Article

Diversity and Potential Multifunctionality of Archaeal CetZ Tubulin-like Cytoskeletal Proteins

Hannah J. Brown [1] and Iain G. Duggin [1,*]

[1] The Australian Institute for Microbiology and Infection, University of Technology Sydney, NSW, 2007 Australia

* Correspondence: Iain.Duggin@uts.edu.au

Abstract: Tubulin superfamily (TSF) proteins are widespread and known for multifaceted roles as cytoskeletal proteins underpinning many basic cellular functions including morphogenesis, division, and motility. In eukaryotes, tubulin assembles into microtubules, a major component of the dynamic cytoskeletal network of fibres, whereas the bacterial homolog FtsZ assembles the division ring at midcell. Functions of the lesser-known archaeal TSF proteins are beginning to be identified, and show surprising diversity, including homologs of tubulin and FtsZ, and a third archaea-specific family, CetZ, implicated in the regulation of cell shape and possibly other unknown functions. In this study, we defined sequence and structural characteristics of the CetZ family and CetZ1 and CetZ2 subfamilies, identified CetZ groups and diversity amongst archaea, and identified potential functional relationships through analysis of the genomic neighbourhoods of *cetZ* genes. At least three subfamilies of orthologous CetZ proteins were identified in the archaeal class Halobacteria, including CetZ1 and CetZ2 and a novel uncharacterized subfamily. CetZ1 and CetZ2 were correlated to one another and to cell shape and motility phenotypes across diverse Halobacteria. Amongst other known CetZ clusters in orders Archaeoglobales, Methanomicrobiales, Methanosarcinales, and Thermococcales, an additional uncharacterized group from Archaeoglobales and Methanomicrobiales affiliated strongly with Halobacteria CetZs, suggesting they originated via horizontal transfer. Subgroups of Halobacteria CetZ2 and Thermococcales CetZ genes were found adjacent to different type IV pili regulons, suggesting a potential utilization of CetZs by type IV systems. More broadly conserved *cetZ* gene neighbourhoods included nucleotide and cofactor biosynthesis (e.g., F_{420}) and predicted cell surface sugar epimerase genes. The findings imply that CetZ subfamilies are involved in multiple functions linked to the cell surface, biosynthesis and motility.

Keywords: archaea; tubulin; FtsZ; CetZ; cytoskeleton

1. Introduction

The cytoskeleton is a dynamic and expansive network of structural proteins, filaments, and polymers necessary for all domains of life. At its core, the function of the cytoskeleton is to provide structure and organisation of the cytoplasm, however the downstream cellular roles and mechanisms of cytoskeletal proteins extends far beyond this, making them fundamentally important for a range of cellular processes, including cell division[1,2], chromosome segregation[2], cell motility and migration[3-6], endocytosis[7], and intracellular transport of many cargo such as signalling molecules, membrane components, and organelles[8,9]. How cytoskeletal proteins contribute to these processes has been extensively studied in bacteria and eukaryotes, but little is known about their functions in the third major grouping of life—Archaea[10]. Archaea share a similar basic cellular organisation and morphology to Bacteria, but archaeal DNA replication and transcription mechanisms resemble those of Eukarya[10], which has raised the question of whether the cytoskeletal functions of archaeal cells are distinct or resemble those of bacteria or eukaryotes.

Tubulin superfamily proteins (TSFs) are widespread across all three domains of life, including in archaea. The tubulin superfamily consists of tubulin, the subunits of which assemble into microtubules, FtsZ, a key cell division protein, and CetZs which are specific to the archaea[4,11]. TSFs possess a critical GTPase active site, where GTP binding between subunits promotes polymerisation, and GTP hydrolysis to GDP promotes depolymerisation[12-19]. It is this dynamic polymerisation which allows tubulin superfamily proteins to mobilize as larger structures that contribute to their diverse array of functions. In eukaryotes, tubulin assembles into cylindrical microtubules that control cell shape, structure, and contribute to chromosome segregation during cell division[2]. In addition, microtubules form tracks for intracellular transport motors such as kinesins and dynein[20] which carry a range of cargo, like organelles[21], vesicles and secretory proteins[22], RNA[23], lipids[24], and other membrane components such as surface adhesins[25]. Microtubules are also involved in cell migration through their facilitation and maintenance of membrane protrusions[26,27] (Reviewed in [3]).

FtsZ is present in bacteria and archaea and is a major contributor to cell division. Bacterial FtsZ does not form microtubules, but instead assembles a multi-protein division ring, or divisome, at mid-cell which constricts to drive cytokinesis through a mechanism involving directed ingrowth of the peptidoglycan cell wall[19,28-30]. Most archaea possess two FtsZ homologues with differing roles in division that appear to be important in archaea that do not have a pseudopeptidoglycan cell wall[29]. While in bacteria there are many characterised divisome proteins[31-34], the functional partners of FtsZs and divisome components in archaea still largely remain unknown[35,36].

How tubulin and its complex assemblies and activities evolved around the time of eukaryogenesis from primordial FtsZ is unknown. Understanding the diversity and evolution of TSFs in archaea, which have a common ancestor with eukaryotes, should help to gain insight into the evolutionary and functional pathways of these proteins in general.

CetZ proteins represent the third major family of tubulin superfamily proteins. They have only been found in archaea, and, interestingly, show specific sequence similarities to both FtsZ and tubulin[4]. Furthermore, CetZs contribute to cell shape and motility but appear to have no direct role in cell division. Current insights into CetZ function come from studies of the halophilic archaeon *Haloferax volcanii*[4,37]. *H. volcanii* cells are pleomorphic, and commonly exhibit irregular flattened plate or rod shapes [38]. Cells transition from plate to rod shapes in several conditions, including during the early stages of growth in batch culture [37,39], when depleted of trace metals[37], or when becoming motile in soft agar[4]. Rod development requires the most highly conserved of the CetZs, CetZ1, and deletion of cetZ1 also results in reduced motility, which was attributed to the inability to form rods[4].

CetZ2 is also conserved across many archaeal species, and appears to form its own orthologous group separate to CetZ1[4], suggesting it could have a distinct role. Consistent with this, deletion of *cetZ2* does not directly impact rod development or motility under the same conditions as CetZ1 does[4]. Currently there is no known phenotype resulting from deletion of *cetZ2*, however overexpression of a CetZ2 GTPase mutant inhibits rod development and motility, implicating it generally in cell shape control and motility[4]. The molecular mechanism through which CetZs contribute to shape control and motility are not yet understood, nor is it clear whether CetZs influence motility or other functions through a mechanism or pathway which is independent to their function in rod development. Here, we explore the diversity and possible roles of the main groups of CetZs in archaea by assessing their distribution across archaeal species and synteny with other genes within their immediate genomic regions.

2. Materials and Methods

2.1. Identification and analysis of archaeal tubulin superfamily homologues

The amino acid sequences of a diverse set of tubulin superfamily proteins, including the partly characterized CetZ1-6, FtsZ1 and FtsZ2 from the archaeon *H. volcanii* were firstly aligned using MUSCLE[40,41]. The alignment was used as a query to search for tubulin superfamily (TSF) homologues in 183 other archaeal species in the UniProt database using JackHMMER[42]. For this analysis, at least one species per family of archaea was selected, prioritising those with whole genomes available (Table S1). The amino acid sequences of 550 significant hits were downloaded and aligned with the search set using MUSCLE. The complete multiple sequence alignment was used to construct a maximum likelihood tree with 100 bootstrap replicates in MEGA X[43]. The labelling of known and novel groups representing clusters of similar sequences was carried out using the characterized FtsZ and CetZ sequences from *H. volcanii* as the reference (Figure 1).

Conserved amino-acid residues in separate alignments of CetZ1 (49 sequences) and CetZ2 (41 sequences) were identified by comparing the consensus sequence and consensus scores for each residue in the multiple alignment and generating a unique residue score, defined as the average of the CetZ1 and CetZ2 consensus scores at that position. Conserved, residues were taken as those that had a unique residue score greater than 90%. A similar analysis comparing conserved FtsZ and CetZ characteristic residues was also performed, using an 80% cut-off for the unique residue score.

3D structural comparisons were carried out in PyMOL[44] v2.5.1, using the super (superimpose) and APBS electrostatics[45] functions.

2.2. Analysis of cetZ genomic regions

A diverse set of archaeal species (20 for *cetZ1*, 22 for *cetZ2*, and 27 for non—Halobacteria *cetZs*) with complete genome sequences available were chosen for analysis of gene content within *cetZ* genomic regions. The DNA sequences and corresponding annotations for the 40 kb region centred on each *cetZ* gene of interest were downloaded from the NCBI genome database (www.ncbi.nlm.nih.gov/genome). The predicted coding sequences in FASTA protein format were also obtained, and used to assign arCOGIDs and arCOG annotations within the region using the eggNOG-mapper v2[46,47]. A summary of the data is provided in Table S3. The arCOGIDs of genes located within the 40 kb *cetZ* genomic regions were counted and compared across species to determine which arCOGs were most often present. Finally, arCOGIDs were mapped onto each genomic region to compare genomic arrangements of the *cetZ* regions between species.

3. Results

3.1. Identification and classification of tubulin superfamily proteins in archaea

To assess the distribution of CetZs amongst archaeal species in greater depth than previously available, tubulin superfamily (TSF) homologues were identified in 183 diverse species selected from across the full breadth of known archaea, and a phylogenetic tree of was then generated from the aligned sequences. Table S1 lists the represented species, the number of identified homologues, and their assigned family where possible. The phylogenetic tree (Fig. 1) was labelled with the three main branches that represent FtsZ1, FtsZ2, and CetZ families based on the functionally characterized *H. volcanii* proteins[4,29,37]; many species possessed at least one homologue from each of these three main families. Several tubulins and non-canonical TSF proteins were also identified, which showed a patchy distribution generally in diverse archaeal species including those belonging to the Thaumarchaeota, Asgard, and DPANN archaea (Table S1).

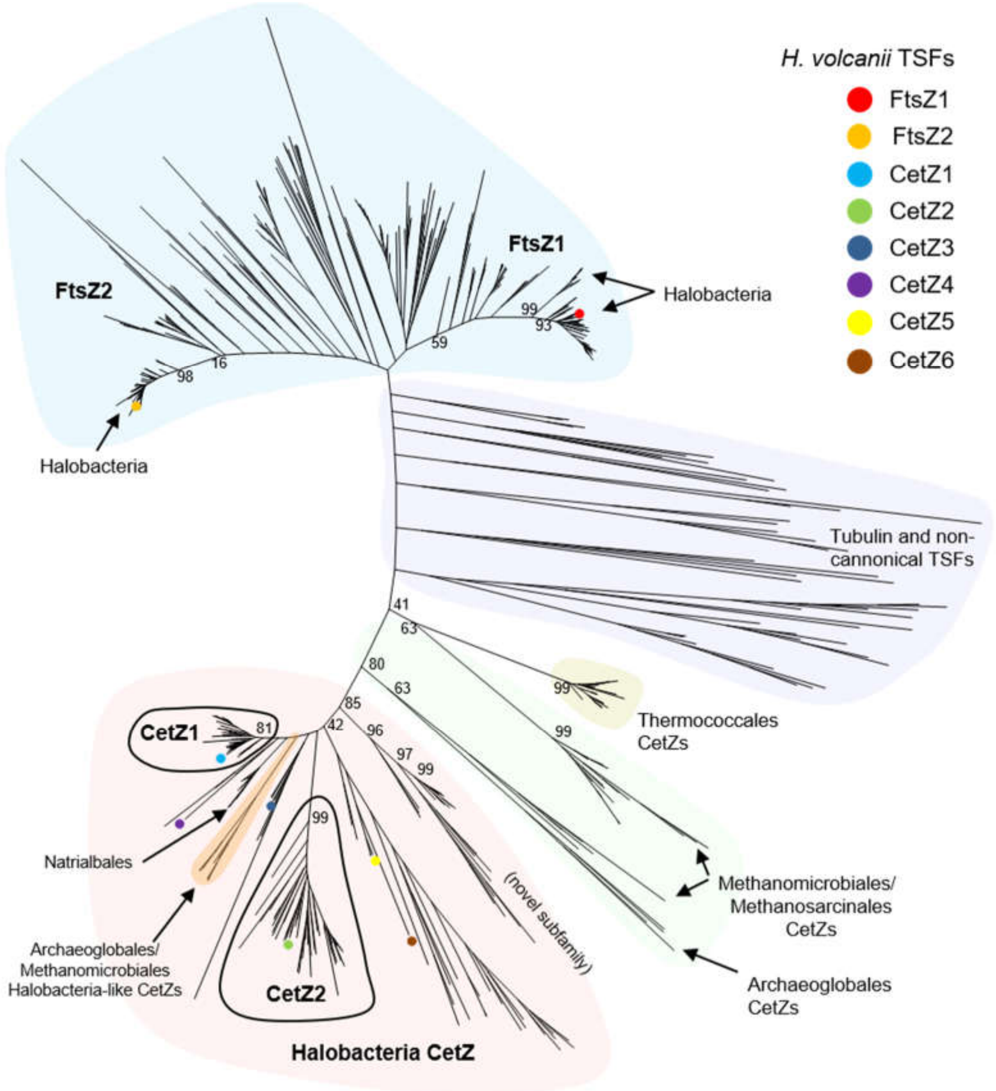
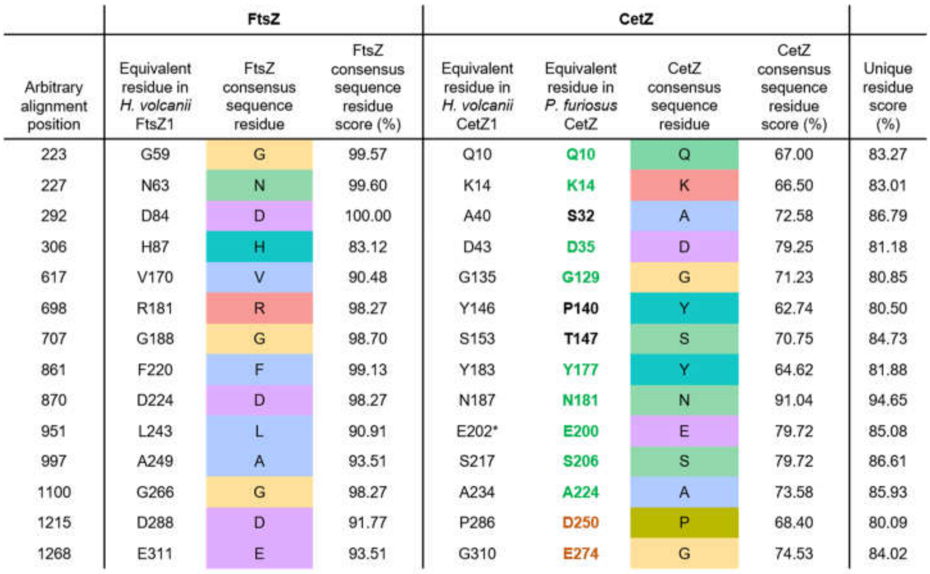


Figure 1. Phylogram of archaeal tubulin superfamily proteins. 550 Identified tubulin superfamily protein sequences from 183 diverse archaea were aligned to generate the phylogram. Archaeal FtsZs are shaded in blue, and tubulin and non-canonical TSFs are shaded in purple. Four main groups of CetZs were identified: Halobacteria CetZs (red); Halobacteria-like CetZs from Archaeoglobales and Methanomicrobiales (orange); non-Halobacteria-like CetZs from Archaeoglobales, Methanomicrobiales, and Methanosarcinales (green), and CetZs from Thermococcales CetZs (yellow). Within Halobacteria, we define two important CetZ clusters, CetZ1 and CetZ2 (circled in black), by comparing bootstrap values from this phylogenetic analysis with unique residue identification (Fig. 3). Selected branch bootstrap percentages are shown.

3.2. Multiple CetZs are abundant in Halobacteria

Proteins from the class Halobacteria form one major branch of the tree (Fig. 1) and many Halobacteria species (of which 60 were included) have multiple CetZs. We identified three distinct subfamilies of CetZ proteins that have representatives in many of the diverse Halobacteria; they formed distinct and strongly supported orthologous groups with relatively short branch lengths (Fig. 1). Two of these subfamilies are named based on whether they grouped with the characterized CetZ1 and CetZ2 proteins from *H. volcanii*[4,48]. The key differences between CetZ1 and CetZ2 will be described further below. Another novel subfamily of CetZs was also identified (Fig. 1), containing uncharacterized proteins from diverse Halobacteria, suggesting these proteins would have a common function in these species. Most of the other CetZs did not sit within clear subgroups, including *H. volcanii* CetZ3-6, suggesting they could have relatively weakly conserved or in some cases potentially redundant roles.



	FtsZ			CetZ				
Arbitrary alignment position	Equivalent residue in *H. volcanii* FtsZ1	FtsZ consensus sequence residue	FtsZ consensus sequence residue score (%)	Equivalent residue in *H. volcanii* CetZ1	Equivalent residue in *P. furiosus* CetZ	CetZ consensus sequence residue	CetZ consensus sequence residue score (%)	Unique residue score (%)
223	G59	G	99.57	Q10	**Q10**	Q	67.00	83.27
227	N63	N	99.60	K14	**K14**	K	66.50	83.01
292	D84	D	100.00	A40	**S32**	A	72.58	86.79
306	H87	H	83.12	D43	**D35**	D	79.25	81.18
617	V170	V	90.48	G135	**G129**	G	71.23	80.85
698	R181	R	98.27	Y146	**P140**	Y	62.74	80.50
707	G188	G	98.70	S153	**T147**	S	70.75	84.73
861	F220	F	99.13	Y183	**Y177**	Y	64.62	81.88
870	D224	D	98.27	N187	**N181**	N	91.04	94.65
951	L243	L	90.91	E202*	**E200**	E	79.72	85.08
997	A249	A	93.51	S217	**S206**	S	79.72	86.61
1100	G266	G	98.27	A234	**A224**	A	73.58	85.93
1215	D288	D	91.77	P286	**D250**	P	68.40	80.09
1268	E311	E	93.51	G310	**E274**	G	74.53	84.02

A further 8 residues were identified to have a unique residue score >80% but had similar chemistry. These were at arbitrary alignment positions: 412 (FtsZ: A; CetZ: V), 578 (FtsZ: T; CetZ: S), 591 (FtsZ: A; CetZ: L), 778 (FtsZ: L; CetZ: W), 923 (FtsZ: I; CetZ: L), 998 (FtsZ: D; CetZ: E), 1016 (FtsZ: M; CetZ: L), and 1328 (FtsZ: A; CetZ: V). *E202 is contained within an unsolved region of the CetZ1 crystal structure.

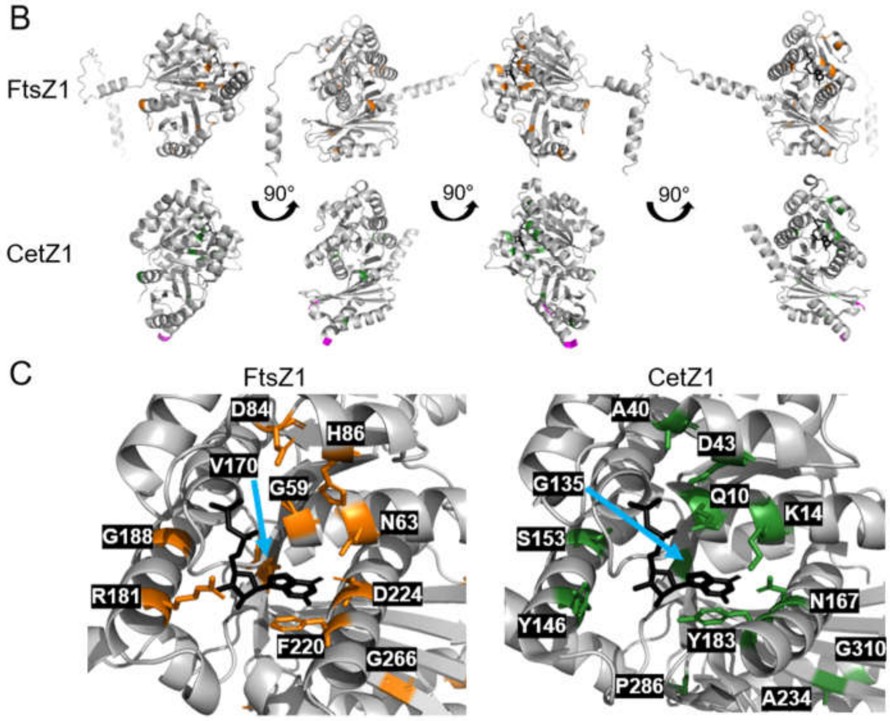


Figure 2. Characterising unique residues in archaeal FtsZ and CetZ-family proteins. A) Unique residues were defined as residues which had a unique residue score of greater than 80%. The consensus sequence residue score represents the frequency at which the indicated amino acid was detected at that position in the alignment of FtsZ and CetZ sequences. The unique residue score is the average of the FtsZ and CetZ consensus sequence residue scores at that alignment position. One letter amino acid codes are used, and consensus residues are coloured according to their chemistry. Red: positively charged; blue: hydrophobic; green: polar; yellow: glycine; purple, negatively charged; turquoise, aromatic; mustard, proline. The equivalent unique residues in CetZ from Pyrococcus furiosus are shown in green if the residue is the same as the CetZ consensus, in orange if the residue is the same as the FtsZ consensus, and in black if the residue does not match either the CetZ or FtsZ consensus. **B)** Three-dimensional structural comparison of the crystal structure of GDP bound CetZ1 from H. volcanii (PDB: 4B46) and the AlphaFold predicted structure of FtsZ1 from *H. volcanii*. The GDP from the CetZ1 crystal structure has been superimposed onto the AlphaFold predicted FtsZ structure. Residues that constitute the M-loop of CetZ1 are represented in magenta. Unique residues listed in A are represented in orange for FtsZ1, and green for CetZ1. **C)** Comparison

of the GTP/GDP binding pockets of FtsZ1 and CetZ1 (containing GDP from the CetZ1 crystal structure), indicating specific identified unique residues and their sidechains.

3.3. Deep branching CetZs in Thermoccales define the CetZ family boundary

All Thermococci species analysed (in the main order Thermococcales and family Thermococcaceae) each encode at least one designated CetZ homologue. Two strains, *Thermococcus AM4* and *Thermococcus gammatolerans EJ3*, each encode an additional, highly divergent TSF protein which weakly branched near other non-cannonical TSF proteins and likely have strain-specific or redundant functions. However, the main Thermococcales CetZs form a tight cluster (Fig. 1), in accordance with the relatively close genomic similarity amongst the known Thermococcaceae. This represents the most deeply branching group we classified as CetZ. We then analysed the multiple sequence alignment (Fig. 2a) and 3D structure predictions (Fig. 2b, Fig. S1) to identify and define CetZ family-specific amino acid residues that differ from FtsZ family-specific residues. Consistent with a previous initial analysis[4], we confirmed that the unique residues were largely clustered around the GTP/GDP binding pocket and GTPase active site (Fig. 2c), which may reflect a fundamental difference in the polymerization properties of FtsZ and CetZ. Two of the key residues in CetZ from *Pyrococcus furiosus* (Thermococcales) were identical to FtsZ (D250 and E274), while three were not consistent with the consensus residues of either CetZ or FtsZ (A40, Y16, S153). However, as most residues were consistent with the CetZ consensus, this supported the inclusion of the Thermococcales proteins in the CetZ family and their use in defining the CetZ family boundary.

3.4. CetZs in Archaeoglobales, Methanomicrobiales, and Methanosarcinales

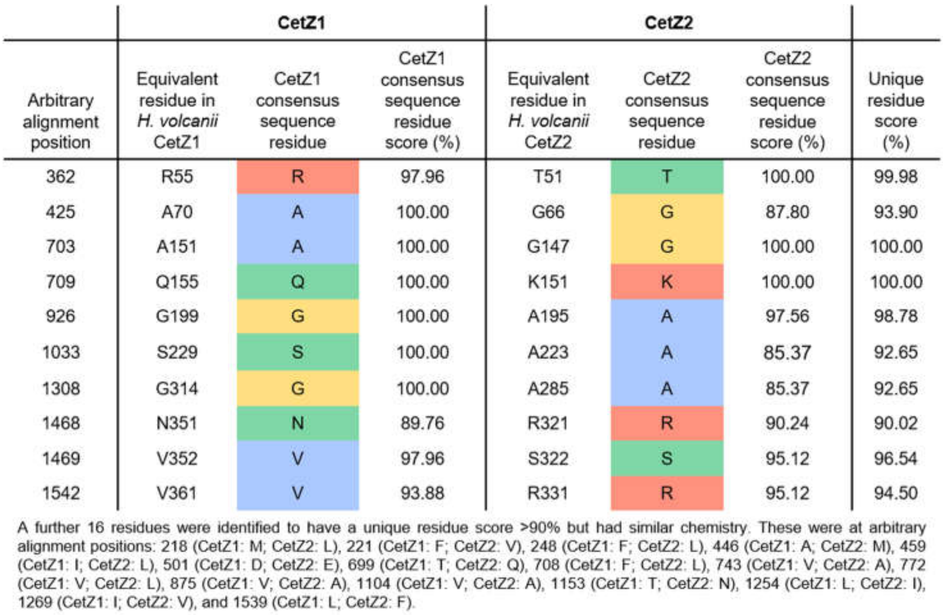
Interestingly, two CetZs were identified in each complete genome analysed from the orders Archaeoglobales and Methanomicrobiales, and they appeared in two corresponding regions of the tree: one formed a single branch within the main CetZ group dominated by Halobacteria sequences and the other formed a more diverse set which also branched more deeply—closer to the deepest classified CetZs from Thermococci (Fig. 1). AlphaFold predicted structures of these CetZs in (Fig. S1) showed that the Halobacteria-like CetZ was structurally more like CetZ1, while the other was akin to CetZs from Thermococci and the crystal structure of CetZ from *Methanosaeta thermophilla*. The two protein subfamilies are therefore likely to have distinct functions in these species, and one appears to be phylogenetically and possibly functionally related to the Halobacteria CetZs. In the order Methanosarcinales, usually only one CetZ was identified per genome which clustered with the non-Halobacteria-like CetZs from Archaeoglobales and Methanomicrobiales.

3.5. Halobacteria CetZ1 and CetZ2 subfamilies show distinct characteristics

Having surveyed the diversity of CetZs across archaea, the strength of the grouping of CetZ1 and CetZ2 subfamilies had become clear. We then sought to identify key characteristics and differences between CetZ1 and CetZ2 by comparing their sequence features and available crystal structures[4]. Ten amino-acid residues with different chemistry between CetZ1 and CetZ2, and 16 residues with similar chemistry, met the 90% conservation criterion (Fig. 3a). When mapped to the crystal structures of CetZ1 and CetZ2 from *H. volcanii*, these residues were generally located on the surface of the proteins, and not in any specific region (Fig. 3b).

Other larger-scale structural differences were also detected. Surface electrostatic analysis showed that the CetZ2 surface was substantially more negatively charged than that of CetZ1 (Fig. 3c). The sequence alignment of CetZ1 and CetZ2 proteins also showed that CetZ1 has a long M-loop (or Microtubule loop, originally assigned a role in tubulin filament lateral association), usually 14-26 residues long, which is unresolved in the crystal structure of CetZ1 from *H. volcanii*. In comparison, CetZ2 has a short M-loop of around 3-6 amino acids. Thermococci CetZs also had a short M-loop, but other CetZs generally had long M-loop regions, including those from Halobacteria, Methanomicrobia, and

Archaeoglobales. Based on the bootstrap values of CetZ1 and CetZ2 branches and by using the above identified characteristics of CetZ1 and CetZ2 proteins, we defined the distinct groupings CetZ1 and CetZ2 homologues as circled in Figure 1. The individual assigned CetZ1 and CetZ2 homologues are listed in Table S2.



	CetZ1			CetZ2			
Arbitrary alignment position	Equivalent residue in *H. volcanii* CetZ1	CetZ1 consensus sequence residue	CetZ1 consensus sequence residue score (%)	Equivalent residue in *H. volcanii* CetZ2	CetZ2 consensus sequence residue	CetZ2 consensus sequence residue score (%)	Unique residue score (%)
362	R55	R	97.96	T51	T	100.00	99.98
425	A70	A	100.00	G66	G	87.80	93.90
703	A151	A	100.00	G147	G	100.00	100.00
709	Q155	Q	100.00	K151	K	100.00	100.00
926	G199	G	100.00	A195	A	97.56	98.78
1033	S229	S	100.00	A223	A	85.37	92.65
1308	G314	G	100.00	A285	A	85.37	92.65
1468	N351	N	89.76	R321	R	90.24	90.02
1469	V352	V	97.96	S322	S	95.12	96.54
1542	V361	V	93.88	R331	R	95.12	94.50

A further 16 residues were identified to have a unique residue score >90% but had similar chemistry. These were at arbitrary alignment positions: 218 (CetZ1: M; CetZ2: L), 221 (CetZ1: F; CetZ2: V), 248 (CetZ1: F; CetZ2: L), 446 (CetZ1: A; CetZ2: M), 459 (CetZ1: I; CetZ2: L), 501 (CetZ1: D; CetZ2: E), 699 (CetZ1: T; CetZ2: Q), 708 (CetZ1: F; CetZ2: L), 743 (CetZ1: V; CetZ2: A), 772 (CetZ1: V; CetZ2: L), 875 (CetZ1: V; CetZ2: A), 1104 (CetZ1: V; CetZ2: A), 1153 (CetZ1: T; CetZ2: N), 1254 (CetZ1: L; CetZ2: I), 1269 (CetZ1: I; CetZ2: V), and 1539 (CetZ1: L; CetZ2: F).

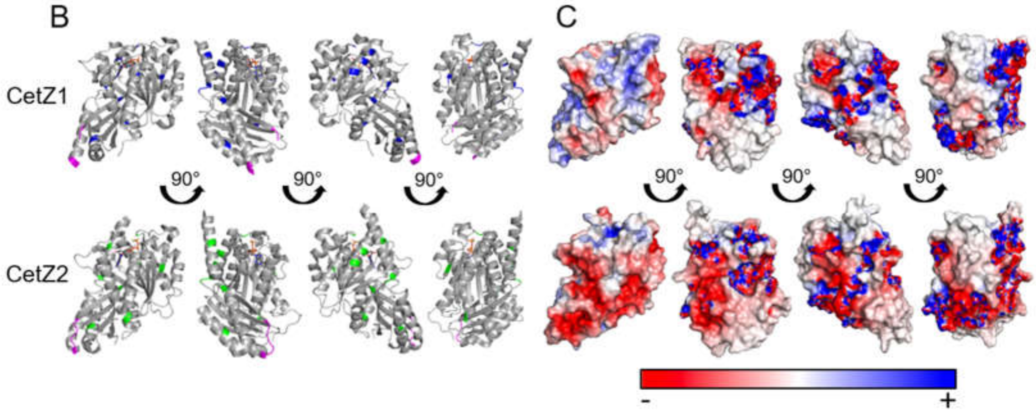


Figure 3. Characterising unique residues in CetZ1 and CetZ2 sub-families from Halobacteria. **A)** Unique residues were defined as residues which had a unique residue score of greater than 90%. The consensus sequence residue score and unique residue score were calculated as in Figure 2. One letter amino acid codes are used, and consensus residues are coloured according to their chemistry. Red: positively charged; blue: hydrophobic; green: polar; yellow: glycine. **B)** Three-dimensional structural comparison of crystal structures of GDP bound CetZ1 (PDB: 4B46) and GTP bound CetZ2 (PDB: 4B45) from *H. volcanii*. Residues that constitute the M-loop of CetZ1 and CetZ2 are represented in magenta. Unique residues listed in A are represented in blue for CetZ1, and green for CetZ2. **C)** Surface charge of CetZ1 and CetZ2, calculated using PyMOL.

3.6. The presence of CetZ1 and CetZ2 in Halobacteria correlates with rod shape and motility

CetZ1 and possibly CetZ2 have been implicated in regulation of cell shape linked to motility in the model archaeon *H. volcanii*. To investigate whether these are likely to be a general function of these subfamilies, the distributions of CetZ1 and CetZ2 across 55 Halobacteria species were compared, and the reported motility and cell shape phenotypes of each species were tabulated (Fig. 4, Fig. S2, Table S1). Note that species that did not fall into the "motility reported" or "rods reported" categories are not necessarily non-motile or non-rod forming species, due to limited observations available for some species or potential conditional phenotypes.

A majority of Halobacteria (45 out of 55 species) were found to have both CetZ1 and CetZ2, with many having other CetZs as well. Forty species were reported to form rods, and 37 of these also had both CetZ1 and CetZ2. A smaller proportion of the 55 Halobacteria were reported as motile (26 species), however 24 of these motile species had both CetZ1 and CetZ2. While there appears to be a strong correlation between the presence of both CetZ1 and CetZ2, rod shape, and motility, the same was not true for the seven species that had CetZ1 but not CetZ2 (CetZ2 was never present without CetZ1); five of the seven were not reported as motile or rod forming, two were rod forming, and one was motile. These observations reinforce the apparent correlation between the presence of both CetZ1 and CetZ2, motility, and rod shape.

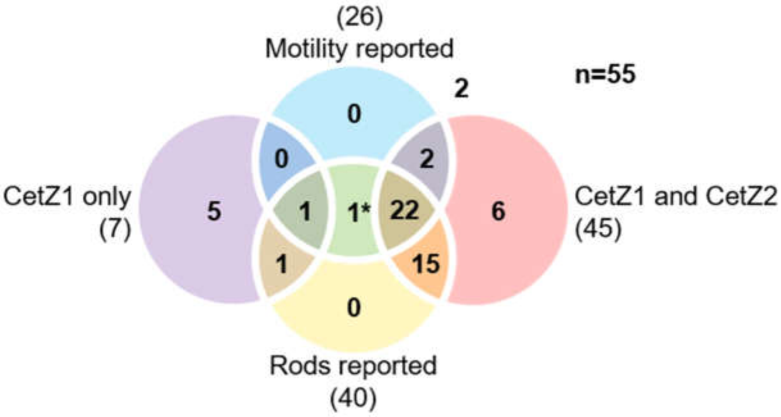


Figure 4. Correlations between CetZ1, CetZ2, rod shape, and motility in Halobacteria. Venn diagram showing the number of Halobacteria species reported as motile or rod forming, having CetZ1 and CetZ2, or having CetZ1 only. Species specific data is provided in Figure S2. *The organism in this category is *Natronomonas pharaonis*. While its genome annotation includes a CetZ1 and a CetZ2, these did not branch within the clear CetZ1 and CetZ2 clusters in Figure 1 (Fig. S3) and are likely to be divergent CetZs specific to this species. As they are uncharacterised in the literature their annotations as CetZ1 and CetZ2 may or may not be correct.

There were 3 species that had no clear CetZ1 nor CetZ2: *Halococcus saccharolyticus*, *Halococcus morrhuae*, and *Natronomonas pharaonis*. Both Halococcus species did not have other CetZs, and were described as coccoid shape with no pleomorphism or motility reported, which appears to be consistent across all species of the Halococcus genus[49,50]. On the other hand, *N. pharaonis* is motile and rod-shaped[51], and was identified to have three CetZs, one that groups with the non-canonical TSF proteins, and two CetZs that clearly lie outside the CetZ1 and CetZ2 branches (Fig. S3a), though have been annotated in the Uniprot database as CetZ1 and CetZ2. Future identification of their potential roles in *N. pharaonis* would be of interest for comparing the function and diversification of CetZs amongst Halobacteria.

The above observations suggested a potential for the conserved groups of CetZ proteins to play multiple complex roles in cell shape and motility in each species. To further investigate the potential multifaceted functions of CetZ1 and CetZ2 in Halobacteria, the 40 kb genomic regions centred on the *cetZ1* and *cetZ2* genes of a diverse set of at least 20 Halobacteria species were analysed to identify local gene content and synteny. The EggNOG v5.0[46,47] database was used to classify genes in these regions based on their homology groups identified in the collection of archaeal Clusters of Orthologous Groups (arCOGs)[52,53], as described below.

3.7. cetZ1 genomic regions are associated with cofactor and nucleotide biosynthesis

Figure 5 lists the arCOGs found in at least half of the analysed *cetZ1* genomic regions and shows maps of the relevant conserved parts of the *cetZ1* genomic regions. Regions within the 40 kb area with no consistent genomic arrangement across species are not shown but are described in Table S3. Twenty *cetZ1* genomic regions were analysed, and the arCOGID for cetZ, arCOG02202, was counted 22 times. This is because in two species, *Halobellus limi* and *Natrialba magadii*, another cetZ gene was located within 20 kb of the cetZ1 gene. The arCOG counted most often was arCOG01117, a transcriptional regulator, with 26 occurrences. Further investigation showed that most often these transcriptional regulators appeared in pairs either side of a gene belonging to arCOG01957 involved in potassium transport, which was identified in 12 genomic regions. These transcriptional regulators were present in 15 of the 20 analysed genomic regions, and belong to the Lrp/AsnC family of transcriptional regulators that are abundant and widespread in archaea[54], and have been implicated in the regulation of amino acid and energy metabolism, translation and DNA repair, and responses to physiological conditions such as growth phase and oxidative stress[55-58]. The next hit, arCOG03015 (*nolA*), was identified in all analysed *cetZ1* genomic regions, and was immediately upstream of *cetZ1* in most. As with most of the genes in the region, the role of *nolA* in Halobacteria is largely unstudied, however, by homology, *nolA* is a predicted NAD-dependent nucleoside-diphosphate-sugar epimerase, and homologs from other species have roles in cell surface polysaccharide biosynthesis[59]. Interestingly, several other genes involved in biosynthesis of coenzyme F_{420} (*cofC*, *cofG* and *cofH*), and nucleotides (*purC*, *purQ*, *purS*, and a thymidylate kinase gene), were also common within the *cetZ1* genomic regions (Fig. 5). The above linkages may suggest a potential functional association or common biological purpose between the role of the CetZ1 cytoskeleton in nutrient-dependent motility and in cellular energy acquisition or biosynthesis, however any potential direct functional significance of the apparent associations is yet to be revealed.

			Order		
arCOGID	arCOG description	All *cetZ1* regions (20 regions)	Haloferacales (8 regions)	Halobacteriales (7 regions)	Natrialbales (5 regions)
arCOG01117	COG 1522 Transcriptional regulators	26	10	7	9
arCOG02202	cetZ, cell shape control	22	9	7	6
arCOG03015	nolA, NAD-dependent epimerase dehydratase	20	8	7	5
arCOG04472	cofC, cofactor biosynthesis	19	8	6	5
arCOG01891	thymidylate kinase	19	8	6	5
arCOG00657	cofG, cofactor biosynthesis	18	8	6	4
arCOG04764	-	15	8	3	4
arCOG04421	purC, purine biosynthesis	13	4	6	3
arCOG00656	cofH, cofactor biosynthesis	13	4	6	3
arCOG00102	purQ, purine biosynthesis	11	3	6	2
arCOG04462	purS, purine biosynthesis	11	3	6	2
arCOG03888	-	11	4	7	0
arCOG01957	trkA2, COG0569 potassium transport systems, NAD-binding component	11	5	2	4
arCOG06366	-	10	7	3	0

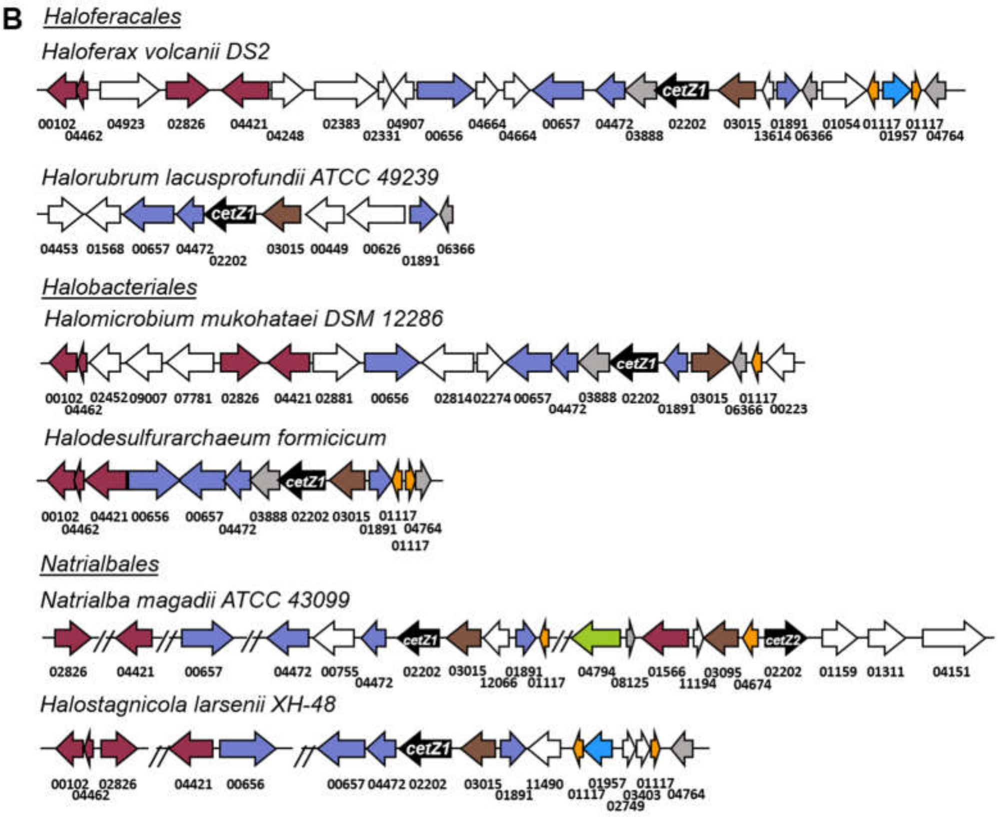
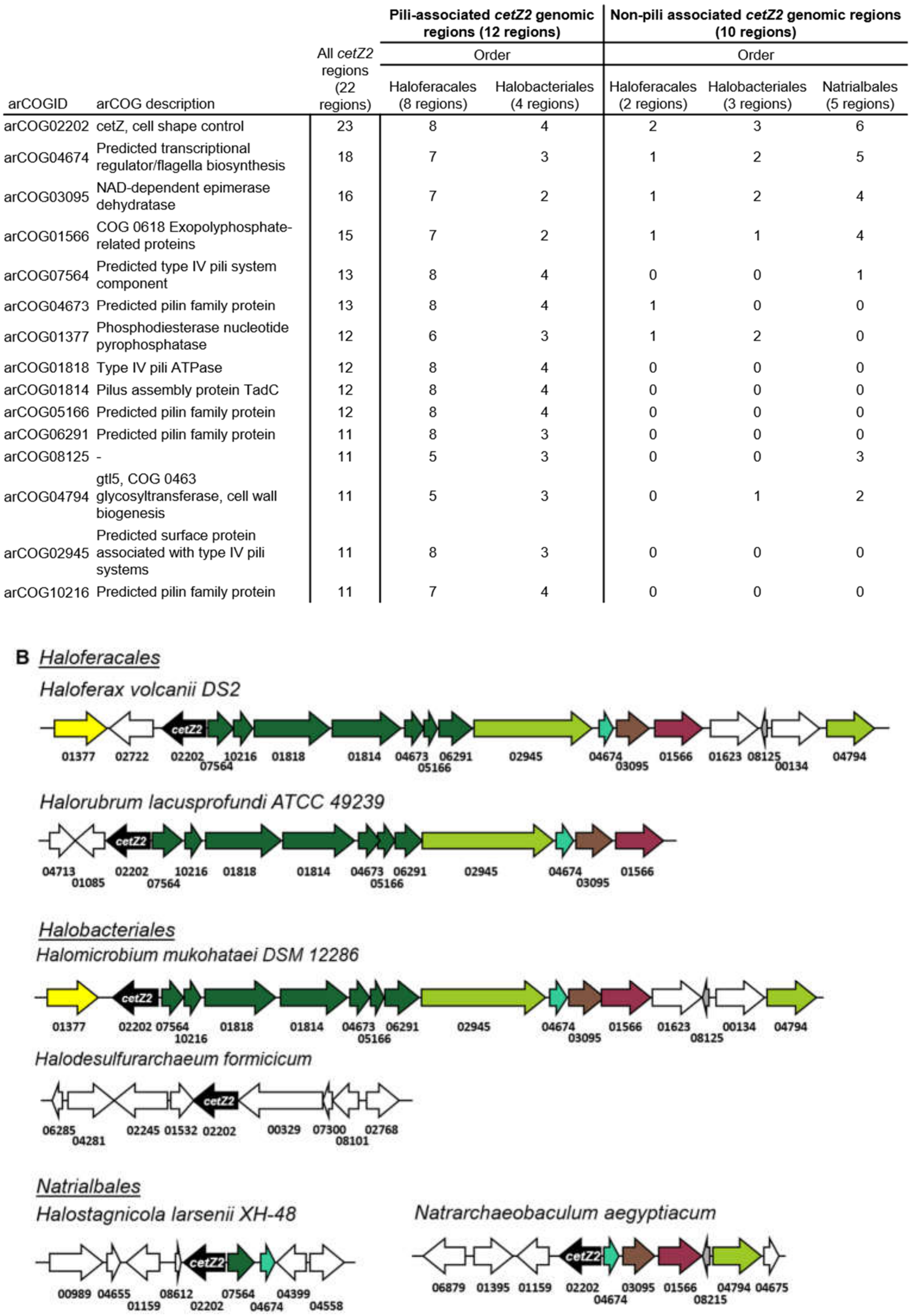


Figure 5. Conserved portions of *cetZ1* genomic regions in selected Halobacteria. A) The 20 kb regions either side of the *cetZ1* gene were analysed from 20 species of the orders Haloferacales (8), Halobacteriales (7), and Natrialbales (5). Table listing the most frequently observed arCOGs and their occurrance in 20 *cetZ1* genomic regions from the order Haloferacales, Halobacteriales, and Natrialbales. **B)** Examples of *cetZ1* genomic regions, showing their conserved residues only. The genes encoding *cetZ1* are represented in black, and genes belonging to arCOGs present within majority (at least 10) of the *cetZ1* genomic regions are coloured according to their COG category as detailed in Table S3, and arCOGIDs are listed beneath each gene. arCOGs encoding uncharacterised proteins which are present in at least half of the *cetZ1* genomic regions are represented in grey, and arCOGs not present in the majority of *cetZ1* genomic regions are represented in white. In Natrialba magadii

ATCC 43099 *cetZ2* was within the 40 kb *cetZ1* genomic region, as well as some arCOGs often conserved within *cetZ2* genomic regions.

3.8. cetZ2 is associated with a type IV pili regulon in Haloferacales and Halobacteriales

The genes surrounding *cetZ2* were next analysed in 22 Halobacteria (Fig. 6). The arCOG for CetZ appeared 23 times in this set—the 22 *cetZ2* genes and one cetZ1 that was located within the 40 kb range of *cetZ2* in *Natrialba magadii*. In most species from the orders Haloferacales and Halobacteriales, *cetZ2* was located adjacent to a *pilB/C* regulon (encoding a predicted type IV pilus system), but not in the third order, Natrialbales. Figure 6a lists the top arCOG hits within pili-associated (12 species) and non-pili-associated (10 species) *cetZ2* regions, examples of which are shown in Figure 6b. The pili-associated *cetZ2* regions showed very similar gene organisation and arCOG conservation within the *pilB/C* regulon, and these arCOGs dominate the *cetZ2* regions overall. Some notable arCOGs within the *cetZ2* pili-associated regions include *pilB*, an ATPase motor which provides the energy required for pilus biogenesis, and *pilC*, the inner membrane component and base of type IV pili. Interestingly, CetZs were frequently adjacent to pili regulons in Halobacteria[60], however that study did not investigate which family the CetZs belonged to (i.e., CetZ1, CetZ2, or other homologues), or whether the CetZ homologues adjacent to pili regulons were consistent between species. Here, we identify that these *pilB/C* regulons are specifically adjacent to *cetZ2*. Their distribution pattern appears to represent a mobile genetic element, as genes beyond the regulon are generally consistent with those found adjacent to *cetZ2* in the non-pili associated regions.



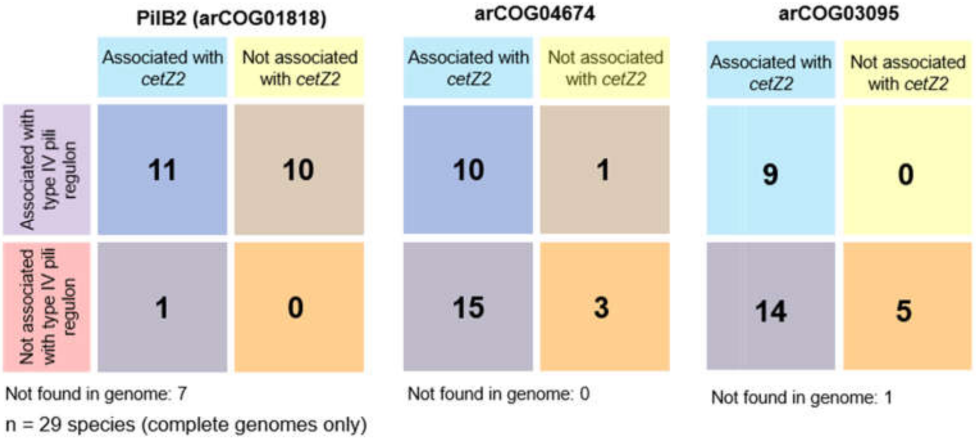
			Pili-associated *cetZ2* genomic regions (12 regions)		Non-pili associated *cetZ2* genomic regions (10 regions)		
		All *cetZ2* regions (22 regions)	Order		Order		
arCOGID	arCOG description		Haloferacales (8 regions)	Halobacteriales (4 regions)	Haloferacales (2 regions)	Halobacteriales (3 regions)	Natrialbales (5 regions)
arCOG02202	cetZ, cell shape control	23	8	4	2	3	6
arCOG04674	Predicted transcriptional regulator/flagella biosynthesis	18	7	3	1	2	5
arCOG03095	NAD-dependent epimerase dehydratase	16	7	2	1	2	4
arCOG01566	COG 0618 Exopolyphosphate-related proteins	15	7	2	1	1	4
arCOG07564	Predicted type IV pili system component	13	8	4	0	0	1
arCOG04673	Predicted pilin family protein	13	8	4	1	0	0
arCOG01377	Phosphodiesterase nucleotide pyrophosphatase	12	6	3	1	2	0
arCOG01818	Type IV pili ATPase	12	8	4	0	0	0
arCOG01814	Pilus assembly protein TadC	12	8	4	0	0	0
arCOG05166	Predicted pilin family protein	12	8	4	0	0	0
arCOG06291	Predicted pilin family protein	11	8	3	0	0	0
arCOG08125	-	11	5	3	0	0	3
arCOG04794	gtl5, COG 0463 glycosyltransferase, cell wall biogenesis	11	5	3	0	1	2
arCOG02945	Predicted surface protein associated with type IV pili systems	11	8	3	0	0	0
arCOG10216	Predicted pilin family protein	11	7	4	0	0	0

Figure 6. Conserved portions of pili-associated and non-pili-associated *cetZ2* genomic regions in Halobacteria. A) The 20 kb regions either side of the *cetZ2* gene were analysed from 22 species of the orders Haloferacales (10), Halobacteriales (7), and Natrialbales (5). Of these 22 regions, 12 were pili-associated and 10 were non-pili-associated. Table listing the most frequently observed arCOGs and their occurrance in 22 *cetZ2* genomic regions from the order Haloferacales, Halobacteriales, and Natrialbales. **B)** Examples of *cetZ2* genomic regions, showing their conserved residues only. The cetZ2 regions from *Haloferax volcanii DS2*, *Halorubrum lacusprofundi ATCC 49239* and *Halomicrobium mukohataei DSM 12286* are pili-associated. The *cetZ2* regions from *Halodesulfurarchaeum formicicum*, *Halostagnicola larsenii XH-48* and *Natrarchaeobaculum aegyptiacum* are non-pili associated. Genes encoding *cetZ2* are represented in black, and genes belonging to arCOGs present within majority (at least 11) of the *cetZ2* genomic regions are coloured according to their COG category as detailed in Table S3. arCOGIDs are listed beneath each gene. arCOGs encoding uncharacterised proteins which are present in at least half of the *cetZ2* genomic regions are represented in grey, and arCOGs not present in majority of *cetZ2* genomic regions are represented in white.

3.9. Non-pili associated genes conserved in cetZ2 regions

No arCOGs were identified to be conserved in only non-pili-associated *cetZ2* regions, consistent with the possibility that the type IV pili regulon was a relatively recent acquisition that has been retained only in some Haloferacales and Halobacteriales. Two arCOGs were almost always present in *cetZ2* regions, whether pili-associated or non-pili-associated. These were arCOG04674 (hypothetical protein), and arCOG03095, an epimerase/dehydratase that is structurally homologous to the epimerase found adjacent to *cetZ1* (Fig. S4). arCOG04674 was previously annotated as a potential transcription factor and is strongly predicted to be structurally homologous to other known transcription factors (Fig. S4). In *H. volcanii*, it is upregulated in response to low and high salinity and low temperature[61]. arCOG04674 from has also been annotated as "COG0630 Type IV secretory pathway, VirB11 components, and related ATPases involved in archaeal flagella biosynthesis", however our analysis would suggest it is not directly associated with the type IV pili system that are sometimes found with *cetZ*. Other *cetZ2*-associated genes included arCOG01566 (predicted exopolyphosphate-related protein), arCOG01377 (predicted phosphodiesterase nucleotide pyrophosphatase), arCOG08125 (uncharacterised protein), and arCOG04794 (predicted glycosyltransferase)

The above results suggested a notable synteny between *cetZ2*, arCOG01818 (pilB2), arCOG04674 and arCOG03095. To further investigate, we expanded the analysis of these associations by including an additional 19 Halobacteria and we identified CetZ2 at the whole-genome level and recorded whether they were proximal to arCOG04674, arCOG03095 or a *pilB2/C2* type IV regulon (Fig. 7, Fig. S5). This confirmed that *cetZ2* was only adjacent to a *pilB2/C2* type IV regulon in the orders Haloferacales and Halobacteriales amongst the 41 total species we analysed. In 13 species, arCOG01818 (*pilB2*) was identified in a *pilB2/C2* type IV regulon that was distant from *cetZ2*. Species from the order Natrialbales typically had *pilB2* not contained within a regulon, or no *pilB2* at all. One exception was *Halobiforma lacisalsi AJ5*, which had a *pilB2/C2* regulon that was not adjacent to *cetZ2*. The hypothetical protein arCOG04674 was found within the *cetZ2* region in 34 of the 41 analysed species, and in the remaining seven species it was positioned elsewhere on the genome, sometimes associated with a *pilB2/C2* type IV regulon. Similarly, 32 species had the predicted sugar epimerase arCOG03095 within the *cetZ2* region, arCOG03095 was detected elsewhere on the genome in eight species and sometimes associated with a *pilB2/C2* regulon, and only one species (*Halodesulfurarchaeum formicicum*) was found to have no gene belonging to this arCOG in its complete genome.



PilB2 (arCOG01818)

	Associated with *cetZ2*	Not associated with *cetZ2*
Associated with type IV pili regulon	11	10
Not associated with type IV pili regulon	1	0

Not found in genome: 7

arCOG04674

	Associated with *cetZ2*	Not associated with *cetZ2*
Associated with type IV pili regulon	10	1
Not associated with type IV pili regulon	15	3

Not found in genome: 0

arCOG03095

	Associated with *cetZ2*	Not associated with *cetZ2*
Associated with type IV pili regulon	9	0
Not associated with type IV pili regulon	14	5

Not found in genome: 1

n = 29 species (complete genomes only)

Figure 7. Synteny of cetZ2 with type IV system components pilB2 (arCOG01818), arCOG04674, and arCOG03095. arCOGs were classified as 'associated with *cetZ2*' if found within 20 kb either side of the *cetZ2* gene. arCOGs 'associated with a type IV pili regulon' were defined as those found within or adjacent to other arCOGs identified in the *H. volcanii PilB2C2* regulon. Species specific data is provided in Figure S5.

3.10. The genomic environments of N. pharaonis cetZ genes

We also examined the gene neighbourhoods of the two *N. pharaonis cetZ* genes that sit outside the currently defined CetZ1 and CetZ2 families in the phylogenetic trees (Fig. 1, Fig. S3a), yet have been annotated as cetZ1 (UniProt accession number Q3IRF0) and *cetZ2* (UniProt accession number Q3IRT7). Interestingly, the less divergent protein, Q3IRF0, had a typical *cetZ1* genomic organisation (Fig. S3d) and was in proximity to many of the arCOGs often conserved in *cetZ1* regions. It shared 5 of the 10 unique residues with CetZ1, and had a long M-loop region (Fig. S3b, c), suggesting that Q3IRF0 may be derived from the CetZ1 subfamily. Conversely, Q3IRT7 was not contained within a genomic region typical for *cetZ2* genes from other species (Fig. S3e) and it only shared 1 of the 10 unique residues with CetZ2, but shared 4 with CetZ1 (Fig. S3b) and had a long M-loop (Fig. S3c) region which is uncharacteristic of other CetZ2 proteins. Hence, Q3IRT7 is unlikely to be a CetZ2, and is more likely an additional or redundant version of Q3IRF0.

3.11. Synteny in cetZ genomic regions of non-Halobacteria Euryarchaeota

Euryarchaea outside of the class Halobacteria that have CetZ belonged to one of four orders: Thermococcales, Archaeoglobales, Methanomicrobiales, and Methanosarcinales. The genomic regions surrounding 27 of these *cetZ* genes (located outside the Halobacteria CetZ branch; Fig. 1) were analysed to search for conservation or synteny. Top arCOG hits for the CetZ regions and exemplary genomic maps are shown in Figure 8. Strikingly, a majority of the top hit arCOGs in *cetZ* regions were predicted pilin family proteins linked to a type IV pili-like system (arCOG05787, 05789, 03821, 03822, 05790, 05786, and 05788), reminiscent of the synteny between the *pilB2/C2* type IV regulon and *cetZ2* in some Halobacteria. Breakdown of the gene association by taxonomic order (Fig. 8) revealed that this strong association was solely present in the Thermococcales; we observed that all analysed *cetZ* regions from Thermococcales were pili-associated, and the genomic organisation of the arCOGs within this region was also well conserved. The tendency we have observed for *cetZ* genes to be associated with type IV systems may indicate that CetZ proteins could have been co-opted by these systems for structural roles in their assembly or function.

		All *cetZ*	Order			
arCOGID	arCOG description	regions (27 regions)	Thermococcales (15 regions)	Archaeoglobales (3 regions)	Methanosarcinales (6 regions)	Methanomicrobiales (3 regions)
arCOG02202	cetZ, cell shape control	27	15	3	6	3
arCOG04346	purP,	21	21	0	0	0
arCOG05787	Predicted surface protein associated with type IV pili like system	15	15	0	0	0
arCOG05789	Predicted pilin family protein	15	15	0	0	0
arCOG03821	Secreted protein, component of type IV pili like system	15	15	0	0	0
arCOG03822	Pilin family protein, contains class III signal peptide	15	15	0	0	0
arCOG05790	Predicted component of type IV pili like system	15	15	0	0	0
arCOG03396	Squaline cyclase	15	15	0	0	0
arCOG00085	guaAB, GMP synthesis	14	14	0	0	0
arCOG00087	guaAA, GMP synthesis	14	14	0	0	0
arCOG05786	Predicted component of type IV pili like system	14	14	0	0	0
arCOG03512	Predicted surface protein	14	14	0	0	0
arCOG05788	Predicted component of type IV like system	13	13	0	0	0
arCOG05785	-	11	11	0	0	0
arCOG04174	taw1, wyosine derivatives biosynthesis	10	10	0	0	0
arCOGs highly conserved in Methanosarcinales *cetZ* regions, or arCOGs which relate to the functions of other highly conserved arCOGs						
arCOG01482	nadC, nicotinate-nucleotide pyrophosphorylase	8	2	0	6	0
arCOG04459	nadA, quinolinate synthase A	7	2	0	5	0
arCOG00476	-	6	1	0	5	0
arCOG00254	nadX, L-aspartate dehydrogenase	6	0	0	6	0
arCOG11330	-	6	6	0	0	0
arCOG04415	purD, purine biosynthesis	6	6	0	0	0
arCOG04462	purS, purine biosynthesis	6	6	0	0	0
arCOG00102	purQ, purine biosynthesis	6	6	0	0	0
arCOG00641	purL, purine biosynthesis	6	6	0	0	0

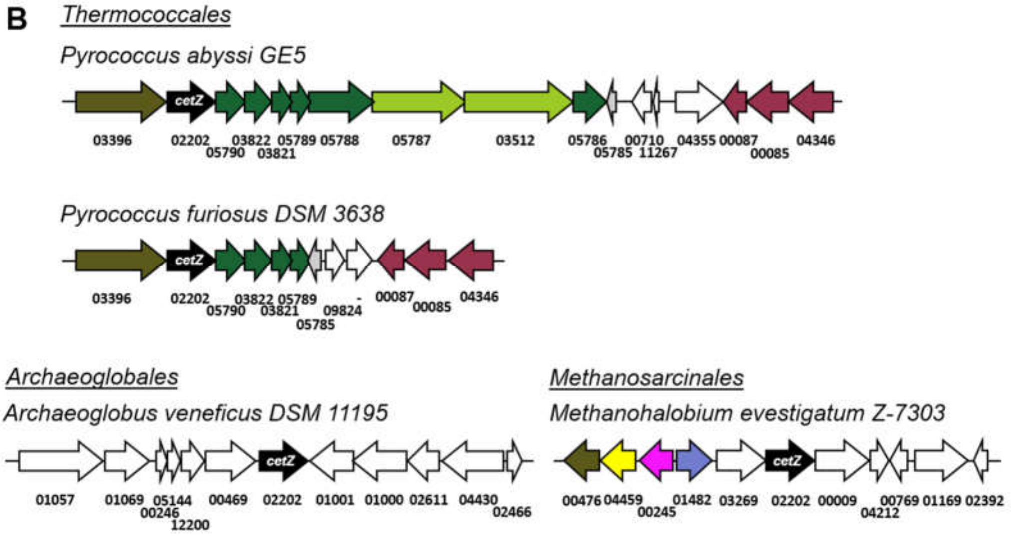


Figure 8. Conserved portions of the 40kb *cetZ* genomic regions from non-Haloarchaea Euryarchaeota. A) The 20 kb regions either side of the *cetZ* gene were analysed from 27 species of the orders Thermococcales (15), Archaeoglobales (3), Methanosarcinales (6), and Methanomicrobiales (3). Table listing the most frequently observed arCOGs and their occurrence in 27 *cetZ* genomic regions from each order. **B)** Examples of *cetZ* genomic regions, showing their conserved residues only (except in the case of *Archaeoglobus veneficus DSM 11195* which has no conserved regions). The genes encoding *cetZ* are represented in black, and genes belonging to arCOGs present within at least 10 of the *cetZ* genomic regions are coloured according to their COG category as detailed in Table S3, and arCOGIDs are listed beneath each gene. arCOGs encoding uncharacterised proteins which are present in at least 10 of the *cetZ* genomic regions are represented in grey, and arCOGs not present in at least 10 of *cetZ* genomic regions are represented in white. Four arCOGs (00476, 04459, 00245, and

doi:10.20944/preprints202211.0451.v1

01482) were not present in majority of all *cetZ* genomic regions, but were highly conserved within the order Methanosarcinales. These arCOGs are also coloured according to their COG category.

The Thermococcales *cetZ* regions beyond the pili-like system were also reminiscent of the Halobacteria *cetZ1* regions in that purine and cofactor biosynthesis genes were often located within the 40 kb *cetZ* region (Fig. 8a). They were also implicated in cell wall/membrane/envelope biogenesis and cell motility/adhesion, like the *cetZ2* genomic regions (Fig. 8, lime green). The analysed *cetZs* in Archaeoglobales, Methanomicrobiales and Methanosarcinales were non-pili associated and showed few similarities within or across taxa. However, four arCOGs were consistently observed in the *cetZ* genomic regions of Methanosarcinales species, which notably included nadC, nadA, and nadX, involved in cofactor (NAD) biosynthesis.

4. Discussion

Despite comprising a major family of the tubulin superfamily, our current knowledge of the biological functions of CetZs is based on a limited number of studies in *H. volcanii*. Previous studies of the distribution of CetZs[4,48] have shown they are diverse, and present across a broad range of archaeal species, but absent in other major groups. Although CetZs have been implicated in the control of archaeal cell shape control and motility, cytoskeletal proteins typically perform a wide range of biological functions and have downstream effects on many biological systems and processes. In this study, we assessed the diversity of tubulin superfamily proteins from 183 archaeal species, defined sequence and structural features of the clearest groups, and searched for genes commonly co-conserved within the *cetZ* genomic regions as a first look into other potential biological functions of CetZs as cytoskeletal proteins.

As expected, Halobacteria were found to have multiple CetZ homologues within the same species. The most conserved of the Halobacteria CetZs were CetZ1 and CetZ2, which formed distinct orthologous groups (Fig. 1). Although both CetZ1 and CetZ2 have previously been reported to function generally in cell shape control and motility, they do not produce the same phenotypes[4], and may have different roles. In this study, we describe key differences between CetZ1 and CetZ2 sub-families, including conserved differences in their amino acid sequences, M-loop regions, and surface charge (Fig. 3). In addition, conserved genes within *cetZ1* and *cetZ2* genomic regions were different, with common genes in *cetZ1* regions having predicted functions in nucleotide and coenzyme biosynthesis (Fig. 6), and common genes in *cetZ2* regions functioning in cell envelope biogenesis and cell motility/adhesion (Fig. 7). CetZ2 was only present in species that also had CetZ1 (Fig. 4, Fig. S3), and the majority of rod forming and motile Halobacteria species had both CetZ1 and CetZ2 (Fig. 4, Fig. S3), suggesting that might be functionally dependent on CetZ1.

Outside of Halobacteria, CetZs were mostly clustered taxonomically, and were only confidently identified in Euryarchaeota, within the orders Thermococcales, Archaeoglobales, Methanomicrobiales, and Methanosarcinales (Fig. 1). The overall CetZ family was compared in sequence and structure to the archaeal FtsZ. Characteristic amino acid differences were concentrated around the GTP/GDP binding pocket (Fig. 2), which may result in fundamental differences in the polymerisation behaviour and function of FtsZs and CetZs.

CetZs from Thermocococcales were strongly clustered in the phylogenetic analysis (Fig. 1) and were structurally comparable to CetZs from other Euryarchaeal species (Fig. 2, Fig. S1), supporting their inclusion in the CetZ family. Thermococcales CetZs also showed consistent genomic organisation with genes involved in nucleotide and coenzyme metabolism (like in Halobacteria *cetZ1* regions), and cell motility/adhesion and cell wall/membrane/envelope biogenesis (like in Halobacteria cetZ2 regions) (Fig. 8). Unlike Halobacteria, Thermococcales species only have one CetZ. It is interesting that the Thermococcales *cetZ* regions contained genes with similar predicted functions to those conserved in both *cetZ1* and *cetZ2* regions from Halobacteria. This is consistent with the notion that there has been divergence and specialization of multiple CetZ functions that could be performed similarly by the sole CetZ in Thermococcales species.

Archaeoglobales and Methanomicrobiales species typically had one Halobacteria-like CetZ and one CetZ which branched closer to CetZs from Thermococcales. Interestingly, sequence and structure predictions of these two subgroups of CetZs from Archaeoglobus fulgidus (Fig. S1) showed that the Halobacteria-like CetZ (UniProt O29053) had a long M-loop like Halobacteria CetZs (except for those belonging to the CetZ2 subfamily), while the A. fulgidus CetZ clustering more closely with Thermococcales CetZs had a short M-loop. The M-loop region thus appears to have variable functions that may be characteristic of CetZ subfamily functions. Perhaps the two sub-families of CetZs in Archaeoglobales and Methanomicrobiales species are functional pairs with separate functions in these species, like CetZ1 and CetZ2 may be in Halobacteria.

The functions of the most common genes found in CetZ regions across Euryarchaeota may point towards biological pathways or mechanisms involving CetZs. We saw that *cetZ* and *cetZ1* genomic regions contained several genes implicated in nucleotide, cofactor, and sugar metabolism. One set of genes, *purP*, *purD*, *purS*, *purQ*, *purL*, *purU*, *purC*, are involved in de novo biosynthesis of purines (inosine 5′-monophosphate from phosphoribosyl pyrophosphate), with pathways feeding into thiamine, histidine and DNA and RNA biosynthesis[62]. Purinosomes (multienzyme complexes of purine biosynthesis enzymes) have been found to functionally associate with microtubules in human cells[63]. Likewise, it may be possible that CetZs contribute a similar role in helping stabilize purine biosynthesis or other multienzyme complexes in archaea. Three cofactor biosynthesis genes were also frequently present in *cetZ1* genomic regions; *cofC*, *cofG*, and *cofH*, all of which are conserved across bacteria and archaea and contribute to cofactor F_{420} biosynthesis. Cofactor F_{420} is an electron carrier and notably required for redox steps in archaeal methanogenesis[64], but the potential significance of the genetic linkage to *cetZ* is currently unknown. Genes belonging to arCOG03015 and 03095 were identified as highly conserved in *cetZ1* and *cetZ2* regions, respectively. These genes are uncharacterised in archaea but are predicted NAD-dependent sugar epimerase/dehydratases by homology to bacterial enzymes. Cytoskeletal proteins including tubulin and FtsZ have been shown to influence and regulate metabolism [65], and are generally known to help localize biosynthetic activities, so the genetic associations noted above might reflect similar functions of CetZs in stabilizing or localizing metabolic activities in archaeal cells.

Several type IV pili related systems were found to be encoded adjacent to all the *cetZ* regions of Thermococcales and a majority of *cetZ2* regions from Halobacteriales and Haloferacales. The biological role of these type IV systems is unknown, although type IV systems generally assemble as a multi-component extracellular dynamic filament embedded in the cell envelope with roles in adhesion and motility. Makarova et al., 2016[60] identified pili regulons in archaea and found CetZs adjacent to PilBC regulons in Thermococci and PilB2 regulons in Halobacteria, but not with other clades of pili regulons. In general, pili-associated *cetZ* regions contained five to six predicted pilins (the extracellular filament subunits), and typically two predicted envelope proteins. Pili-associated *cetZ2* regions usually had seven predicted pilins, including one ATPase (PilB2) and one transmembrane component (PilC2), as well as one predicted surface protein. Other cytoskeletal proteins including FtsZ are known to act as scaffolds that direct the biosynthesis of the bacterial cell envelope and its substructures [1,33,34,66], and potentially in archaea as well[29,67]. Similarly, it seems possible that CetZs may have been adopted to act as scaffolds for the assembly of type IV pili systems in some archaea.

In summary, we have shown that the CetZ family is comprised of multiple diverse subfamilies across a subset of the Euryarchaeota. In many archaeal species, multiple CetZs from different subfamilies are likely to have separated, specialized functions that may work together in cytoskeletal roles, potentially akin to the known multifunctionality, specialization and coordination of tubulin subfamily members in eukaryotes. The gene association analyses suggested CetZs may act in ways that promote the assembly or

doi:10.20944/preprints202211.0451.v1

localization of biosynthetic and cell-surface associated complexes, akin to the function of the well characterized bacterial cytoskeletal proteins.

Supplementary Materials: Figure S1: Structures of archaeal TSF proteins; Figure S2: The distribution of CetZ1 and CetZ2 in Halobacteria compared to motility and cell morphologies; Figure S3: Natronomonous cetZs; Figure S4: AlphaFold predictions of highly conserved genes within *cetZ1* and *cetZ2* genomic regions; Figure S5: Synteny of *cetZ2*, *pilB2* (arCOG01818), arCOG04674, and arCOG0305; Table S1: Species and TSF extended data for phylogenetic analysis; Table S2: List of *cetZ1* and *cetZ2* homologues; Table S3: *cetZ* genomic region analysis extended data.

Author Contributions: Conceptualization, I.G.D. and H.J.B.; methodology, I.G.D. and H.J.B.; formal analysis, I.G.D. and H.J.B.; investigation, H.J.B.; data curation, H.J.B.; writing—original draft preparation, H.J.B.; writing—review and editing, I.G.D. and H.J.B.; supervision, I.G.D.; project administration, I.G.D.; funding acquisition, I.G.D.. All authors have read and agreed to the published version of the manuscript.

Acknowledgments: This work was supported by the Australian Research Council (IGD) and an Australian Government research training scholarship (HJB). The authors acknowledge use of facilities of the Structural Biology Facility, Mark Wainwright Analytical Centre, UNSW, Australia, and the public databases utilized in this study.

References

1. Adams, D.W.; Errington, J. Bacterial cell division: assembly, maintenance and disassembly of the Z ring. Nature Reviews Microbiology 2009, 7, 642.
2. Barton, N.R.; Goldstein, L. Going mobile: microtubule motors and chromosome segregation. Proceedings of the National Academy of Sciences 1996, 93, 1735-1742.
3. Garcin, C.; Straube, A. Microtubules in cell migration. Essays in biochemistry 2019, 63, 509-520.
4. Duggin, I.G.; Aylett, C.H.; Walsh, J.C.; Michie, K.A.; Wang, Q.; Turnbull, L.; Dawson, E.M.; Harry, E.J.; Whitchurch, C.B.; Amos, L.A. CetZ tubulin-like proteins control archaeal cell shape. Nature 2015, 519, 362.
5. Vicente-Manzanares, M.; Choi, C.K.; Horwitz, A.R. Integrins in cell migration–the actin connection. Journal of cell science 2009, 122, 199-206.
6. Kaverina, I.; Straube, A. Regulation of cell migration by dynamic microtubules. In Proceedings of the Seminars in cell & developmental biology, 2011; pp. 968-974.
7. Qualmann, B.; Kessels, M.M.; Kelly, R.B. Molecular links between endocytosis and the actin cytoskeleton. The Journal of cell biology 2000, 150, F111-F116.
8. Appert-Rolland, C.; Ebbinghaus, M.; Santen, L. Intracellular transport driven by cytoskeletal motors: General mechanisms and defects. Physics Reports 2015, 593, 1-59.
9. Hirokawa, N.; Noda, Y.; Tanaka, Y.; Niwa, S. Kinesin superfamily motor proteins and intracellular transport. Nature reviews Molecular cell biology 2009, 10, 682-696.
10. Woese, C.R.; Kandler, O.; Wheelis, M.L. Towards a natural system of organisms: proposal for the domains Archaea, Bacteria, and Eucarya. Proceedings of the National Academy of Sciences 1990, 87, 4576-4579.
11. Ettema, T.J.; Lindås, A.C.; Bernander, R. An actin-based cytoskeleton in archaea. Molecular microbiology 2011, 80, 1052-1061.
12. Nogales, E.; Downing, K.H.; Amos, L.A.; Löwe, J. Tubulin and FtsZ form a distinct family of GTPases. Nature Structural and Molecular Biology 1998, 5, 451.
13. Desai, A.; Mitchison, T.J. Microtubule polymerization dynamics. Annual review of cell and developmental biology 1997, 13, 83-117.
14. Mitchison, T. Localization of an exchangeable GTP binding site at the plus end of microtubules. Science 1993, 261, 1044-1047.
15. Mitchison, T.; Kirschner, M. Dynamic instability of microtubule growth. nature 1984, 312, 237.
16. Nogales, E. Structural insights into microtubule function. Annual review of biophysics and biomolecular structure 2000, 30, 397-420.
17. Mukherjee, A.; Lutkenhaus, J. Guanine nucleotide-dependent assembly of FtsZ into filaments. Journal of Bacteriology 1994, 176, 2754-2758.
18. Mukherjee, A.; Lutkenhaus, J. Dynamic assembly of FtsZ regulated by GTP hydrolysis. The EMBO journal 1998, 17, 462-469.
19. Erickson, H.P.; Taylor, D.W.; Taylor, K.A.; Bramhill, D. Bacterial cell division protein FtsZ assembles into protofilament sheets and minirings, structural homologs of tubulin polymers. Proceedings of the National Academy of Sciences 1996, 93, 519-523.
20. Goldstein, L.S.; Yang, Z. Microtubule-based transport systems in neurons: the roles of kinesins and dyneins. Annual review of neuroscience 2000, 23, 39-71.
21. Kulić, I.M.; Brown, A.E.; Kim, H.; Kural, C.; Blehm, B.; Selvin, P.R.; Nelson, P.C.; Gelfand, V.I. The role of microtubule movement in bidirectional organelle transport. Proceedings of the National Academy of Sciences 2008, 105, 10011-10016.
22. Lawson, M.A.; Maxfield, F.R. Ca 2+-and calcineurin-dependent recycling of an integrin to the front of migrating neutrophils. Nature 1995, 377, 75-79.

23. Johnsson, A.-K.; Karlsson, R. Microtubule-dependent localization of profilin I mRNA to actin polymerization sites in serum-stimulated cells. European journal of cell biology 2010, 89, 394-401.
24. Bergmann, J.E.; Kupfer, A.; Singer, S. Membrane insertion at the leading edge of motile fibroblasts. Proceedings of the National Academy of Sciences 1983, 80, 1367-1371.
25. Ezratty, E.J.; Partridge, M.A.; Gundersen, G.G. Microtubule-induced focal adhesion disassembly is mediated by dynamin and focal adhesion kinase. Nature cell biology 2005, 7, 581-590.
26. Laan, L.; Husson, J.; Munteanu, E.L.; Kerssemakers, J.W.; Dogterom, M. Force-generation and dynamic instability of microtubule bundles. Proceedings of the National Academy of Sciences 2008, 105, 8920-8925.
27. Inoué, S.; Salmon, E.D. Force Generation by Microtubule Assembly/Disassembly in Mitosis and Related Movements. Molecular Biology of the Cell 1995, 6, 1619-1640, doi:10.1091/mbc.6.12.1619.
28. Liao, Y.; Ithurbide, S.; de Silva, R.T.; Erdmann, S.; Duggin, I.G. Archaeal cell biology: diverse functions of tubulin-like cytoskeletal proteins at the cell envelope. Emerging Topics in Life Sciences 2018, 2, 547-559.
29. Liao, Y.; Ithurbide, S.; Evenhuis, C.; Löwe, J.; Duggin, I.G. Cell division in the archaeon Haloferax volcanii relies on two FtsZ proteins with distinct functions in division ring assembly and constriction. Nature Microbiology 2021, 6, 594-605.
30. Bi, E.; Dai, K.; Subbarao, S.; Beall, B.; Lutkenhaus, J. FtsZ and cell division. Research in microbiology 1991, 142, 249-252.
31. Goehring, N.W.; Beckwith, J. Diverse paths to midcell: assembly of the bacterial cell division machinery. Current Biology 2005, 15, R514-R526.
32. Haeusser, D.P.; Margolin, W. Splitsville: structural and functional insights into the dynamic bacterial Z ring. Nature Reviews Microbiology 2016, 14, 305-319.
33. Yang, X.; Lyu, Z.; Miguel, A.; McQuillen, R.; Huang, K.C.; Xiao, J. GTPase activity–coupled treadmilling of the bacterial tubulin FtsZ organizes septal cell wall synthesis. Science 2017, 355, 744-747.
34. Osawa, M.; Erickson, H.P. Liposome division by a simple bacterial division machinery. Proceedings of the National Academy of Sciences 2013, 110, 11000-11004.
35. Nußbaum, P.; Gerstner, M.; Dingethal, M.; Erb, C.; Albers, S.-V. The archaeal protein SepF is essential for cell division in Haloferax volcanii. Nature communications 2021, 12, 1-15.
36. Liao, Y.; Vogel, V.; Hauber, S.; Bartel, J.; Alkhnbashi, O.S.; Maaß, S.; Schwarz, T.S.; Backofen, R.; Becher, D.; Duggin, I.G. CdrS is a global transcriptional regulator influencing cell division in Haloferax volcanii. bioRxiv 2021.
37. de Silva, R.T.; Abdul-Halim, M.F.; Pittrich, D.A.; Brown, H.J.; Pohlschroder, M.; Duggin, I.G. Improved growth and morphological plasticity of Haloferax volcanii. Microbiology (Reading) 2021, doi:10.1099/mic.0.001012.
38. Mullakhanbhai, M.F.; Larsen, H. Halobacterium volcanii spec. nov., a Dead Sea halobacterium with a moderate salt requirement. Archives of Microbiology 1975, 104, 207-214.
39. Li, Z.; Kinosita, Y.; Rodriguez-Franco, M.; Nußbaum, P.; Braun, F.; Delpech, F.; Quax, T.E.; Albers, S.-V. Positioning of the motility machinery in halophilic archaea. MBio 2019, 10, e00377-00319.
40. Madeira, F.; Park, Y.M.; Lee, J.; Buso, N.; Gur, T.; Madhusoodanan, N.; Basutkar, P.; Tivey, A.R.; Potter, S.C.; Finn, R.D. The EMBL-EBI search and sequence analysis tools APIs in 2019. Nucleic acids research 2019, 47, W636-W641.
41. Edgar, R.C. MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic acids research 2004, 32, 1792-1797.
42. Potter, S.C.; Luciani, A.; Eddy, S.R.; Park, Y.; Lopez, R.; Finn, R.D. HMMER web server: 2018 update. Nucleic acids research 2018, 46, W200-W204.
43. Kumar, S.; Stecher, G.; Li, M.; Knyaz, C.; Tamura, K. MEGA X: molecular evolutionary genetics analysis across computing platforms. Molecular biology and evolution 2018, 35, 1547.
44. Schrödinger, L The PyMOL Molecular Graphics System, Version 1.8, 2015.
45. Jurrus, E.; Engel, D.; Star, K.; Monson, K.; Brandi, J.; Felberg, L.E.; Brookes, D.H.; Wilson, L.; Chen, J.; Liles, K. Improvements to the APBS biomolecular solvation software suite. Protein Science 2018, 27, 112-128.
46. Cantalapiedra, C.P.; Hernández-Plaza, A.; Letunic, I.; Bork, P.; Huerta-Cepas, J. eggNOG-mapper v2: Functional Annotation, Orthology Assignments, and Domain Prediction at the Metagenomic Scale. Molecular Biology and Evolution 2021.
47. Huerta-Cepas, J.; Szklarczyk, D.; Heller, D.; Hernández-Plaza, A.; Forslund, S.K.; Cook, H.; Mende, D.R.; Letunic, I.; Rattei, T.; Jensen, L.J. eggNOG 5.0: a hierarchical, functionally and phylogenetically annotated orthology resource based on 5090 organisms and 2502 viruses. Nucleic acids research 2019, 47, D309-D314.
48. Aylett, C.H.; Duggin, I.G. The tubulin superfamily in archaea. In Prokaryotic Cytoskeletons; Springer: 2017; pp. 393-417.
49. Kocur, M.; Hodgkiss, W. Taxonomic status of the genus Halococcus Schoop. International Journal of Systematic and Evolutionary Microbiology 1973, 23, 151-156.
50. Montero, C.G.; Ventosa, A.; Rodriguez-Valera, F.; Kates, M.; Moldoveanu, N.; Ruiz-Berraquero, F. Halococcus saccharolyticus sp. nov., a new species of extremely halophilic non-alkaliphilic cocci. Systematic and applied microbiology 1989, 12, 167-171.
51. Gonzalez, O.; Oberwinkler, T.; Mansueto, L.; Pfeiffer, F.; Mendoza, E.; Zimmer, R.; Oesterhelt, D. Characterization of growth and metabolism of the haloalkaliphile Natronomonas pharaonis. PLoS computational biology 2010, 6, e1000799.
52. Makarova, K.S.; Koonin, E.V.; Kelman, Z. The CMG (CDC45/RecJ, MCM, GINS) complex is a conserved component of the DNA replication system in all archaea and eukaryotes. Biology direct 2012, 7, 1-10.
53. Makarova, K.S.; Wolf, Y.I.; Koonin, E.V. Archaeal clusters of orthologous genes (arCOGs): an update and application for analysis of shared features between Thermococcales, Methanococcales, and Methanobacteriales. Life 2015, 5, 818-840.

54. Pérez-Rueda, E.; Janga, S.C. Identification and genomic analysis of transcription factors in archaeal genomes exemplifies their functional architecture and evolutionary origin. Molecular biology and evolution 2010, 27, 1449-1459.
55. Lemmens, L.; Maklad, H.R.; Bervoets, I.; Peeters, E. Transcription regulators in archaea: homologies and differences with bacterial regulators. Journal of molecular biology 2019, 431, 4132-4146.
56. Schwaiger, R.; Schwarz, C.; Furtwängler, K.; Tarasov, V.; Wende, A.; Oesterhelt, D. Transcriptional control by two leucine-responsive regulatory proteins in Halobacterium salinarum R1. BMC molecular biology 2010, 11, 1-15.
57. Ettema, T.J.; Brinkman, A.B.; Tani, T.H.; Rafferty, J.B.; Van Der Oost, J. A novel ligand-binding domain involved in regulation of amino acid metabolism in prokaryotes. Journal of Biological Chemistry 2002, 277, 37464-37468.
58. Peeters, E.; Albers, S.V.; Vassart, A.; Driessen, A.J.; Charlier, D. Ss-LrpB, a transcriptional regulator from Sulfolobus solfataricus, regulates a gene cluster with a pyruvate ferredoxin oxidoreductase-encoding operon and permease genes. Molecular microbiology 2009, 71, 972-988.
59. Islam, R.; Brown, S.; Taheri, A.; Dumenyo, C.K. The gene encoding NAD-dependent epimerase/dehydratase, wcaG, affects cell surface properties, virulence, and extracellular enzyme production in the soft rot phytopathogen, Pectobacterium carotovorum. Microorganisms 2019, 7, 172.
60. Makarova, K.S.; Koonin, E.V.; Albers, S.-V. Diversity and evolution of type IV pili systems in archaea. Frontiers in microbiology 2016, 7, 667.
61. Jevtić, Ž.; Stoll, B.; Pfeiffer, F.; Sharma, K.; Urlaub, H.; Marchfelder, A.; Lenz, C. The Response of Haloferax volcanii to Salt and Temperature Stress: A Proteome Study by Label-Free Mass Spectrometry. Proteomics 2019, 19, 1800491.
62. Brown, A.M.; Hoopes, S.L.; White, R.H.; Sarisky, C.A. Purine biosynthesis in archaea: variations on a theme. Biology direct 2011, 6, 1-21.
63. An, S.; Deng, Y.; Tomsho, J.W.; Kyoung, M.; Benkovic, S.J. Microtubule-assisted mechanism for functional metabolic macromolecular complex formation. Proceedings of the National Academy of Sciences 2010, 107, 12872-12876.
64. Grinter, R.; Greening, C. Cofactor F420: an expanded view of its distribution, biosynthesis and roles in bacteria and archaea. FEMS Microbiology Reviews 2021, 45, fuab021.
65. Cassimeris, L.; Silva, V.C.; Miller, E.; Ton, Q.; Molnar, C.; Fong, J. Fueled by microtubules: does tubulin dimer/polymer partitioning regulate intracellular metabolism? Cytoskeleton 2012, 69, 133-143.
66. Szwedziak, P.; Wang, Q.; Bharat, T.A.; Tsim, M.; Löwe, J. Architecture of the ring formed by the tubulin homologue FtsZ in bacterial cell division. Elife 2014, 3, e04601.
67. Abdul Halim, M.F.; Schulze, S.; DiLucido, A.; Pfeiffer, F.; Bisson Filho, A.W.; Pohlschroder, M. Lipid Anchoring of Archaeosortase Substrates and Midcell Growth in Haloarchaea. mBio 2020, 11, e00349-00320, doi:10.1128/mBio.00349-20.