Molecular and structural innovations of the stator motor complexes during the origin of the bacterial flagellum

Caroline Puente-Lelievre[1,2], Pietro Ridone[3], Jordan Douglas[2,4], Kaustubh Amritkar[5], Betül Kaçar[5], Matthew Baker[3], Nicholas Matzke[1,2]

[1]School of Biological Sciences, University of Auckland, New Zealand
[2]Centre for Computational Evolution, University of Auckland, New Zealan
[3]School of Biotechnology and Biomolecular Sciences, University of New South Wales, Sydney, Australia
[4]Department of Physics, University of Auckland, New Zealand
[5]Department of Bacteriology, University of Wisconsin-Madison, Madison, USA

Corresponding author: matthew.baker@unsw.edu.au

Abstract
Understanding the origin and evolution of the bacterial flagellum remains challenging questions due to its age and complexity. The MotA5B2 protein oligomer serves as the flagellar stator and converts ion motive force into flagellar rotation. Using 193 genomes sampled across bacterial diversity, we conduct a comprehensive phylogenetic analysis of MotAB and their nonflagellar homologs. We combine this with AlphaFold protein structure prediction to map the distribution and evolution of structural innovations. Our results indicate that while all proteins share a similar scaffold, the Flagellar Ion Transporters Oligomers (FITO) cluster together and are structurally conserved, while the Generic Ion Transporters Oligomers (GITO) form a separate group and are structurally and functionally diverse. We found a unique structural feature in the FITO MotAs: a transmembrane square fold. Ancestral sequence reconstruction analyses suggest that the transmembrane square fold arose with the function of flagellar motility. Finally, motility assays in *E. coli* indicate that these structural elements play an important role in flagellar motility, ~~and modeling suggests that TGI5 interacts with YcgR, a flagellar "break" protein which may assist biofilm formation~~.

Introduction

Motility is one of the most ancient features of life[1]. Bacterial flagellar motility is a complex biomolecular mechanism that dates to the Last Common Ancestor of bacteria[2], and rotation of the bacterial filament is driven by a sophisticated rotary nanomachine powered by ion motive force harnessed by stator complexes. Stators are multimeric complexes embedded within the cell wall and inner membrane. Each is made of five A subunits (motA for proton-powered flagella, PomA for sodium-powered flagella) that form a ring around two B subunits (motB and PomB, respectively). The B subunit acts as a stopper to regulate ion flux and, ultimately, to modulate the movement of the filament.

Beeby et al. (2020) proposed that the stator proteins arose by exapting pre-existing ion-channel proteins. Evidence for this proposition comes from the structurally similar MotAB homologues ExbBD from the Ton complex, and TolQR from the Tol-Pal system[3]. These proteins use proton motive force (pmf) to energise active transport across the outer membrane and maintain membrane stability and integrity.

Inferring the phylogenetic relationship between the MotAB complex and its nonflagellar homologs and characterizing the distribution structural innovations within and between these groups, should give insight into the nature of the ancestral stator, later innovations, and how they have shaped bacterial diversity and adaptation to different environments. The only previous effort at a phylogeny of this complex [ref: Livingstone Marmon (2013). Elucidating the origin of the ExbBD components of the TonB system through Bayesian inference and maximum-likelihood phylogenies. Molecular Phylogenetics and Evolution 69(3),

DOI:10.1016/j.ympev.2013.07.010] was limited to 17 MotA, 15 ExbB, 15 MotB, and 14 ExbD sequences.
Mapping innovations in protein structure onto a phylogeny can inform the evolutionary history of ancient proteins because they tend to be conserved and rarely change[4]; therefore, determining the order in which structural innovations occurred is possible. Until recently, protein 3D data was restricted to experimentally solved structures, and the structures available were predominantly from model organisms and far from representing biological diversity. The emergence of AlphaFold provides the opportunity for investigating protein diversity.

We investigated the evolution of the stator motor complex motAB and the structural innovations related to bacterial flagellar motility. We developed a pipeline to detect and extract remote homologs from bacterial genomes, build protein sequence alignments informed by protein structure and gene order, and estimate deep phylogenies using Bayesian inference (BEAST 2). We then used this phylogenetic framework to systematically characterize structural diversity across the bacterial ion transporters and identify structural traits that were unique to the flagellar stator complexes. We then experimentally confirmed that these structural elements were essential to achieving motility. Finally, we performed Ancestral Sequence Reconstruction (ASR) to determine ithe flagellar specific structural innovations arose and evolved along with motility.

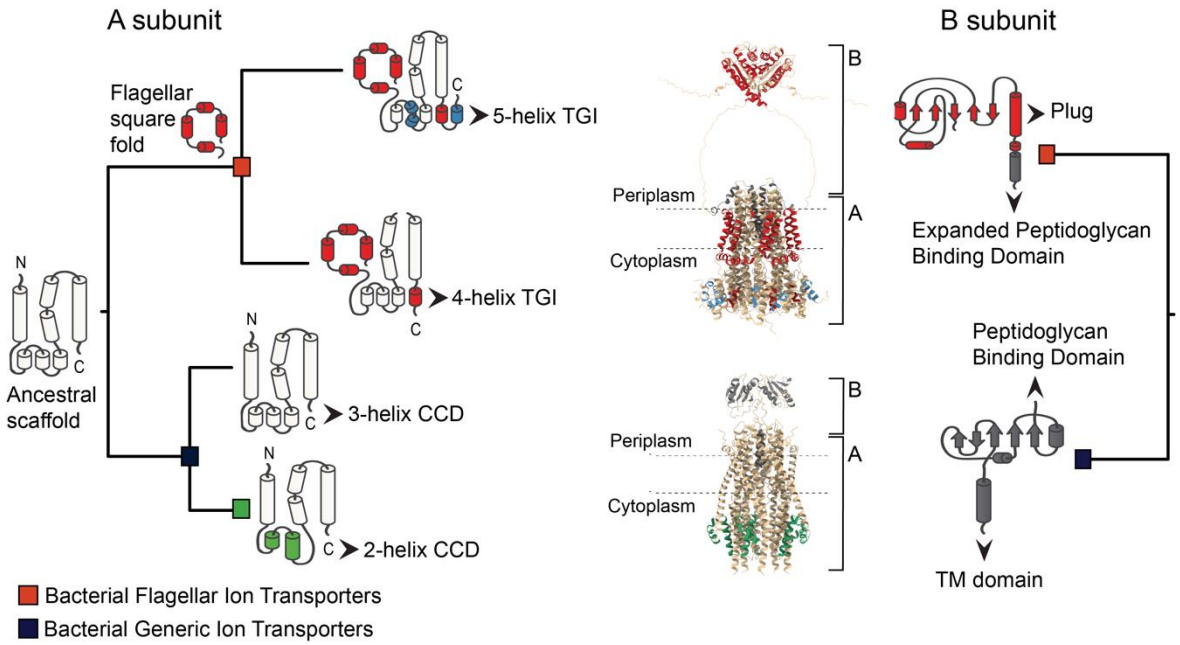

Figure 1. Phylogenetic relationships of the bacterial ion transporters and the structural innovations associated with flagellar motility.

Materials and Methods

Homology search and phylogenetic analyses
One hundred and ninety-three fully annotated and complete bacterial genomes were sampled across 27 phyla (Supplementary Information 1). When genome annotation was not available from NCBI[5], genomes were annotated using Prokka[6]. The MotA sequence dataset was assembled using HMMER 3.3.2 (Nov 2020); http://hmmer.org/ to perform homology searches with MotA from *Escherichia coli* K12 as a reference (Accession No. AAC74960.1.) Available from: https://www.ncbi.nlm.nih.gov/protein/AAC74960). Jackhammer was run for five iterations with the default search parameters. Since the A and B subunits work as an operon,

the B subunits were located in the genome immediately downstream A subunits and assembled using gene order. Protein sequences were initially aligned against a tailored HMMER profile that was generated from these datasets. Alignments were then refined with MAFFT[7]. Alignments included only the conserved transmembrane domain and the C-terminus domain. The A and B subunits were analysed separately and then concatenated in a partitioned alignment. Phylogenetic inference was performed using BEAST 2.7.6[8]. BEAST 2 analyses were conducted using an (uncalibrated) optimised relaxed clock (from the ORC 1.2.0 package[9]), the Yule skyline tree prior (BICEPS 1.1.2[10]), and the OBAMA substitution model (OBAMA 1.1.1[11]). The Yule skyline model assumes that diversification rates vary through time in a smooth piecewise fashion, providing a model-based method of root placement. The OBAMA method averages across amino acid substitution models and selected the VT[12] model with four gamma rate heterogeneity categories, with 100% posterior support in all three analyses (A, B, and AB concatenated). Markov chain Monte Carlo was run until all parameters had an effective sample size over 200, according to Tracer 1.7[13]. The CCD0 tree was used to summarise the posterior distribution of trees[14]. Tree entropy was measured to estimate phylogenetic information content in each data set[15]. Trees were visualised using Densitree[16, 17] v3.0.3 and Figtree v1.4.4 (http://tree.bio.ed.ac.uk/software/figtree/).

Ancestral Sequence Reconstruction[18]

ModelFinder[19] from IQ-Tree[20] was implemented to identify the best evolutionary model for the MotA sequence alignment (LG+F+R10). Ancestral sequence inference for the MotA phylogenetic tree was performed using Paml4.9[21] (LG+F model). Sequence gaps were reconstructed by calculating the probability of a gap for each position based on a presence-absence sequence alignment and treating all positions with a gap_probability >= 0.5 as a gap[22].

Sequence conservation analysis

Flagellar and non-flagellar subunits were distinguished based on the clade distribution in the concatenated phylogenetic tree and reference to experimentally characterized systems (**https://doi.org/10.2210/pdb6ykm/pdb**, https://doi.org/10.2210/pdb8GQY/pdb, https://doi.org/10.2210/pdb5sv0/pdb, https://doi.org/10.2210/pdb8ODT/pdb). Sequence alignments for the flagellar and non-flagellar MotA homologs were constructed using MAFFT with iteration refinement over 1000 cycles. Residues falling within a defined conservation range, as determined by sequence alignment, were allocated transparency levels using UCSF-Chimera[23].

E. coli strains, plasmids, and culture media for swim assays

Stator variants were tested in stator-deleted derivative strains of E. coli RP437 (E. coli ΔmotAB[24]. The pDB108 (pBAD33 backbone, Cm+) plasmid encoding motA and motB[25] was used as the vector to clone and express all constructs. Deletions in the plasmid-encoded motA gene were generated using inverse-deletion PCR. Primers are listed in the Supplementary Information. PCR was used to linearize the pDB108 plasmid excluding nucleotides coding for the regions to be deleted. Linear PCR products were then re-ligated using T4 ligase (M0202S, New England Biolabs) and T4 Polynucleotide Kinase (Genesearch Pty Ltd) in T4 ligase buffer for 2 hr at 37°C followed by overnight incubation at 16°C. Every ligation reaction was then transformed into NEB10β competent cells (New England Biolabs). Plasmids were extracted from successful transformants for verification. Cloned constructs were confirmed by Sanger sequencing (Ramaciotti Centre for Genomics, University of New South Wales, Kensington, Sydney, Australia). Liquid cell culturing was done using LB broth (NaCl or KCl, 0.5% yeast extract (70161, Sigma-Aldrich), and 1% Bacto tryptone). Cells were cultured in agar plates composed of LB broth and 1% Bacto agar (BD Biosciences, USA). Swim plate cultures were grown in the same substrates adjusted for agar content (0.25% Bacto agar) and NaCL (85 mM).

Structural modelling

AlphaFold predictions were generated for all the proteins included in the phylogenetic analyses and for the ancestral proteins estimated with ASR. Models of stator monomers and heptameric complexes were produced with AlphaFold2[26] supported by the Australian AlphaFold Service (https://www.biocommons.org.au/alphafold) on the public server at UseGalaxy.org.au[27],. The software was run in multimer mode using FASTA files as input[26, 28]. Models were visualized using PyMOL version 2.5.4[29]. Protein secondary structure was annotated using DSSP (https://doi.org/10.1002/bip.360221211, https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3013697/) and structural features were identified by aligning multiple structures using 3DCOMB (https://www.nature.com/articles/srep01448).

Results

Remote homology search

Jackhmmer searches recovered 746 potential remote homologues for motA. However, the final alignments included 379 sequences because proteins with sequence similarity under 10% were excluded. These sequences were challenging to align and their homology was uncertain. They were included in preliminary phylogenetic analyses, but were identified as outliers, as they resultedin low posterior clade supports values and long branches. Thus, they were excluded from the final alignments. The length of the MotA alignment was 295 residues, and 308 for MotB. Overall pairwise identity in the final A subunit sequence alignment was 17.7%, 10.6% for the B subunits, and 13.6% for the concatenated AB subunits.

Homology searches for either the complete protein sequence or for the conserved transmembrane (TM) domain were unsuccessful for the B subunits. Only proteins annotated as motB or OmpA were recovered. The well-known structural homologues ExbD and TolR[3] could not be recovered with this approach. Therefore, the gene immediately downstream of the *motA* homolog was used, after a manual check, to build the B-subunit data set. In all cases (except when the A-subunit gene was duplicated), the B-subunit was found next to the A-subunit (Supplementary Information).

Phylogenetic characterisation

Congruence and phylogenetic signal between the tree topologies produced from the A and B subunits separately were measured by estimating clade support entropy; a measure that describes the amount of uncertainty in a set of phylogenetic trees [15]. BEAST 2 analyses recovered 3449 clades from the A dataset, with an entropy of 105.8; 71518 clades from the B dataset, with an entropy of 256.7; and 1998 clades, from the concatenated AB dataset and entropy = 90.8. The number of clades recovered from the B alignments was significantly higher than the A and AB alignments (20x and 36x, respectively). Densitree visualisations confirmed that clade distribution in the B subunit tree is diffused, particularly in the deeper branches of the tree with no statistical support (Figure 1, Supplementary Information). Even though the B subunit tree contained lower phylogenetic information, concatenating the A and B subunit datasets increased the phylogenetic information content (90.8) and the posterior probability values (Figure 2, Supplementary Information).

Two major clades can be identified in the resulting phylogeny: the first clade contains the well-characterised motAB complex from *E. coli,* and the second clade includes its well-known structural homologs ExbBD and TolQR[3] from *E. coli* (Figure 1; Figure 2, Supplementary Information). Since the flagellar proteins in *E. coli* are some of the better studied and characterised, they are used here as a reference to name each clade. Accordingly, the group that contains *E. coli* motAB is called the Bacterial Flagellar Ion Transporters (FITO) clade, and the clade that harbours TolQR and ExbBD from *E. coli* is called the Bacterial Generic Ion Transporters (GITO) clade (Figure 1). Of the 379 proteins included in the phylogenies, 107 belonged to the FITO clade, and 272 to the GITO clade. Each of these clades is subdivided

into two well-supported groups. The FITO proteins were further divided into two subclades named after their diagnostic structural elements: TGI4 and TGI5 (see structural characterisation below). The TGI4 clade comprises Gram-positive and Gram-negative bacteria from the following phylla: Aquificae, Proteobacteria, Bacillota, Spirochaetes, Planctomycetota, Acidobacteria, Deferribacteres, Chloroflexi, and Nitrospirae. The TGI5 clade includes only Gram-negative bacteria, predominantly Proteobacteria, but also Planctomycetota, Verrucomicrobia, Armatimonadetes, Gemmatimonadetes, Acidobacteria, and Nitrospirae. Within the GITO proteins, one subclade contains the well-characterised structural homologues ExBD and TolQR; and the second subclade includes a diverse set of proteins of unknown function. The GITO clade comprises the highest protein diversity, with most members, including many large clades, having unknown function.

Structural characterisation

AlphaFold predictions were generated for all 379 tips of the phylogenetic trees (Supplementary Information). The diagnostic structural traits for the FITO proteins are 1) an Expanded Cytoplasmic Domain (ECD) with at least four helices, which work as a Torque Generating Interface (TGI), in the A-subunit, 2) four helices that form a four-helix structure in the shape of a square (the square fold) across the transmembrane layer in the A-subunit, 3) a Plug domain in the periplasmic layer of the B-subunit, and 4) an Expanded Peptidoglycan Binding (EPGB) domain in the periplasmic layer of the B-subunit (Figures 1 and 3). The structural difference between the two FITO subclades is the number of short helices in the cytoplasmic layer. One clade includes proteins with 4 short helices (TGI4), and the proteins in the second clade have 5 short helices (TGI5). The FITO proteins are conserved and display a limited number of possible structural combinations (squared fold + TGI4 or squared fold + TGI5).

Conversely, the GITO proteins show greater structural and functional diversity (when the function is known), notably in their C and N-terminus elements of the A-subunit. They also lack a plug domain in the B subunits (Figures 1 and 2). These proteins are characterised by having a Condensed Cytoplasmic Domain (CCD), which can have two (CCD2) or three (CCD3) short helices. Unlike the FITO proteins, they can have very large domains in the periplasmic layer composed of up to 305 residues (Protein [Internet]. Bethesda (MD): National Library of Medicine (US), National Center for Biotechnology Information; Accession No. QTD50492.1: MotA/TolQ/ExbB proton channel family protein [Sulfidibacter corallicola] [2024/05/12]. Available from: https://www.ncbi.nlm.nih.gov/protein/QTD50492) and at least thirty different configurations of the periplasmic, transmembrane and cytoplasmic layers (Supplementary Information).

Despite the structural differences between the FITO and GITO proteins, they share a common scaffold. Sequence conservation analyses indicate that the most conserved regions are the inner shell of the A subunits and the TM domain in the B subunits. This is where the interaction between the A and the B subunits during ion flux occurs (ref) (Figure 2).

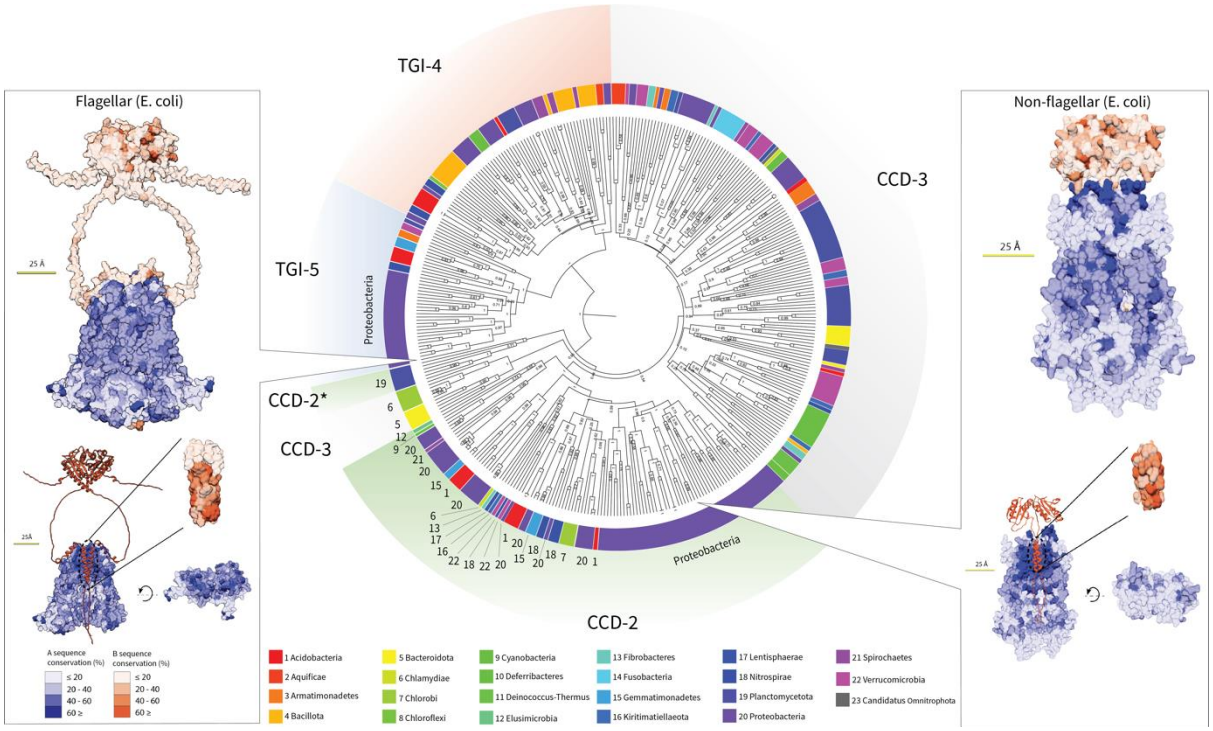

Figure 2. Taxonomic diversity mapped into the AB concatenated phylogeny and comparison of residue conservation between the flagellar and the non-flagellar A and B subunits. CCD-2* proteins form a separate clade from the rest of the CCD2 in the concatenated tree, but belong to the same clade in the A phylogeny. These proteins have a B subunit that with a unique N-terminal domain.

Two main structural groups were also identified in the B subunit phylogeny. Although these two groups share a homologous TM domain, the FITO proteins have an additional Plug + Linker domain that regulates ion flow. The degree of sequence conservation also seems to vary between MotB-like (proton powered) vs PomB (sodium powered) Plug domains. The PomB-like Plug domains (only found in the TGI4 subclade) showed significantly lower sequence conservation in comparison to the MotB-like Plug domains (found in both, the TGI4 and TGI5 subclades).

AlphaFold structural predictions of reconstructed ancestral protein sequences (Supplementary Information) suggest that the ancestor to all FITO proteins had a square fold in the cytoplasmic domain and a 3-helix CCD. Additional helices were gradually acquired and expanded the cytoplasmic domain to become TGI4 and TGI5. Conversely, the ancestor to all the GITO proteins appears to have had a linear transmembrane domain without a square fold. Unlike the FITO proteins, the CCD domain further reduced in the CCD2 lineage (eg. ExbB and TolQ). Although sequence similarity with the extant *E. coli* ExbB and TolQ is low (below 10%), they are structurally very alike. None of the structural elements that are characteristic of the extant FITO proteins was recovered in any of the nodes in the GITO clade, which suggests that these structural features have only arisen in the flagellar stator motor complexes.

Testing the function of flagellar clade-specific structural elements
Flagellar-specific structural domains are conserved xxxxxx
To investigate the function of a flagellar clade-specific conserved domain within the TGI region of MotA, eight variants of *Ec*MotA with partially deleted TGI5 domains were generated and expressed. Deletions in MotA spanned residues N103 – D124, a region of the protein encompassing the TGI5-specific structural elements as suggested by sequence alignments. Each of these variants was cloned in an arabinose-inducible, bicistronic plasmid (pDB108, pBAD33 backbone) together with WT *E. coli motB* and expressed in a *motAB* knock-out *E.*

coli strain (*E. coli* Δ*motAB*) , then tested on soft-agar swim plates for its ability to rescue motility and generate swim rings (Fig. 3). All tested MotA variants were unable to rescue motility, while clones expressing WT *motAB* displayed swim rings typical of soft-agar motility assays. These results suggest that the TGI domain defined by residues N103-D124 is essential for motility in *E. coli*.

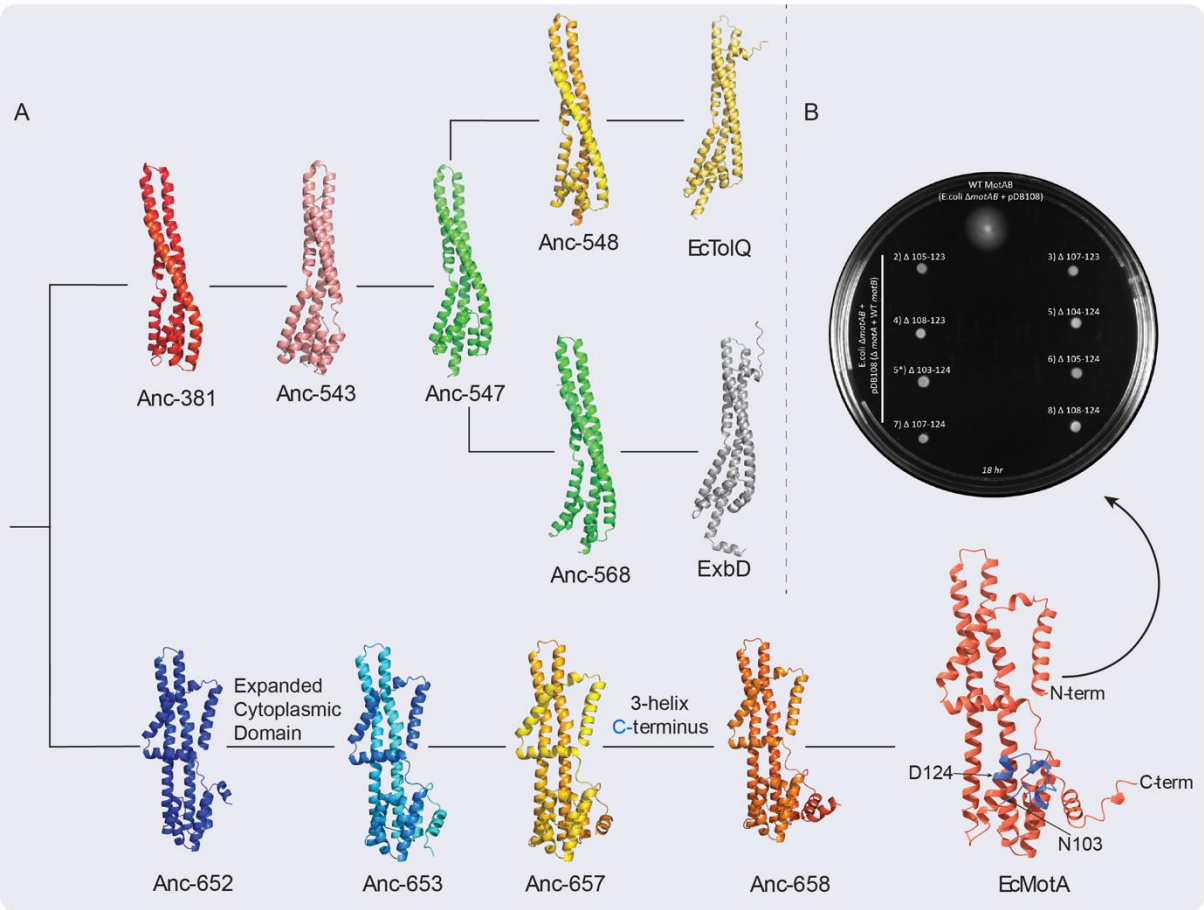

Figure 3. A) Alphafold structural predictions of the ancestral proteins in key nodes. B) Soft-agar motility assay of TGI-deleted MotA variants in *E. coli* Δ*motAB*. Left) AF2 model of *Ec*MotA. The region ranging from N103 to D124 (yellow) of the TGI domain was targeted for deletions of various length. Right) Soft-agar motility assay on 85mM Na^+ LB agar (0.25%) plate (Ø=10cm) inoculated with *E. coli* Δ*motAB* expressing WT (top, motile) or deleted variants as labelled (2-9, non-motile). The plate was incubated for 18 hr at 30°C.

Discussion

This study systematically surveyed and characterised the phylogenetic and structural diversity of the flagellar stator motor complex and their homologues across a broad range of bacterial lineages. Structural characterisation using modelling, sequence conservation analyses, and ASR suggests a conserved multimeric arrangement between the A and B subunits. The A subunit acts as a channel by which ions travel across to generate pmf while the B subunit modulates the ion flux. The interaction between the two subunits occurs in the TM domains, which is the most conserved region in both of them (Figure 2). In the A subunits, the TM domain is located in the inner shell of the ion channel, facing the B subunits. Additionally, the neighbouring position of the A and B subunits across all genomes sampled further implies that the two subunits are necessary to maintain the function of the Bacterial Ion Transporter Oligomers (BITO).

Similarly, the position across the cell membrane appears to be consistent between the FITO and GITO proteins. Three layers can be identified in all complexes: periplasmic, transmembrane (TM), and cytoplasmic layers. In the A subunit, the N-terminus is periplasmic and the C-terminus is cytoplasmic. In the B subunit, the N-terminus is in the cytoplasmic layer, and the C-terminus is in the periplasmic layer. The structural differences between the FITO and the GITO clades happen in the periplasmic and cytoplasmic layers, while the TM layer is homologous across all proteins. Therefore, the functional versatility of the BITO proteins seems to be confined to these two layers. In the case of the A subunits, they are located in the outer shell, facing the rest of the cell (periplasm, cell membrane or cytoplasm).

The structural diversity of the outer shell in the GITO proteins is highlighted by a plethora of N-terminal appendages typical of the CCD2 clade that extend into the periplasmic space. The FITO proteins, on the other hand, showcase a squared arrangement of helices at their N-terminal domain, with the N-terminus pointing towards the cytoplasm. This results in a large diversity of functions for non-flagellar homologs compared to flagellar ones, which, in contrast, do not exhibit any recognizable protein domain preceding the transmembrane, lipid facing, squared domain at their N-terminus.

Notably, the high structural diversity observed in the periplasmic layer, where additional domains of unknown function (DUF) were often identified (Supplementary Information), suggesting that the BITO complexes are involved in a broad range of biological functions, most of which remain unknown. It is likely that these additional periplasmic domains evolved independently from the conserved scaffold, then coupled to the TM domain and adjusted to achieve a specific function. Future research focusing on characterising the functions of these DUF domains could help discover new biological systems and functions, such as the recently characterised Zorya anti-phage defence system[30].

The molecular and structural innovations related to flagellar motility are also found in the periplasmic and cytoplasmic layers. In the A subunit of the FITO proteins, the ECD is the point of contact with the rotor and works as a TGI. The TGI is found in the cytoplasmic layer and can have 4 (TGI4) or 5 (TGI5) helices (Figures 1 and 4). Bacteria in the TGI5 clade are Gram-negative and predominantly Proteobacteria, including *E. coli*. Bacteria in the TGI4 clade belong to different lineages and include Gram-positive (Bacillota/Firmicutes) and Gram-negative (Figure 2). Motility assays in EcMotA showed that the lack of the TGI5 region results in the loss of motility in taxa that naturally have it. A possible explanation for this would be that the absence of TGI5 would alter the spatial configuration of the residues (i.e. R90) that are critical for the motA-FliG interaction, resulting in loss of flagellar rotation. Alternatively, the TGI5 form could be involved in a different type of regulatory mechanism of flagellar rotation. AlphaFold protein-protein simulations show a high affinity between the TGI5 region and the YcgR protein (Supplementary Information). YcgR has been reported to interact with MotA and FliG to downregulate and redirect flagellar rotation for biofilm formation in *E. coli* and related species[31].

The two well-defined subclades in the FITO clade signal that there could be two different mechanisms by which the stator interacts with the rotor, each mechanism determined by either the TGI4 or TGI5 variant. Whether TGI4 and TGI5 interact differently with FliG or regulate rotation in various ways remains to be confirmed experimentally. Future research could aim to experimentally investigate how the rotor-stator interactions take place in TGI4 and TGI5 bacteria and whether they are related to directional switching mechanisms[32]. Determining the underlying mechanisms can improve our understanding of how different flagellar systems are regulated and their potential association or impact on bacterial lifestyle.

Clade-specific differences are seen also in inner shell homologues. Strikingly, it appears that non-flagellar homologues lack a defined plug domain as observed in members of the flagellar clade and a more compact peptidoglycan-binding domain that shares little homology with the

bulkier one displayed by flagellar homologues. PGB domains in the B-subunit are so structurally divergent that an analogous origin can be hypothesised (refs).
PGB domains in the B subunit have distinct structural topologies and an undetectable level of sequence similarity, and therefore most likely have distinct origins (https://pubmed.ncbi.nlm.nih.gov/11722743/, https://pubmed.ncbi.nlm.nih.gov/17927700/).

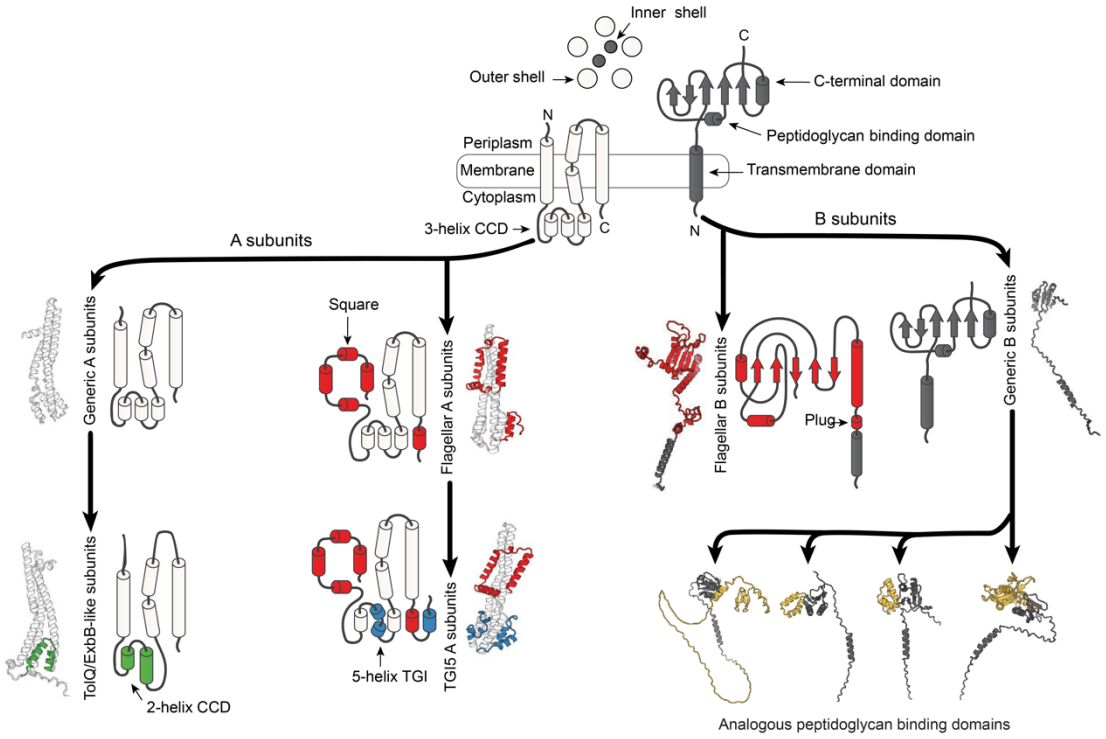

Figure 4. Simplified model of structure evolution of the A and B subunits of the multimeric ion transporters.

Given the limitations inherent to sequence-based phylogenetics of proteins in the Twilight Zone below 25% identity[33], and the difficulty in identifying even more remote outgroup homologs that also share sufficient sequence for reliable phylogenetic inference, the qutestion of what group of proteins came first, FITO or GITO, remains formally unresolved. However, there is a reasonable case that the FITO and the GITO proteins are sister groups descending from a common ancestor, as suggested by both the BEAST2 analysis which samples the position of the root in a relative-dating framework, and the fact that midpoint rooting of undated phylogenies from IQtree puts the root between the groups. This hypothesis also provides the simplest hypothesis for structural evolution, where the ancestral protein would have had a CCD3 A-subunit and generic peptidoglycan binding and transmembrane domains in the B-subunit (Figure 4). Subsequently, the gain of a square fold domain and an ECD in the A subunit, and a plug domain and an EPDB in the B subunit, would have led to the emergence of a specialised flagellar stator motor complex. An alternative is that the structural traits related to flagellar motility were the ancestral states and were lost across multiple lineages thereafter, but we provisionally favour the first hypothesis as the most parsimonious explanation.

It is possible that proteins giving further information on the ancestral protein were not included in this study because of low (under 10%) sequence similarity and inability to be aligned against the other proteins. This limitation could be overcome with a tertiary structure-based phylogenetic approach [ref: Puente-Lelievre, Caroline; Malik, Ashar J.; Douglas, Jordan; Ascher, David; Baker, Matthew; Allison, Jane; Poole, Anthony; Lundin, Daniel; Fullmer, Matthew; Bouckert, Remco; Kim, Hyunbin; Obara, Masafumi; Steinegger, Martin; Matzke, Nicholas J. (2024). Tertiary-interaction characters enable fast, model-based structural phylogenetics beyond the twilight zone. bioRxiv, 571181. https://www.biorxiv.org/content/10.1101/2023.12.12.571181v2] that would enable

a more extensive sampling and include remote homologues with minimal sequence similarity, such as the Zorya proteins[30]. Despite the limitations intrinsic to using a sequence-only phylogenetic approach, this study has produced the most comprehensive, state-of-the-art phylogeny of the flagellar stator motor complex proteins and alike. This roadmap lays the foundation for future research on flagellar motility.

Acknowledgements
CPL, PR, KA, BK, MABB and NJM are supported by Human Frontier Science Program Grant No. RGY0072/2021. NJM was additionally supported by NZ RSTA grants 21-UOA-040 & 18-UOA-034. JD is supported by the Alfred P. Sloan Foundation Matter-to-Life program Grant number G2021-16944. BK and KA thank the Center for High Throughput Computing at the UW-Madison. This work was supported by the Australian BioCommons which is enabled by NCRIS via Bioplatforms Australia funding.

Supplementary Information

1. List of genomes sampled, taxonomic data and accession numbers.
2. Gene order spreadsheet.
3. Tree files
4. Densitrees visualizations
5. Table with primers used for the motility assays
6. AlphaFold predictions and confidence scores

References

1. Miyata, M. et al. Tree of motility – A proposed history of motility systems in the tree of life. *Genes to Cells* **25**, 6-21 (2020).
2. Coleman, G.A. et al. A rooted phylogeny resolves early bacterial evolution. *Science* **372**, eabe0511 (2021).
3. Cascales, E., Lloubès, R. & Sturgis, J.N. The TolQ–TolR proteins energize TolA and share homologies with the flagellar motor proteins MotA–MotB. *Molecular Microbiology* **42**, 795-807 (2001).
4. Illergard, K., Ardell, D.H. & Elofsson, A. Structure is three to ten times more conserved than sequence--a study of structural response in protein cores. *Proteins* **77**, 499-508 (2009).
5. Sayers, E.W. et al. Database resources of the national center for biotechnology information. *Nucleic Acids Res* **50**, D20-d26 (2022).
6. Seemann, T. Prokka: rapid prokaryotic genome annotation. *Bioinformatics* **30**, 2068-2069 (2014).
7. Katoh, K., Misawa, K., Kuma, K.i. & Miyata, T. MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. *Nucleic Acids Research* **30**, 3059-3066 (2002).
8. Bouckaert, R. et al. BEAST 2.5: An advanced software platform for Bayesian evolutionary analysis. *PLOS Computational Biology* **15**, e1006650 (2019).
9. Douglas, J., Zhang, R. & Bouckaert, R. Adaptive dating and fast proposals: Revisiting the phylogenetic relaxed clock model. *PLOS Computational Biology* **17**, e1008322 (2021).
10. Bouckaert, R.R. An Efficient Coalescent Epoch Model for Bayesian Phylogenetic Inference. *Systematic Biology* **71**, 1549-1560 (2022).
11. Bouckaert, R.R. OBAMA: OBAMA for Bayesian amino-acid model averaging. *PeerJ* **8**, e9460 (2020).

12. Muller, T. & Vingron, M. Modeling Amino Acid Replacement. *Journal of Computational Biology* **7**, 761-776 (2000).
13. Rambaut, A., Drummond, A.J., Xie, D., Baele, G. & Suchard, M.A. Posterior Summarization in Bayesian Phylogenetics Using Tracer 1.7. *Systematic Biology* **67**, 901-904 (2018).
14. Berling, L. et al. A tractable tree distribution parameterized by clade probabilities and its application to Bayesian phylogenetic point estimation. *bioRxiv*, 2024.2002.2020.581316 (2024).
15. Lewis, P.O. et al. Estimating Bayesian Phylogenetic Information Content. *Systematic Biology* **65**, 1009-1023 (2016).
16. Bouckaert, R.R. & Heled, J. DensiTree 2: Seeing Trees Through the Forest. *bioRxiv*, 012401 (2014).
17. Bouckaert, R.R. DensiTree: making sense of sets of phylogenetic trees. *Bioinformatics* **26**, 1372-1373 (2010).
18. Garcia, A.K. & Kaçar, B. How to resurrect ancestral proteins as proxies for ancient biogeochemistry. *Free Radical Biology and Medicine* **140**, 260-269 (2019).
19. Kalyaanamoorthy, S., Minh, B.Q., Wong, T.K.F., von Haeseler, A. & Jermiin, L.S. ModelFinder: fast model selection for accurate phylogenetic estimates. *Nature Methods* **14**, 587-589 (2017).
20. Nguyen, L.-T., Schmidt, H.A., von Haeseler, A. & Minh, B.Q. IQ-TREE: A Fast and Effective Stochastic Algorithm for Estimating Maximum-Likelihood Phylogenies. *Molecular Biology and Evolution* **32**, 268-274 (2014).
21. Yang, Z. PAML 4: Phylogenetic Analysis by Maximum Likelihood. *Molecular Biology and Evolution* **24**, 1586-1591 (2007).
22. Aadland, K., Pugh, C. & Kolaczkowski, B. in Computational Methods in Protein Evolution. (ed. T. Sikosek) 135-170 (Springer New York, New York, NY; 2019).
23. Pettersen, E.F. et al. UCSF Chimera—A visualization system for exploratory research and analysis. *Journal of Computational Chemistry* **25**, 1605-1612 (2004).
24. Ridone, P. & Baker, M.A.B. Allosteric adaptation in the stator complex rescues bacterial motility in Exb/Mot chimeras. *bioRxiv*, 2024.2003.2012.584617 (2024).
25. Ishida, T. et al. Sodium-powered stators of the bacterial flagellar motor can generate torque in the presence of phenamil with mutations near the peptidoglycan-binding region. *Molecular Microbiology* **111**, 1689-1699 (2019).
26. Jumper, J. et al. Highly accurate protein structure prediction with AlphaFold. *Nature* **596**, 583-589 (2021).
27. Afgan, E. et al. The Galaxy platform for accessible, reproducible and collaborative biomedical analyses: 2016 update. *Nucleic Acids Research* **44**, W3-W10 (2016).
28. Evans, R. et al. Protein complex prediction with AlphaFold-Multimer. *bioRxiv*, 2021.2010.2004.463034 (2022).
29. Schrodinger, L. (2010).
30. Hu, H. et al. Structure and mechanism of Zorya anti-phage defense system. *bioRxiv*, 2023.2012.2018.572097 (2023).
31. Han, Q. et al. Flagellar brake protein YcgR interacts with motor proteins MotA and FliG to regulate the flagellar rotation speed and direction. *Frontiers in Microbiology* **14** (2023).

32. Johnson, S. et al. Structural basis of directional switching by the bacterial flagellum. *Nature Microbiology* **9**, 1282-1292 (2024).
33. Chung, S.Y. & Subbiah, S. A structural explanation for the twilight zone of protein sequence homology. *Structure* **4**, 1123-1127 (1996).