frontiers
in Microbiology

ORIGINAL RESEARCH
published: 18 August 2021
doi: 10.3389/fmicb.2021.721392

A Unique Gene Module in *Thermococcales* Archaea Centered on a Hypervariable Protein Containing Immunoglobulin Domains

Kira S. Makarova, Yuri I. Wolf, Svetlana Karamycheva and Eugene V. Koonin*

National Center for Biotechnology Information, National Library of Medicine, Bethesda, MD, United States

Molecular mechanisms involved in biological conflicts and self vs nonself recognition in archaea remain poorly characterized. We apply phylogenomic analysis to identify a hypervariable gene module that is widespread among *Thermococcales*. These loci consist of an upstream gene coding for a large protein containing several immunoglobulin (Ig) domains and unique combinations of downstream genes, some of which also contain Ig domains. In the large Ig domain containing protein, the C-terminal Ig domain sequence is hypervariable, apparently, as a result of recombination between genes from different *Thermococcales*. To reflect the hypervariability, we denote this gene module VARTIG (VARiable *Thermococcales* IG). The overall organization of the VARTIG modules is similar to the organization of Polymorphic Toxin Systems (PTS). Archaeal genomes outside *Thermococcales* encode a variety of Ig domain proteins, but no counterparts to VARTIG and no Ig domains with comparable levels of variability. The specific functions of VARTIG remain unknown but the identified features of this system imply three testable hypotheses: (i) involvement in inter-microbial conflicts analogous to PTS, (ii) role in innate immunity analogous to the vertebrate complement system, and (iii) function in self vs nonself discrimination analogous to the vertebrate Major Histocompatibility Complex. The latter two hypotheses seem to be of particular interest given the apparent analogy to the vertebrate immunity.

Keywords: archaea, *Thermococcales*, hypervariability, gene shuffling, immunoglobulin, polymorphic toxins

OPEN ACCESS

Edited by:
Yulong Shen,
Shandong University, China

Reviewed by:
Roger Garrett,
University of Copenhagen, Denmark
Sandip Paul,
JIS Institute of Advanced Studies and Research, India

Correspondence:
Eugene V. Koonin
koonin@ncbi.nlm.nih.gov

Specialty section:
This article was submitted to Biology of Archaea, a section of the journal Frontiers in Microbiology

Received: *07 June 2021*
Accepted: *22 July 2021*
Published: *18 August 2021*

Citation:
Makarova KS, Wolf YI, Karamycheva S and Koonin EV (2021) A Unique Gene Module in Thermococcales Archaea Centered on a Hypervariable Protein Containing Immunoglobulin Domains. Front. Microbiol. 12:721392. doi: 10.3389/fmicb.2021.721392

INTRODUCTION

Most archaea and bacteria inhabit multispecies environments where they are involved in complex interactions with each other and other community members (Huber et al., 2002; Keller and Surette, 2006; Schopf et al., 2008; Stewart and Franklin, 2008; Nadell et al., 2009; Castro et al., 2016; Flemming and Wuertz, 2019; Imachi et al., 2020). Even in pure cultures, cells interact and, under certain conditions, form biofilms, densely packed microbial communities attached to surfaces and surrounded by extracellular polymers (Lopez et al., 2010; Frols et al., 2012; Jachlewski et al., 2015; Flemming and Wuertz, 2019). In the case of archaea, the current understanding of the key molecular pathways and proteins involved in cell aggregation is limited although archaea known to form complex cell aggregates and biofilms (Chimileski et al., 2014; van Wolferen et al., 2018; Flemming and Wuertz, 2019). Furthermore, relatively little is known about the composition and functions of surface proteins and structures in archaea. Notable exceptions are the archaellum, a motility system, that has been studied in detail and shown to participate in cell adhesion and several other type IV pili structures also shown to be involved in cell adhesion in model organisms (Albers and Jarrell, 2015; Makarova et al., 2016; Chaudhury et al., 2018; Pohlschroder et al., 2018). Even the identity of the major S-layer protein comprising the archaeal cell envelope is known only for several model archaea, whereas the majority of the numerous lipoproteins encoded in archaeal genomes remain completely uncharacterized (Rodrigues-Oliveira et al., 2017; Pohlschroder et al., 2018).

In addition to the widespread biofilm formation, some specialized cell contacts have been described in prokaryotes (Konovalova and Sogaard-Andersen, 2011; Chaudhury et al., 2018; Peterson et al., 2020; Ruhe et al., 2020). In archaea, these include intercellular DNA exchange through different mechanisms, such as conjugation, which is mediated by dedicated systems typically encoded on plasmids that have been characterized, primarily, in Sulfolobales (Stedman et al., 2000; Erauso et al., 2006). Thermococci can form tubular structures with a row of internal vesicles carrying proteins and DNA (Marguet et al., 2013), and Halobacteria form intercellular bridges for mating (Sivabalasarma et al., 2020). Mechanistic details of the formation of such structures in archaea remain poorly understood. It also has been predicted that many archaea possess contact-dependent toxin delivery systems, known as polymorphic toxin systems (PTS) (Makarova et al., 2019), but so far, these have not been experimentally characterized.

Thermococcales are hyperthermophilic motile cocci that typically grow in media containing peptides, chitin, starch, or pyruvate, using sulfur or hydrogen as terminal electron acceptors (Rashid and Aslam, 2020). *Termococcus kodakarensis* and *Pyrococcus furiosus* are most commonly used as models (Leigh et al., 2011; Rashid and Aslam, 2020). These organisms are known to form contacts between cells of the same or even different species through their archaeallum filaments (Nather et al., 2006; Schopf et al., 2008).

Only a few surface proteins and structures of *Thermococcales* have been functionally characterized, including the components of archaeallum (Makarova et al., 2016; Daum et al., 2017) and the main S-layer protein (Goda et al., 2018). Altogether, the genome of *T. kodakarensis*, a model organism for this archaeal lineage, encodes 144 proteins (6.3%) predicted to be secreted, most of which remain uncharacterized (Pohlschroder et al., 2018).

We are interested in the functions and evolution of highly variable genes in bacteria and archaea, particularly, those that are subject to frequent horizontal gene transfer (HGT). During the systematic analysis of such genes, we identified a unique, hypervariable gene module in *Thermococcales*, which is centered on immunoglobulin domain-containing proteins predicted to be secreted. We discuss potential functions of this system, which we denote VARTIG (VARiable *Thermococcales* IG), in the light of its remarkable variability.

MATERIALS AND METHODS

Comparative Genomic Framework and Evolutionary Reconstructions

Genome sequences of 526 archaea with complete or nearly complete genomes were downloaded from the NCBI FTP site.[1] Sequences were assigned to the 2014 arCOGs using PSI-BLAST (Altschul et al., 1997) with the arCOG alignments used to generate position-specific scoring matrices (PSSM) sources as previously described (Makarova et al., 2015). Phyletic patterns, that is, patterns of presence-absence of arCOGs, were derived from the respective arCOGs assignments with a few patterns manually corrected after detailed sequence analysis.

Construction of COGs for *Thermococcales*

Protein complements of 42 *Thermococcales* genomes were clustered using UCLUST (Edgar, 2010) with sequence similarity threshold of 0.5; sequences within clusters were aligned using MUSCLE (Edgar, 2004). Cluster sequence alignments (as well as singletons) were compared as follows: each alignment was used as a query in a search against the database of alignment consensus sequences using PSI-BLAST (Altschul et al., 1997), with a cut-off *e*-value of 0.0001 (composition-based statistics on, low complexity filtering off). Matches covering at least 75% of the query length were considered full-length; when an alignment with more sequences had a full-length match to a longer alignment with fewer sequences, the longer alignment was cut into segments (putative domains) corresponding to the footprint of the shorter alignment. Alignments that had full-length matches to each other were merged using HHALIGN (Soding et al., 2005). Approximate Maximum Likelihood trees were reconstructed from alignments using FastTree (Price et al., 2010) with the WAG evolutionary model and gamma-distributed site rates; trees were rooted at midpoint. Clades (subtrees), that offer the best trade-off between the representation of species and the number of paralogs, calculated as the value of $S_C^2/(S_T N_C)$ index (where N_C is the number of leaves in a clade and S_C and S_T are the numbers of species in the clade and in the entire tree, respectively), were iteratively extracted from the tree as clusters of orthologs, and the alignments were partitioned accordingly. The procedure of merging the clusters that produce full-length alignments and splitting clusters with many paralogs was performed iteratively.

Reconstruction of Gene Gains and Losses

The scaffold phylogenetic tree for *Thermococcales* genomes was obtained from 16S rRNA sequences, aligned using MUSCLE program (Edgar, 2004). Approximate ML phylogenetic tree was constructed using FastTree (Price et al., 2010) (GTR evolutionary model, 20 site rate categories). The GLOOME program (Cohen et al., 2010) with independent gamma-distributed gain and loss rates was used to reconstruct the posterior probabilities for ancestral states for all *Thermococcales* COGs in all internal nodes

[1] https://ftp.ncbi.nlm.nih.gov/genomes/

of the scaffold tree, based on the COG phyletic pattern. Gene gains and losses were inferred from changes in the posterior probabilities ($P_{descendant}$ - $P_{ancestral}$ > 0.5 was interpreted as a gain; $P_{ancestral}$ - $P_{descendant}$ > 0.5 as a loss). Chromosomal segments of extant genomes with contiguous stretches of genes gained at the terminal branch of the phylogenetic tree, were defined as recently acquired islands.

Sequence Analysis

Iterative profile searches using PSI-BLAST (Altschul et al., 1997), with a cut-off *e*-value of 0.0001, and composition-based statistics and low complexity filtering turned off, were employed to search for similar sequences in either the NR (non-redundant) database or the protein sequence database of 524 archaeal genomes, unless indicated otherwise in a **Supplementary Figure** legend. HHsearch was used to identify proteins matching selected profiles from the PFAM database (Soding, 2005). Additionally, other sensitive methods for distant sequence similarity detection were employed including CDD-search (Marchler-Bauer et al., 2009), with cut-off *e*-value of 0.01 and low complexity filtering turned off, and HHpred search with default parameters against PDB, Pfam and CDD profile databases (Zimmermann et al., 2018). Transmembrane helices were predicted using TMHMM v. 2.0c with default parameters (Krogh et al., 2001). Signal peptides were predicted using SignalP v. 5.0 (Almagro Armenteros et al., 2019). Protein secondary structure was predicted using Jpred 4 (Drozdetskiy et al., 2015). Approximate Maximum Likelihood phylogenetic trees were constructed using FastTree with default parameters (Price et al., 2010).

RESULTS

Gene Content of Hypervariable, Recently Acquired Loci in Different Lineages of *Thermococcales*

We explored large genomic islands that appeared to have been recently acquired by different lineages of *Thermococcales* using information on the point of gene acquisition during the evolution of this group of archaea (see section "Materials and Methods" for details). Two such genomic islands containing more than 20 genes each were identified in the genome of *Thermococcus sp.* AM4 (TAM4_RS07620-TAM4_RS07720 and TAM4_RS09285-TAM4_RS09405) and were predicted to have been acquired by this organism only (**Supplementary Table 1**). These two islands did not share any genes from same Thermococci-specific COGs, but both encoded proteins of arCOG07710 and arCOG14200, initially annotated as "Cell surface protein" and "Uncharacterized protein," respectively. Both these arCOGs are represented in *Thermococcales* only (**Supplementary Table 2**). Further examination of the neighborhoods of arCOG07710 and arCOG14200 revealed 31 islands in 22 of the 42 analyzed thermococcal genomes. Almost all genes in these islands were predicted to be acquired at the respective terminal branches, with the exception of two strains of *P. furiosus*, where acquisition of such an island was mapped to the common ancestor of these strains (**Supplementary Tables 1, 2**). The length of the islands varied from 1 to 25 genes (**Figure 1**). Altogether, these islands contained 239 genes assigned to 120 thermococcal COGs including singletons, 68 of which are represented only in a single genome (**Supplementary Table 2**). None of these genes was detected in any genome location other than these islands, suggesting that they all belong to the same functional system. The arCOG07710 proteins are most common, being represented in 28 islands. However, the arCOG07710 proteins were assigned to 17 thermococcal COGs because, due to their extreme divergence, they were not merged in a single COG by our procedure and because several of the respective genes appear to be fragments of pseudogenes (**Figure 1** and **Supplementary Table 1**). Of the 120 thermococcal COGs represented in these islands, 82 consist of proteins predicted to contain one or more transmembrane segments, and in three more COGs, at least half of the proteins contain predicted signal peptides (**Supplementary Table 1**). The arCOG07710 proteins are always encoded at the 5′ end of a putative operon, followed by genes coding for smaller proteins, most of which contain predicted transmembrane segments (**Figure 1**). These observations strongly suggest that the putative functional system encoded in the islands is membrane-associated, and that its activity involves protein secretion.

Sequence Analysis of Proteins From the Hypervariable Islands

Given that the proteins encoded in the identified variable islands apparently evolve fast, we ran PSI-BLAST for up to three iterations against proteins from the arCOG database using a representative of each thermococcal COG as the query to determine whether some of these proteins could be further combined into broader groups of homologs. These searches allowed us to delineate 44 homology groups (**Supplementary Table 2**). The largest group combined members of arCOG07710 and arCOG14200, and a variety of other proteins encoded in the islands. Most of the arCOG07710 members are large proteins (>500 aa), whereas arCOG14200 members and other proteins of this homology group are smaller and align with the C-terminal region of the arCOG07710 protein (**Supplementary Table 1** and **Supplementary Figure 1A**). A signal peptide was predicted for 19 of the 30 arCOG07710 proteins, and for 8 proteins from this arCOG, a transmembrane domain was predicted at the C-terminus (**Supplementary Table 1** and **Supplementary Figure 1A**).

All strong matches (probability >90%) detected by HHpred were confined to the C-terminal half of the large arCOG07710 proteins, which aligns with other members of this homology group, as exemplified by PAP_07455 (WP_048165390.1) protein from *Palaeococcus pacificus* DY20341 (**Figure 2A**). In particular, the bacterial cell adhesion related domain CARDB (PF07705.13) was identified as a homolog (**Figure 2A**). CARDB belongs to PFAM clan CL0159 termed the "early" (E)-set and includes multiple protein families with the immunoglobulin (Ig) fold found in bacteria, archaea and eukaryotes, including fibronectin type-III, PKD repeats, bacterial Ig-like domains, and many

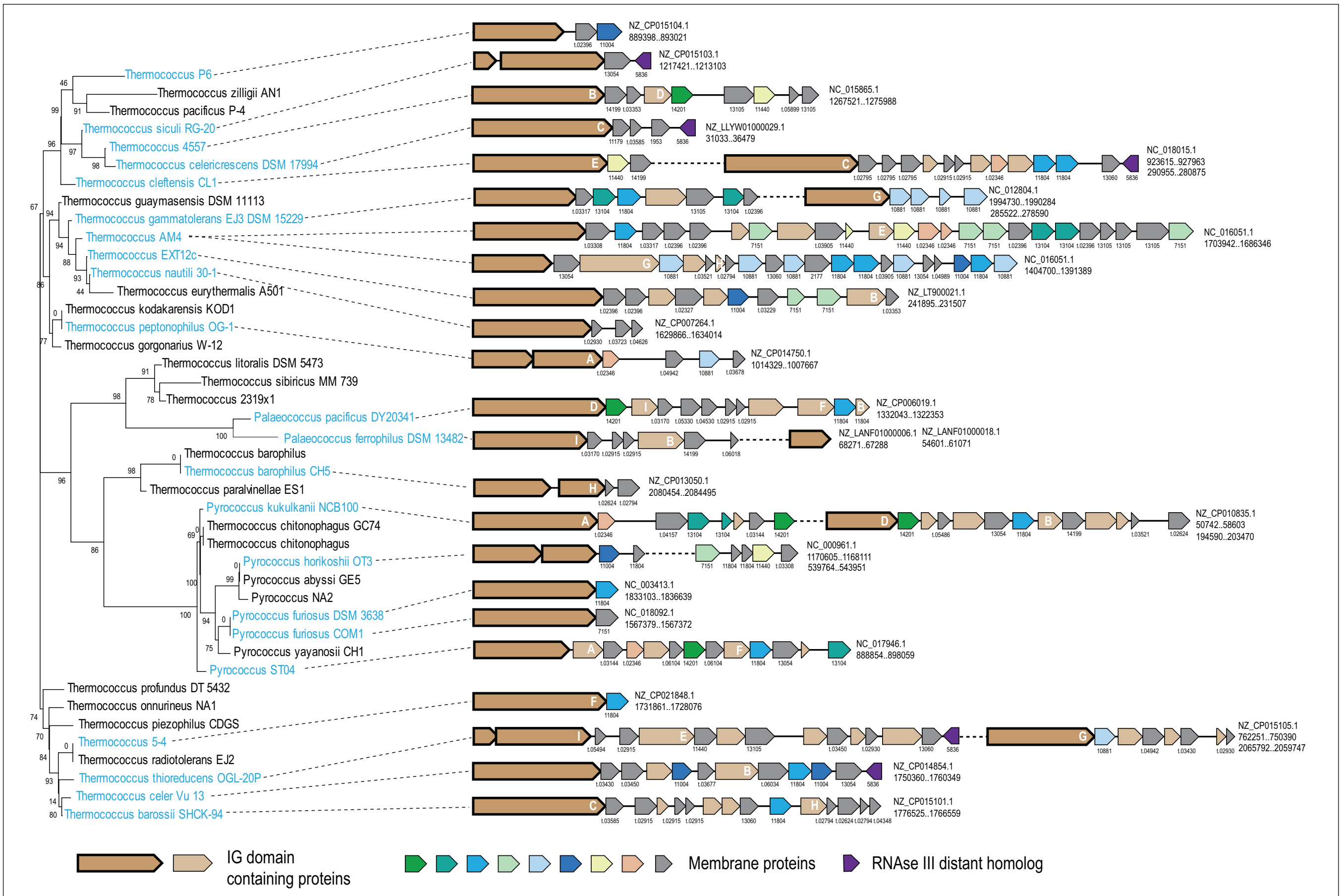
Thermococcus P6
Thermococcus zilligii AN1
Thermococcus pacificus P-4
Thermococcus siculi RG-20
Thermococcus 4557
Thermococcus celericrescens DSM 17994
Thermococcus cleftensis CL1
Thermococcus guaymasensis DSM 11113
Thermococcus gammatolerans EJ3 DSM 15229
Thermococcus AM4
Thermococcus EXT12c
Thermococcus nautili 30-1
Thermococcus eurythermalis A501
Thermococcus kodakarensis KOD1
Thermococcus peptonophilus OG-1
Thermococcus gorgonarius W-12
Thermococcus litoralis DSM 5473
Thermococcus sibiricus MM 739
Thermococcus 2319x1
Palaeococcus pacificus DY20341
Palaeococcus ferrophilus DSM 13482
Thermococcus barophilus
Thermococcus barophilus CH5
Thermococcus paralvinellae ES1
Pyrococcus kukulkanii NCB100
Thermococcus chitonophagus GC74
Thermococcus chitonophagus
Pyrococcus horikoshii OT3
Pyrococcus abyssi GE5
Pyrococcus NA2
Pyrococcus furiosus DSM 3638
Pyrococcus furiosus COM1
Pyrococcus yayanosii CH1
Pyrococcus ST04
Thermococcus profundus DT 5432
Thermococcus onnurineus NA1
Thermococcus piezophilus CDGS
Thermococcus 5-4
Thermococcus radiotolerans EJ2
Thermococcus thioreducens OGL-20P
Thermococcus celer Vu 13
Thermococcus barossii SHCK-94
NZ_CP015104.1 889398..893021
NZ_CP015103.1 1217421..1213103
NC_015865.1 1267521..1275988
NZ_LLYW01000029.1 31033..36479
NC_018015.1 923615..927963 290955..280875
NC_012804.1 1994730..1990284 285522..278590
NC_016051.1 1703942..1686346
NC_016051.1 1404700..1391389
NZ_LT900021.1 241895..231507
NZ_CP007264.1 1629866..1634014
NZ_CP014750.1 1014329..1007667
NZ_CP006019.1 1332043..1322353
NZ_LANF01000006.1 68271..67288
NZ_LANF01000018.1 54601..61071
NZ_CP013050.1 2080454..2084495
NZ_CP010835.1 50742..58603 194590..203470
NC_000961.1 1170605..1168111 539764..543951
NC_003413.1 1833103..1836639
NC_018092.1 1567379..1567372
NC_017946.1 888854..898059
NZ_CP021848.1 1731861..1728076
NZ_CP015105.1 762251..750390 2065792..2059747
NZ_CP014854.1 1750360..1760349
NZ_CP015101.1 1776525..1766559
IG domain containing proteins
Membrane proteins
RNAse III distant homolog

FIGURE 1 | Genomic organization of hypervariable VARTIG loci in *Thermococcales*. The loci are mapped to the 16S rRNA phylogenetic tree of *Thermococcales*. Names of species at the tree leaves where the loci are present are colored light blue. For each gene neighborhood, the genbank accession and coordinates of the locus are indicated on the right. Genes are shown by block arrows, roughly to scale. Genes are colored according to the homology groups (only if found five or more times in the loci) and the key shown at the bottom. Arrows corresponding to gene assigned to arCOG07710 are shown by a thick outline. Cluster number is indicated for all genes, except those coding for Ig-like domain containing proteins. For the genes that are in respective arCOGs the cluster number corresponds to the respective arCOG number. If the proteins are not assigned to any arCOG, the cluster number correspond to Thermococcus specific COGs and indicated with prefix "t." The letters A-H inside the genes coding for Ig-like domain containing proteins correspond to the respective branches in the tree built for the C-terminal region of these proteins (**Supplementary Figure 2**). Detailed information on these neighborhoods is available in **Supplementary Table 1**.

others[2] (Bork et al., 1994). Furthermore, a structure of a closely related single CARDB domain was solved for another cell surface protein from *Thermococcales* (PDB: 2KL6; PF1109 from *P. furiosus*). However, these proteins (arCOG07086) are encoded neither in the variable islands described here (hereafter VARTIG islands or loci) nor in any other identified variable loci although our reconstructions indicate a relatively high HGT propensity for these genes (**Supplementary Table 2**). The region from 556 to 824 amino acid (here and after mapped to WP_048165390.1) positions corresponds to the PDB: 4E9L (invasin-like protein FdeC) profile and, more specifically, to region B in FdeC (Nesta et al., 2012; **Figure 2A**). In the FdeC structure, region B accommodates three Ig-like domains, and the CARDB similarity region roughly corresponds to a single Ig unit. The TIGR03900, carboxyl-terminal-processing protease, profile can be mapped to the same region of arCOG07710 as FdeC but also displays similarity to two more regions covering 527 amino acids altogether (**Figure 2A**). The longest region of similarity corresponds to PDB:2PN5, insect thioester-containing protein (TEP) (**Figure 2A**), which contains five Ig domains (Baxter et al., 2007). Thus, we predict that arCOG07710 proteins contain five Ig-like domains that cover the region from 422 to 952 amino acid positions (**Figure 2A**).

We next examined the multiple alignment of the arCOG07710 proteins to identify any distinct shared features and detected a highly conserved N-terminal region containing a GN motif located close to the Ig domain (**Figure 2A** and **Supplementary Figure 1A**). Considering parallels with PTS and other surface proteins involved in immunity, such as bacteriocins and archaeocins, the GN motif might comprise a protease cleavage site resembling several known cleavage sites that all include a glycine residue (Dirix et al., 2004; Shen, 2010; Makarova et al., 2019). Many polymorphic toxins contain a protease domain that is involved in the protein maturation (Prochazkova et al., 2009;

[2] http://pfam.xfam.org/clan/E-set

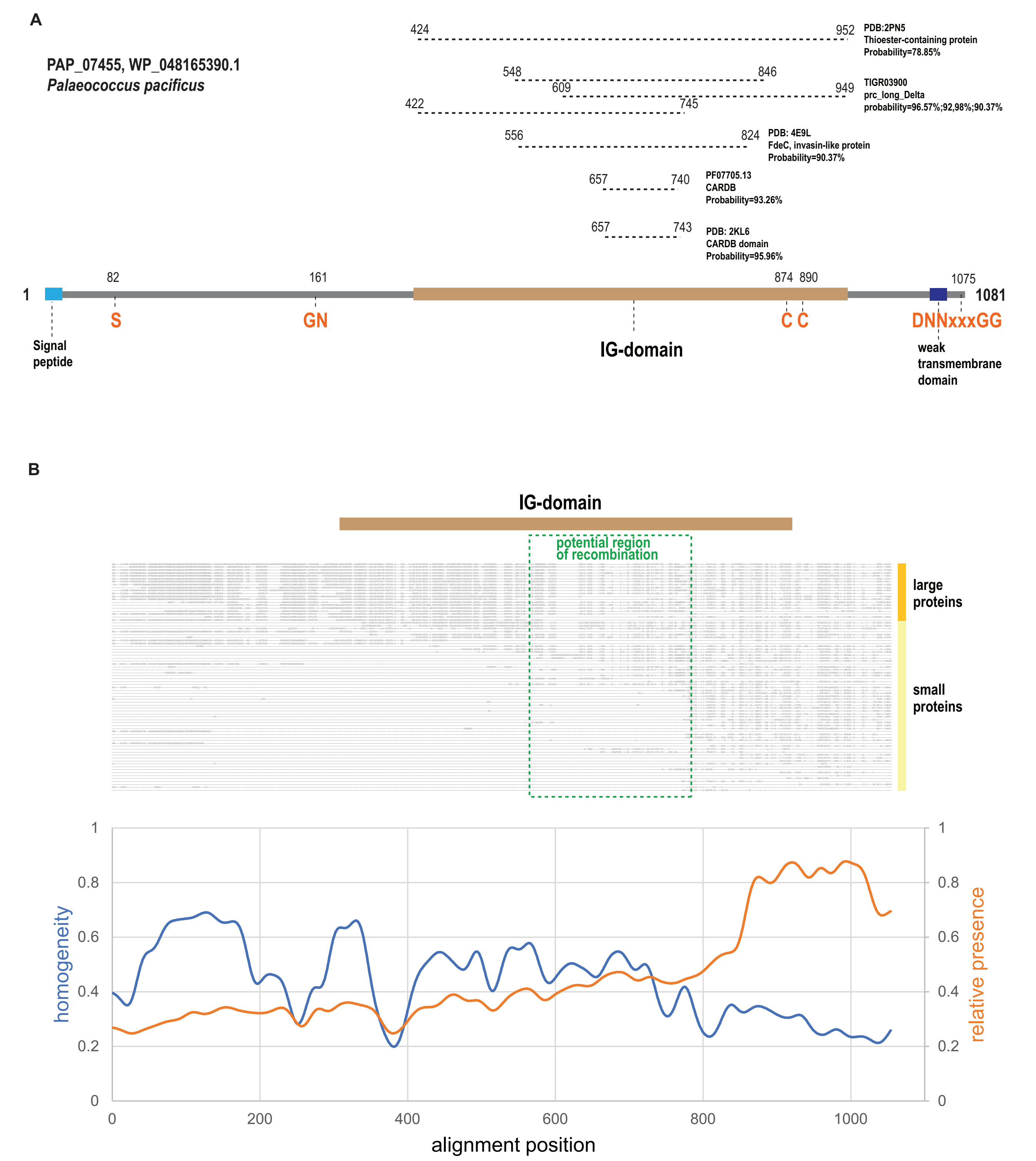

FIGURE 2 | Domain organization, sequence motifs and C-terminal variability of Ig-domain containing proteins encoded in VARTIG loci. **(A)** Domains and sequence motifs discussed in the text are mapped on a representative protein of arCOG07710. Coordinates of the motifs are indicated. Regions of similarity to several Ig-domain containing proteins or profiles identified using HHpred are indicated above the domain scheme. Coordinates for the respective regions are indicated. **(B)** Schematic multiple alignment of Ig domain-containing proteins (**Supplementary Figure 1**) and sequence conservation plot. The Ig region and the potential area of recombination are mapped to the alignment. The smoothed curves show the fraction of sequences in the corresponding position of the alignment (orange, calculated for sequences shorter than 700 amino acids) and the alignment homogeneity (Esterman et al., 2021) (blue, calculated for sequences longer than 700 amino acids; homogeneity of 1 corresponds to perfect conservation, homogeneity of 0 corresponds to a random amino acid mixture).

Shen, 2010). Depending on the type of the protease, conserved histidines, cysteines, aspartates and/or serines that are directly involved in catalysis can be found in such protease domains. Thus, we checked the conserved region for the presence of such residues. The only strictly conserved residue we identified was a serine, suggesting that these proteins might contain a serine endoprotease (**Figure 2A** and **Supplementary Figure 1A**).

Next, we noticed two conserved cysteines in the region corresponding to the most distal Ig-domain (**Figure 2A** and **Supplementary Figure 1A**). In extracellular proteins including the Ig domains, cysteines often form disulfide bonds that additionally stabilize the mature protein (Wiedemann et al., 2020). We further noticed that 31 arCOG07710 proteins and smaller homologous proteins containing predicted Ig domains contained a DNNxxxGG (where x is any amino acid) motif at the C-terminal end (**Figure 2A** and **Supplementary Figure 1A**). This motif is often preceded by a hydrophobic region that in some cases is predicted to form a transmembrane segment. A short C-terminal conserved motif preceded by a hydrophobic region might be a recognition signal for a sortase, a dedicated enzyme that covalently attaches proteins to the cell surface, although this particular signal has not been described previously (Haft et al., 2012; Malik and Kim, 2019). All known sortases are membrane transpeptidases with a catalytic triad that includes an active cysteine, typically, along with a histidine and an asparagine (Haft et al., 2012; Malik and Kim, 2019). Sortases are often encoded in the immediate vicinity of the protein containing the cognate signal because most are specific for only a few or even a single target (Haft et al., 2012). Because we did not find the D[ND]NxxxG[GD] (x is any amino acid) motif in the 15 C-terminal amino acids of any other proteins of *Thermococcales*, it could be expected that a dedicated sortase (if any) would be encoded in the same loci. However, we failed to identify any protein family encoded within these loci that would possess a conserved cysteine along with the other conserved residues that might contribute to the catalytic triad of a sortase. The possibility remains that arCOG07710 proteins are processed by a sortase that is encoded elsewhere in the thermococcal genomes. Proteins of one small family (arCOG14201) represented in the variable islands contained several conserved histidines suggesting that these could be metalloenzymes (**Supplementary Figure 1B**). The function of these putative enzymes remains unclear.

Evidence of Frequent Recombination in the C-Terminal Ig Domain-Containing Region of arCOG07710 Proteins

Examination of the alignment of the Ig-domain containing proteins from VARTIG (**Figure 2B**) showed that the smaller proteins aligned with the C-terminal region of the larger (arCOG07710) proteins. The plot of amino acid residue conservation for the shorter proteins shows a prominent plateau corresponding to the last 200 amino acids of the longer proteins (**Figure 2B**). This region contains the most distal Ig domain (∼100 aa), including conserved cysteines and the C-terminal region with the putative sorting signal. In contrast, the conservation plot for the large proteins only (20 proteins >700 aa) shows that this C-terminal region is the most variable portion of these proteins (**Figure 2B**). The likely explanation of such diversity is frequent recombination generating many variants of the distal Ig domain, which remains in the putative mature protein after the cleavage of the sorting signal.

To further analyze this apparent recombination, we constructed a phylogenetic tree for the ∼200 aa C-terminal region of arCOG07710 only and mapped 9 well-supported branches (A–I) each consisting of closely similar sequences on the respective gene neighborhoods (**Figure 1** and **Supplementary Figure 2**). This mapping showed that the level of sequence conservation did not follow the species tree, with the most similar sequences scattered among the tree branches. No indications of recombination within the loci were detected. Instead, according to the species tree, all exchanges occurred between different species, at least on several occasions, *in situ*, next to a gene encoding a distant homolog of RNAse III (arCOG05836) (**Figure 1**). This observation strongly suggests that the smaller genes recombine with the long genes in apparently random combinations and that this shuffling might occur during DNA exchange between cells from different populations or even species.

Given the hypervariability of the C-terminal Ig domain of the arCOG07710 proteins, we denoted this gene module VARTIG (after VARiable *Thermococcales* IG). Although we did not detect intra-gene recombination events in other gene families in the VARTIG loci, the fact that all these loci have different gene arrangements (**Figure 1**) strongly suggests that they undergo extensive shuffling during the evolution of *Thermococcales*.

Search for Analogous, Hypervariable Gene Modules in Other Archaea

In order to test for the presence of analogous, hypervariable systems in other archaea, we employed three distinct strategies. First, we performed a PSI-BLAST search against the arCOG database using the sequence of PAP_07455 (WP_048165390.1) protein from *P. pacificus*, a typical arCOG07710 member, as the query. This search converged after 5 iterations and identified, in addition to the homologous proteins from *Thermococcales*, 139 proteins from other groups of archaea, mostly Halobacteria (83) and Methanomicrobia (45) (**Supplementary Table 3**). According to the automatic arCOG assignments, these proteins possess 110 distinct domain architectures, the most frequent one (combination of arCOGs 2540, 3256, and 7560) identified in only 9 proteins from Halobacteria (e.g., WP_082230105.1). These are large (predicted) secreted proteins containing several CARDB domains. The arCOGs identified most frequently in these proteins are arCOG07560, multidomain secreted protein, often associated with PTS (71), and arCOG02508, secreted proteins, containing PKD repeats (68). The genomic neighborhoods (5 genes upstream and downstream) of these genes were examined in an attempt to identify variable loci that might have similar organization to VARTIG but would be characteristic of other major archaeal lineages (**Supplementary Table 3**).

In the second approach, we examined the neighborhoods (five genes upstream and downstream) of genes encoding

CARDB domain-containing proteins identified by similarity to the respective PFAM profile (pfam07705). Altogether, we analyzed 858 neighborhoods of genes encoding 914 non-redundant CARDB-containing proteins (**Supplementary Table 3**). According to the automatic arCOG assignments, these proteins display 489 distinct domain architectures, the most frequent one being a single CARDB domain containing protein (arCOG02532).

The third strategy was to search for any Ig-domain containing proteins encoded in a close vicinity of each other. To this end, we used assignments of all proteins from the arCOGs to the 238 PFAMs profiles within the E-set clan (see text footnote 2, **Supplementary Table 3**). We identified those loci that encoded two or more Ig-containing proteins separated by no more than five other genes (**Supplementary Table 3**). Altogether, we visually examined 38 sets of such loci, in search for organizations similar to that of VARTIG. In particular, we searched for loci specific to a major archaeal lineage and encoding a large Ig-domain containing protein (1000 aa or more) with a variable C-terminal region, followed by a variable number of other genes, at least one of which would also encode an Ig-domain containing protein.

Neither of these approaches identified any gene modules closely resembling VARTIG in other archaea. However, we detected many loci encoding several large Ig-domain containing proteins, often with different but overlapping domain organizations (**Supplementary Table 3**). In particular, Methanomicrobiales encompass numerous loci that consist of multiple genes encoding putative surface proteins containing pfam18911, PKD repeats of the Ig fold (for example, WP_011024165.1-WP_011024196.1 in *Methanosarcina acetivorans* C2A). It could be expected that genes in such loci would frequently recombine, resulting in new domain combinations and/or mosaic protein containing several domains of different origins. Indeed, we identified several examples of apparent mosaic proteins in Methanosarcinales (**Supplementary Figure 3**). Unlike the case of VARTIG, however, in many of these loci, the C-terminal Ig domain is either conserved or homogenized by recombination.

Archaeal Gene Neighborhoods Encoding α2-Macroglobulin Homologs

Considering the pronounced similarity of the domain organizations and sequences between the Ig domain-containing regions of the arCOG07710 proteins and TEP, several distant homologs from other archaea appeared of interest. Specifically, we identified a group of archaeal proteins with an even greater similarity to TEPs and other α2-macroglobulin family members from several arCOGs (**Supplementary Tables 2, 3**) that are encoded in archaeal genomic neighborhoods sharing genes with PTS (Makarova et al., 2019). Several of these proteins, primarily in Halobacteria and Methanomicrobia, were picked up by our third search strategy because they are encoded in the immediate vicinity of other Ig domain-containing proteins (**Figure 3A** and **Supplementary Table 3**). Sequence analysis of these proteins (for example, WP_050459478.1 from *Haloferax gibbonsii* ARA6) showed that, in addition to the Ig-domain, they contain the thioester-containing alpha helical domain (TED), in an arrangement similar to TEPs and alpha 2-macroglobulins (**Figure 3A**). Furthermore, these archaeal proteins contain the motif "CxEQ," where the catalytic cysteine forms a covalent bond with the target (Blandin and Levashina, 2004; Baxter et al., 2007; Rehman et al., 2013).

Next, we searched the arCOG database using PSI-BLAST and the TED domain from WP_050459478.1 as a query and identified the TED domain with intact "CxEQ" motif in 27 proteins from 24 archaeal genomes (**Supplementary Figure 1C** and **Supplementary Tables 2, 3**). Several proteins from Methanomicrobia are closely related to bacterial α2-macroglobulin homologs (Garcia-Ferrer et al., 2015), for example, WP_015053894.1 from *Methanolobus psychrophilus*, and thus, most likely, were acquired from bacteria via HGT (**Figure 3A**). The remaining α2-macroglobulin homologs are specific to archaea and typically contain a predicted signal peptide, but otherwise, have different domain organizations in different archaea (**Figure 3A**). In some archaeal genomes, genes encoding α2-macroglobulin homologs are located in complex neighborhoods along with other putative surface proteins, some of which contain Ig domains, antimicrobial peptide transporters, endoproteases and uncharacterized proteins. These loci are not as diverse as the VARTIG loci, and we found no evidence of shuffling in the C-terminal regions of these proteins (**Figure 3B** and **Supplementary Table 2**). Notably, however, in Halobacteria, the α2-macroglobulin domain-containing genes are often followed by genes of arCOG14565 encoding an uncharacterized protein that is also encoded next to many PTS (Makarova et al., 2019; **Figure 3B** and **Supplementary Table 3**). Additionally, the transcriptional regulator from arCOG02808 associated with Halobacterial alpha 2-macroglobulin modules were detected in the vicinity of genes encoding archaeocins (Makarova et al., 2019). Taken together, these observations suggest that archaeal immune systems and systems involved in inter-microbial conflicts share components and, potentially, the mechanisms for active molecule delivery and domain shuffling.

DISCUSSION

Here, we describe the VARTIG modules that are unique to *Thermococcales* and consist of genes encoding multiple, highly variable Ig domain-containing proteins along with heterogeneous sets of additional proteins. We obtained strong indications of recombination occurring in the C-terminal region of the large VARTIG protein resulting in its extreme variability. Surprisingly, recombination appears to occur not between different VARTIG loci within the same genome, but rather, between loci from different strains and species of Thermococci. In addition, the VARTIG loci seem to be subject to extensive shuffling. Recombination and rearrangement in the VARTIG loci likely represent dedicated diversification mechanisms that are, in principle, analogous to the intragene rearrangements that occur in genes encoding various surface proteins in bacteria and archaea, such as PTS (Hill et al., 1994; Makarova et al., 2019), adhesins (Martin-Galiano, 2017),

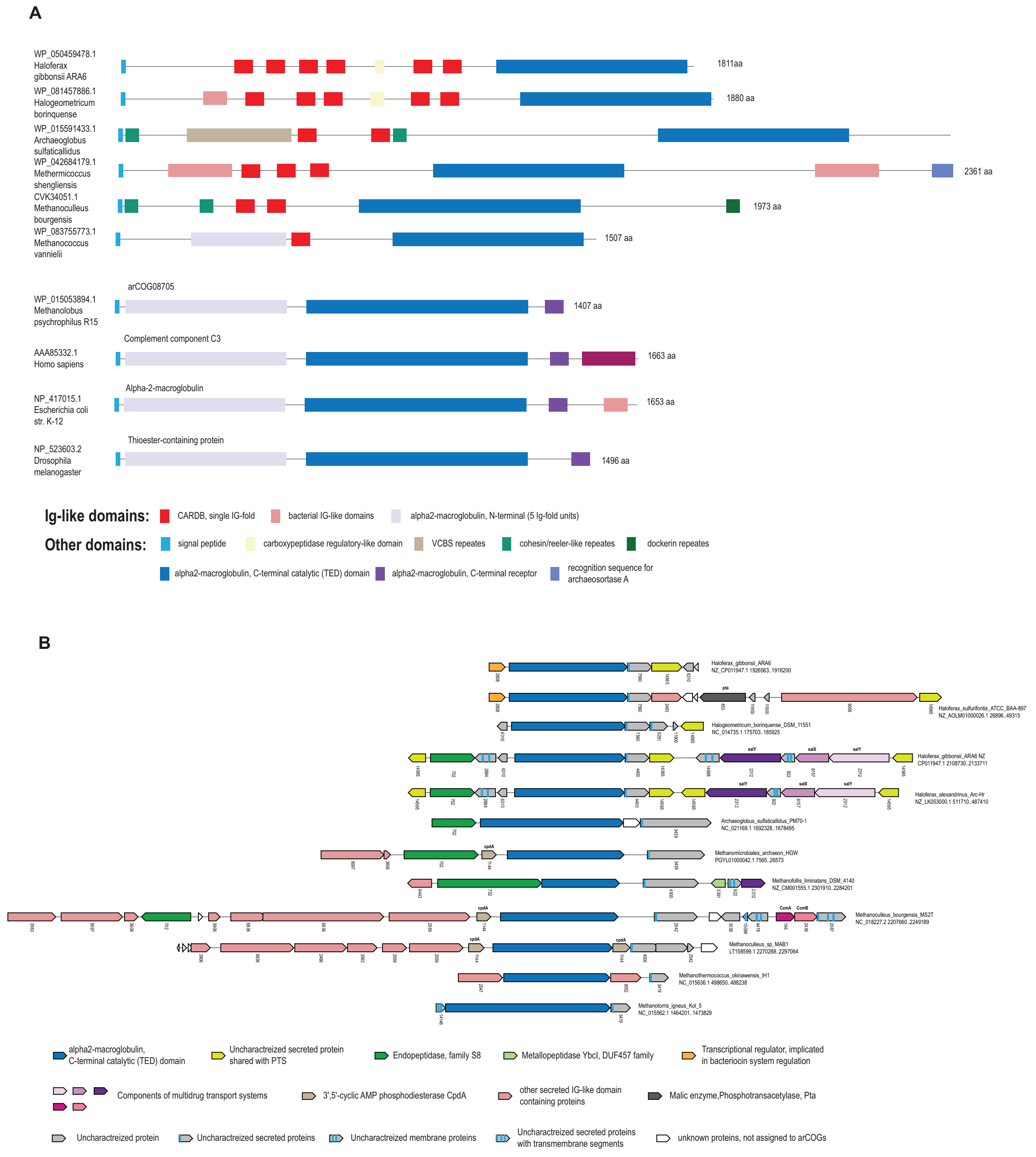
A
WP_050459478.1 Haloferax gibbonsii ARA6
1811aa
WP_081457886.1 Halogeometricum borinquense
1880 aa
WP_015591433.1 Archaeoglobus sulfaticallidus
WP_042684179.1 Methermicoccus shengliensis
2361 aa
CVK34051.1 Methanoculleus bourgensis
1973 aa
WP_083755773.1 Methanococcus vannielii
1507 aa
arCOG08705
WP_015053894.1 Methanolobus psychrophilus R15
1407 aa
Complement component C3
AAA85332.1 Homo sapiens
1663 aa
Alpha-2-macroglobulin
NP_417015.1 Escherichia coli str. K-12
1653 aa
Thioester-containing protein
NP_523603.2 Drosophila melanogaster
1496 aa
Ig-like domains: CARDB, single IG-fold; bacterial IG-like domains; alpha2-macroglobulin, N-terminal (5 Ig-fold units)
Other domains: signal peptide; carboxypeptidase regulatory-like domain; VCBS repeates; cohesin/reeler-like repeates; dockerin repeates
alpha2-macroglobulin, C-terminal catalytic (TED) domain; alpha2-macroglobulin, C-terminal receptor; recognition sequence for archaeosortase A
B
Haloferax_gibbonsii_ARA6 NZ_CP011947.1 1926563..1916200
2808 7560 14565 6310
pta
2808 7560 2483 853 11900 11900 9058 14565
Haloferax_sulfurifontis_ATCC_BAA-897 NZ_AOLM01000026.1 26896..49315
Halogeometricum_borinquense_DSM_11551 NC_014735.1 175703..185925
6310 7560 9291 11900 14565
salY salX salY
14565 702 2884 6310 4400 14565 14688 2312 922 8157 2312 14565
Haloferax_gibbonsii_ARA6 NZ CP011947.1 2108730..2133711
salY salX salY
14565 702 2884 6310 4400 14565 14565 2312 922 8157 2312 14565
Haloferax_alexandrinus_Arc-Hr NZ_LK053000.1 511710..487410
702 3439
Archaeoglobus_sulfaticallidus_PM70-1 NC_021169.1 1692328..1678495
cpdA
9597 3606 702 1144 3439
Methanomicrobiales_archaeon_HGW PGYL01000042.1 7565..26573
9443 702 4500 3391 922 2312
Methanofollis_liminatans_DSM_4140 NZ_CM001555.1 2301910..2284201
cpdA CcmA CcmB
3563 9597 3606 702 3906 6836 6836 2559 1144 2542 3038 15099 4478 196 2436 2087
Methanoculleus_bourgensis_MS2T NC_018227.2 2207660..2249189
cpdA cpdA
3906 6836 2486 3563 2559 2559 1144 1144 4500 2542
Methanoculleus_sp_MAB1 LT158599.1 2270288..2297064
2547 9552 3419
Methanothermococcus_okinawensis_IH1 NC_015636.1 498650..488238
14146 3419
Methanotorris_igneus_Kol_5 NC_015562.1 1464201..1473829
alpha2-macroglobulin, C-terminal catalytic (TED) domain
Uncharactreized secreted protein shared with PTS
Endopeptidase, family S8
Metallopeptidase YbcI, DUF457 family
Transcriptional regulator, implicated in bacteriocin system regulation
Components of multidrug transport systems
3',5'-cyclic AMP phosphodiesterase CpdA
other secreted IG-like domain containing proteins
Malic enzyme,Phosphotransacetylase, Pta
Uncharactreized protein
Uncharactreized secreted proteins
Uncharactreized membrane proteins
Uncharactreized secreted proteins with transmembrane segments
unknown proteins, not assigned to arCOGs

FIGURE 3 | Alpha 2-macroglobulin homologs in archaea. **(A)** The domain organizations of selected alpha 2-macroglobulin homologs. Domains are shown by colored rectangles according to the color key shown beneath the protein schematics. The length of the domain scheme and the domain boundaries are shown roughly to scale. For each protein, the GenBank accession and the species name are indicated. The length of each protein is indicated on the right. **(B)** Selected neighborhoods of genes coding for alpha 2-macroglobulin homologs. Designations are mostly the same as in the **Figure 1**. The arCOG numbers are indicated below the respective arrows. Gene names are indicated for a few known protein families. The color code is shown underneath the schematics.

and virulence factors (Dashper et al., 2017). Another, perhaps distant, but relevant analogy is the diversification of Ig in the vertebrate immune system that occurs via a combination of hypermutation and V(D)J recombination (Schatz and Swanson, 2011). However, diversification via inter-strain or even inter-species recombination appears to be a unique feature of VARTIG that implies the possibility of community-level adaptation.

The molecular mechanisms of recombination and rearrangement in the VARTIG loci remain unclear. There are not enough closely related genomes of *Thermococcales* currently available to align the nucleotide sequences of the genes coding for long Ig domain containing proteins and identify recombination sites precisely, especially, because these sites are likely spread across a region of about 600 nucleotides (**Figure 2B**). These genes seem to be too far diverged for homologous recombination, and the other genes in the VARTIG loci are even more variable. It is unclear if any integrase-like proteins are involved in VARTIG shuffling in a manner similar to the phase variation mechanism of restriction-modification systems (De Ste Croix et al., 2017), but no identifiable integrases are encoded in the VARTIG neighborhoods. Considering the parallels in the organizations of the VARTIG loci and PTS, it appears likely that the mechanisms of shuffling are similar as well. Apart from direct experimental study, genome sequencing of closely related strains should at least allow the identification of the sequence features promoting recombination.

The function of VARTIG remains enigmatic. Considering the patchy distribution and variability of VARTIG in *Thermococcales*, it is highly unlikely to be essential. We could not identify any consistent difference between the lifestyles of those Thermococci that encode VARTIG modules and those that lack them, at least, with the current state of the knowledge of the biology of these archaea. This pattern of phyletic distribution, and especially, the extensive diversification seems to be compatible with the involvement of VARTIG in inter-microbial conflicts (Aravind et al., 2012). Again, parallels with PTS are prominent. Both types of systems typically encompass a long protein encoded by the first gene in an array of co-directed genes. In PTS, the gene arrays include a variable number of genes encoding proteins homologous to the C-terminal regions of the longer protein that encompass diverse toxin domains (Zhang et al., 2012; Makarova et al., 2019). These loci also often encode immunity proteins, and typically, a dedicated immunity protein is encoded in the immediate vicinity of the respective toxin (Zhang et al., 2012; Makarova et al., 2019). A similar trend was observed in the VARTIG modules: genes encoding a distinct set of (predicted) membrane proteins often follow the same subfamily of genes coding for Ig domain proteins (**Figure 1**). For example, membrane proteins from cluster 14199 are encoded next to Ig domain protein from branch B, whereas membrane proteins from cluster 11440 are encoded next to a Ig domain protein from branch E. Furthermore, PTS containing Ig domains within the effector protein also have been identified in archaea (**Figure 3**; Makarova et al., 2019). The major caveat of the hypothesis that VARTIG is a distinct variety of PTS is that we could not identify any potential toxins, in particular, DNase or RNase domains, which are the most common toxin moieties in archaeal PTS (Makarova et al., 2019).

The second hypothesis is that VARTIG is an innate immunity system. The similarity between the Ig domain-containing proteins of VARTIG and the Ig-like N-terminal domain of TEP and complement components, which are a key animal innate immunity systems (Hajishengallis et al., 2017; Shokal and Eleftherianos, 2017), is best compatible with this hypothesis (**Figure 2A**). Although VARTIG proteins do not contain TED domains, the conserved cysteines potentially could be involved in the formation of covalent bonds with the targets. Thus, considering the diversity of the C-terminal region of the large Ig domain-containing protein, VARTIG might be analogous to the vertebrate complement factors that can bind different targets (Ricklin et al., 2016). The identification of TEP homologs in archaea and the fact that they also share some features with PTS further suggest that archaea evolved immune systems that generate an extensive repertoire of immunity proteins via gene shuffling and recombination (**Figure 3A**). An intriguing possibility is that the α2-macroglobulin domain-containing module is ancestral to VARTIG, which evolved into a more sophisticated complement-like innate immunity system capable of targeting many different antimicrobial molecules.

The third possibility is that VARTIG is involved in cell-cell adhesion and/or self-nonself recognition. This hypothesis is based on the fact that many surface proteins in bacteria and eukaryotes are known to be involved in cellular aggregation or in attachment to the host cells, in the case of parasites (Trunk et al., 2018; Honig and Shapiro, 2020). In particular, the Ig domain of PAP_07455 protein shows a highly significant sequence similarity with *E. coli* FdeC, the adherence factor which mediates *E. coli* adhesion to mammalian cells (Nesta et al., 2012; **Figure 2**). In archaea, so far, only type IV pili have been shown to be involved in cell-cell adhesion (Chaudhury et al., 2018; Pohlschroder et al., 2018; and references therein). Recently, it has been shown that UV-inducible type IV pili in *Sulfolobus acidocaldarius* mediate formation of cellular aggregates, ensuring specific self vs nonself recognition (van Wolferen et al., 2020). This specificity is primarily determined by the variable region of the major pilin UpsA (van Wolferen et al., 2020). Notably, archaeal type IV major pilins contain Ig-fold domains (Braun et al., 2016) that are highly variable and are often encoded in multiple copies in some archaeal genomes (Makarova et al., 2016). Another thoroughly characterized case of the involvement of Ig domain in self vs nonself discrimination are the major histocompatibility complex (MHC) proteins in vertebrates (Wieczorek et al., 2017). Numerous allelic variants of MHC (more than 200 in humans) recombine in meiosis producing a unique combination of alleles for each individual (Abualrous et al., 2021). Each MHC molecule displays an epitope, a small peptide derived from a degraded protein on the cell surface. The presented peptides derived from self proteins prevent the immune system from self-targeting. This specificity is achieved during the early development of the T lymphocytes, which are selected to recognize self MHC molecules, but not any self antigens. The MHC also appears to affect the mate choice (Boehm and Zufall, 2006). The predicted ability of VARTIG to generate numerous genetic variants of the surface domains ("epitopes"), possibly, by intragenic shuffling during DNA exchange, is compatible with the hypothesis that VARTIG is involved in self vs nonself recognition, and perhaps, in the mate choice for DNA exchange.

Given the extreme variability of the VARTIG modules, it does not seem surprising that their spread is currently limited to Thermococci. Conceivably, whatever the specific biological functions of VARTIG are, these are limited to interactions among Thermococci and perhaps their viruses, so that HGT to other

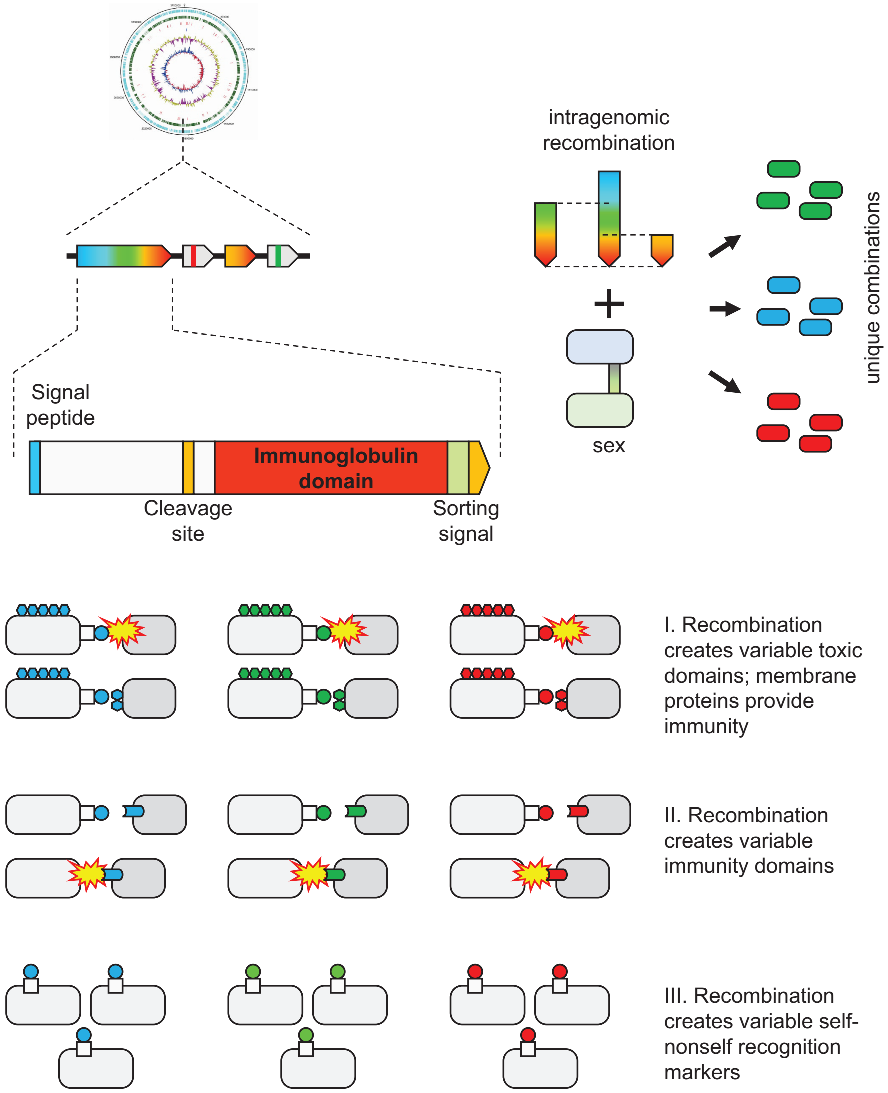
intragenomic recombination
+
sex
unique combinations
Signal peptide
Immunoglobulin domain
Cleavage site
Sorting signal
I. Recombination creates variable toxic domains; membrane proteins provide immunity
II. Recombination creates variable immunity domains
III. Recombination creates variable self-nonself recognition markers

FIGURE 4 | Three hypotheses on the function of VARTIG.

organisms would not result in fixation of VARTIG in their genomes. Although Ig domain-containing proteins encoded in other currently available archaeal genomes, are not found in such highly variable loci, some of these might also contribute to cell-cell adhesion, especially those that are unlikely to be involved in inter-microbial conflicts or immunity because they consist entirely of domains frequently found in adhesins and S-layer proteins (**Figure 3A**).

CONCLUSION

We describe here a gene module that is widespread among *Thermococcales* and is characterized by variability that is so far unprecedented in archaea. These loci contain an upstream gene coding for a large protein containing several Ig domains and unique combinations of downstream genes, some of which also encode Ig domains. In the large Ig domain-containing proteins (arCOG07710), the C-terminal Ig domain sequence is hypervariable, apparently, as a result of recombination between genes from different *Thermococcales*. To reflect this hypervariability, we named this gene module VARTIG, after VARiable *Thermococcales* IG. The overall organization of the VARTIG modules is similar to the organization of PTS. We searched all archaeal genomes for Ig-containing variable proteins and identified numerous such proteins, but no counterpart to VARTIG outside *Thermococcales* and no domains with a comparable level of variability. The specific functions of VARTIG remain unknown. However, the identified features of this system suggest three testable hypotheses: (i) involvement in inter-microbial conflicts analogous to PTS, (ii) role in innate immunity analogous to the vertebrate complement system, and (iii) function in self vs nonself recognition analogous to the vertebrate MHC (**Figure 4**). The latter two hypotheses appear to be of particular interest due to the unsuspected analogies to the vertebrate immune mechanisms.

DATA AVAILABILITY STATEMENT

The data used in the study are publicly available at the FTP site (https://ftp.ncbi.nlm.nih.gov/genomes/). Detailed information for specific accessions is provided in **Supplementary Tables 1–3**.

AUTHOR CONTRIBUTIONS

KM initiated the study. KM and EK wrote the manuscript. All authors analyzed the data, edited, and approved the manuscript.

FUNDING

This research was supported by the NIH Intramural Research Program at the National Library of Medicine, United States Department of Health and Human Services.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fmicb.2021.721392/full#supplementary-material

Supplementary Figure 1 | Supporting alignments for key components of thermococcal hypervariable system and for alpha 2-macroglobulin catalytic domain.

Supplementary Figure 2 | Phylogenetic tree of the ∼200 last amino acids of Ig-like domain containing proteins associated with thermococcal hypervariable system.

Supplementary Figure 3 | Mosaic proteins containing Ig-like domain. In each scheme is shown high similarity region for each blast hit which is mapped on the query protein (first line). The length of the line is proportional for respective protein length. For query sequence the domain organization is schematically shown above the respective line. The BLAST search was performed with the parameters modified as follows: "-seg no -comp_based_stats 0 -dbsize 20000000 -gapopen 32767 -gapextend 32767." This modification allows to identify ungapped regions of a very high similarity, which might be suggestive of intraprotein recombination (or mosaicism).

Supplementary Table 1 | Detailed information for the genes encoded in thermococcal variable system loci.

Supplementary Table 2 | Two worksheets. (1) Phyletic patterns (protein presence/absence) for *Thermococcales* COGs associated with VARTIG system; (2) Phyletic patterns (protein presence/absence) for the arCOGs mentioned in the text.

Supplementary Table 3 | Five worksheets. (1) Results of PSI-BLAST search for homologs of PAP_07455 protein. (2) Detailed information for the neighborhoods of the genes found using first strategy of the search for analogous systems in other archaea. (3) Detailed information for the genes found using second strategy of the search for analogous systems in other archaea. (4) Detailed information for the genes found using third strategy of the search for analogous systems in other archaea. (5) Detailed information for the neighborhoods of archaeal gene coding for alpha 2-macroglobulin homologs.

REFERENCES

Abualrous, E. T., Sticht, J., and Freund, C. (2021). Major histocompatibility complex (MHC) class I and class II proteins: impact of polymorphism on antigen presentation. *Curr. Opin. Immunol.* 70, 95–104. doi: 10.1016/j.coi.2021.04.009

Albers, S. V., and Jarrell, K. F. (2015). The archaellum: how Archaea swim. *Front. Microbiol.* 6:23. doi: 10.3389/fmicb.2015.00023

Almagro Armenteros, J. J., Tsirigos, K. D., Sonderby, C. K., Petersen, T. N., Winther, O., Brunak, S., et al. (2019). SignalP 5.0 improves signal peptide predictions using deep neural networks. *Nat. Biotechnol.* 37, 420–423. doi: 10.1038/s41587-019-0036-z

Altschul, S. F., Madden, T. L., Schaffer, A. A., Zhang, J., Zhang, Z., Miller, W., et al. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* 25, 3389–3402. doi: 10.1093/nar/25.17.3389

Aravind, L., Anantharaman, V., Zhang, D., De Souza, R. F., and Iyer, L. M. (2012). Gene flow and biological conflict systems in the origin and evolution of eukaryotes. *Front. Cell Infect. Microbiol.* 2:89. doi: 10.3389/fcimb.2012.00089

Baxter, R. H., Chang, C. I., Chelliah, Y., Blandin, S., Levashina, E. A., and Deisenhofer, J. (2007). Structural basis for conserved complement factor-like function in the antimalarial protein TEP1. *Proc. Natl. Acad. Sci. U.S.A.* 104, 11615–11620. doi: 10.1073/pnas.0704967104

Blandin, S., and Levashina, E. A. (2004). Thioester-containing proteins and insect immunity. *Mol. Immunol.* 40, 903–908. doi: 10.1016/j.molimm.2003.10.010

Boehm, T., and Zufall, F. (2006). MHC peptides and the sensory evaluation of genotype. *Trends Neurosci.* 29, 100–107. doi: 10.1016/j.tins.2005.11.006

Bork, P., Holm, L., and Sander, C. (1994). The immunoglobulin fold. Structural classification, sequence patterns and common core. *J. Mol. Biol.* 242, 309–320. doi: 10.1016/s0022-2836(84)71582-8

Braun, T., Vos, M. R., Kalisman, N., Sherman, N. E., Rachel, R., Wirth, R., et al. (2016). Archaeal flagellin combines a bacterial type IV pilin domain with an Ig-like domain. *Proc. Natl. Acad. Sci. U.S.A.* 113, 10352–10357. doi: 10.1073/pnas.1607756113

Castro, C., Zhang, R., Liu, J., Bellenberg, S., Neu, T. R., Donati, E., et al. (2016). Biofilm formation and interspecies interactions in mixed cultures of thermo-acidophilic archaea *Acidianus* spp. and *Sulfolobus metallicus*. *Res. Microbiol.* 167, 604–612. doi: 10.1016/j.resmic.2016.06.005

Chaudhury, P., Quax, T. E. F., and Albers, S. V. (2018). Versatile cell surface structures of archaea. *Mol. Microbiol.* 107, 298–311. doi: 10.1111/mmi.13889

Chimileski, S., Franklin, M. J., and Papke, R. T. (2014). Biofilms formed by the archaeon *Haloferax volcanii* exhibit cellular differentiation and social motility, and facilitate horizontal gene transfer. *BMC Biol.* 12:65. doi: 10.1186/s12915-014-0065-5

Cohen, O., Ashkenazy, H., Belinky, F., Huchon, D., and Pupko, T. (2010). GLOOME: gain loss mapping engine. *Bioinformatics* 26, 2914–2915. doi: 10.1093/bioinformatics/btq549

Dashper, S. G., Mitchell, H. L., Seers, C. A., Gladman, S. L., Seemann, T., Bulach, D. M., et al. (2017). Porphyromonas gingivalis uses specific domain rearrangements and allelic exchange to generate diversity in surface virulence factors. *Front. Microbiol.* 8:48. doi: 10.3389/fmicb.2017.00048

Daum, B., Vonck, J., Bellack, A., Chaudhury, P., Reichelt, R., Albers, S. V., et al. (2017). Structure and in situ organisation of the *Pyrococcus furiosus* archaellum machinery. *Elife* 6:e27470.

De Ste Croix, M., Vacca, I., Kwun, M. J., Ralph, J. D., Bentley, S. D., Haigh, R., et al. (2017). Phase-variable methylation and epigenetic regulation by type I restriction-modification systems. *FEMS Microbiol. Rev.* 41, S3–S15.

Dirix, G., Monsieurs, P., Dombrecht, B., Daniels, R., Marchal, K., Vanderleyden, J., et al. (2004). Peptide signal molecules and bacteriocins in gram-negative bacteria: a genome-wide in silico screening for peptides containing a double-glycine leader sequence and their cognate transporters. *Peptides* 25, 1425–1440. doi: 10.1016/j.peptides.2003.10.028

Drozdetskiy, A., Cole, C., Procter, J., and Barton, G. J. (2015). JPred4: a protein secondary structure prediction server. *Nucleic Acids Res.* 43, W389–W394.

Edgar, R. C. (2004). MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.* 32, 1792–1797. doi: 10.1093/nar/gkh340

Edgar, R. C. (2010). Search and clustering orders of magnitude faster than BLAST. *Bioinformatics* 26, 2460–2461. doi: 10.1093/bioinformatics/btq461

Erauso, G., Stedman, K. M., Van De Werken, H. J. G., Zillig, W., and Van Der Oost, J. (2006). Two novel conjugative plasmids from a single strain of *Sulfolobus*. *Microbiology (Reading)* 152, 1951–1968. doi: 10.1099/mic.0.28861-0

Esterman, E. S., Wolf, Y. I., Kogay, R., Koonin, E. V., and Zhaxybayeva, O. (2021). Evolution of DNA packaging in gene transfer agents. *Virus Evol.* 7:veab015.

Flemming, H. C., and Wuertz, S. (2019). Bacteria and archaea on Earth and their abundance in biofilms. *Nat. Rev. Microbiol.* 17, 247–260. doi: 10.1038/s41579-019-0158-9

Frols, S., Dyall-Smith, M., and Pfeifer, F. (2012). Biofilm formation by haloarchaea. *Environ. Microbiol.* 14, 3159–3174. doi: 10.1111/j.1462-2920.2012.02895.x

Garcia-Ferrer, I., Arede, P., Gomez-Blanco, J., Luque, D., Duquerroy, S., Caston, J. R., et al. (2015). Structural and functional insights into *Escherichia coli* alpha2-macroglobulin endopeptidase snap-trap inhibition. *Proc. Natl. Acad. Sci. U.S.A.* 112, 8290–8295. doi: 10.1073/pnas.1506538112

Goda, S., Koga, T., Yamashita, K., Kuriura, R., and Ueda, T. (2018). A novel carbohydrate-binding surface layer protein from the hyperthermophilic archaeon *Pyrococcus horikoshii*. *Biosci. Biotechnol. Biochem.* 82, 1327–1334. doi: 10.1080/09168451.2018.1460571

Haft, D. H., Payne, S. H., and Selengut, J. D. (2012). Archaeosortases and exosortases are widely distributed systems linking membrane transit with posttranslational modification. *J. Bacteriol.* 194, 36–48. doi: 10.1128/jb.06026-11

Hajishengallis, G., Reis, E. S., Mastellos, D. C., Ricklin, D., and Lambris, J. D. (2017). Novel mechanisms and functions of complement. *Nat. Immunol.* 18, 1288–1298. doi: 10.1038/ni.3858

Hill, C. W., Sandt, C. H., and Vlazny, D. A. (1994). Rhs elements of *Escherichia coli*: a family of genetic composites each encoding a large mosaic protein. *Mol. Microbiol.* 12, 865–871. doi: 10.1111/j.1365-2958.1994.tb01074.x

Honig, B., and Shapiro, L. (2020). Adhesion protein structure, molecular affinities, and principles of cell-cell recognition. *Cell* 181, 520–535. doi: 10.1016/j.cell.2020.04.010

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

Imachi, H., Nobu, M. K., Nakahara, N., Morono, Y., Ogawara, M., Takaki, Y., et al. (2020). Isolation of an archaeon at the prokaryote-eukaryote interface. *Nature* 577, 519–525.

Jachlewski, S., Jachlewski, W. D., Linne, U., Brasen, C., Wingender, J., and Siebers, B. (2015). Isolation of extracellular polymeric substances from biofilms of the thermoacidophilic archaeon *Sulfolobus acidocaldarius*. *Front. Bioeng. Biotechnol.* 3:123. doi: 10.3389/fbioe.2015.00123

Keller, L., and Surette, M. G. (2006). Communication in bacteria: an ecological and evolutionary perspective. *Nat. Rev. Microbiol.* 4, 249–258. doi: 10.1038/nrmicro1383

Konovalova, A., and Sogaard-Andersen, L. (2011). Close encounters: contact-dependent interactions in bacteria. *Mol. Microbiol.* 81, 297–301. doi: 10.1111/j.1365-2958.2011.07711.x

Krogh, A., Larsson, B., Von Heijne, G., and Sonnhammer, E. L. (2001). Predicting transmembrane protein topology with a hidden markov model: application to complete genomes. *J. Mol. Biol.* 305, 567–580. doi: 10.1006/jmbi.2000.4315

Leigh, J. A., Albers, S. V., Atomi, H., and Allers, T. (2011). Model organisms for genetics in the domain Archaea: methanogens, halophiles, Thermococcales and Sulfolobales. *FEMS Microbiol. Rev.* 35, 577–608. doi: 10.1111/j.1574-6976.2011.00265.x

Lopez, D., Vlamakis, H., and Kolter, R. (2010). Biofilms. *Cold Spring Harb. Perspect. Biol.* 2:a000398.

Makarova, K. S., Koonin, E. V., and Albers, S. V. (2016). Diversity and evolution of type IV pili systems in Archaea. *Front. Microbiol.* 7:667. doi: 10.3389/fmicb.2016.00667

Makarova, K. S., Wolf, Y. I., Karamycheva, S., Zhang, D., Aravind, L., and Koonin, E. V. (2019). Antimicrobial peptides, polymorphic toxins, and self-nonself recognition systems in archaea: an untapped armory for intermicrobial conflicts. *mBio* 10, e00715–e00719.

Makarova, K. S., Wolf, Y. I., and Koonin, E. V. (2015). Archaeal clusters of orthologous genes (arCOGs): an update and application for analysis of shared features between Thermococcales, Methanococcales, and Methanobacteriales. *Life (Basel)* 5, 818–840. doi: 10.3390/life5010818

Malik, A., and Kim, S. B. (2019). A comprehensive in silico analysis of sortase superfamily. *J. Microbiol.* 57, 431–443. doi: 10.1007/s12275-019-8545-5

Marchler-Bauer, A., Anderson, J. B., Chitsaz, F., Derbyshire, M. K., Deweese-Scott, C., Fong, J. H., et al. (2009). CDD: specific functional annotation with the conserved domain database. *Nucleic Acids Res.* 37, D205–D210.

Marguet, E., Gaudin, M., Gauliard, E., Fourquaux, I., Le Blond Du Plouy, S., Matsui, I., et al. (2013). Membrane vesicles, nanopods and/or nanotubes produced by hyperthermophilic archaea of the genus *Thermococcus*. *Biochem. Soc. Trans.* 41, 436–442. doi: 10.1042/bst20120293

Martin-Galiano, A. J. (2017). The MiiA motif is a common marker present in polytopic surface proteins of oral and urinary tract invasive bacteria. *Infect. Genet. Evol.* 49, 283–292. doi: 10.1016/j.meegid.2017.02.002

Nadell, C. D., Xavier, J. B., and Foster, K. R. (2009). The sociobiology of biofilms. *FEMS Microbiol. Rev.* 33, 206–224. doi: 10.1111/j.1574-6976.2008.00150.x

Nather, D. J., Rachel, R., Wanner, G., and Wirth, R. (2006). Flagella of *Pyrococcus furiosus*: multifunctional organelles, made for swimming, adhesion to various surfaces, and cell-cell contacts. *J. Bacteriol.* 188, 6915–6923. doi: 10.1128/jb.00527-06

Nesta, B., Spraggon, G., Alteri, C., Moriel, D. G., Rosini, R., Veggi, D., et al. (2012). FdeC, a novel broadly conserved *Escherichia coli* adhesin eliciting protection against urinary tract infections. *mBio* 3:e00010-12.

Peterson, S. B., Bertolli, S. K., and Mougous, J. D. (2020). The central role of interbacterial antagonism in bacterial life. *Curr. Biol.* 30, R1203–R1214.

Pohlschroder, M., Pfeiffer, F., Schulze, S., and Abdul Halim, M. F. (2018). Archaeal cell surface biogenesis. *FEMS Microbiol. Rev.* 42, 694–717. doi: 10.1093/femsre/fuy027

Price, M. N., Dehal, P. S., and Arkin, A. P. (2010). FastTree 2–approximately maximum-likelihood trees for large alignments. *PLoS One* 5:e9490. doi: 10.1371/journal.pone.0009490

Prochazkova, K., Shuvalova, L. A., Minasov, G., Voburka, Z., Anderson, W. F., and Satchell, K. J. (2009). Structural and molecular mechanism for autoprocessing of MARTX toxin of *Vibrio cholerae* at multiple sites. *J. Biol. Chem.* 284, 26557–26568. doi: 10.1074/jbc.m109.025510

Rashid, N., and Aslam, M. (2020). An overview of 25 years of research on *Thermococcus kodakarensis*, a genetically versatile model organism for archaeal research. *Folia Microbiol. (Praha)* 65, 67–78. doi: 10.1007/s12223-019-00730-2

Rehman, A. A., Ahsan, H., and Khan, F. H. (2013). alpha-2-macroglobulin: a physiological guardian. *J. Cell. Physiol.* 228, 1665–1675. doi: 10.1002/jcp.24266

Ricklin, D., Reis, E. S., Mastellos, D. C., Gros, P., and Lambris, J. D. (2016). Complement component C3–the "Swiss Army Knife" of innate immunity and host defense. *Immunol. Rev.* 274, 33–58. doi: 10.1111/imr.12500

Rodrigues-Oliveira, T., Belmok, A., Vasconcellos, D., Schuster, B., and Kyaw, C. M. (2017). Archaeal S-layers: overview and current state of the art. *Front. Microbiol.* 8:2597. doi: 10.3389/fmicb.2017.02597

Ruhe, Z. C., Low, D. A., and Hayes, C. S. (2020). Polymorphic toxins and their immunity proteins: diversity, evolution, and mechanisms of delivery. *Annu. Rev. Microbiol.* 74, 497–520. doi: 10.1146/annurev-micro-020518-115638

Schatz, D. G., and Swanson, P. C. (2011). V(D)J recombination: mechanisms of initiation. *Annu. Rev. Genet.* 45, 167–202. doi: 10.1146/annurev-genet-110410-132552

Schopf, S., Wanner, G., Rachel, R., and Wirth, R. (2008). An archaeal bi-species biofilm formed by *Pyrococcus furiosus* and *Methanopyrus kandleri*. *Arch. Microbiol.* 190, 371–377. doi: 10.1007/s00203-008-0371-9

Shen, A. (2010). Autoproteolytic activation of bacterial toxins. *Toxins (Basel)* 2, 963–977. doi: 10.3390/toxins2050963

Shokal, U., and Eleftherianos, I. (2017). Evolution and function of thioester-containing proteins and the complement system in the innate immune response. *Front. Immunol.* 8:759. doi: 10.3389/fimmu.2017.00759

Sivabalasarma, S., Wetzel, H., Nussbaum, P., Van Der Does, C., Beeby, M., and Albers, S. V. (2020). Analysis of cell-cell bridges in *Haloferax volcanii* using electron cryo-tomography reveal a continuous cytoplasm and S-layer. *Front. Microbiol.* 11:612239. doi: 10.3389/fmicb.2020.612239

Soding, J. (2005). Protein homology detection by HMM-HMM comparison. *Bioinformatics* 21, 951–960. doi: 10.1093/bioinformatics/bti125

Soding, J., Biegert, A., and Lupas, A. N. (2005). The HHpred interactive server for protein homology detection and structure prediction. *Nucleic Acids Res.* 33, W244–W248.

Stedman, K. M., She, Q., Phan, H., Holz, I., Singh, H., Prangishvili, D., et al. (2000). pING family of conjugative plasmids from the extremely thermophilic archaeon *Sulfolobus islandicus*: insights into recombination and conjugation in Crenarchaeota. *J. Bacteriol.* 182, 7014–7020. doi: 10.1128/jb.182.24.7014-7020.2000

Stewart, P. S., and Franklin, M. J. (2008). Physiological heterogeneity in biofilms. *Nat. Rev. Microbiol.* 6, 199–210. doi: 10.1038/nrmicro1838

Trunk, T., Khalil, H. S., and Leo, J. C. (2018). Bacterial autoaggregation. *AIMS Microbiol.* 4, 140–164. doi: 10.3934/microbiol.2018.1.140

van Wolferen, M., Orell, A., and Albers, S. V. (2018). Archaeal biofilm formation. *Nat. Rev. Microbiol.* 16, 699–713. doi: 10.1038/s41579-018-0058-4

van Wolferen, M., Shajahan, A., Heinrich, K., Brenzinger, S., Black, I. M., Wagner, A., et al. (2020). Species-specific recognition of sulfolobales mediated by UV-inducible pili and S-layer glycosylation patterns. *mBio* 11:e03014-19.

Wieczorek, M., Abualrous, E. T., Sticht, J., Alvaro-Benito, M., Stolzenberg, S., Noe, F., et al. (2017). Major histocompatibility complex (MHC) class I and MHC class II proteins: conformational plasticity in antigen presentation. *Front. Immunol.* 8:292. doi: 10.3389/fimmu.2017.00292

Wiedemann, C., Kumar, A., Lang, A., and Ohlenschlager, O. (2020). Cysteines and disulfide bonds as structure-forming units: insights from different domains of life and the potential for characterization by NMR. *Front. Chem.* 8:280. doi: 10.3389/fchem.2020.00280

Zhang, D., De Souza, R. F., Anantharaman, V., Iyer, L. M., and Aravind, L. (2012). Polymorphic toxin systems: comprehensive characterization of trafficking modes, processing, mechanisms of action, immunity and ecology using comparative genomics. *Biol. Direct.* 7:18. doi: 10.1186/1745-6150-7-18

Zimmermann, L., Stephens, A., Nam, S. Z., Rau, D., Kubler, J., Lozajic, M., et al. (2018). A completely reimplemented MPI bioinformatics toolkit with a new HHpred server at its core. *J. Mol. Biol.* 430, 2237–2243. doi: 10.1016/j.jmb.2017.12.007

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.