-ribitylamino)uracil–L-tyrosine 4-hydroxyphenyl transferase) and thiamine metabolism (K00878, hydroxyethylthiazole kinase; K00788, thiamine-phosphate pyrophosphorylase);* ***B****, pyrimidine metabolism (K22026, nucleoside kinase; K18678, phytol kinase) and methane metabolism (K11781, 5-amino-6-(D-ribitylamino)uracil–L-tyrosine 4-hydroxyphenyl transferase).*

The analysis of gene synteny for spoVAE and spoVAD genes (Fig. 6) demonstrated that sporulation genes are grouped in operons (K06405, K06406, K06407). Moreover, the flanking regions around sporulation genes in the archaeal genomes include genes with key archaeal and even methanogenic functions (Fig. 6). In addition, the flanking regions of both spoVAE and spoVAD genes are also encoded by the genome of the representative *M. smithii* isolate DSM 861 (Fig. 7). However, in contrast to our MAGs, the isolate's genome did not encode the spoVAE and spoVAD genes.

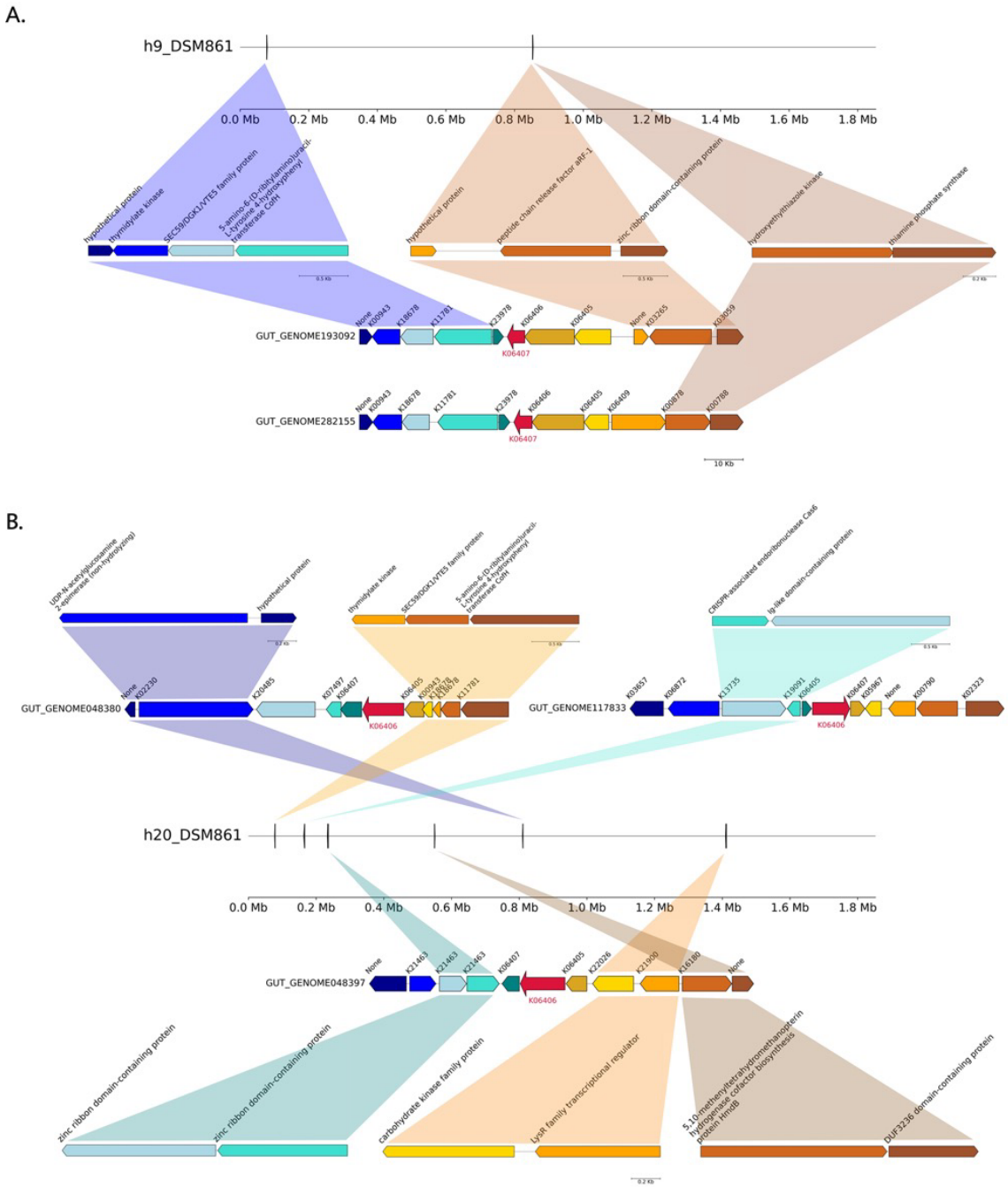

Figure 7. *Genomic context of the archaeal flanking regions up- and downstream of the A, spoVAE and B, spoVAD gene clusters in the M. smithii strain DSM 861.*

As we speculate that the spoVAE and spoVAD genes were acquired by the archaea via HGT, we also analyzed the syntenies with bacterial genomes from the clusters *h9* and *h20* derived from the phylogenetic trees (Supp. Fig. 3-4).

The flanking regions of spoVAE and spoVAD genes in the bacterial genomes include genes mediating and facilitating HGT, such as site-specific DNA recombinase (K06400) encoded upstream from spoVAE and type IV pilus assembly proteins (K02662, K02664) encoded downstream from spoVAD. Moreover, flanking regions around spoVAD in archaeal and bacterial genomes encode a common gene K05967 (Supp. Fig. 3-4). Finally, bacterial genomes encoding genes from clusters *h9* and *h20* are organisms from the Firmicutes phylum: Clostridium sp. CAG-302 and CAG-269 for *h9* and *h20*, respectively.

Discussion

In this work, we demonstrate how the usage of both sequence- and latest structure-based annotations benefits the functional assignment of archaeal proteins. This approach incorporates state-of-the-art sequence- and structure-annotation tools with a clear strategy and criteria for the selection of annotations. Indeed, methods based on sequence similarity prove effective when running annotations on a pair of proteins with high sequence similarity. Given that little to nothing is known about archaeal proteins; high sequence similarity may not fully represent the reported function. Moreover, traditional sequence-based methods are not reliable when proteins with generally low sequence identities (<~30% or in the “twilight zone” [58]) are analyzed, as at such levels functional similarity cannot be inferred. Unfortunately, public databases lack experimentally validated archaeal sequences with known functions (2157 archaeal vs ~40,000 bacterial and ~253,000 eukaryotic proteins, UniProtKB Nov 2022). Therefore sequence-based protein annotations can only be applied on a limited number of archaeal proteins. At the same time, we demonstrated here that recently developed structure prediction methods allow protein structure prediction and annotation of proteins without reference sequences with high sequence similarities (i.e., templates in template-based modelling by trRosetta, and search against existing PDB structures by ProFunc). Thus, protein annotation based on structure prediction is advantageous compared to sequence similarity-based approaches as it can capture similarity between proteins even beyond close sequence similarity.

A combination of sequence- and structure-based methods for protein functional assignment allows the discovery of functions for proteins with no close homologs in public databases. Having estimated the accuracy of our structure-based strategy, we conclude that such an approach is more likely to report correct archaeal protein functional assignments rather than conventional sequence-based methods. Moreover, our analysis shows that correct archaeal annotations are more likely to be assigned to the group of unique proteins. However, in our analysis, for only ~15% of archaeal protein clusters sequence- and structure-based annotations are matching. Majority of these have bacterial homologs and only 1 unique archaeal protein clusters has matching annotations when annotated with sequence- and structure-based methods. This result once again points at the importance of the protein structure prediction and annotation approach when working with uncharacterized archaeal proteins. This is not only due to the lack of experimental characterization of protein function in general but also because of insufficient functional annotation of experimentally validated proteins. However, even structure-based functional assignment is not always possible. In our work, a substantial portion of relevant proteins remains unannotated structurally (27 out of 73, ~37%), however we find KEGG annotations for all the 27 structure-unannotated proteins. However, literature search suggested that these KEGG annotations do not provide reasonable or explainable functional assignments for the majority of such unannotated proteins. For instance, protein *h39* is annotated as Mitochondrial import receptor subunit TOM40 with KEGG; however, this is a eukaryotic protein essential for protein import into mitochondria. On the other hand, with our structure-based strategy the very same protein appears to be an intimin/invasin-like protein whose role is evident in the context of archaeal biology. Another example is protein *u8* which is annotated as Endophilin-A, a eukaryotic protein with membrane curvature-sensing and -inducing properties [59], while based on the predicted structure it is similar to PilC, type IVa pilus, a subunit of a prokaryotic adhesion filament present in many prokaryotic domains [60]. One could argue that presence of a eukaryotic protein is not surprising in archaea as the former could be an early representation of that eukaryotic protein, especially assuming the fact that these two proteins modulate structural recognition and attachment. However, annotation of a protein to its evolutionary homolog from a different kingdom is not precise enough when one’s

goal is the most possible comprehensive resolution of protein functions. Here, our structure-based annotation benefits the goal and provides the closest possible functional assignment in the context of intestinal archaea. Therefore, we analyzed archaeal functions based on available results to the extent possible, however, keeping in mind the unavailability of functions for ~37% of proteins.

Due to the unavailability of extensive information on annotations, we focus on understanding the extent to which some of the identified genes were conserved. We specifically assess this in the most abundant human gut archaea, *M. smithii*, due to its reported role in health and disease [14], [61], [62]. Interestingly, the available MAGs showing HGT events were sampled in different locations (Austria, France, UK, US). Given the *in silico* analyses performed in our study for HGT events, their propensity to exchange parts of their genomes may be explained by the similarity of their respective local environments (individuals' diet, lifestyle etc.). On the one hand, if these MAGs were exposed to similar diets or stresses, it is likely that their evolution via HGT events may have followed similar trajectories. On the other hand, moderate to none HGT events among the majority (97.4%) of available *M. smithii* genomes points to their overall genomic conservation and stability. Low HGT could be explained by the fact that respective MAGs were sampled from people living in diet- and lifestyle-wise similar conditions, *i.e.*, so called "Westernized microbiomes". In order to further investigate it is important to consider microbiomes of individuals from non-Westernized sites in future analyses.

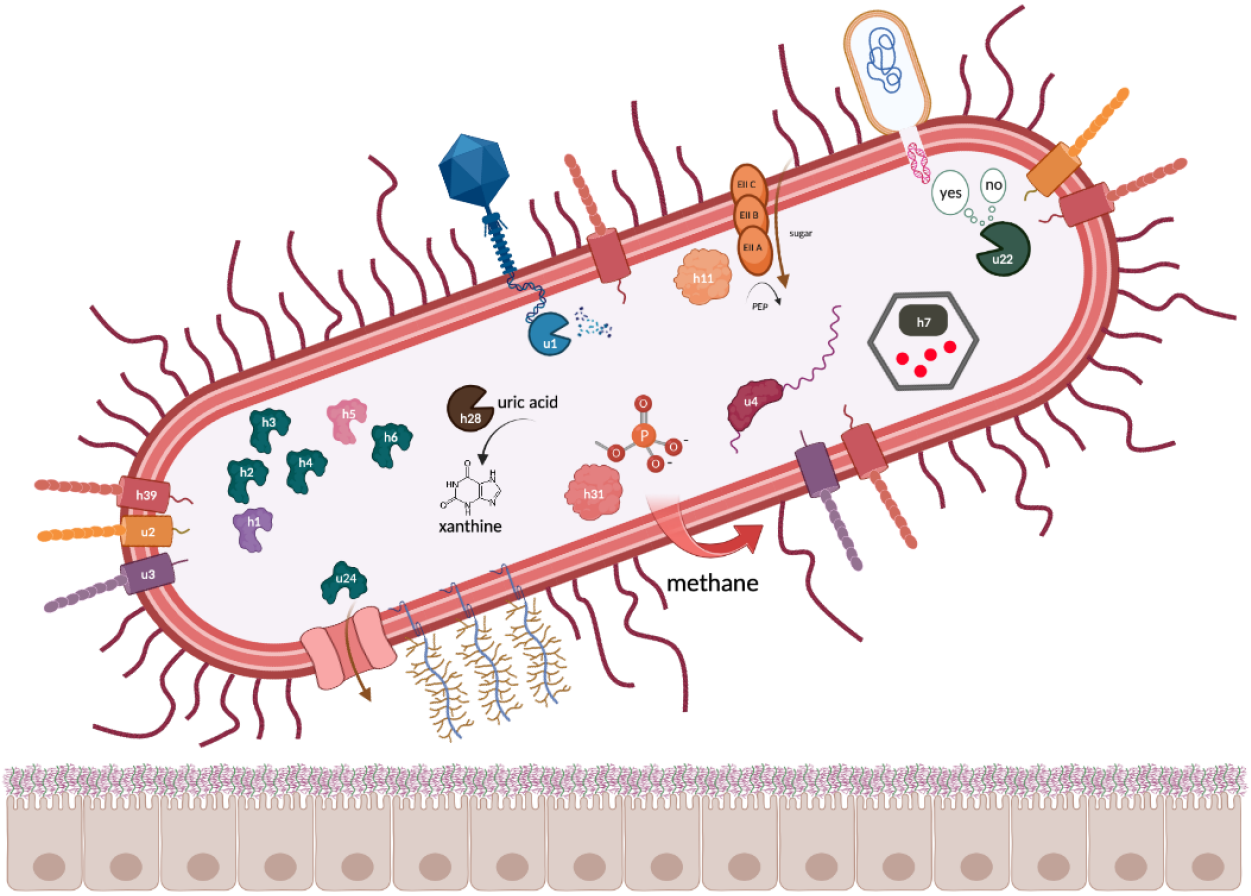

Figure 8. *Diagram highlighting gut-specific archaeal proteins with described functions:* ***u1*** *- Type II restriction endonuclease BglII,* ***u2*** *- Intimin/invasin-like protein with a Ig-like domain,* ***u3*** *- Intimin/invasin-like protein,* ***u22*** *- Type I restriction-modification EcoKI enzyme, specificity subunit,* ***h1*** *- 4-amino-4-deoxy-L-arabinose transferase or related glycosyltransferases of PMT family,* ***h2,3,4,6*** *- Dolichyl-phosphate-mannose – protein mannosyltransferase 1,* ***h5*** *- Dolichyl-diphosphooligosaccharide-protein glycosyltransferase subunit STT3B,* ***h7*** *- Propanediol utilisation protein pduA,* ***h11*** *- Phosphoenolpyruvate-dependent PTS system, IIA component,* ***h28*** *- Transthyretin-like protein,* ***h31*** *- 2-aminoethylphosphonate-pyruvate aminotransferase.*

We describe functions of the available proteins and distinguish the most frequently occurring processes. Our results demonstrate that homologous proteins from clusters *h1, h2, h3, h4, h5* and *h6* constitute the most prevalent archaeal gut-specific process and thus deserve particular attention (Supp. Tab. 2). All proteins of this clusters are types of glycosyltransferases (GTs) acting on various sugar donors via transferring sugar residues from donors to acceptor molecules through establishing natural glycosidic linkages. GTs are also described to be the most common enzyme class in archaea [63], yet hyperthermophilic archaeal species encode fewer glycosyltransferases than nonhyperthermophilic archaea [64]. Gene synteny analysis reveals that the flanking regions of the abovementioned genes also contain various transferases (see Supplementary Material, *h1, h2, h3, h4, h5, h6*). In our work, the most prevalent unique archaeal protein is from the group of GTs from cluster *h1* and is a *4-amino-4-deoxy-L-arabinose (L-Ara4N) transferase* (Supp. Tab. 2). The mentioned protein was predicted as L-Ara4N transferase with both tools with *very high* confidence score, therefore we can assume that the homology behind this annotation is robust. This protein is reported to transfer a L-Ara4N unit to the lipid A, a complex glycolipid on the surface of most Gram-negative bacteria, which is essential for protection from environmental stress, symbiosis, virulence and resistance against antimicrobial

448 activity [65], [66]. Whether it encodes for similar functions in archaea, given their unique cell wall composition [67],
449 which is unlike Gram-negative bacteria, needs to be validated in future studies. Moreover, one of the six GTs is a
450 dolichyl-diphosphooligosaccharide – protein glycosyltransferase subunit STT3B (*h5*) and functions as an accessory
451 protein in N-glycosylation and provides its maximal efficiency [68]. Indeed, N-linked glycosylation is known as a
452 ubiquitous posttranslational modification in archaea, however this fact has only been confirmed in a limited number of
453 species. For example, N-linked glycans have been observed in methanogens such as *M. voltae [69], M. maripaludis*
454 [70], *M. fervidus* [71]. Although there is yet no evidence about N-linked glycosylation pathways in *M. smithii* or other
455 *Methanobrevibacter* species, multiple related studies point to its probable existence. Indeed, archaeal N-glycosylation
456 plays an important role for the viability and adaptivity of archaeal cells to external conditions such as high salinity [72],
457 [73], elevated temperatures [71] acidic environment [74], it also maintains structural integrity of cells [75], [76]. Similarly,
458 N-glycosylation of archaeal proteins might benefit species' viability and adaptation to the intestinal conditions.

459 Three out of six identified GTs (clusters *h2, h3, h4, h6*) are dolichyl-phosphate-mannose – protein mannosyltransferases
460 1 (POMT1), which are responsible for O-linked glycosylation of proteins in eukaryotes. However, O-glycosylation
461 occurring in archaea is not well understood. There is limited evidence about O-glycosylation in archaea, yet it is known
462 that threonine-rich regions of S-layer glycoproteins supplement galactose–glucose disaccharides in *Hbt. salinarum* and
463 *Hfx. Volcanii* [77], [78]. However, another O-glycosylation associated protein, *polypeptide N-*
464 *acetylgalactosaminyltransferase*, was found in the subset of unique archaeal proteins (unique cluster u24). O-
465 glycosylation is mainly found in eukaryotic cells. In particular, mucins are examples of O-glycosylated glycoproteins.
466 Mucins are the main structural components of the intestinal mucus and contribute to the interaction between gut
467 microorganisms and epithelial surfaces [79]. On the other hand, Samuel *et al.* demonstrated that *M. smithii* decorates
468 its cellular surface with sugar residues mimicking those present in the glycan landscape of the intestinal environment
469 [62]. Taken together, the fact that *M. smithii* has human mucus- and epithelial cell surface-associated glycans and that
470 archaeal gut species possess coding potential for enzymes which function in O-linked glycosylation suggests that *M.*
471 *smithii* cells could indeed mimic surfaces of eukaryotic cells of the intestinal mucus. O-glycans do not only play a role
472 on the structural level of individual proteins, but also O-glycans might modulate protein interactions between other
473 proteins as well as cells, which in turn benefits cell trafficking and environmental recognition [80]. O-glycosylated
474 proteins are traditionally known to be localized extracellularly including cell surfaces and luminal compartments [81].
475 However, Torres and Hart demonstrated that *O*-linked modification is dynamic, reversible and short lived and O-linked
476 GlcNAc residues often occur intracellularly [82]. Therefore, O-glycosylated proteins might also have regulatory
477 functions. With regards to the former statement, Van den Steen *et al.* report that two main proteins implicated in
478 Alzheimer's disease – tau and beta-amyloid precursor protein – are known to be O-glycosylated and increased in
479 disease [83].

480 Besides POMT1 and related GTs, there are other archaeal proteins which reflect adaptation to external conditions and
481 cell survival (homologous clusters *h7, h12, h15, h21, h25, h35* and *h38*). Interestingly, all these proteins are
482 considerably enriched in archaea rather than bacteria based on their relative occurrence. For example, the second most
483 frequently occurring archaeal protein (h7) in the group of homologous proteins is structurally similar to propanediol
484 utilization protein *pduA*. *PduA* has yet been described only in bacteria. *PduA* is a major shell protein that is essential
485 for bacterial microcompartment assembly and required for the formation of polyhedral organelles involved in coenzyme
486 B12-dependent 1,2-propanediol degradation. The majority of bacterial microcompartments (MCs) are metabolosomes.
487 It is hypothesized that the primary role of MCs involved in 1,2-propanediol degradation might be the regulation of
488 aldehyde production to prevent toxicity and DNA damage [84]–[86]. Metabolosomes are also found in archaea including
489 those inhabiting human intestine [87]. Even though 1,2-propanediol degrading MCs have not been described in archaea
490 up to date, detection of proteins involved in such a process suggests that gut archaea might somehow protect
491 themselves from toxicity exhibited by surrounding organisms and even pathogens. For instance, in the human intestine
492 1,2-propanediol is excreted by some saccharolytic bacteria when anaerobically grown on monosaccharides, such as
493 fucose and rhamnose [88], [89].

494 Protein cluster that are relatively equally represented by archaeal and bacterial proteins (clusters *h11*, *h27* and *h31*)
495 are all involved in energy metabolism and related processes. Indeed, archaeal metabolism resembles in its complexity
496 the metabolism of bacteria and lower Eukaryotes [90]. For instance, cluster *h31* is represented by 2-
497 aminoethylphosphonate-pyruvate (2-AEP) aminotransferase, a very common bacterial enzyme playing a role in
498 phosphonate degradation [91]. Phosphonate degradation has also been shown as an important source and a common
499 production pathway for methane [92]. More specifically, Sosa *et al.* demonstrated that phosphonate degradation is a
500 source and production pathway of methane and ethylene in surface waters of western North Atlantic Ocean [93]. We
501 hypothesize *M. smithii* might degrade 2-AEP produced by bacteria [94], pointing at symbiotic relationships between
502 archaea and bacteria, however the role of 2-AEP aminotransferase in archaea is yet not described in the literature [95].
503 In general, the exact biological function of phosphonate macromolecules has not been comprehensively described [96].
504 Cluster *h11* represents a phosphoenolpyruvate-dependent sugar phosphotransferase system (PTS) system, IIA
505 component. Until recently, it was assumed that PTS proteins are absent from archaeal species except for their presence
506 in *H. marismortui* [97]. However, today it is widely accepted that archaea possess PTS, and it is an essential player for
507 key cellular processes. In addition, several proteins performing metabolic functions are more specific to archaea
508 (homologous clusters h19, h28), for example, the protein from cluster *h28* is a transthyretin-like protein. It has been
509 shown that modern transthyretins evolved from their ancient ancestors with a functional role in uric acid catabolism
510 through an event of duplication followed by divergence [98], [99]. By assuming that the described evolutionary history
511 of transthyretin (TTR) and TTR-related proteins is indeed correct, we speculate that the archaeal protein identified in
512 our work might have uricase activity and therefore participate in urate catabolism. Functional uricase enzymes are found
513 in eukaryotes and prokaryotes. There exist only limited studies describing uricase activity in archaea, however, it was
514 shown that the strict anaerobe, *M. vannielii*, among certain purines can use uric acid as a sole nitrogen source for
515 growth [100]. According to other studies, uric acid fermentation might be a triggering step for methane production in
516 geese [101].

517 The prevalence of unique archaeal protein clusters is significantly lower in the gut compared to archaeal-bacterial
518 homologs. However, the most frequent ones also play a role in adaptation and host cell defense processes. The most
519 frequent protein cluster is the type II restriction endonuclease BglII (unique cluster *u1*), whose primary function is the
520 protection of the host genome against foreign DNA, for example invading viral DNA. Another example of such a defense
521 mechanism is demonstrated by the presence of the specificity unit of type I restriction-modification (R-M) EcoKI enzyme
522 (*u22*). Roer *et al.* report that EcoKI R-M enzyme is involved in regulation of rejection/acceptance of genetic material
523 transferred upon conjugation of microbial cells [102]. Conjugation is known as a mechanism of horizontal gene transfer
524 which plays an important role in the evolution of archaeal species and provides means of adaptation to new ecological
525 niches [103]–[105]. Indeed, archaea are known to encode genes of bacterial origin, moreover the flow of genes from
526 bacteria to archaea occurs several-fold more often than in the reverse direction [106].

527 In our analysis, we observe two genes that play a role in sporulation: sporulation stage V protein AE (*h9)* and AD *(h20*).
528 Knowledge about the genes involved in sporulation is sparse. However, sporulation has only been described for spore-
529 forming bacteria and has never been observed in archaea. Nevertheless, sporulation genes are known to be encoded
530 by non-sporulating species. In these bacterial taxa, they are likely to encode regulatory proteins or enzymes involved
531 in peptidoglycan (PPG) turnover and, thus, playing a role in cell division and/or development [107], [108]. Archaea lack
532 PPG but methanogenic archaea, including species of the genus Methanobrevibacter, use pseudopeptidoglycan
533 (pseudo-PPG) instead, which functions similarly to PPG in a bacterial cell and results in Gram positive staining [109].
534 Certain structural similarities between methanogens and bacteria described above leave open the question of whether
535 methanogenic archaea could use sporulation proteins in processes involving pseudo-PPG turnover in an analogous
536 way to non-sporulating bacteria. Although it is not possible to decipher the exact function of the abovementioned genes
537 without further detailed experimental analysis, this finding constitutes great interest, especially in the context of the work
538 by Nelson Sathi *et al.* who reported that methanogens in particular tend to contain horizontally transferred genes from
539 bacteria and many of them are functionally active [110]. Phylogenetic analysis of sporulation stage V proteins AE and
540 AD (*h9* and *h20*, respectively) has demonstrated that sequences from the abovementioned clusters are compositionally

homogeneous and phylogenetically distant from other known similar sequences from Uniprot, and therefore might be unique to the human gut environment. Moreover, archaeal and bacterial sequences from gut SCs *h9* and *h20* branch out together, which suggests that sporulation genes encoded in archaea are indeed the result of horizontal gene transfer from bacteria to archaea. This speculation is also corroborated by the fact that in archaeal genomes sporulation genes are encoded in operons with other sporulation genes and are surrounded by genes having archaeal-specific functions such pyrimidine, thiamine and methane metabolism. Moreover, genes in flanking regions up- and downstream from spoVAE and spoVAD genes are indeed encoded in the representative *M. smithii* isolate DSM 861. As bacteria encoding similar spoVAE and spoVAD proteins and bacterial sequences from clusters *h9* and *h20* belong to various species of the *Clostridium* genus, HGT probably occurred in the direction from the abovementioned species to *M. smithii*. In addition to this, genes *h9* and *h20* as well as their surrounding genes are expressed by the archaeal genomes sampled from human faecal samples. Taking together our findings and evidence for the presence of sporulation genes in archaea, their functional potential has to be further investigated.

Unique cluster *u4* is represented by the Unr protein involved in stress adaptation. Cold-shock domains comprising this protein provide adaptation to external conditions through stress-induced transcriptional and post-transcriptional control of gene expression. Several frequently occurring archaeal proteins were annotated as adhesin-like, and such proteins were found in both bacterial homologous and unique groups, however the unique archaeal counterparts are more frequent in the gut (homologous cluster *h39*, unique clusters *u2, u3).* It was shown that *M. smithii* possess a variety of adhesin-like proteins used in different repertoires with different substrates. This strategy allows *M. smithii* strains developing adaptive responses to different metabolic niches in the gut, as well as establishing syntrophic relationships with bacterial neighbors with other metabolic potentials [111]. Taken together, these mechanisms contribute to the viability of archaea in the human gut.

Conclusion

In this work we have used the most recent methods for protein functional assignment in application to the annotation of gut-specific unique archaeal proteins. We demonstrate that the usage of protein structure prediction methods improves the annotation as they are more sensitive to detecting remote relationships between proteins beyond searching for sequence similarity. We furthermore emphasize the importance of structure-based approaches in characterizing previously uncharacterized proteins.

We identify major processes occurring in the gut which involve archaeal gut-specific proteins and unravel unexpected functions of common methanogenic archaeon in the human intestine *M. smithii*. Firstly, in our work we highlight the presence of a variety of archaeal glycosyl transferases participating in both N- and O-linked glycosylation. This is an important finding, since N-linked glycosylation has been described in several methanogens, while O-linked glycosylation remains not well understood and currently lacks observation of this phenomenon in archaea. Furthermore, we identify archaeal gut-specific proteins which function to ensure maximal efficiency of glycosylation, which in turn contribute not only to viability and adaptability to external conditions, but also to environmental recognition and benefits cell trafficking. Among these, several types of GTs common in gut-specific archaeal species provide *M. smithii* with the ability to construct O-glycosylated structures on their surface proteins. We hypothesize, this property may be used by these organisms to mimic glycan structures similar to the ones found on the cell surfaces of the intestinal mucus. Via structural mimicking of the surface cells of the human gut archaea might accomplish environmental recognition, vitality, and proliferation.

We discover that archaeal species possess proteins hinting at their possible unique metabolic features such as phosphonate degradation and uric acid catabolism. Considering limited knowledge about the mentioned processes, our findings highlight the importance of further investigation of the details of archaeal metabolism.

M. smithii possess a number of characteristic functionalities allowing for nutritional and existential advantages. L-Ara4N transferase is the most frequent defense system archaeal enzyme and crucial for antimicrobial defense and virulence. The most often occurring unique archaeal protein is a BglII restriction endonuclease playing a role in host genome defense against invading foreign DNA. Prevalence of proteins responsible for various types of defense mechanisms and virulence highlights the importance of such behavior for archaeal species. Syntrophic relationships are accentuated by the availability of proteins playing a role in the 1,2-propanediol utilization pathway. This process might be an example of how archaeal species protect themselves from surrounding pathogens in the human gut. Additionally, adhesin-like proteins contribute to organisms' settlement in different metabolic niches in the intestine and the formation of syntrophic relationships with bacterial neighbors.

Our analysis provides evidence for horizontal gene transfer occurring between gut bacteria and *M. smithii*, as we identify presence and activity of proteins not only modulating the exchange of genetic information between microorganisms, but also the ones being the possible result of HGT from bacteria to archaea, *i.e.*, Stage V sporulation proteins AE and AD. These findings collectively once again highlight that archaeal species are very adaptive and benefit from the interaction with surrounding species.

We would like to emphasize that our strategy provided comprehensive *in silico* characterization of gut-specific archaeal functions and prepared a target list for future experimental validation. Indeed, the work presented herein remains purely computational, and therefore additional experimental work is required to confirm exact functions of identified genes and/or reveal additional functions of interest.

Acknowledgments

All authors proof-read and approved of the content in this research paper. The authors declare that they have no competing interests. Experiments presented in this work were carried out using the HPC facility of the University of Luxembourg [112]. This project has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation program (grant agreement No. 863664).

References

[1] C. R. Woese and G. E. Fox, ‘Phylogenetic structure of the prokaryotic domain: The primary kingdoms’, *Proc. Natl. Acad. Sci.*, vol. 74, no. 11, pp. 5088–5090, Nov. 1977, doi: 10.1073/pnas.74.11.5088.

[2] W. E. Balch, L. J. Magrum, G. E. Fox, R. S. Wolfe, and C. R. Woese, ‘An ancient divergence among the bacteria’, *J. Mol. Evol.*, vol. 9, no. 4, pp. 305–311, Aug. 1977, doi: 10.1007/BF01796092.

[3] G. E. Fox, L. J. Magrum, W. E. Balch, R. S. Wolfe, and C. R. Woese, ‘Classification of methanogenic bacteria by 16S ribosomal RNA characterization’, *Proc. Natl. Acad. Sci. U. S. A.*, vol. 74, no. 10, pp. 4537–4541, Oct. 1977, doi: 10.1073/pnas.74.10.4537.

[4] C. R. Woese, O. Kandler, and M. L. Wheelis, ‘Towards a natural system of organisms: proposal for the domains Archaea, Bacteria, and Eucarya’, *Proc. Natl. Acad. Sci. U. S. A.*, vol. 87, no. 12, pp. 4576–4579, Jun. 1990, doi: 10.1073/pnas.87.12.4576.

[5] A. Spang *et al.*, ‘Complex archaea that bridge the gap between prokaryotes and eukaryotes’, *Nature*, vol. 521, no. 7551, pp. 173–179, May 2015, doi: 10.1038/nature14447.

[6] M. Könneke, A. E. Bernhard, J. R. de la Torre, C. B. Walker, J. B. Waterbury, and D. A. Stahl, ‘Isolation of an autotrophic ammonia-oxidizing marine archaeon’, *Nature*, vol. 437, no. 7058, pp. 543–546, Sep. 2005, doi: 10.1038/nature03911.

[7] M. Pester, C. Schleper, and M. Wagner, ‘The Thaumarchaeota: an emerging view of their phylogeny and ecophysiology’, *Curr. Opin. Microbiol.*, vol. 14, no. 3, pp. 300–306, Jun. 2011, doi: 10.1016/j.mib.2011.04.007.

[8] E. F. DeLong, ‘Everything in moderation: archaea as “non-extremophiles”’, *Curr. Opin. Genet. Dev.*, vol. 8, no. 6, pp. 649–654, Dec. 1998, doi: 10.1016/s0959-437x(98)80032-4.

[9] C. Schleper, ‘Ammonia oxidation: different niches for bacteria and archaea?’, *ISME J.*, vol. 4, no. 9, pp. 1092–1094, Sep. 2010, doi: 10.1038/ismej.2010.111.

[10] D. L. Valentine, ‘Adaptations to energy stress dictate the ecology and evolution of the Archaea’, *Nat. Rev. Microbiol.*, vol. 5, no. 4, pp. 316–323, Apr. 2007, doi: 10.1038/nrmicro1619.

[11] C. Hoegenauer, H. F. Hammer, A. Mahnert, and C. Moissl-Eichinger, ‘Methanogenic archaea in the human gastrointestinal tract’, *Nat. Rev. Gastroenterol. Hepatol.*, vol. 19, no. 12, Art. no. 12, Dec. 2022, doi: 10.1038/s41575-022-00673-z.

[12] C. M. Thomas, E. Desmond-Le Quéméner, S. Gribaldo, and G. Borrel, ‘Factors shaping the abundance and diversity of the gut archaeome across the animal kingdom’, *Nat. Commun.*, vol. 13, no. 1, Art. no. 1, Jun. 2022, doi: 10.1038/s41467-022-31038-4.

[13] J. Y. Kim *et al.*, ‘The human gut archaeome: identification of diverse haloarchaea in Korean subjects’, *Microbiome*, vol. 8, no. 1, p. 114, Aug. 2020, doi: 10.1186/s40168-020-00894-x.

[14] T. L. Miller, M. J. Wolin, E. Conway de Macario, and A. J. Macario, ‘Isolation of Methanobrevibacter smithii from human feces’, *Appl. Environ. Microbiol.*, vol. 43, no. 1, pp. 227–232, Jan. 1982, doi: 10.1128/aem.43.1.227-232.1982.

641 [15] G. Borrel, J.-F. Brugère, S. Gribaldo, R. A. Schmitz, and C. Moissl-Eichinger, 'The host-associated archaeome',
642 *Nat. Rev. Microbiol.*, vol. 18, no. 11, Art. no. 11, Nov. 2020, doi: 10.1038/s41579-020-0407-y.

643 [16] G. Borrel *et al.*, 'Genomics and metagenomics of trimethylamine-utilizing Archaea in the human gut microbiome',
644 *ISME J.*, vol. 11, no. 9, pp. 2059–2074, Sep. 2017, doi: 10.1038/ismej.2017.72.

645 [17] C. Bang, K. Weidenbach, T. Gutsmann, H. Heine, and R. A. Schmitz, 'The Intestinal Archaea Methanosphaera
646 stadtmanae and Methanobrevibacter smithii Activate Human Dendritic Cells', *PLoS ONE*, vol. 9, no. 6, p. e99411,
647 Jun. 2014, doi: 10.1371/journal.pone.0099411.

648 [18] Z. Lyu and W. B. Whitman, 'Transplanting the pathway engineering toolbox to methanogens', *Curr. Opin.*
649 *Biotechnol.*, vol. 59, pp. 46–54, Oct. 2019, doi: 10.1016/j.copbio.2019.02.009.

650 [19] A. M. Lee, J. R. Sevinsky, J. L. Bundy, A. M. Grunden, and J. L. Stephenson, 'Proteomics of Pyrococcus furiosus,
651 a hyperthermophilic archaeon refractory to traditional methods', *J. Proteome Res.*, vol. 8, no. 8, pp. 3844–3851,
652 Aug. 2009, doi: 10.1021/pr801119h.

653 [20] A. Tebbe *et al.*, 'Analysis of the cytosolic proteome of Halobacterium salinarum and its implication for genome
654 annotation', *Proteomics*, vol. 5, no. 1, pp. 168–179, Jan. 2005, doi: 10.1002/pmic.200400910.

655 [21] C. J. Castelle *et al.*, 'Genomic Expansion of Domain Archaea Highlights Roles for Organisms from New Phyla in
656 Anaerobic Carbon Cycling', *Curr. Biol.*, vol. 25, no. 6, pp. 690–701, Mar. 2015, doi: 10.1016/j.cub.2015.01.014.

657 [22] K. Zaremba-Niedzwiedzka *et al.*, 'Asgard archaea illuminate the origin of eukaryotic cellular complexity', *Nature*,
658 vol. 541, no. 7637, pp. 353–358, Jan. 2017, doi: 10.1038/nature21031.

659 [23] P. Bork, 'Powers and Pitfalls in Sequence Analysis: The 70% Hurdle', *Genome Res.*, vol. 10, no. 4, pp. 398–400,
660 Jan. 2000, doi: 10.1101/gr.10.4.398.

661 [24] J. C. Wooley, A. Godzik, and I. Friedberg, 'A Primer on Metagenomics', *PLOS Comput. Biol.*, vol. 6, no. 2, p.
662 e1000667, Feb. 2010, doi: 10.1371/journal.pcbi.1000667.

663 [25] K. W. Ellens, N. Christian, C. Singh, V. P. Satagopam, P. May, and C. L. Linster, 'Confronting the catalytic dark
664 matter encoded by sequenced genomes', *Nucleic Acids Res.*, vol. 45, no. 20, pp. 11495–11514, Nov. 2017, doi:
665 10.1093/nar/gkx937.

666 [26] K. S. Makarova, Y. I. Wolf, and E. V. Koonin, 'Towards functional characterization of archaeal genomic dark matter',
667 *Biochem. Soc. Trans.*, vol. 47, no. 1, pp. 389–398, Feb. 2019, doi: 10.1042/BST20180560.

668 [27] V. Márquez *et al.*, 'Proteomic characterization of archaeal ribosomes reveals the presence of novel archaeal-
669 specific ribosomal proteins', *J. Mol. Biol.*, vol. 405, no. 5, pp. 1215–1232, Feb. 2011, doi:
670 10.1016/j.jmb.2010.11.055.

671 [28] P. Wu, J. S. Brockenbrough, M. R. Paddy, and J. P. Aris, 'NCL1, a novel gene for a non-essential nuclear protein
672 in Saccharomyces cerevisiae', *Gene*, vol. 220, pp. 109–117, Oct. 1998, doi: 10.1016/s0378-1119(98)00330-8.

673 [29] J. D. Watson *et al.*, 'Towards Fully Automated Structure-Based Function Prediction In Structural Genomics: A Case
674 Study', *J. Mol. Biol.*, vol. 367, no. 5, pp. 1511–1522, Apr. 2007, doi: 10.1016/j.jmb.2007.01.063.

675 [30] C. Chothia and A. M. Lesk, 'The relation between the divergence of sequence and structure in proteins', *EMBO J.*,
676 vol. 5, no. 4, pp. 823–826, Apr. 1986, doi: 10.1002/j.1460-2075.1986.tb04288.x.

[31] J. Skolnick, M. Gao, H. Zhou, and S. Singh, 'AlphaFold 2: Why It Works and Its Implications for Understanding the Relationships of Protein Sequence, Structure, and Function', *J. Chem. Inf. Model.*, vol. 61, no. 10, pp. 4827–4831, Oct. 2021, doi: 10.1021/acs.jcim.1c01114.

[32] J. Jumper *et al.*, 'Highly accurate protein structure prediction with AlphaFold', *Nature*, vol. 596, no. 7873, pp. 583–589, Aug. 2021, doi: 10.1038/s41586-021-03819-2.

[33] Z. Du *et al.*, 'The trRosetta server for fast and accurate protein structure prediction', *Nat. Protoc.*, vol. 16, no. 12, pp. 5634–5651, Dec. 2021, doi: 10.1038/s41596-021-00628-9.

[34] S. Nayfach *et al.*, 'A genomic catalog of Earth's microbiomes', *Nat. Biotechnol.*, vol. 39, no. 4, pp. 499–509, Apr. 2021, doi: 10.1038/s41587-020-0718-6.

[35] A. Almeida *et al.*, 'A unified catalog of 204,938 reference genomes from the human gut microbiome', *Nat. Biotechnol.*, vol. 39, no. 1, pp. 105–114, Jan. 2021, doi: 10.1038/s41587-020-0603-3.

[36] 'Prodigal: prokaryotic gene recognition and translation initiation site identification | BMC Bioinformatics | Full Text'. https://bmcbioinformatics.biomedcentral.com/articles/10.1186/1471-2105-11-119 (accessed Jan. 19, 2022).

[37] M. Steinegger and J. Söding, 'MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets', *Nat. Biotechnol.*, vol. 35, no. 11, pp. 1026–1028, Nov. 2017, doi: 10.1038/nbt.3988.

[38] M. Steinegger and J. Söding, 'Clustering huge protein sequence sets in linear time', *Nat. Commun.*, vol. 9, no. 1, p. 2542, Jun. 2018, doi: 10.1038/s41467-018-04964-5.

[39] N. T. Pierce, L. Irber, T. Reiter, P. Brooks, and C. T. Brown, 'Large-scale sequence comparisons with sourmash', *F1000Research*, vol. 8, p. 1006, Jul. 2019, doi: 10.12688/f1000research.19675.1.

[40] C. T. Brown and L. Irber, 'sourmash: a library for MinHash sketching of DNA', *J. Open Source Softw.*, vol. 1, no. 5, p. 27, Sep. 2016, doi: 10.21105/joss.00027.

[41] P. Queirós, F. Delogu, O. Hickl, P. May, and P. Wilmes, 'Mantis: flexible and consensus-driven genome annotation', *GigaScience*, vol. 10, no. 6, p. giab042, Jun. 2021, doi: 10.1093/gigascience/giab042.

[42] Moult, J., Fidelis, K., Kryshtafovych, A., Schwede, T. & Topf, M., 'Critical assessment of techniques for protein structure prediction, fourteenth round.', *ASP 14 Abstract Book*, 2020, [Online]. Available: https://www.predictioncenter.org/casp14/doc/CASP14_Abstracts.pdf

[43] R. A. Laskowski, J. D. Watson, and J. M. Thornton, 'ProFunc: a server for predicting protein function from 3D structure', *Nucleic Acids Res.*, vol. 33, no. Web Server issue, pp. W89-93, Jul. 2005, doi: 10.1093/nar/gki414.

[44] V. Gligorijević *et al.*, 'Structure-based protein function prediction using graph convolutional networks', *Nat. Commun.*, vol. 12, no. 1, p. 3168, May 2021, doi: 10.1038/s41467-021-23303-9.

[45] A. Heintz-Buschart *et al.*, 'Integrated multi-omics of the human gut microbiome in a case study of familial type 1 diabetes', *Nat. Microbiol.*, vol. 2, no. 1, Art. no. 1, Oct. 2016, doi: 10.1038/nmicrobiol.2016.180.

[46] H. Li, 'Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM', 2013, doi: 10.6084/M9.FIGSHARE.963153.V1.

711 [47] H. Li *et al.*, 'The Sequence Alignment/Map format and SAMtools', *Bioinformatics*, vol. 25, no. 16, pp. 2078–2079,
712 Aug. 2009, doi: 10.1093/bioinformatics/btp352.

713 [48] B. S. Pedersen and A. R. Quinlan, 'Mosdepth: quick coverage calculation for genomes and exomes', *Bioinforma.*
714 *Oxf. Engl.*, vol. 34, no. 5, pp. 867–868, Mar. 2018, doi: 10.1093/bioinformatics/btx699.

715 [49] W. Song, B. Wemheuer, S. Zhang, K. Steensen, and T. Thomas, 'MetaCHIP: community-level horizontal gene
716 transfer identification through the combination of best-match and phylogenetic approaches', *Microbiome*, vol. 7,
717 no. 1, p. 36, Mar. 2019, doi: 10.1186/s40168-019-0649-y.

718 [50] Y. Shimoyama, 'pyGenomeViz: A genome visualization python package for comparative genomics'. Jun. 2022.
719 Accessed: Jul. 23, 2022. [Online]. Available: https://github.com/moshi4/pyGenomeViz

720 [51] R. Apweiler *et al.*, 'UniProt: the Universal Protein knowledgebase', *Nucleic Acids Res.*, vol. 32, no. Database issue,
721 pp. D115-119, Jan. 2004, doi: 10.1093/nar/gkh131.

722 [52] 'BLAST: Basic Local Alignment Search Tool'. https://blast.ncbi.nlm.nih.gov/Blast.cgi (accessed Dec. 06, 2022).

723 [53] K. Katoh, K. Misawa, K. Kuma, and T. Miyata, 'MAFFT: a novel method for rapid multiple sequence alignment
724 based on fast Fourier transform', *Nucleic Acids Res.*, vol. 30, no. 14, pp. 3059–3066, Jul. 2002, doi:
725 10.1093/nar/gkf436.

726 [54] A. Criscuolo and S. Gribaldo, 'BMGE (Block Mapping and Gathering with Entropy): a new software for selection of
727 phylogenetic informative regions from multiple sequence alignments', *BMC Evol. Biol.*, vol. 10, p. 210, Jul. 2010,
728 doi: 10.1186/1471-2148-10-210.

729 [55] L.-T. Nguyen, H. A. Schmidt, A. von Haeseler, and B. Q. Minh, 'IQ-TREE: A Fast and Effective Stochastic Algorithm
730 for Estimating Maximum-Likelihood Phylogenies', *Mol. Biol. Evol.*, vol. 32, no. 1, pp. 268–274, Jan. 2015, doi:
731 10.1093/molbev/msu300.

732 [56] G. Yu, 'Using ggtree to Visualize Data on Tree-Like Structures', *Curr. Protoc. Bioinforma.*, vol. 69, no. 1, p. e96,
733 Mar. 2020, doi: 10.1002/cpbi.96.

734 [57] M. Kanehisa, Y. Sato, M. Kawashima, M. Furumichi, and M. Tanabe, 'KEGG as a reference resource for gene and
735 protein annotation', *Nucleic Acids Res.*, vol. 44, no. D1, pp. D457–D462, Jan. 2016, doi: 10.1093/nar/gkv1070.

736 [58] B. Rost, 'Twilight zone of protein sequence alignments', *Protein Eng.*, vol. 12, no. 2, pp. 85–94, Feb. 1999, doi:
737 10.1093/protein/12.2.85.

738 [59] S. Gowrisankaran *et al.*, 'Endophilin-A coordinates priming and fusion of neurosecretory vesicles via intersectin',
739 *Nat. Commun.*, vol. 11, no. 1, p. 1266, Mar. 2020, doi: 10.1038/s41467-020-14993-8.

740 [60] M. Pohlschroder and R. N. Esquivel, 'Archaeal type IV pili and their involvement in biofilm formation', *Front.*
741 *Microbiol.*, vol. 6, 2015, Accessed: Feb. 07, 2022. [Online]. Available:
742 https://www.frontiersin.org/article/10.3389/fmicb.2015.00190

743 [61] R. Mohammadzadeh, A. Mahnert, S. Duller, and C. Moissl-Eichinger, 'Archaeal key-residents within the human
744 microbiome: characteristics, interactions and involvement in health and disease', *Curr. Opin. Microbiol.*, vol. 67, p.
745 102146, Jun. 2022, doi: 10.1016/j.mib.2022.102146.

[62] B. S. Samuel *et al.*, 'Genomic and metabolic adaptations of Methanobrevibacter smithii to the human gut', *Proc. Natl. Acad. Sci. U. S. A.*, vol. 104, no. 25, pp. 10643–10648, Jun. 2007, doi: 10.1073/pnas.0704189104.

[63] H. Magidovich and J. Eichler, 'Glycosyltransferases and oligosaccharyltransferases in Archaea: putative components of the N-glycosylation pathway in the third domain of life', *FEMS Microbiol. Lett.*, vol. 300, no. 1, pp. 122–130, Nov. 2009, doi: 10.1111/j.1574-6968.2009.01775.x.

[64] H. Magidovich and J. Eichler, 'Glycosyltransferases and oligosaccharyltransferases in Archaea: putative components of the N-glycosylation pathway in the third domain of life', *FEMS Microbiol. Lett.*, vol. 300, no. 1, pp. 122–130, Nov. 2009, doi: 10.1111/j.1574-6968.2009.01775.x.

[65] A. Anandan and A. Vrielink, 'Structure and function of lipid A-modifying enzymes', *Ann. N. Y. Acad. Sci.*, vol. 1459, no. 1, pp. 19–37, Jan. 2020, doi: 10.1111/nyas.14244.

[66] S. D. Breazeale, A. A. Ribeiro, and C. R. H. Raetz, 'Origin of Lipid A Species Modified with 4-Amino-4-deoxy-l-arabinose in Polymyxin-resistant Mutants of Escherichia coli: AN AMINOTRANSFERASE (ArnB) THAT GENERATES UDP-4-AMINO-4-DEOXY-l-ARABINOSE *', *J. Biol. Chem.*, vol. 278, no. 27, pp. 24731–24739, Jul. 2003, doi: 10.1074/jbc.M304043200.

[67] A. Mahnert, M. Blohs, M.-R. Pausan, and C. Moissl-Eichinger, 'The human archaeome: methodological pitfalls and knowledge gaps', *Emerg. Top. Life Sci.*, vol. 2, no. 4, pp. 469–482, Dec. 2018, doi: 10.1042/ETLS20180037.

[68] A. Dell, A. Galadari, F. Sastre, and P. Hitchen, 'Similarities and Differences in the Glycosylation Mechanisms in Prokaryotes and Eukaryotes', *Int. J. Microbiol.*, vol. 2010, p. e148178, Jan. 2011, doi: 10.1155/2010/148178.

[69] S. Voisin *et al.*, 'Identification and characterization of the unique N-linked glycan common to the flagellins and S-layer glycoprotein of Methanococcus voltae', *J. Biol. Chem.*, vol. 280, no. 17, pp. 16586–16593, Apr. 2005, doi: 10.1074/jbc.M500329200.

[70] D. J. VanDyke *et al.*, 'Identification of a putative acetyltransferase gene, MMP0350, which affects proper assembly of both flagella and pili in the archaeon Methanococcus maripaludis', *J. Bacteriol.*, vol. 190, no. 15, pp. 5300–5307, Aug. 2008, doi: 10.1128/JB.00474-08.

[71] U. Kärcher *et al.*, 'Primary structure of the heterosaccharide of the surface glycoprotein of Methanothermus fervidus.', *J. Biol. Chem.*, vol. 268, no. 36, pp. 26821–26826, Dec. 1993, doi: 10.1016/S0021-9258(19)74185-4.

[72] M. Abu-Qarn and J. Eichler, 'Protein N-glycosylation in Archaea: defining Haloferax volcanii genes involved in S-layer glycoprotein glycosylation', *Mol. Microbiol.*, vol. 61, no. 2, pp. 511–525, Jul. 2006, doi: 10.1111/j.1365-2958.2006.05252.x.

[73] B. Chaban, S. Voisin, J. Kelly, S. M. Logan, and K. F. Jarrell, 'Identification of genes involved in the biosynthesis and attachment of Methanococcus voltae N-linked glycans: insight into N-linked glycosylation pathways in Archaea', *Mol. Microbiol.*, vol. 61, no. 1, pp. 259–268, Jul. 2006, doi: 10.1111/j.1365-2958.2006.05226.x.

[74] U. Zähringer, H. Moll, T. Hettmann, Y. A. Knirel, and G. Schäfer, 'Cytochrome b558/566 from the archaeon Sulfolobus acidocaldarius has a unique Asn-linked highly branched hexasaccharide chain containing 6-sulfoquinovose', *Eur. J. Biochem.*, vol. 267, no. 13, pp. 4144–4149, Jul. 2000, doi: 10.1046/j.1432-1327.2000.01446.x.

[75] M. F. Mescher and J. L. Strominger, 'Purification and characterization of a prokaryotic glycoprotein from the cell envelope of Halobacterium salinarium', *J. Biol. Chem.*, vol. 251, no. 7, pp. 2005–2014, Apr. 1976.

[76] A. Tamir and J. Eichler, 'N-Glycosylation Is Important for Proper Haloferax volcanii S-Layer Stability and Function', *Appl. Environ. Microbiol.*, vol. 83, no. 6, pp. e03152-16, Mar. 2017, doi: 10.1128/AEM.03152-16.

[77] M. F. Mescher and J. L. Strominger, 'Structural (shape-maintaining) role of the cell surface glycoprotein of Halobacterium salinarium', *Proc. Natl. Acad. Sci. U. S. A.*, vol. 73, no. 8, pp. 2687–2691, Aug. 1976, doi: 10.1073/pnas.73.8.2687.

[78] M. Sumper, E. Berg, R. Mengele, and I. Strobel, 'Primary structure and glycosylation of the S-layer protein of Haloferax volcanii', *J. Bacteriol.*, vol. 172, no. 12, pp. 7111–7118, Dec. 1990, doi: 10.1128/jb.172.12.7111-7118.1990.

[79] L. E. Tailford, E. H. Crost, D. Kavanaugh, and N. Juge, 'Mucin glycan foraging in the human gut microbiome', Mar. 2015, Accessed: Mar. 24, 2022. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4365749/

[80] H. H. Wandall, M. A. I. Nielsen, S. King-Smith, N. de Haan, and I. Bagdonaite, 'Global functions of O-glycosylation: promises and challenges in O-glycobiology', *FEBS J.*, vol. 288, no. 24, pp. 7183–7212, 2021, doi: 10.1111/febs.16148.

[81] M. Bektas and D. S. Rubenstein, 'The role of intracellular protein O-glycosylation in cell adhesion and disease', *J. Biomed. Res.*, vol. 25, no. 4, pp. 227–236, Jul. 2011, doi: 10.1016/S1674-8301(11)60031-6.

[82] C. R. Torres and G. W. Hart, 'Topography and polypeptide distribution of terminal N-acetylglucosamine residues on the surfaces of intact lymphocytes. Evidence for O-linked GlcNAc', *J. Biol. Chem.*, vol. 259, no. 5, pp. 3308–3317, Mar. 1984.

[83] P. V. den Steen, P. M. Rudd, R. A. Dwek, and G. Opdenakker, 'Concepts and Principles of O-Linked Glycosylation', *Crit. Rev. Biochem. Mol. Biol.*, vol. 33, no. 3, pp. 151–208, Jan. 1998, doi: 10.1080/10409239891204198.

[84] G. D. Havemann, E. M. Sampson, and T. A. Bobik, 'PduA is a shell protein of polyhedral organelles involved in coenzyme B(12)-dependent degradation of 1,2-propanediol in Salmonella enterica serovar typhimurium LT2', *J. Bacteriol.*, vol. 184, no. 5, pp. 1253–1261, Mar. 2002, doi: 10.1128/JB.184.5.1253-1261.2002.

[85] N. W. Kennedy, S. P. Ikonomova, M. Slininger Lee, H. W. Raeder, and D. Tullman-Ercek, 'Self-assembling Shell Proteins PduA and PduJ have Essential and Redundant Roles in Bacterial Microcompartment Assembly', *J. Mol. Biol.*, vol. 433, no. 2, p. 166721, Jan. 2021, doi: 10.1016/j.jmb.2020.11.020.

[86] E. M. Sampson and T. A. Bobik, 'Microcompartments for B12-dependent 1,2-propanediol degradation provide protection from DNA and cellular damage by a reactive metabolic intermediate', *J. Bacteriol.*, vol. 190, no. 8, pp. 2966–2971, Apr. 2008, doi: 10.1128/JB.01925-07.

[87] L.-N. Liu, M. Yang, Y. Sun, and J. Yang, 'Protein stoichiometry, structural plasticity and regulation of bacterial microcompartments', *Curr. Opin. Microbiol.*, vol. 63, pp. 133–141, Oct. 2021, doi: 10.1016/j.mib.2021.07.006.

[88] B. Dogan *et al.*, 'Inflammation-associated adherent-invasive Escherichia coli are enriched in pathways for use of propanediol and iron and M-cell translocation', *Inflamm. Bowel Dis.*, vol. 20, no. 11, pp. 1919–1932, Nov. 2014, doi: 10.1097/MIB.0000000000000183.

[89] C. Fan, S. Cheng, S. Sinha, and T. A. Bobik, 'Interactions between the termini of lumen enzymes and shell proteins mediate enzyme encapsulation into bacterial microcompartments', *Proc. Natl. Acad. Sci. U. S. A.*, vol. 109, no. 37, pp. 14995–15000, Sep. 2012, doi: 10.1073/pnas.1207516109.

[90] C. Bräsen, D. Esser, B. Rauch, and B. Siebers, 'Carbohydrate Metabolism in Archaea: Current Insights into Unusual Enzymes and Pathways and Their Regulation', *Microbiol. Mol. Biol. Rev. MMBR*, vol. 78, no. 1, pp. 89–175, Mar. 2014, doi: 10.1128/MMBR.00041-13.

[91] E. Zangelmi, T. Stanković, M. Malatesta, D. Acquotti, K. Pallitsch, and A. Peracchi, 'Discovery of a New, Recurrent Enzyme in Bacterial Phosphonate Degradation: (R)-1-Hydroxy-2-aminoethylphosphonate Ammonia-lyase', *Biochemistry*, vol. 60, no. 15, pp. 1214–1225, Apr. 2021, doi: 10.1021/acs.biochem.1c00092.

[92] W. W. Metcalf *et al.*, 'Synthesis of Methylphosphonic Acid by Marine Microbes: A Source for Methane in the Aerobic Ocean', *Science*, vol. 337, no. 6098, pp. 1104–1107, Aug. 2012, doi: 10.1126/science.1219875.

[93] O. A. Sosa, T. J. Burrell, S. T. Wilson, R. K. Foreman, D. M. Karl, and D. J. Repeta, 'Phosphonate cycling supports methane and ethylene supersaturation in the phosphate-depleted western North Atlantic Ocean', *Limnol. Oceanogr.*, vol. 65, no. 10, pp. 2443–2459, 2020, doi: 10.1002/lno.11463.

[94] W. W. Metcalf and W. A. van der Donk, 'Biosynthesis of Phosphonic and Phosphinic Acid Natural Products', *Annu. Rev. Biochem.*, vol. 78, pp. 65–94, 2009, doi: 10.1146/annurev.biochem.78.091707.100215.

[95] S. Li and G. P. Horsman, 'An inventory of early branch points in microbial phosphonate biosynthesis', Biochemistry, preprint, Apr. 2021. doi: 10.1101/2021.04.07.438883.

[96] K.-S. Ju, J. R. Doroghazi, and W. W. Metcalf, 'Genomics-enabled discovery of phosphonate natural products and their biosynthetic pathways', *J. Ind. Microbiol. Biotechnol.*, vol. 41, no. 2, pp. 345–356, Feb. 2014, doi: 10.1007/s10295-013-1375-2.

[97] I. Comas, F. González-Candelas, and M. Zúñiga, 'Unraveling the evolutionary history of the phosphoryl-transfer chain of the phosphoenolpyruvate:phosphotransferase system through phylogenetic analyses and genome context', *BMC Evol. Biol.*, vol. 8, no. 1, p. 147, May 2008, doi: 10.1186/1471-2148-8-147.

[98] A. C. Keebaugh and J. W. Thomas, 'The Evolutionary Fate of the Genes Encoding the Purine Catabolic Enzymes in Hominoids, Birds, and Reptiles', *Mol. Biol. Evol.*, vol. 27, no. 6, pp. 1359–1369, Jun. 2010, doi: 10.1093/molbev/msq022.

[99] L. Carrijo de Oliveira *et al.*, 'Reenacting the Birth of a Function: Functional Divergence of HIUases and Transthyretins as Inferred by Evolutionary and Biophysical Studies', *J. Mol. Evol.*, vol. 89, no. 6, pp. 370–383, Jul. 2021, doi: 10.1007/s00239-021-10010-8.

[100] E. DeMoll and L. Tsai, 'Conversion of purines to xanthine by Methanococcus vannielii', *Arch. Biochem. Biophys.*, vol. 250, no. 2, pp. 440–445, Nov. 1986, doi: 10.1016/0003-9861(86)90747-2.

[101] Y.-H. Chen, S.-Y. Wang, and J.-C. Hsu, 'Effect of Caecectomy on Body Weight Gain, Intestinal Characteristics and Enteric Gas Production in Goslings', *Asian-Australas. J. Anim. Sci.*, vol. 16, no. 7, pp. 1030–1034, Jan. 2003, Accessed: Feb. 15, 2022. [Online]. Available: https://www.animbiosci.org/journal/view.php?doi=10.5713/ajas.2003.1030

[102] L. Roer, F. M. Aarestrup, and H. Hasman, 'The EcoKI Type I Restriction-Modification System in Escherichia coli Affects but Is Not an Absolute Barrier for Conjugation', *J. Bacteriol.*, vol. 197, no. 2, pp. 337–342, Jan. 2015, doi: 10.1128/JB.02418-14.

[103] R. T. Papke, J. E. Koenig, F. Rodríguez-Valera, and W. F. Doolittle, 'Frequent recombination in a saltern population of Halorubrum', *Science*, vol. 306, no. 5703, pp. 1928–1929, Dec. 2004, doi: 10.1126/science.1103289.

[104] H. Cadillo-Quiroz *et al.*, ‘Patterns of gene flow define species of thermophilic Archaea’, *PLoS Biol.*, vol. 10, no. 2, p. e1001265, Feb. 2012, doi: 10.1371/journal.pbio.1001265.

[105] A. Wagner *et al.*, ‘Mechanisms of gene flow in archaea’, *Nat. Rev. Microbiol.*, vol. 15, no. 8, pp. 492–501, Aug. 2017, doi: 10.1038/nrmicro.2017.41.

[106] P. Deschamps, Y. Zivanovic, D. Moreira, F. Rodriguez-Valera, and P. López-García, ‘Pangenome Evidence for Extensive Interdomain Horizontal Transfer Affecting Lineage Core and Shell Genes in Uncultured Planktonic Thaumarchaeota and Euryarchaeota’, *Genome Biol. Evol.*, vol. 6, no. 7, pp. 1549–1563, Jul. 2014, doi: 10.1093/gbe/evu127.

[107] D. J. Rigden and M. Y. Galperin, ‘Sequence analysis of GerM and SpoVS, uncharacterized bacterial “sporulation” proteins with widespread phylogenetic distribution’, *Bioinformatics*, vol. 24, no. 16, pp. 1793–1797, Aug. 2008, doi: 10.1093/bioinformatics/btn314.

[108] R. U. Onyenwoke, J. A. Brill, K. Farahi, and J. Wiegel, ‘Sporulation genes in members of the low G+C Gram-type-positive phylogenetic branch (Firmicutes)’, *Arch. Microbiol.*, vol. 182, no. 2, pp. 182–192, Oct. 2004, doi: 10.1007/s00203-004-0696-y.

[109] ‘BacMap’. http://bacmap.wishartlab.com/organisms/525 (accessed Jan. 25, 2023).

[110] S. Nelson-Sathi *et al.*, ‘Origins of major archaeal clades correspond to gene acquisitions from bacteria’, *Nature*, vol. 517, no. 7532, pp. 77–80, Jan. 2015, doi: 10.1038/nature13805.

[111] E. E. Hansen *et al.*, ‘Pan-genome of the dominant human gut-associated archaeon, Methanobrevibacter smithii, studied in twins’, *Proc. Natl. Acad. Sci. U. S. A.*, vol. 108 Suppl 1, pp. 4599–4606, Mar. 2011, doi: 10.1073/pnas.1000071108.

[112] S. Varrette, P. Bouvry, H. Cartiaux, and F. Georgatos, ‘Management of an academic HPC cluster: The UL experience’, in *2014 International Conference on High Performance Computing Simulation (HPCS)*, Jul. 2014, pp. 959–967. doi: 10.1109/HPCSim.2014.6903792.