ISME Communications, 2024, 4(1), ycad014
https://doi.org/10.1093/ismeco/ycad014
Advance access publication: 10 January 2024
Original Article

Functional prediction of proteins from the human gut archaeome

Polina V. Novikova[1,*], Susheel Bhanu Busi[1,2], Alexander J. Probst[3], Patrick May[4], Paul Wilmes[1,5,*]

[1]Systems Ecology, Luxembourg Centre for Systems Biomedicine, University of Luxembourg, Esch-sur-Alzette L-4362, Luxembourg
[2]UK Centre for Ecology and Hydrology, Wallingford, OX10 8 BB, United Kingdom
[3]Environmental Metagenomics, Department of Chemistry, Research Center One Health Ruhr of the University Alliance Ruhr, for Environmental Microbiology and Biotechnology, University Duisburg-Essen, Duisburg 47057, Germany
[4]Bioinformatics Core, Luxembourg Centre for Systems Biomedicine, University of Luxembourg, Esch-sur-Alzette L-4362, Luxembourg
[5]Department of Life Sciences and Medicine, Faculty of Science, Technology and Medicine, University of Luxembourg, Esch-sur-Alzette L-4362, Luxembourg

*Corresponding authors: Paul Wilmes and Polina V. Novikova, Systems Ecology, Luxembourg Centre for Systems Biomedicine, University of Luxembourg, Esch-sur-Alzette L-4362, Luxembourg. Email: paul.wilmes@uni.lu, polina.novikova@uni.lu

Abstract

The human gastrointestinal tract contains diverse microbial communities, including archaea. Among them, *Methanobrevibacter smithii* represents a highly active and clinically relevant methanogenic archaeon, being involved in gastrointestinal disorders, such as inflammatory bowel disease and obesity. Herein, we present an integrated approach using sequence and structure information to improve the annotation of *M. smithii* proteins using advanced protein structure prediction and annotation tools, such as AlphaFold2, trRosetta, ProFunc, and DeepFri. Of an initial set of 873 481 archaeal proteins, we found 707 754 proteins exclusively present in the human gut. Having analysed archaeal proteins together with 87 282 994 bacterial proteins, we identified unique archaeal proteins and archaeal–bacterial homologs. We then predicted and characterized functional domains and structures of 73 unique and homologous archaeal protein clusters linked the human gut and *M. smithii*. We refined annotations based on the predicted structures, extending existing sequence similarity-based annotations. We identified gut-specific archaeal proteins that may be involved in defense mechanisms, virulence, adhesion, and the degradation of toxic substances. Interestingly, we identified potential glycosyltransferases that could be associated with *N*-linked and *O*-glycosylation. Additionally, we found preliminary evidence for interdomain horizontal gene transfer between *Clostridia* species and *M. smithii*, which includes *sporulation Stage V proteins AE and AD*. Our study broadens the understanding of archaeal biology, particularly *M. smithii*, and highlights the importance of considering both sequence and structure for the prediction of protein function.

Keywords: protein structure, archaea, methanogens, gut microbiome

Introduction

In 1977, Woese and Fox, and colleagues discovered the kingdom of *Archaebacteria*, later renamed Archaea, revealing a new branch in the tree of life [1-4]. The discovery of the *Asgard* superphylum and its close relationship with the eukaryotic branch supports the notion of an archaeal origin for eukaryotes, yet ongoing debates continue regarding whether the archaeal ancestor of eukaryotes belongs within the *Asgard* superphylum or represents a sister group to all other archaea [5, 6]. Historically, archaea were associated with extreme environments but have since been recognized for their general importance and prevalence [7, 8]. Their ability to thrive in extreme environments and to resist chemicals is attributed, in part, to their unique cell envelope structures. In nature, archaea perform distinctive biogeochemical functions, such as methanogenesis, anaerobic methane oxidation, and ammonia oxidation [9, 10]. By employing diverse ecological strategies for energy production, archaea can inhabit a wide variety of environments [11]. Archaea are also host-associated, such as on plants, in human and animal gastrointestinal tracts [12, 13], on human skin [14, 15], in respiratory airways [16], and in the oral cavity [17]. Based on recent estimates, archaea comprise up to 10% of the human gut microbiota [18].

Methanobrevibacter smithii, a ubiquitous and active methanogen in the human gut microbiome, has remarkable clinical relevance and is relatively well annotated [19]. It plays an important role in the degradation of complex carbohydrates, leading to the production of methane, which has significant physiological effects on human physiology. Imbalances in the population of *M. smithii* have been implicated as factors contributing to gastrointestinal disorders such as inflammatory bowel disease (IBD) [20, 21] and obesity [22-24]. Given the prevalence of *M. smithii* in the gut, further research aimed at *M. smithii* is key to understanding their role in disease. Archaeal proteins, including those of *M. smithii*, play a crucial role in adapting to diverse environments and showcase their unique biology. The knowledge about diverse archaea, including novel species, in the human gut microbiome has expanded, underscoring their significance [25]. Some host-associated taxa, like *Methanomassiliicoccales*, have potential beneficial effects on human health [26], while others like *Methanosphaera stadtmanae* have been linked to proinflammatory immune processes [27]. Given the current interest in the role of archaea in

Received 12 December 2023. **Revised:** 16 December 2023. **Accepted:** 19 December 2023

human health and disease, understanding the archaeal proteome is crucial for understanding the functional potential of archaea.

Studying archaeal proteins presents challenges both in experimental and computational aspects. Previous research has highlighted the potential for biotechnological applications in various archaeal genera [28]. However, genetic toolboxes for targeted genomic modifications are currently limited to mesophilic *Methanococcus* and *Methanosarcina* genera [29]. Although alternative methods like mass spectrometry-based searches exist, difficulties arise from inaccurate predictions of protein coding sequences (CDSs) due to limited knowledge of ribosomal binding sites and promoter consensus sequences [30]. Another unresolved challenge lies in the isolation and cultivation of archaea under laboratory conditions, although recent progress has been made [31, 32]. To overcome these challenges, metagenomic sequencing has emerged as a promising approach to study archaea and their ecological relationships. Metagenomics has enhanced our understanding of the archaeal branches within the tree of life [31-33], whereby assembled sequences allow prediction of protein CDSs and their functional characterization *in silico*. However, metagenome-assembled genomes (MAGs) face challenges in functional assignment due to incomplete sequences and difficulties in predicting and annotating open-reading frames (ORFs) [34, 35]. Sequence-based protein function annotation, commonly used but limited in cases of distant protein homologies, proves to be not particularly effective [36]. Moreover, the databases containing information about archaeal proteins and functions are not consistently updated, creating a 2-fold challenge in the sequence-based annotation of archaeal proteins. On one hand, Makarova *et al.* [37] report that *archaeal ribosomal proteins L45 and L47*, experimentally identified in 2011 [38], and pre-rRNA processing and ribosome biogenesis proteins of the *NOL1/NOP2/fmu* family, characterized in 1998 [39], were not added to annotation pipelines by 2019 and were labelled as "hypothetical." On the other hand, sequence similarity–based approaches fail to capture relationships between highly divergent proteins when aligned with a known database protein [40-42]. Archaea, the least characterized domain of life, suffer from incorrect protein annotations due to insufficient experimental data and outdated databases [43]. Furthermore, the study by Makarova *et al.* indicates that a substantial proportion of genes within archaeal genomes (30%–80%) have not been thoroughly characterized, leading to their classification as archaeal "dark matter" [37]. Poorly annotated proteins limit our study of microbial functionality and their roles in biological processes. However, protein structure prediction represents an alternative strategy addressing the gap in sequence–function annotation [44]. It complements sequence-based approaches, particularly when annotations are limited or conflicting across databases, by utilizing the conservation of tertiary structure to infer functional roles [45, 46]. Advanced computational techniques, such as AlphaFold2 (AF) [47] and trRosetta (TR) [48], offer accurate predictions of 3D structures, providing valuable functional insights.

Here, we present an integrated *in silico* approach to enhance protein functional characterization and improve accuracy of protein annotations in archaeon *M. smithii*. Having compared archaeal gut–specific proteins to bacterial gut proteins, we found 73 unique and homologous archaeal protein clusters. Our approach incorporates advanced protein structure prediction and annotation tools, such as AlphaFold2 (AF), trRosetta (TR), ProFunc (PF), and DeepFri (DF), into a comprehensive workflow. We predict and characterize the functional domains and structures of 73 gut-specific archaeal protein clusters. The predicted functions are linked to the adaptation to changing environments, survival, and nutritional capabilities of *M. smithii* within the human gut microbiome. We additionally identified sporulation-related archaeal proteins, presumably horizontally transferred to archaea from *Clostridium* species.

Materials and methods

Selection of gut-specific archaeal proteins

To select specific proteins of gut-associated archaea, we utilized archaeal MAGs obtained from the Genomes from Earth's Microbiomes (GEM) catalog [49] and the Unified Human Gastrointestinal Genome (UHGG) collection [50], along with bacterial MAGs from the UHGG collection (accessed in November 2020). Genomes were extracted based on available metadata and filtered by taxonomy to specifically target archaea.

Gene prediction was performed using Prodigal (V2.6.3) [51] on the archaeal and bacterial MAGs from the UHGG collection, while CDSs from the GEM catalog were downloaded from the provided source (https://portal.nersc.gov/GEM). Archaeal and bacterial proteins were further separately clustered using MMseqs2 (MM2) (v12.113e3-2) [52, 53] (Fig. 1) with the following parameters: –cov-mode 0 –min-seq-id 0.9 -c 0.9.

To identify unique functions of gut-associated archaea, we selected proteins specific to the human gut and encoded by gut-associated archaea. MAGs were selected based on available metadata indicating their sampling location. First, we included protein clusters containing at least one protein from a MAG sampled in the human gut. We then excluded protein clusters that had proteins from MAGs sampled in other environments. The final selection included protein clusters where all proteins were encoded by MAGs sampled exclusively from the human gut.

From the selected gut-specific protein clusters, only those with complete KEGG annotations were included. Fully annotated archaeal and bacterial MM2 clusters were additionally clustered together with Sourmash (v4.0.0) [54, 55]. Archaeal protein clusters were categorized into two groups: those sharing KEGG Orthology identifiers (KOs) with bacterial proteins (prefix *h*) and those with unique KOs (prefix *u*) (Fig. 1).

Protein function annotation

Archaeal and bacterial proteins were annotated with KEGG orthologs (KOs) using Mantis (1.5.4) [56] (Fig. 1). AF [47, 57] and TR [48] were used as structure prediction tools. For each tool, the predicted protein structure was then annotated separately. The TR-based model was annotated using templates with the highest *identity* and *coverage* features. TR used a template for prediction if it met the criteria of confidence >0.6, *E*-value <0.001, and coverage >0.3. The protein model generated by AF was submitted to the PF [58] web server for structure-based annotation. "Sequence search vs existing PDB entries" and 3D functional template searches sections from the PF report were used for structure-based protein annotation. 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 [59] was used as an auxiliary tool, providing broad and general descriptions to verify or refute suggestions from AF and/or TR. DeepFri predictions with a certainty score > 0.7 were considered. Our combined approach integrates multiple methods to enhance the resolution of functional

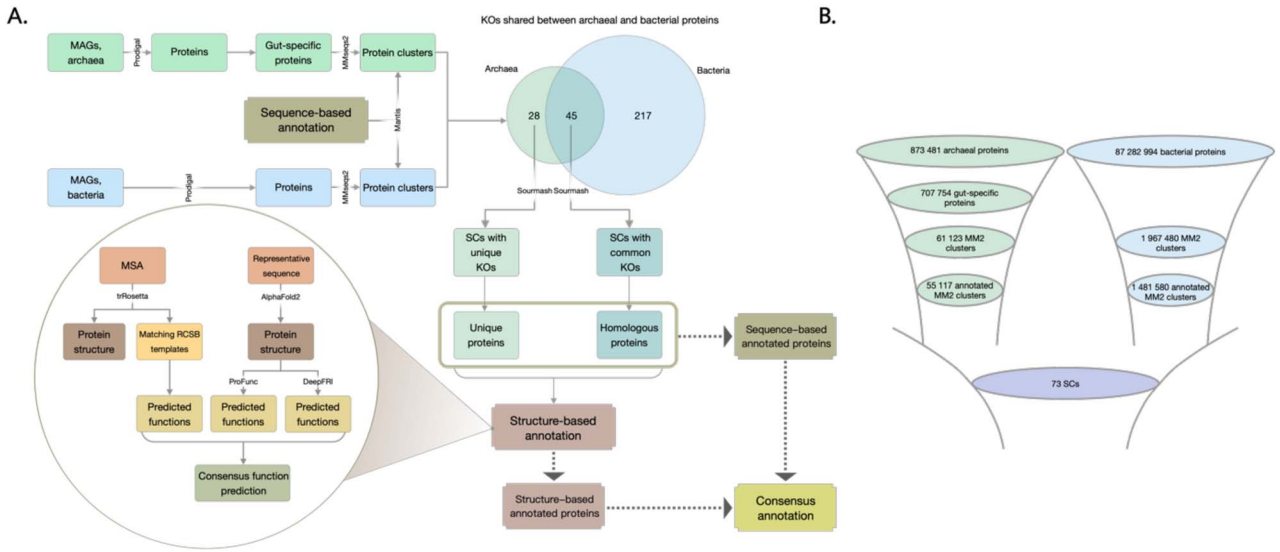

Figure 1. (A) Flowchart demonstrating major steps of the analysis; the Venn diagram demonstrates the number of shared KOs assigned to archaeal and bacterial sourmash clusters; (B) funnels illustrating the protein count at each stage of protein selection; MM2, MMseqs2 clusters; SCs, sourmash clusters.

Table 1. Relationships between PF likelihood and TR TM-scores.

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

annotation, particularly for challenges faced by traditional methods.

When 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 different results, we prioritized the result with highest confidence. For instance, when the confidence of the model predicted by TR was *very high* and template matches were provided, and AF-based PF reported a match with a lower confidence (anything but *certain match*), the template hit by TR was used as the primary source for the annotation. The relationship between PF likelihood and TR template modeling scores (TM-scores) generated in our analysis is shown in 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 where there were no 3D functional hits, TR annotation 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* was chosen. Here, we define coverage as *coverage* feature in TR and the ratio $\frac{\text{longest fitted segment}}{\text{query sequence length}}$ as in PF, and for identity, we take *identity* as in TR and *percentage sequence identity* as in PF.

The appropriateness of an annotation was determined based on the extent to which the assigned function of a protein was found to be directly relevant to archaea and supported by relevant literature. Any other annotations were classified as incorrect. Following this initial step, *sensitivity* was calculated as $\textbf{sensitivity} = \frac{N_{str}}{N_{str}+N_{seq}}$, *specificity* as $\textbf{specificity} = \frac{N_{seq}}{N_{seq}+N_{str}}$, *positive likelihood ratio* as $PLR = \frac{sensitivity}{1-specificity}$, negative likelihood ratio as $NLR = \frac{1-sensitivity}{specificity}$, where N_{seq} and N_{str} are the numbers of correct sequence- and structure-based annotations, respectively.

Protein relative occurrence calculation

Relative occurrence or frequency of protein functions in the groups of unique and homologous proteins was calculated. The measure was calculated as the ratio of the number of proteins with a specific KO to the total number of proteins of bacterial or archaeal proteins. For example, the relative occurrence of unique archaeal proteins annotated as *K20411* (sourmash Cluster 1) is $\frac{N_{select}}{N_{total}} * 10^6$, where N_{select} is the amount of proteins annotated with K20411 and N_{total} is the total number of archaeal proteins. The reason for using a constant factor of 10^6 in the equation is to scale the values and generate numbers better suited for graphical representation.

Gene expression analysis

To comprehensively assess the expression of archaeal proteins in the context of human health and disease, gene expression was verified using a dataset, which we previously published, by mapping metatranscriptomic reads of fecal samples of healthy individuals and patients with Type 1 diabetes mellitus (T1DM) [60] to nucleotide sequences of genes of interest using bwa mem [61]. Mapping files were processed with SAMtools (v1.6) [62]. Mosdepth (v0.3.3) [63] was used to calculate mean read coverage per gene of interest.

Horizontal gene transfer analysis

To assess the stability of gene structures in *M. smithii* genomes, we conducted a horizontal gene transfer (HGT) analysis using metaCHIP (v1.10.12) [64] on all *M. smithii* MAGs available in the included datasets. One *Methanobrevibacter_A oralis* MAG derived from UHGG were also included for the comparison of the number of HGT events.

Gene synteny analysis

pyGenomeViz (v0.3.2) [65] 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 10 kb 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 [56]. Here, we exclusively focused on *M. smithii*, as our analysis revealed that all the gut-specific proteins encoded by gut-associated archaea were encoded by *M. smithii*, and thus, this taxon was considered representative for our analyses. The *M. smithii*–type strain DSM 861 was used to assess the presence of genes from flanking regions of specific genes in an archaeal culture.

Phylogenetic analysis

To build phylogenetic trees for selective sourmash clusters, additional similar sequences were added from Uniprot [66] using BLAST (v2.0.15.153) [67] with default parameters on the consensus sequences representing sourmash clusters of interest, namely *h9* and *h20*. Furthermore, Uniprot sequences and sourmash cluster sequences were used to build trees. Multiple sequence alignments were built using MAFFT (v7) [68] and trimmed with BMGE (v1.12) [69] using BLOSUM95 similarity matrix and the default cut-off 0.5. Maximum likelihood phylogenetic trees were built with IQ-TREE (v1.6.12) [70] and visualized using the R library *ggtree* (v3.6.2) [71].

Results and discussion

Our study aimed to analyze the gut-specific proteins encoded by *M. smithii* in the human gastrointestinal tract. As we focused on identifying archaeal unique proteins and archaeal–bacterial homologs, we analysed gut-specific archaeal and gut bacterial proteins together. Having compared the two subsets based on their sequence-based annotation, we categorized archaeal gut–specific proteins into two groups: unique and homologous proteins. To annotate them, we used KEGG KOs due to their consistent functional annotations across organisms and widespread usage. For structure-based functional assignment, we utilized a combination of structure prediction and annotation tools (Fig. 1), leveraging the higher prediction accuracy of AlphaFold2 and the rapid and accurate *de novo* predictions obtained via TR. Our central goal is to enhance the accuracy and reliability of protein structure predictions through the integration of these two approaches. Utilizing representative sequences of unique and homologous proteins, AF produced protein structures, and subsequent functional annotations were accomplished by integrating PF and DeepFRI. TR was employed to predict structures of unique and homologous proteins showing detectable homologous matches in the Protein Data Bank, which were subsequently used for further structure annotation.

It is important to note that our methodology includes semi-manual tools, making it most suitable for a limited number of select proteins. The primary design intent of our workflow was to facilitate the further refinement of functions for specific proteins of interest. Although alternative tools such as ESMFold [72] or EMBER3D [73] are available and hold promise for augmenting the potential of the described pipeline, our approach remains specialized and well-suited for in-depth protein analysis.

Enhancing annotations of proteins encoded by *M. smithii*

To explore the uncharted functional space of *M. smithii*, we first selected gut-specific proteins of gut-associated archaea. We collected the encoded proteins of a total of 1190 archaeal and 285 835 bacterial MAGs, resulting in 873 481 archaeal proteins and 87 282 994 bacterial proteins (Fig. 1). We focused on proteins associated with archaea of the human gut microbiome, which represented 37% (707 754 proteins) of all predicted archaeal proteins. These proteins were grouped into 61 123 MM2 clusters for archaea (≥2 proteins per cluster) and 1 967 480 MM2 clusters for bacteria (≥10 proteins per cluster). By retaining fully annotated protein clusters, we obtained 55 117 archaeal MM2 clusters and 1 481 580 bacterial MM2 clusters. Using our proposed functional prediction strategy (Fig. 1A), we analyzed the gut-associated archaeal proteins alongside bacterial proteins, resulting in 45 homologous sourmash clusters, *i.e.* shared between archaea and bacteria, and 28 unique sourmash clusters, *i.e.* composed exclusively of archaeal proteins. The bacterial data served as a reference to distinguish unique proteins encoded and transcribed by archaea, as well as archaeal proteins with homologs to bacterial ORFs. A summary of the annotations as well as comparison of annotations by structure-based tools is provided in Supplementary Tables 1–3.

All archaeal proteins from the abovementioned sourmash clusters were classified as *M. smithii*. We thus sought to extend our knowledge of *M. smithii* by exploring functions that could have implications for human health and disease. The investigation of the relative occurrence of identified proteins and their associated processes revealed distinct types of functions in unique and homologous protein clusters (Fig. 2). The most frequently identified functions in the unique sourmash clusters were related to adaptation to changing environments and protection mechanisms, e.g. defense against foreign DNA and oxidative stress, while processes such as RNA and DNA regulation, energy metabolism, and cell wall integrity and maintenance were less represented (Supplementary Table 4). Homologous sourmash clusters showed frequent functions related to adaptation, various protection mechanisms, energy metabolism, and cell structural integrity (Supplementary Table 5). Analysis of fecal metatranscriptomic data confirmed the transcription of the majority of encoded genes, with some unique and homologous genes exhibiting higher expression levels (Fig. 2). Two unique and 19 homologous sourmash clusters with relatively high expression levels were identified, including genes associated with adaptation to changing environments, defense against foreign DNA and oxidative stress, DNA/RNA regulation, and energy metabolism, while the rest were unannotated (Fig. 2).

Our analysis demonstrated disparity in annotations between sequence- and structure-based approaches. Notably, 46% (13 out of 28) and 31% (14 out of 45) of the unique and homologous sourmash clusters, respectively, lacked structure-based annotations, suggesting a reliance on sequence information for their functional annotation thus far. Literature searches suggest that the KEGG annotations may not provide reasonable or meaningful functional assignments for most of these unannotated proteins. For instance, a protein annotated as *mitochondrial import receptor subunit TOM40* by KEGG is predicted to be a *putative intimin/invasin-like protein* based on its structure, which is more relevant in the context of archaeal biology than being a eukaryotic protein involved in mitochondrial protein import. Similarly, a protein annotated as *Endophilin-A*, a eukaryotic protein involved in membrane curvature, shows structural similarity to *PilC, a Type IVa pilus subunit* of a prokaryotic adhesion filament. Although the presence of eukaryotic proteins in archaea is not surprising from an evolutionary perspective, the assignment of a protein to its evolutionary homolog from a different kingdom may not provide precise functional assignment of protein function. Moreover, examining the

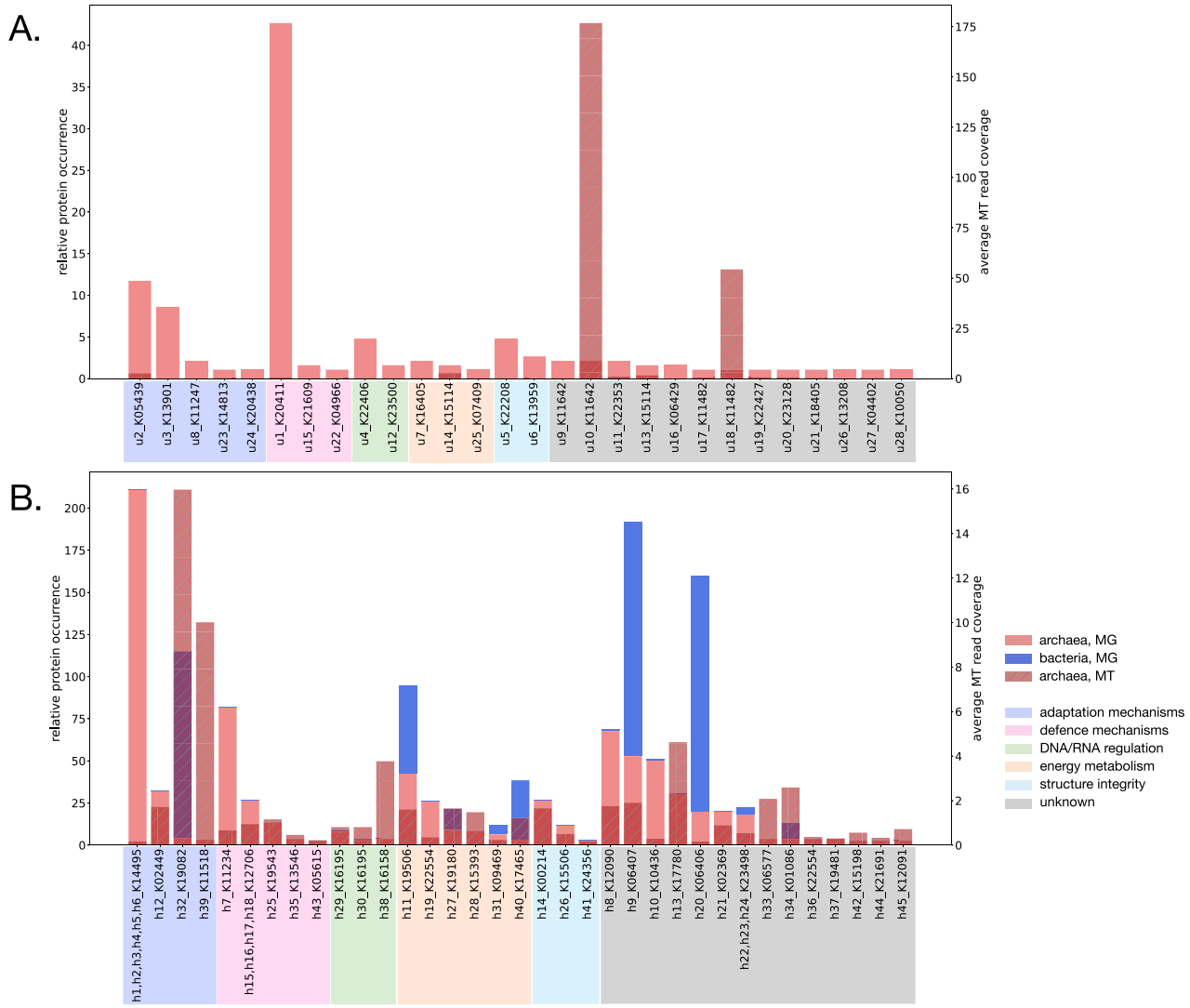

Figure 2. Relative metagenomic occurrence and average metatranscriptomic read coverage of proteins in the (A) unique and (B) homologous groups of clusters with archaeal proteins; MG, metagenomics; MT, metatranscriptomics.

sequence identities between protein clusters annotated through sequence-based methods and the corresponding sequences in UniProt, it is evident that the majority of proteins lack any discernible similarity with those in UniProt. Furthermore, for those instances where some degree of sequence identity is observed, they do not surpass 70% for archaea-specific, unique and 49% for homologous protein clusters (Supplementary Tables 6 and 7).

In general, the agreement between the sequence- and structure-based methods was limited, with 4% (1 out of 28) and 25% (11 out of 45) of the unique and homologous proteins showing consistent annotations, respectively (Supplementary Tables 4–5 and 8). The rest of the proteins exhibited disparity between sequence- and structure-based annotations, which was assessed by comparing their reported functions. For example, unique sourmash cluster *u24* yielded different annotations using EGGNOG, KEGG, and Pfam databases, which we used to potentially resolve disparities in the annotations (Supplementary Table 4). However, a consensus structure-based annotation identified it as *polypeptide N-acetylgalactosaminyltransferase*, providing additional annotation beyond sequence analysis. Similarly, the homologous protein clusters *h15*–*h18* had the same functional assignments as *novobiocin biosynthesis protein NovC* using KEGG, but structure-based annotation revealed further distinctions: *h16* and *h18* were classified as members of the *LytR-Cps2A-Psr protein family*, *h15* was annotated as *78 kDa glucose-regulated protein*, and *h17* remained unannotated (Supplementary Table 5). The incorporation of structural information in protein annotation enables the distinction between closely related sequences, offering additional insights into protein function, which highlights the crucial role of structural data in understanding protein functionality. In addition, the observed disparity between sequence and structure-based annotations, coupled with low sequence identities between sequence-based annotations and corresponding UniProt sequences, underscores the complementarity of structure-based methods to the abovementioned approach for protein function annotation.

We further identified glycosyltransferases responsible for N- and O-linked glycosylation from clusters *h1*–*h6* as prevalent archaeal gut-specific proteins. These proteins may contribute to the viability and adaptability of archaeal cells in the gut. For instance, the most prevalent unique archaeal glycosyltransferase is *4-amino-4-deoxy-L-arabinose (L-Ara4N) transferase*, which is essential for the protection from environmental stress, symbiosis, virulence, and resistance against antimicrobial activity [74, 75]. Moreover, one of the six glycosyltransferases is a

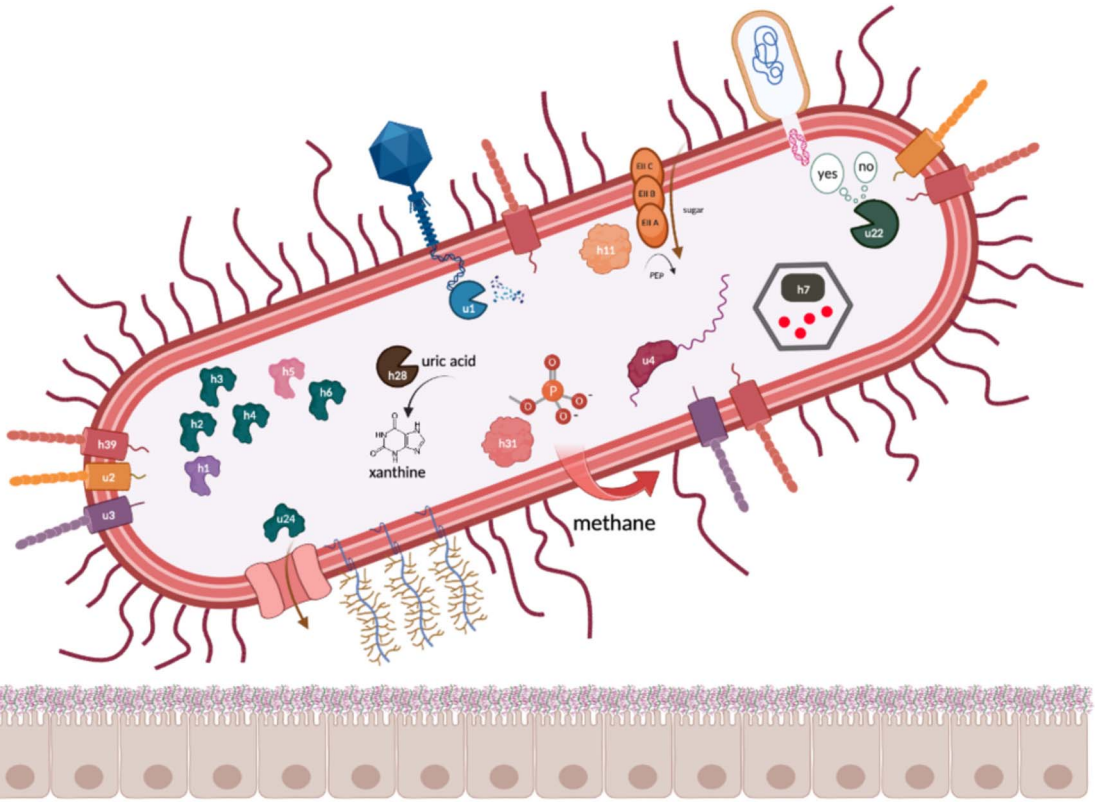

Figure 3. Schematic proposal highlighting proteins specific to gut-associated archaea with described functions: *u1, Type II restriction endonuclease BglII; u2, intimin/invasin-like protein with a Ig-like domain; u3, intimin/invasin-like protein; u4, Unr protein; u22, Type I restriction–modification EcoKI enzyme, specificity subunit; u24, polypeptide N-acetylgalactosaminyltransferase; 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 utilization protein pduA; h11, phosphoenolpyruvate-dependent PTS system, IIA component; h28, transthyretin-like protein; h31, 2-AEP aminotransferase.*

dolichyl-diphosphooligosaccharide–protein glycosyltransferase subunit STT3B (h5), which functions as an accessory protein in *N*-glycosylation and provides its maximal efficiency [76]. Archaeal *N*-glycosylation is known to play an important role in the viability and adaptivity of archaeal cells to external conditions such as high salinity [77], elevated temperatures [78], and an acidic environment [79] while also maintaining the structural integrity of cells [80, 81]. Four out of the six identified glycosyltransferases are *dolichyl-phosphate-mannose–protein mannosyltransferases 1 (POMT1)*, which are responsible for O-linked glycosylation of proteins in eukaryotes. Another *O*-glycosylation–associated protein, *polypeptide N-acetylgalactosaminyltransferase*, was found in the subset of unique archaeal proteins (u24). *M. smithii* has been found to decorate its cellular surface with sugar residues mimicking those present in the glycan landscape of the intestinal environment [82]. The presence of human mucus– and epithelial cell surface–associated glycans in *M. smithii*, along with the coding potential for enzymes involved in O-linked glycosylation in archaeal gut species, suggests that *M. smithii* cells might have the capability to emulate the surfaces of eukaryotic cells in the intestinal mucus. Beyond their structural role in proteins, *O*-glycans can also act as regulators of protein interactions, influencing both interprotein and cell-to-cell communication processes involved in cell trafficking and environmental recognition [83].

Further findings suggest that *2-aminoethylphosphonate-pyruvate (2-AEP) aminotransferase, transthyretin-like protein* and *phosphoenolpyruvate-dependent sugar phosphotransferase system system* encoded by *M. smithii* contribute to energy metabolism. 2-AEP is an enzyme commonly found in bacteria and is known to play a critical role in phosphonate degradation, which serves as an important source and production pathway for methane [84]. Additionally, cold-shock domains of *Unr protein* potentially provide *M. smithii* with adaptation strategies through stress-induced control of gene expression [85]. Furthermore, the predicted involvement of proteins such as the *specificity subunit of Type I restriction–modification EcoKI enzyme* [86] and *Type II restriction endonuclease BglII* [87] suggests their potential role in host defense strategies employed by *M. smithii* to protect themselves in the gut environment. Additionally, it is conceivable that archaeal proteins may play a role in protecting against toxicity from other organisms in the gut using *propanediol utilization protein pduA* [88-90], as well as acquiring genes of bacterial origin through HGT. If this is the case, the presence of adhesin-like proteins in archaea could potentially enable them to form symbiotic relationships with bacterial neighbors with diverse metabolic potentials [91]. Figure 3 provides a schematic representation emphasizing specific proteins identified in this study, which could potentially play a significant role in the functional dynamics of archaea within the human intestine. A more detailed description of all identified *M. smithii* proteins is provided in Supplementary Materials.

Characterization of select proteins and gene structures in *M. smithii* genomes

To elucidate the level of conservation among the identified genes recovered in our analyses, we assessed the level of genomic conservation within genomes of two strains of *M. smithii*, two strains *of Ca. Methanobrevibacter intestini* and the related species *Methanobrevibacter_A oralis* as a reference. *Ca. M. intestini* has been recently classified as an independent species within the *M. smithii* clade. We analysed HGT events and evaluated gene

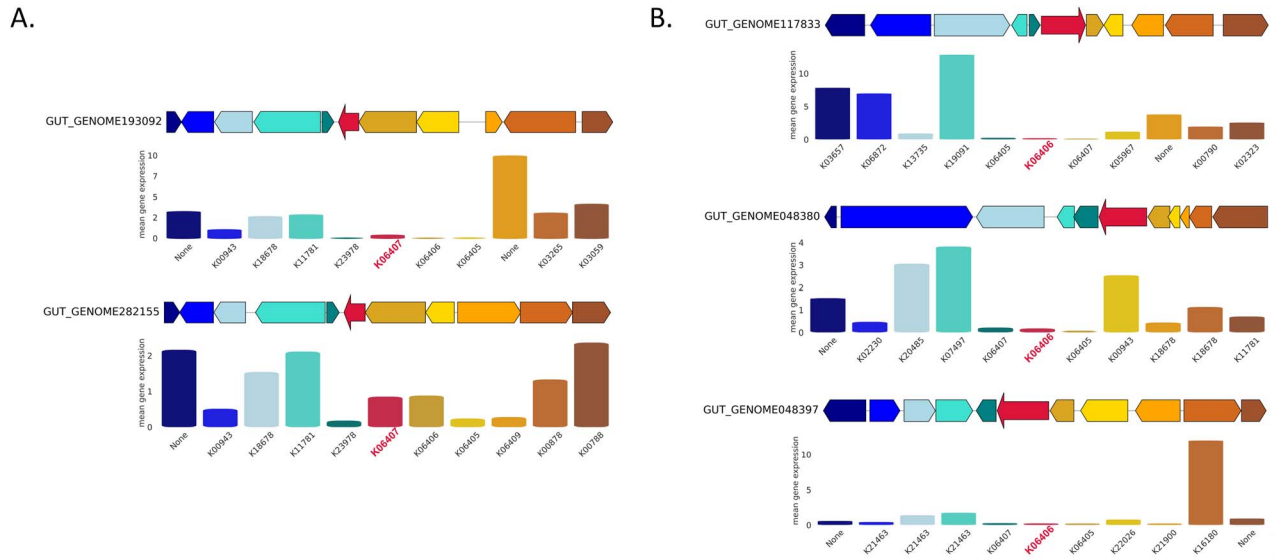

Figure 4. Gene synteny for *sporulation stage V genes AE* and *AD* from their respective sourmash clusters (A) *h9* and (**B**) *h20*; gene expression of target genes (*spoVAE* and *spoVAD*) as well as genes from flanking regions are demonstrated below each sequence and are colored correspondingly. Genes 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*).

structure stability. Using 1022 available MAGs, we noted an increase in HGT events between 319 genomes of two *M. smithii* strains: *Methanobrevibacter_A smithii* and *GCF_000016525.1* (based on GTDB classification) (Supplementary Fig. 1). Specifically, 2.6% of the MAGs ($n = 27$) exhibited HGT events involving the transfer of $\sim 10 \pm 3$ genes to other MAGs. Intriguingly, MAGs exhibiting HGT events were sampled in diverse geographical locations such as Austria, France, the UK, and the USA. Our results suggest that the propensity of these MAGs to exchange genomic segments may be attributed to similarities in their respective local environments [92], including dietary and lifestyle factors of the individuals. Thus, it is plausible that exposure to similar diets or stresses may have influenced the evolution of these MAGs via HGT along comparable trajectories. Conversely, the low occurrence of HGT events among the majority (97.4%) of available *M. smithii* genomes indicates their overall genomic conservation and stability. This could be explained by the fact that these MAGs were sampled from individuals living under similar dietary and lifestyle conditions. Importantly, our findings support the concept of genomic stability in *M. smithii*, as we observed a high degree of conservation in the flanking regions of the genes of interest across various *M. smithii* genomes. Through synteny analyses, we found compelling evidence of conserved synteny for genes encoded in *M. smithii* genomes (https://doi.org/10.5281/zenodo.8024791).

Among the proteins specific for gut-associated archaea, we identified *Stage V sporulation proteins AE* (*spoVAE*) and *AD* (*spoVAD*) (*h9* and *h20*). Using BLAST searches, we extracted 250 bacterial protein sequences for SpoVAE and SpoVAD from Uniprot, including 12 *spoVAE* and 38 *spoVAD* proteins from environmental samples and the rest from isolate bacterial genomes belonging to the *Firmicutes* phylum. Phylogenetic trees demonstrated that proteins from *h9* and *h20* are phylogenetically and compositionally distinct from other sequences and form separate branches (Supplementary Figs 2 and 3). Gene synteny analyses revealed that sporulation genes are grouped in operons (K06405, K06406, and K06407; Fig. 4). Moreover, the flanking regions around sporulation genes include genes with key archaeal as well as methanogenic functions. In addition, the flanking regions of both *spoVAE* and *spoVAD* genes are also encoded in the *M. smithii* isolate DSM 861 genome (Fig. 5). This particular isolate served as the representative strain for our research. Furthermore, to further validate the representativeness of DSM 861, we also computed the average nucleotide identity (ANI) between the type strain DSM 861 and two other available strains, DSM 2374 and DSM 2375. The ANI calculations yielded estimates of 98.3 between *M. smithii* strains DSM 861 and DSM 2374, and 98.2 between DSM 861 and DSM 2375, respectively. However, in contrast to our MAGs, the isolate's genome did not encode the *spoVAE* and *spoVAD* genes. To assess whether *spoVAE* and *spoVAD* genes were acquired by *M. smithii* via HGT, we performed synteny analysis of bacterial sequences obtained from our human gut dataset that shared similarities with the archaeal sequences in clusters *h9* and *h20*. This analysis revealed that in the bacterial genomes found in the human intestine, the flanking regions of *spoVAE* and *spoVAD* genes include genes mediating and facilitating HGT, such as a site-specific DNA recombinase (K06400) encoded upstream from *spoVAE* and *Type IV pilus assembly proteins* (K02662, K02664) encoded downstream from *spoVAD* (Supplementary Figs 4 and 5). Genes originating from clusters *h9* and *h20* are found within bacterial genomes of *Firmicutes* phylum members, specifically *Clostridium* sp. CAG-302 and CAG-269, which highlights their association with known bacterial taxa in the gut and indicates HGT between these distantly related taxa.

Although sporulation has been primarily observed in spore-forming bacteria and not in archaea, it is known that non-sporulating bacterial species also encode sporulation genes. In these bacterial taxa, the genes likely encode regulatory proteins involved in peptidoglycan (PPG) turnover, thereby playing a role in cell division and/or development [93, 94]. Archaea lack PPG but methanogenic archaea, including *Methanobrevibacter* species, use pseudopeptidoglycan (pseudo-PPG) instead, which functions similarly to PPG in a bacterial cell and results in Gram-positive staining certain structural similarities between methanogens and bacteria described above leave open the question of whether

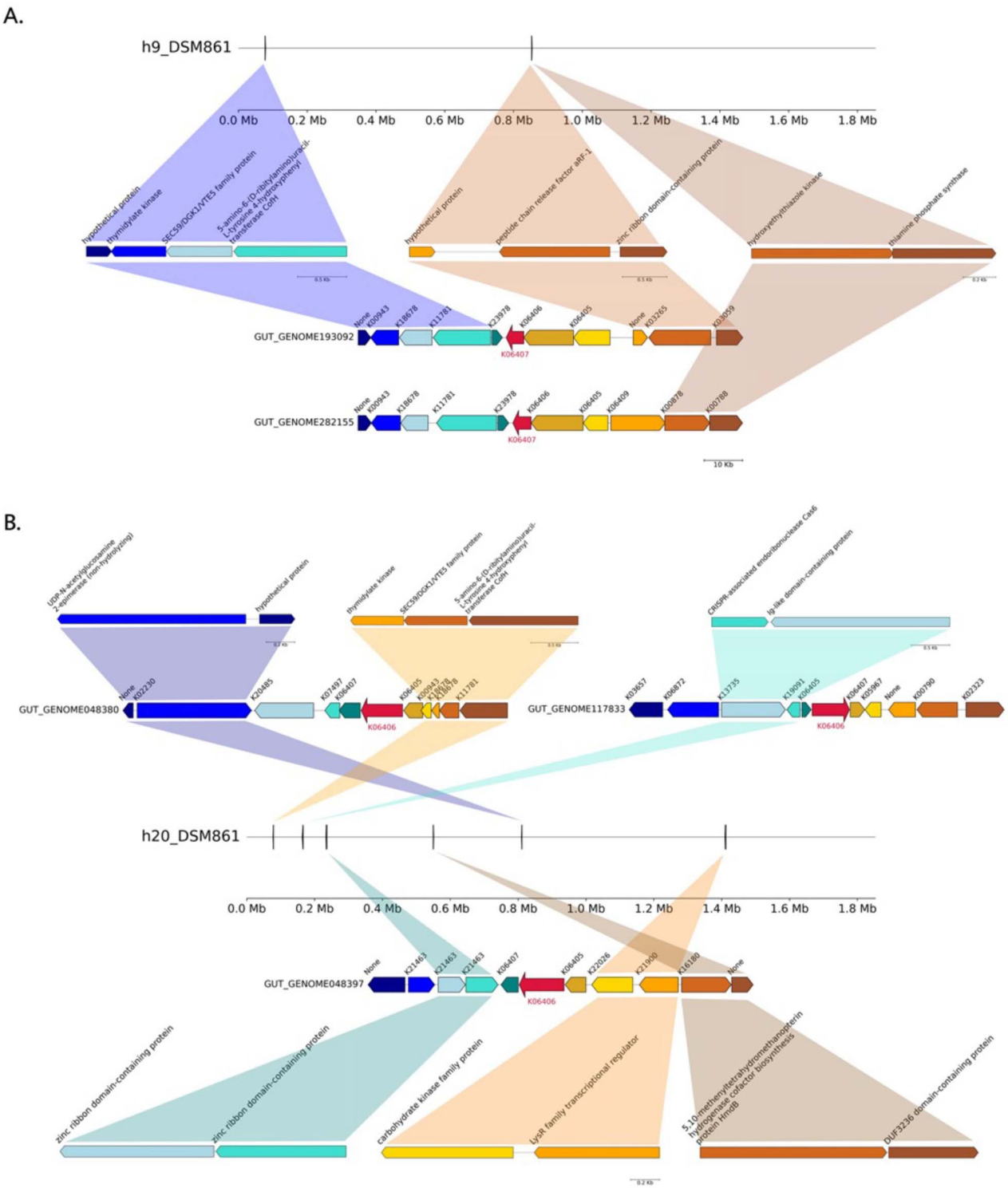

Figure 5. Genomic context of the archaeal flanking regions up- and downstream of the (A) *spoVAE* and (B) *spoVAD* gene clusters in the *M. smithii* strain DSM 861.

sporulation proteins could play a similar role in pseudo-PPG turnover in methanogenic archaea, analogous to their function in non-sporulating bacteria. The identification of these genes holds significant interest, especially in light of the work by Nelson Sathi *et al.*, suggesting that methanogens frequently acquire functionally active genes through horizontal transfer from bacteria. Comprehensive experimental analysis is required to determine their specific functions, but these findings present an exciting opportunity for further exploration. Phylogenetic analysis of *spoVAE* and *spoVAD* has demonstrated that sequences from the abovementioned clusters are compositionally homogeneous but phylogenetically distant from other known similar sequences in Uniprot and therefore might be unique to the human gut environment. Moreover, archaeal and bacterial sequences from sourmash clusters *h9* and *h20* branch out together, which suggests that sporulation genes encoded in archaea might be

the result of HGT from bacteria to archaea. This study provides evidence that archaeal genomes exhibit clustered sporulation genes surrounded by genes linked to archaea-specific functions like pyrimidine, thiamine, and methane metabolism. Moreover, genes in flanking regions up- and downstream of *spoVAE* and *spoVAD* genes are indeed encoded in the representative *M. smithii* isolate DSM 861. The study's intended scope did not include experimental investigations in the wet-lab, such as the application of a protocol using antibiotics, to confirm *M. smithii*'s sporulation capability [95, 96]. Such work represents a logical extension of our reported *in silico* results but goes beyond the scope of the present study. 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*. Moreover, Ruaud, Esquivel-Elizondo, de la Cuesta-Zuluaga *et al.* have provided evidence of a syntrophic relationship between *Firmicutes* bacteria and *M. smithii*. The co-occurrence of these microorganisms is likely facilitated by physical and metabolic interactions. In addition to this, genes *h9* and *h20* as well as their surrounding genes are expressed by the archaeal genomes sampled from human fecal samples.

Conclusion

Our study aimed to uncover the potential functions of archaeal proteins, particularly those encoded by *M. smithii*, in the human gut. Sequence similarity–based methods, while effective for highly similar proteins (>70%–80% identity), may not accurately represent the functions of archaeal proteins due to the lack of experimental validation. More specifically, publicly available databases have limited experimentally validated archaeal sequences compared to bacterial and eukaryotic proteins (~7 000 000 archaeal, ~166 000 000 bacterial, and ~70 000 000 eukaryotic proteins, UniProtKB Jun 2023) making sequence-based protein annotations applicable to only a subset of archaeal proteins. In contrast, recent deep learning-based methods enable protein structure prediction and annotation without relying on high sequence similarity, allowing for functional similarity beyond close sequence matches. We used structural methods to improve the annotation of archaeal proteins, gaining better insights into their functions compared to traditional sequence-based methods. This approach allowed us to refine some existing annotations and discover new functions for others, giving us valuable insights into the roles of archaeal genes in the human gut. Our findings focus on the characterization of human-associated and gut-specific proteins identified in *M. smithii*, a metabolically proficient and clinically relevant methanogenic archaeon known to be linked to gastrointestinal disorders, including IBD and obesity. In upcoming research, the primary focus should be on improving the accuracy of determining translation initiation and termination sites through the integration of additional specialized tools [97, 98], as this holds significant promise for enhancing structural predictions. Furthermore, the refinement of our computational efforts with experimental approaches holds the key to elucidating the predicted protein structures and their corresponding functions.

Acknowledgements

All authors proof-read and approved of the content in this research paper. We are especially grateful for the critical feedback and suggestions provided by Dr Cedric Laczny. Experiments presented in this work were carried out using the HPC facility of the University of Luxembourg.

Supplementary material

Supplementary material is available at *The ISME Journal* online.

Conflicts of interest

None declared.

Funding

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).

Data availability

Microbial MAGs from UHGG collection are available from the MGnify FTP site at http://ftp.ebi.ac.uk/pub/databases/metagenomics/mgnify_genomes/; MAGs from the GEM catalog are accessible at https://portal.nersc.gov/GEM/. Metatranscriptomic sequencing reads are available from NCBI BioProject PRJNA289586 and assembled contigs can be assessed at MG-RAST (submission IDs are indicated in MT_assembly_RAST_ids.xlsx). A description of the analyses including pre-processing steps along with the scripts for the main analysis, archaeal gut–specific unique and homologous sourmash clusters, and synteny plots can be found at GitLab: https://gitlab.lcsb.uni.lu/polina.novikova/archaea-in-gut.

References

1. Woese CR, Fox GE. Phylogenetic structure of the prokaryotic domain: the primary kingdoms. *Proc Natl Acad Sci U S A* 1977;**74**:5088–90. https://doi.org/10.1073/pnas.74.11.5088
2. Balch WE, Magrum LJ, Fox GE *et al.* An ancient divergence among the bacteria. *J Mol Evol* 1977;**9**:305–11. https://doi.org/10.1007/BF01796092
3. Fox GE, Magrum LJ, Balch WE *et al.* Classification of methanogenic bacteria by 16S ribosomal RNA characterization. *Proc Natl Acad Sci U S A* 1977;**74**:4537–41. https://doi.org/10.1073/pnas.74.10.4537
4. Woese CR, Kandler O, Wheelis ML. Towards a natural system of organisms: proposal for the domains archaea, bacteria, and eucarya. *Proc Natl Acad Sci U S A* 1990;**87**:4576–9. https://doi.org/10.1073/pnas.87.12.4576
5. Liu Y, Makarova KS, Huang WC *et al.* Expanded diversity of Asgard archaea and their relationships with eukaryotes. *Nature* 2021;**593**:553–7. https://doi.org/10.1038/s41586-021-03494-3
6. Williams TA, Cox CJ, Foster PG *et al.* Phylogenomics provides robust support for a two-domains tree of life. *Nat Ecol Evol* 2020;**4**:138–47. https://doi.org/10.1038/s41559-019-1040-x
7. Könneke M, Bernhard AE, de la Torre *et al.* Isolation of an autotrophic ammonia-oxidizing marine archaeon. *Nature* 2005;**437**:543–6. https://doi.org/10.1038/nature03911
8. Pester M, Schleper C, Wagner M. The Thaumarchaeota: an emerging view of their phylogeny and ecophysiology. *Curr Opin Microbiol* 2011;**14**:300–6. https://doi.org/10.1016/j.mib.2011.04.007

9. DeLong EF. Everything in moderation: archaea as "non-extremophiles". *Curr Opin Genet Dev* 1998;**8**:649–54. https://doi.org/10.1016/s0959-437x(98)80032-4
10. Schleper C. Ammonia oxidation: different niches for bacteria and archaea? *ISME J* 2010;**4**:1092–4. https://doi.org/10.1038/ismej.2010.111
11. Valentine DL. Adaptations to energy stress dictate the ecology and evolution of the archaea. *Nat Rev Microbiol* 2007;**5**:316–23. https://doi.org/10.1038/nrmicro1619
12. Hoegenauer C, Hammer HF, Mahnert A *et al.* Methanogenic archaea in the human gastrointestinal tract. *Nat Rev Gastroenterol Hepatol* 2022;**19**:805–13. https://doi.org/10.1038/s41575-022-00673-z
13. Thomas CM, Desmond-Le Quéméner E, Gribaldo S *et al.* Factors shaping the abundance and diversity of the gut archaeome across the animal kingdom. *Nat Commun* 2022;**13**:3358. https://doi.org/10.1038/s41467-022-31038-4
14. Moissl-Eichinger C, Probst AJ, Birarda G *et al.* Human age and skin physiology shape diversity and abundance of archaea on skin. *Sci Rep* 2017;**7**:4039. https://doi.org/10.1038/s41598-017-04197-4
15. Probst AJ, Auerbach AK, Moissl-Eichinger C. Archaea on human skin. *PLoS One* 2013;**8**:e65388. https://doi.org/10.1371/journal.pone.0065388
16. Kumpitsch C, Koskinen K, Schöpf V *et al.* The microbiome of the upper respiratory tract in health and disease. *BMC Biol* 2019;**17**:87. https://doi.org/10.1186/s12915-019-0703-z
17. Sogodogo E, Doumbo O, Aboudharam G *et al.* First characterization of methanogens in oral cavity in Malian patients with oral cavity pathologies. *BMC Oral Health* 2019;**19**:232. https://doi.org/10.1186/s12903-019-0929-8
18. Kim JY, Whon TW, Lim MY *et al.* The human gut archaeome: identification of diverse haloarchaea in Korean subjects. *Microbiome* 2020;**8**:114. https://doi.org/10.1186/s40168-020-00894-x
19. Eckburg PB, Bik EM, Bernstein CN *et al.* Diversity of the human intestinal microbial flora. *Science* 2005;**308**:1635–8. https://doi.org/10.1126/science.1110591
20. Ghavami SB, Rostami E, Sephay AA *et al.* Alterations of the human gut Methanobrevibacter smithii as a biomarker for inflammatory bowel diseases. *Microb Pathog* 2018;**117**:285–9. https://doi.org/10.1016/j.micpath.2018.01.029
21. Houshyar Y, Massimino L, Lamparelli LA *et al.* Going beyond bacteria: uncovering the role of archaeome and mycobiome in inflammatory bowel disease. *Front Physiol* 2021;**12**:783295. https://doi.org/10.3389/fphys.2021.783295
22. Basseri RJ, Basseri B, Pimentel M *et al.* Intestinal methane production in obese individuals is associated with a higher body mass index. *Gastroenterol Hepatol* 2012;**8**:22–8
23. Samuel BS, Gordon JI. A humanized gnotobiotic mouse model of host–archaeal–bacterial mutualism. *Proc Natl Acad Sci U S A* 2006;**103**:10011–6. https://doi.org/10.1073/pnas.0602187103
24. Mathur R, Amichai M, Chua KS *et al.* Methane and hydrogen positivity on breath test is associated with greater body mass index and body fat. *J Clin Endocrinol Metab* 2013;**98**:E698–702. https://doi.org/10.1210/jc.2012-3144
25. Borrel G, Brugère J-F, Gribaldo S *et al.* The host-associated archaeome. *Nat Rev Microbiol* 2020;**18**:622–36. https://doi.org/10.1038/s41579-020-0407-y
26. Borrel G, McCann A, Deane J *et al.* Genomics and metagenomics of trimethylamine-utilizing archaea in the human gut microbiome. *ISME J* 2017;**11**:2059–74. https://doi.org/10.1038/ismej.2017.72
27. Bang C, Weidenbach K, Gutsmann T *et al.* The intestinal archaea *Methanosphaera stadtmanae* and *Methanobrevibacter smithii* activate human dendritic cells. *PLoS One* 2014;**9**:e99411. https://doi.org/10.1371/journal.pone.0099411
28. Lyu Z, Whitman WB. Transplanting the pathway engineering toolbox to methanogens. *Curr Opin Biotechnol* 2019;**59**:46–54. https://doi.org/10.1016/j.copbio.2019.02.009
29. Thomsen J, Weidenbach K, Metcalf WW *et al.* Genetic methods and construction of chromosomal mutations in methanogenic archaea. *Methods Mol Biol* 2022;**2522**:105–17. https://doi.org/10.1007/978-1-0716-2445-6_6
30. Tebbe A, Klein C, Bisle B *et al.* Analysis of the cytosolic proteome of *Halobacterium salinarum* and its implication for genome annotation. *Proteomics* 2005;**5**:168–79. https://doi.org/10.1002/pmic.200400910
31. Zaremba-Niedzwiedzka K, Caceres EF, Saw JH *et al.* Asgard archaea illuminate the origin of eukaryotic cellular complexity. *Nature* 2017;**541**:353–8. https://doi.org/10.1038/nature21031
32. Spang A, Saw JH, Jørgensen SL *et al.* Complex archaea that bridge the gap between prokaryotes and eukaryotes. *Nature* 2015;**521**:173–9. https://doi.org/10.1038/nature14447
33. Castelle CJ, Wrighton KC, Thomas BC *et al.* Genomic expansion of domain archaea highlights roles for organisms from new phyla in anaerobic carbon cycling. *Curr Biol* 2015;**25**:690–701. https://doi.org/10.1016/j.cub.2015.01.014
34. Bork P. Powers and pitfalls in sequence analysis: the 70% hurdle. *Genome Res* 2000;**10**:398–400. https://doi.org/10.1101/gr.10.4.398
35. Wooley JC, Godzik A, Friedberg I. A primer on metagenomics. *PLoS Comput Biol* 2010;**6**:e1000667. https://doi.org/10.1371/journal.pcbi.1000667
36. Ellens KW, Christian N, Singh C *et al.* Confronting the catalytic dark matter encoded by sequenced genomes. *Nucleic Acids Res* 2017;**45**:11495–514. https://doi.org/10.1093/nar/gkx937
37. Makarova KS, Wolf YI, Koonin EV. Towards functional characterization of archaeal genomic dark matter. *Biochem Soc Trans* 2019;**47**:389–98. https://doi.org/10.1042/BST20180560
38. Márquez V, Fröhlich T, Armache JP *et al.* Proteomic characterization of archaeal ribosomes reveals the presence of novel archaeal-specific ribosomal proteins. *J Mol Biol* 2011;**405**:1215–32. https://doi.org/10.1016/j.jmb.2010.11.055
39. Wu P, Brockenbrough JS, Paddy MR *et al.* NCL1, a novel gene for a non-essential nuclear protein in *Saccharomyces cerevisiae*. *Gene* 1998;**220**:109–17. https://doi.org/10.1016/s0378-1119(98)00330-8
40. Vakirlis N, Carvunis A-R, McLysaght A, 'Synteny-based analyses indicate that sequence divergence is not the main source of orphan genes', *eLife* 2020;**9**:e53500. https://doi.org/10.7554/eLife.53500.
41. Weisman CM, Murray AW, Eddy SR. Many, but not all, lineage-specific genes can be explained by homology detection failure. *PLoS Biol* 2020;**18**:e3000862. https://doi.org/10.1371/journal.pbio.3000862
42. Sinha S, Lynn AM, Desai DK. Implementation of homology based and non-homology based computational methods for the identification and annotation of orphan enzymes: using *Mycobacterium tuberculosis* H37Rv as a case study. *BMC Bioinformatics* 2020;**21**:466. https://doi.org/10.1186/s12859-020-03794-x
43. Mahnert A, Blohs M, Pausan M-R *et al.* The human archaeome: methodological pitfalls and knowledge gaps. *Emerg Top Life Sci* 2018;**2**:469–82. https://doi.org/10.1042/ETLS20180037
44. Watson JD, Sanderson S, Ezersky A *et al.* Towards fully automated structure-based function prediction in structural genomics: a case study. *J Mol Biol* 2007;**367**:1511–22. https://doi.org/10.1016/j.jmb.2007.01.063

45. Chothia C, Lesk AM. The relation between the divergence of sequence and structure in proteins. *EMBO J* 1986;**5**:823–6. https://doi.org/10.1002/j.1460-2075.1986.tb04288.x
46. Skolnick J, Gao M, Zhou H *et al.* AlphaFold 2: why it works and its implications for understanding the relationships of protein sequence, structure, and function. *J Chem Inf Model* 2021;**61**:4827–31. https://doi.org/10.1021/acs.jcim.1c01114
47. Jumper J, Evans R, Pritzel A *et al.* Highly accurate protein structure prediction with AlphaFold. *Nature* 2021;**596**:583–9. https://doi.org/10.1038/s41586-021-03819-2
48. Du Z, Su H, Wang W *et al.* The trRosetta server for fast and accurate protein structure prediction. *Nat Protoc* 2021;**16**:5634–51. https://doi.org/10.1038/s41596-021-00628-9
49. Nayfach S, Roux S, Seshadri R *et al.* A genomic catalog of Earth's microbiomes. *Nat Biotechnol* 2021;**39**:499–509. https://doi.org/10.1038/s41587-020-0718-6
50. Almeida A, Nayfach S, Boland M *et al.* A unified catalog of 204,938 reference genomes from the human gut microbiome. *Nat Biotechnol* 2021;**39**:105–14. https://doi.org/10.1038/s41587-020-0603-3
51. Hyatt D, Chen GL, LoCascio PF *et al.* Prodigal: prokaryotic gene recognition and translation initiation site identification. *BMC Bioinformatics* 2010;**11**:119. https://doi.org/10.1186/1471-2105-11-119.
52. Steinegger M, Söding J. MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets. *Nat Biotechnol* 2017;**35**:1026–8. https://doi.org/10.1038/nbt.3988
53. Steinegger M, Söding J. Clustering huge protein sequence sets in linear time. *Nat Commun* 2018;**9**:2542. https://doi.org/10.1038/s41467-018-04964-5
54. Brown CT, Irber L. Sourmash: a library for MinHash sketching of DNA. *J Open Source Softw* 2016;**1**:27. https://doi.org/10.21105/joss.00027
55. Pierce NT, Irber L, Reiter T *et al.* Large-scale sequence comparisons with sourmash. *F1000Res* 2019;**8**:1006. https://doi.org/10.12688/f1000research.19675.1
56. Queirós P, Delogu F, Hickl O *et al.* Mantis: flexible and consensus-driven genome annotation. *GigaScience* 2021;**10**:giab042. https://doi.org/10.1093/gigascience/giab042
57. Kryshtafovych A, Schwede T, Topf M, Fidelis K, Moult, J. Critical assessment of methods of protein structure prediction (CASP)-Round XIV. *Proteins* 2021;**89**(12):1607–17. https://doi.org/10.1002/prot.26237.
58. Laskowski RA, Watson JD, Thornton JM. ProFunc: a server for predicting protein function from 3D structure. *Nucleic Acids Res* 2005;**33**:W89–93. https://doi.org/10.1093/nar/gki414
59. Gligorijević V, Renfrew PD, Kosciolek T *et al.* Structure-based protein function prediction using graph convolutional networks. *Nat Commun* 2021;**12**:3168. https://doi.org/10.1038/s41467-021-23303-9
60. Heintz-Buschart A, May P, Laczny C et al. Integrated multi-omics of the human gut microbiome in a case study of familial type 1 diabetes. *Nat Microbiol* 2017;**2**:16180. https://doi.org/10.1038/nmicrobiol.2016.180.
61. Li H. *Aligning Sequence Reads, Clone Sequences and Assembly Contigs with BWA-MEM*. 2013, doi: https://doi.org/10.6084/M9.FIGSHARE.963153.V1.
62. Li H, Handsaker B, Wysoker A *et al.* The sequence alignment/map format and SAMtools. *Bioinformatics* 2009;**25**:2078–9. https://doi.org/10.1093/bioinformatics/btp352
63. Pedersen BS, Quinlan AR. Mosdepth: quick coverage calculation for genomes and exomes. *Bioinformatics* 2018;**34**:867–8. https://doi.org/10.1093/bioinformatics/btx699
64. Song W, Wemheuer B, Zhang S *et al.* MetaCHIP: community-level horizontal gene transfer identification through the combination of best-match and phylogenetic approaches. *Microbiome* 2019;**7**:36. https://doi.org/10.1186/s40168-019-0649-y
65. Shimoyama Y. *pyGenomeViz: A Genome Visualization Python Package for Comparative Genomics*. Jun. 2022. 23 July 2022, date last accessed. Available: https://github.com/moshi4/pyGenomeViz
66. Apweiler R, Bairoch A, Wu CH *et al.* UniProt: the universal protein knowledgebase. *Nucleic Acids Res* 2004;**32**:115D–19. https://doi.org/10.1093/nar/gkh131
67. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. *Journal of molecular biology* 1990;**215**(3):403–10. https://doi.org/10.1016/S0022-2836(05)80360-2.
68. Katoh K, Misawa K, Kuma K *et al.* MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. *Nucleic Acids Res* 2002;**30**:3059–66. https://doi.org/10.1093/nar/gkf436
69. 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* 2010;**10**:210. https://doi.org/10.1186/1471-2148-10-210
70. Nguyen L-T, Schmidt HA, von Haeseler A *et al.* IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. *Mol Biol Evol* 2015;**32**:268–74. https://doi.org/10.1093/molbev/msu300
71. Yu G. Using ggtree to visualize data on tree-like structures. *Curr Protoc Bioinform* 2020;**69**:e96. https://doi.org/10.1002/cpbi.96
72. Lin Z, Akin H, Rao R *et al.* Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science* 2023;**379**:1123–30. https://doi.org/10.1126/science.ade2574
73. Weissenow K, Heinzinger M, Steinegger M *et al.* Ultra-fast protein structure prediction to capture effects of sequence variation in mutation movies. bioRxiv. p. 2022.11.14.516473, 18 November 2022. https://doi.org/10.1101/2022.11.14.516473
74. Anandan A, Vrielink A. Structure and function of lipid A-modifying enzymes. *Ann N Y Acad Sci* 2020;**1459**:19–37. https://doi.org/10.1111/nyas.14244
75. Breazeale SD, Ribeiro AA, Raetz CRH. 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* 2003;**278**:24731–9. https://doi.org/10.1074/jbc.M304043200
76. Dell A, Galadari A, Sastre F *et al.* Similarities and differences in the glycosylation mechanisms in prokaryotes and eukaryotes. *Int J Microbiol* 2011;**2010**:e148178. https://doi.org/10.1155/2010/148178
77. Abu-Qarn M, Eichler J. Protein N-glycosylation in archaea: defining *Haloferax volcanii* genes involved in S-layer glycoprotein glycosylation. *Mol Microbiol* 2006;**61**:511–25. https://doi.org/10.1111/j.1365-2958.2006.05252.x
78. Kärcher U, Schröder H, Haslinger E *et al.* Primary structure of the heterosaccharide of the surface glycoprotein of *Methanothermus fervidus*. *J Biol Chem* 1993;**268**:26821–6. https://doi.org/10.1016/S0021-9258(19)74185-4
79. Zähringer U, Moll H, Hettmann T *et al.* Cytochrome b558/566 from the archaeon *Sulfolobus acidocaldarius* has a unique Asn-linked highly branched hexasaccharide chain containing 6-sulfoquinovose. *Eur J Biochem* 2000;**267**:4144–9. https://doi.org/10.1046/j.1432-1327.2000.01446.x
80. Mescher MF, Strominger JL. Purification and characterization of a prokaryotic glycoprotein from the cell envelope of

Halobacterium salinarium. *J Biol Chem* 1976;**251**:2005–14. https://doi.org/10.1016/S0021-9258(17)33647-5
81. Tamir A, Eichler J. N-glycosylation is important for proper *Haloferax volcanii* S-layer stability and function. *Appl Environ Microbiol* 2017;**83**:e03152–16. https://doi.org/10.1128/AEM.03152-16
82. Samuel BS, Hansen EE, Manchester JK *et al.* Genomic and metabolic adaptations of *Methanobrevibacter smithii* to the human gut. *Proc Natl Acad Sci U S A* 2007;**104**:10643–8. https://doi.org/10.1073/pnas.0704189104
83. Wandall HH, Nielsen MAI, King-Smith S *et al.* Global functions of O-glycosylation: promises and challenges in O-glycobiology. *FEBS J* 2021;**288**:7183–212. https://doi.org/10.1111/febs.16148
84. Metcalf WW, Griffin BM, Cicchillo RM *et al.* Synthesis of methylphosphonic acid by marine microbes: a source for methane in the aerobic ocean. *Science* 2012;**337**:1104–7. https://doi.org/10.1126/science.1219875
85. Dormoy-Raclet V, Markovits J, Malato Y *et al.* Unr, a cytoplasmic RNA-binding protein with cold-shock domains, is involved in control of apoptosis in ES and HuH7 cells. *Oncogene* 2007;**26**:2595–605. https://doi.org/10.1038/sj.onc.1210068
86. Roer L, Aarestrup FM, Hasman H. The EcoKI type I restriction-modification system in *Escherichia coli* affects but is not an absolute barrier for conjugation. *J Bacteriol* 2015;**197**:337–42. https://doi.org/10.1128/JB.02418-14
87. Pingoud A, Fuxreiter M, Pingoud V *et al.* Type II restriction endonucleases: structure and mechanism. *Cell Mol Life Sci* 2005;**62**:685–707. https://doi.org/10.1007/s00018-004-4513-1
88. Havemann GD, Sampson EM, Bobik TA. 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* 2002;**184**:1253–61. https://doi.org/10.1128/JB.184.5.1253-1261.2002
89. Kennedy NW, Ikonomova SP, Slininger Lee M *et al.* Self-assembling shell proteins PduA and PduJ have essential and redundant roles in bacterial microcompartment assembly. *J Mol Biol* 2021;**433**:166721. https://doi.org/10.1016/j.jmb.2020.11.020
90. Sampson EM, Bobik TA. Microcompartments for B12-dependent 1,2-propanediol degradation provide protection from DNA and cellular damage by a reactive metabolic intermediate. *J Bacteriol* 2008;**190**:2966–71. https://doi.org/10.1128/JB.01925-07
91. Hansen EE, Lozupone CA, Rey FE *et al.* Pan-genome of the dominant human gut-associated archaeon, *Methanobrevibacter smithii*, studied in twins. *Proc Natl Acad Sci U S A* 2011;**108**:4599–606. https://doi.org/10.1073/pnas.1000071108
92. Acar Kirit H, Bollback JP, Lagator M. The role of the environment in horizontal gene transfer. *Mol Biol Evol* 2022;**39**:msac220. https://doi.org/10.1093/molbev/msac220
93. Rigden DJ, Galperin MY. Sequence analysis of GerM and SpoVS, uncharacterized bacterial "sporulation" proteins with widespread phylogenetic distribution. *Bioinformatics* 2008;**24**:1793–7. https://doi.org/10.1093/bioinformatics/btn314
94. Onyenwoke RU, Brill JA, Farahi K *et al.* Sporulation genes in members of the low G+C Gram-type-positive phylogenetic branch (Firmicutes). *Arch Microbiol* 2004;**182**:182–92. https://doi.org/10.1007/s00203-004-0696-y
95. Pschorn W, Paulus H, Hansen J *et al.* Induction of sporulation in *Bacillus brevis*. *Eur J Biochem* 1982;**129**:403–7. https://doi.org/10.1111/j.1432-1033.1982.tb07064.x
96. Suárez JM, Edwards AN, McBride SM. The *Clostridium difficile* cpr locus is regulated by a noncontiguous two-component system in response to type a and B lantibiotics. *J Bacteriol* 2013;**195**:2621–31. https://doi.org/10.1128/JB.00166-13
97. Gleason AC, Ghadge G, Chen J *et al.* Machine learning predicts translation initiation sites in neurologic diseases with nucleotide repeat expansions. *PLoS One* 2022;**17**:e0256411. https://doi.org/10.1371/journal.pone.0256411
98. Zhang S, Hu H, Jiang T *et al.* TITER: predicting translation initiation sites by deep learning. *Bioinformatics* 2017;**33**:i234–42. https://doi.org/10.1093/bioinformatics/btx247