Easy come, easier go: mapping the loss of flagellar motility across the tree of life

Jamiema Sara Philip[1], Sehhaj Grewal[1], Jacob Scadden[1], Caroline Puente-Lelievre[2], Nicholas J. Matzke[2], Luke McNally[3] and Matthew AB Baker[1]

[1]School of Biotechnology and Biomolecular Science, University of New South Wales, Sydney, Australia
[2]School of Biological Sciences, University of Auckland, Auckland, New Zealand
[3]School of Biological Sciences, University of Edinburgh, Edinburgh, United Kingdom

Abstract

Most bacterial swimming is powered by the bacterial flagellar motor, a nanomachine that self-assembles from up to 45 proteins into a membrane-spanning complex. The number and types of proteins involved in the flagellar motor vary widely. Predicting flagellar motility from genomic data can facilitate large-scale genomic studies where experimental validation may not be feasible. Using sequence and structural homology, we conducted a homology searches for 54 flagellar pathway genes across 11,365 bacterial genomes. We developed and validated a classifier to predict whether a specific genome was motile and mapped the evolution of flagellar motility across the microbial tree of life. We determined that the ancestral state was motile, and the rate of loss of motility was 4 times the rate of gain.

Introduction

Flagellar motility is widespread among bacteria, with the flagellar apparatus exhibiting diverse structures and arrangements across species (1). Despite these structural differences, all flagellated bacteria share a common rotary nanomachine known as the bacterial flagellar motor (BFM), which generates thrust through coupling ion flux to torque generation (1,2). The motors exhibit significant structural diversity across species, with conserved core structures and novel divergent structures surrounding them (3–5). Additional novel components and modifications in the motor structure reveal its adaptation to diverse environments, increasing bacterial fitness (6–8).

Flagellar assembly and function are regulated by a complex network of genes that have evolved considerably across different bacterial lineages (8). While much is known about the structure and function of the BFM, little is known about the evolutionary origins of the gene clusters that define these diverse organelles. Early work identified 20 essential genes for flagellar assembly and function in *Escherichia coli* (9). Over time, research revealed the involvement of nearly 40 genes in *E. coli* and *Salmonella* Typhimurium (10), with some species like *Vibrio parahaemolyticus* possessing dual flagellar systems—each regulated by distinct gene sets comprising around 50 polar and 40 lateral flagellar genes (11). A broader comparative study of 41 flagellated bacterial species identified a core set of 24 structural genes thought to have been present in the common bacterial ancestor (8,12). However, while much research has shed light on flagellar motors in model species, no comprehensive search for flagellar genes across all domains has yet been conducted.

Traditional sequence-based approaches often struggle to capture flagellar protein-encoding genes across diverse species due to low sequence homology (8). However, protein structure searches offer greater sensitivity in detecting homology and resolving deep evolutionary

relationships (13). While structural data exists for many flagellar components, including the basal body, filament, export apparatus, and regulatory proteins, some gaps still remain particularly in understudied species with additional proteins (14). Extending comparative genomic analysis to a wide range of bacterial species by integrating both sequence and structural information for all flagellar proteins could enhance our understanding of flagellar diversity.

Here, we utilized a dataset of 11,365 complete and non-redundant bacterial genomes from the PATRIC database, ensuring representation across diverse bacterial taxa. We employed a combination of sequence and structure-based methods to identify and quantify flagellar protein-encoding genes within these genomes. Our analysis revealed a bimodal distribution in the number of flagellar genes, which allowed us to develop a classifier that can classify genomes as motile or non-motile. By integrating structural homology with sequence-based searches, we improved the detection of distant flagellar homologs, reducing false positives and uncovering previously uncharacterised proteins. We further combined these predictions with phylogenetic methods to understand the gain and loss of motility in bacterial evolutionary history.

Results

We gathered 11,365 genomes from the PATRIC database and curated a search-query set of 54 flagellar protein-encoding genes associated with key components of the flagellar motor from the flagellar assembly reference pathway [*map03040*] in KEGG (15) (Fig. 1A). This set includes genes related to the C-ring, M,S,P and L rings, rod and hook, H and T rings, filament, stator, Type III secretion system, and various regulatory genes. We then examined the presence or absence of these 54 flagellar protein-encoding genes in all genomes in our dataset.

We observed a bimodal distribution of flagellar gene counts across the genomes indicating the presence of two distinct groups within the dataset. Accordingly, we tested thresholds of 15 and 32 genes to separate these groups (Fig. 1B). Specifically, 47% of the genomes had fewer than 15 protein-encoding genes, while 49% had more than 32. This distinct separation in flagellar protein-encoding gene counts allowed us to categorize the genomes into two groups: those with low counts (non-motile) and those with high counts (motile). The remaining 4% of the genomes that fall between these two groups were classified as partially motile.

Subsequently, we examined the type of flagellar protein-encoding genes in both the motile and non-motile groups (Fig. 1C). The filament was found exclusively in the motile group (100% motile, 0% non-motile). Similar patterns were observed for the Rod and Hook (96% motile, 4% non-motile), C-ring (96% motile, 4% non-motile), MSP and L rings (94% motile, 6% non-motile), and Type-III secretion system (85% motile, 15% non-motile). For the stators, motile species still predominated (73% motile, 36% non-motile). On the other hand, the genes associated with other functional categories, such as transporters, sigma factors (59% motile, 41% non-motile), and regulators (60% motile, 40% non-motile), were more evenly distributed between the two groups. The most prevalent genes in this group included *flrA, fliI, flrC, rpoD, and fliY.*

We further investigated the distribution of flagellar protein-encoding genes across the genomes though clustering and dimensionality reduction. UMAP analysis identified six

distinct clusters with different compositions (Fig. 1D). Cluster I included 48% of the genomes and consisted entirely of non-motile genomes from phyla such as Terrabacteria, Pseudomonadota, and the FCB group. Clusters II, III, and IV represented 27.4%, 6.6%, and 15.2% of the genomes, respectively, and were all classified as motile. Notably, Cluster III was primarily composed of genomes from the Pseudomonadota phylum, while Cluster IV included genomes from the Terrabacteria group. Cluster V and VI, contained genomes classified as partially motile, primarily from phyla such as the PVC group and Pseudomonadota. The genomes classified as partially motile were nearer to the motile cluster than the non-motile cluster.

We then validated our classifier using published phenotypic data (16–18). Observations of motility for Firmicutes curated from Bergey's Manual (18) demonstrated an accuracy rate of 95%, with 159 accurate predictions and nine mismatches (Supplementary Fig. S1). We further compared 1937 species phenotypic data curated by Madin et al., 2020 (17) and correctly classified 808 species as motile and 867 as non-motile with an overall accuracy of 86.5% (Supplementary Fig. S2).

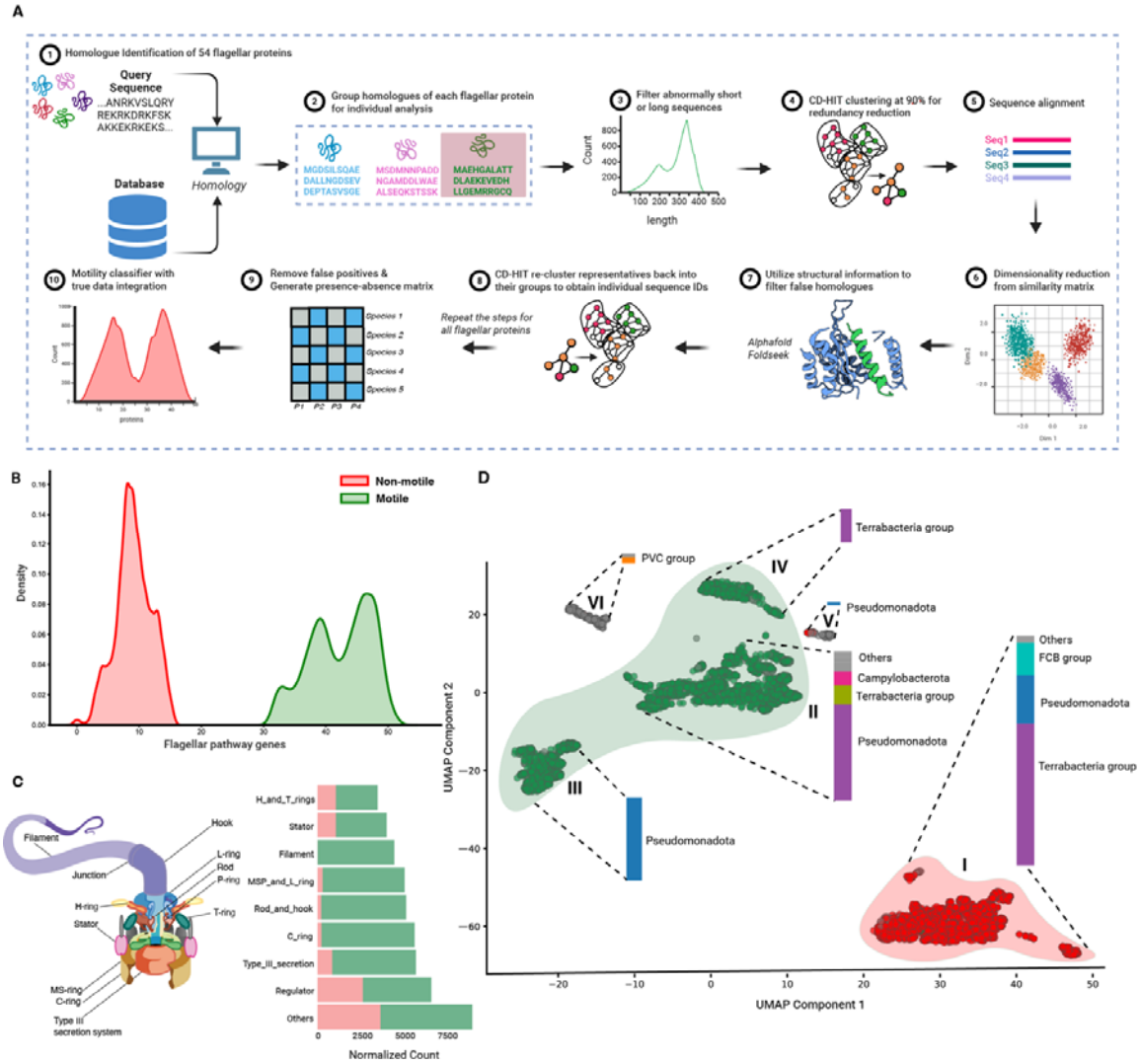

Fig. 1: Bimodal distribution of flagellar gene number across genomes. (A) Schematic overview of the flagellar protein-encoding gene homologue search in bacterial genomes. The query proteins assembled from six bacterial species were searched against the PATRIC database. The sequence homologues were further processed for downstream analysis such as length filtering and structural similarity checks to improve the confidence of the homology search. **(**B) The density plot depicts the bimodal distribution of flagellar protein-encoding across 11,365 genomes. Genomes with $\leq$ 15 flagellar pathway genes were categorised as non-motile, whereas those with $\geq$ 32 were considered motile. (C) Horizontal stacked bar plots showing the counts of flagellar pathway genes in the motile and non-motile groups are shown in Figure 1B. The bars are color-coded, with red representing non-motile and green representing motile groups. The bars correspond to various components and systems involved in the structure and function of the bacterial flagellum, illustrating the comparative abundance of these genes between the two groups. The Other category corresponds to the transporters and sigma factors involved in flagellar gene pathways. (D) UMAP visualisation of 11,365 genomes clustered based on the presence-absence matrix of flagellar proteins. Each dot represents a genome, with colors indicating motility classification from Figure 1B: green for motile, red for non-motile, and grey for genomes that do not fall into either group. Stacked bar plots placed on each cluster depict the relative abundance of major bacterial phyla within those clusters.

To investigate the evolution of flagellar motility traits we refined the reference bacterial phylogeny from iTOL (19) at the genus level and mapped the distribution of flagellar gene components (Fig. 2A, Supplementary Fig. S6). This enabled us to condense the motility data from 4,312 species into the 74 genera of the reference tree, revealing a distinct distribution of motile and non-motile traits. Out of the 74 genera, 55 were clearly classified as either motile or non-motile, while 20 genera exhibited a mixed distribution of traits. Most taxa within the phyla Pseudomonadota, Desulfobacterota, and Spirochaetota were classified as motile. In contrast, phyla such as Cyanobacteriota, Actinobacteriota, and Deinococcota were largely composed of non-motile species, reflecting ecological adaptations where motility may have played a less critical role. Interestingly, the genera within the Bacillota phylum displayed a mix of motile and non-motile traits.

Next, we employed trait analysis using BayesTraits (20) to infer the ancestral states and transition rates between motile and non-motile traits across a pruned phylogenetic tree comprising 4,125 bacterial species from GTDB matching our species list (Fig. 2B). Among the models tested, the multistate model with qMN ≠ qNM rates provided the best fit (Bayes factor (BF)= 76.4) for our data, Supplementary Table S2). The multistate model yielded a mean root probability of 0.99931 for the motile trait and 0.00069 for the non-motile trait. Across the phylogeny, transitions from motile to non-motile states (qMN) were more frequent, with a mean transition rate of 0.8058, than transitions from non-motile to motile (qNM) which had a lower transition rate of 0.21987. This indicates a tendency for motility to be lost over evolutionary time (Supplementary Fig. S3). We observed that 54% of the internal nodes exhibited high ancestral probabilities for the motile state, indicating a strong ancestral association with motility, while 45% of the nodes (predominantly Actinomycetota, Bacteroidota, Pseudomonadota, Bacillota, and Cyanobacteriota) displayed high probabilities for the non-motile state, suggesting shifts toward non-motility in these lineages.

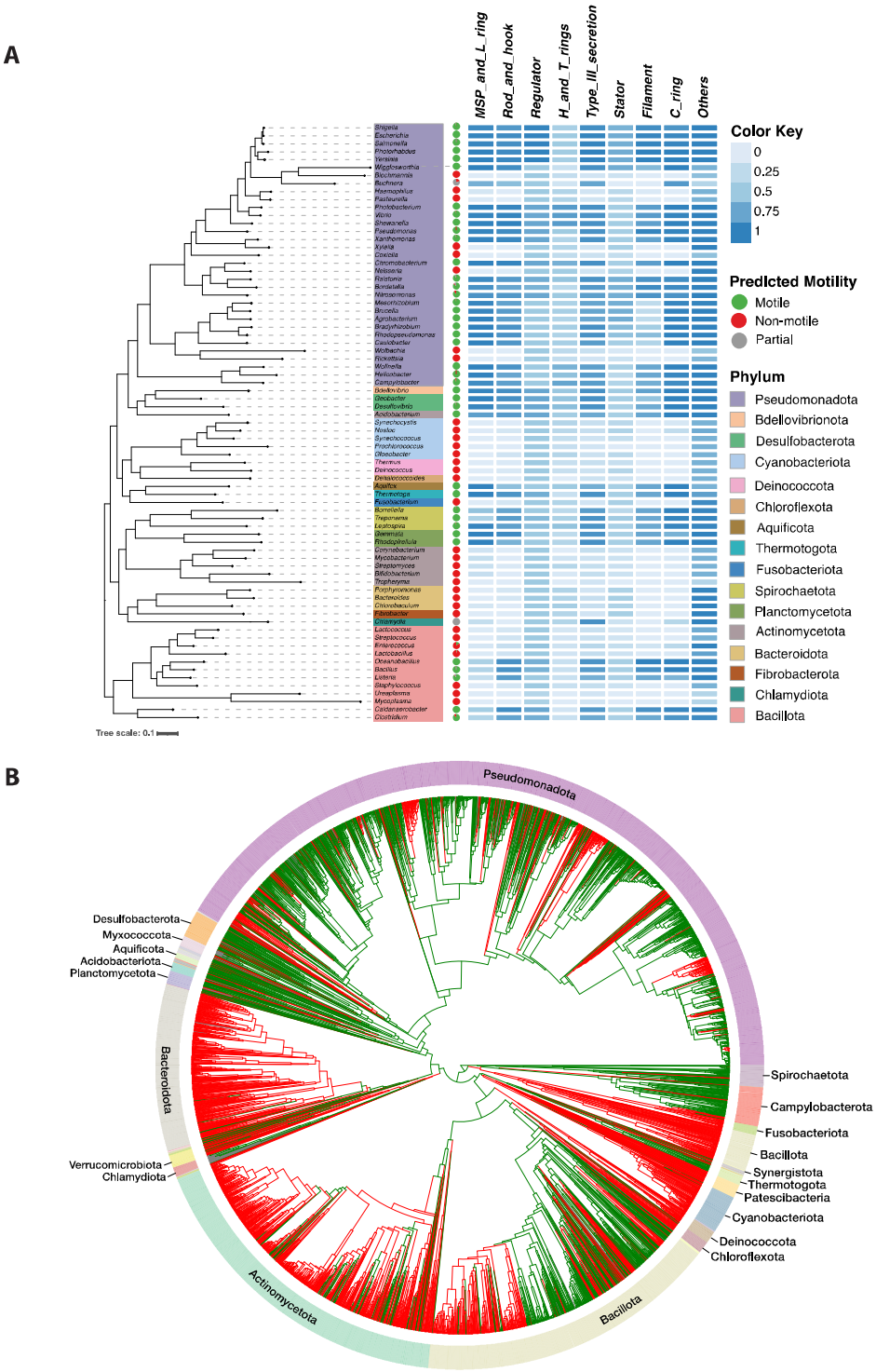

Fig. 2: Flagellar protein-encoding gene distribution across bacterial tree of life. (A) The phylogenetic tree on the left is a pruned version of the Tree of Life phylogeny from iTOL, depicts the taxonomic relationships among the genera in the study dataset. The heatmap shows the normalised abundance of flagellar pathway genes at the genus level. Presence-absence counts were aggregated by genera and normalized based on the number of species within each genus. The 54 genes were categorized into 9 categories, and further normalized by the number of genes in each substructure. Color intensity corresponds to the proportion of genes present,

highlighting their distribution and abundance across genera. The pie chart at each node shows the proportions of predicted motile, partial, and non-motile species per genus. The colours indicate the following states: red for non-motile and green for motile. (B) Bayesian reconstruction of ancestral motility traits. The phylogenetic tree depicts the relationships among 4,125 bacterial species common to our dataset and the Genome Taxonomy Database (GTDB). Tree is pruned from the Tree of Life phylogeny from iTOL. Branch colours represent the posterior probabilities ($P_{mot}(\theta|X)$) of the motility states estimated for each node using the Markov Chain Monte Carlo (MCMC) multistate model in BayesTraits. The colours indicate the following states: red for non-motile ($P_{mot}(\theta|X)>0.5$), green for motile ($P_{mot}(\theta|X)<0.5$), and grey for nodes having equal probabilities for both traits ($P_{mot}(\theta|X)=0.5$).

Discussion

A classifier for motility

In this work we demonstrated a classification system that can accurately predict whether a taxon is motile by quantifying the presence and absence of flagellar genes in a given genome. We validated this system across a large dataset of motility data using published datasets and Bergey's manual (18). Our system performs well: it is 95% accurate across *Firmicutes* (Supplementary Fig. S1) using annotations in Bergey's Manual (16,18) and displays 87 % sensitivity (true-positive rate) and 86% specificity (true-negative rate) across all prokaryotes using curated metadata from Madin et al., 2020 (17) (Supplementary Fig. S2).

Common bacterial ancestor was motile.

We confirmed, as reported elsewhere (21), that the last bacterial common ancestor (LBCA) likely contained a flagellar motor (posterior probability of 0.99). Our approach is an improvement from previous attempts due to the size of the dataset and through our use of a combination of methods for sequence homology search (Jackhmmer) for structural homology (Foldseek, AlphaFold), dimensionality reduction (UMAP, PCA) and well-curated databases (KEGG/PATRIC). We quantified gain and loss rates as well as to estimate flagellar motility across each node of the phylogeny to show that, overall, flagellar motility was only lost in a few clades (predominantly Actinomycetota, Bacteroidota, Pseudomonadota, Bacillota, and Cyanobacteriota) and never gained.

Filament genes are informative in simple motility assessment.

By far the best, and most simple determinant for flagellar motility is the presence of absence of a filament. We note that were our classification system reduced to a search only for *fliC* it would still be 99.5% accurate (Supplementary Fig. S4, Supplementary Fig. S5) . Thus a search for FliC is a good, coarse assessment of whether a species is motile or not.

Cases where there were some, but not all, of the flagellar genes are worthy of further consideration. To paraphrase Mivart 'what use is half a motor?' (22,23). Our model, and experimental experience(24,25), asserts that half a motor is not motile, and we validate this, but the question remains how did a half motor come to be? If an essential rotor gene such as FliG was lost, are the other genes only remnants of a motor long since lost? By and large these loss events were long ago. Non functional remnants of flagellar genes have been observed in several species (26,27). We observed that no species had a filament gene without any rotor genes. In contrast, we saw that 8% of species had at least one rotor gene (FliG, FliM, FliN) but no filament genes. Filament proteins like FliC make up the vast mass of the flagella and thus demand a high energy cost for synthesis (28). If a bacterium invests such energy to make a filament, it requires a motor and stator complex to make that cost worthwhile and confer a selective advantage. However, if the filament gene is lost in a loss event, then the energy cost to maintain, for example, a few rotor proteins, is comparatively low.

Horizontal gene transfer (HGT) is known to play a significant role in the acquisition of flagellar systems. For instance, *Rhodobacter sphaeroides* possess two flagellar systems, with one acquired through HGT from a γ-proteobacterium, demonstrating that entire flagellar gene sets can be transferred between species (29,30). Therefore, there is also the possibility that key flagellar genes were lost rendering a species non-motile yet maintaining the other genes within the cluster.

Exceptions to the rule

We can generate hypotheses surrounding the few strains that appear misclassified (≤13%, Supplementary Table S3). In cases such as *Bacillus selenitireducens* (31), the motility claim is not directly validated by a swim plate measurement. In the case of the *Firmicutes*, only two species, *Bacillus mycoides* and *Bacillus methanolicus*, are directly confirmed to be non-motile, with *B. mycoides* having only 20 flagellar genes (excluding M, S, P and L rings). The presence of vacuoles in *B. methanolicus* may imply a role for the vacuole in motility, such as via buoyancy which can be biased to affect diffusive behaviour. We considered experimentally testing specific strains to execute our own single-cell motility tests, however many of these strains are hard to obtain, difficult to culture, and it is uncertain whether motility would be observed using laboratory assays. Some bacteria can swim only in liquid media, whereas others may require specific environmental stimuli to express their motility genes (32). Nevertheless, we identified *Lactobacillus curvatus DSM-20019* (33), *Geobacillus thermoleovorans CCB_US3_UF5 and G. thermoleovorans ARTRW1* as the most accessible and promising strains for any future experimental optimisation of our classifier to rule out systematic errors in our approach.

Conclusions

It is known that filament expression is down-regulated in host-associated pathogens as the presence of a filament frequently activates the host immune response (34,35). We calculated the presence and absence of FliC along with motility traits and reconstructed it across our reference GTDB tree of 4125 species. We found strong evidence for a dependent model of trait evolution (Bayes factor (BF) = 1993.88) (Supplementary Table S2) indicating a close interconnection between the flagellar motility and the presence/absence of the FliC. This dependency suggests that the presence/absence of FliC strongly influences the motility traits in bacteria. Given the energetic cost in filament assembly, the benefit to host-associated pathogens to lose the filament, and the strong predictive power of filament determining motility, we hypothesise that the filament protein was lost first, then the rotor proteins, but this remains to be tested across a larger reference tree and given the loss events are in the distant past these analyses will be subject to substantial noise. It may be necessary to consider exactly when additional genes such as those present in epsilon-proteobacteria arose (36,37). Better reference phylogenies and reconciliation of gene trees combined with ecological data will help to quantify the ecological shifts in flagellar motility throughout history.

Methods

Data collection

Protein sequences from 11365 complete and non-redundant bacterial genomes were downloaded from the PATRIC database.

As a core set of flagellar genes for subsequent querying, 54 proteins from the flagellar assembly-reference pathway of the Kyoto Encyclopaedia or Genes and Genomes (KEGG) database (www.genome.jp/kegg/pathway/eco/eco02040.html) were selected. As many of these proteins appear in different combinations and are highly diverse in different species, we assembled query datasets of flagellar protein encoding genes from six model species: *Escherichia coli*, *Bacillus subtilis*, *Pseudomonas aeruginosa*, *Vibrio alginolyticus*, *Shewanella oneidensis*, and *Vibrio alginolyticus* .

In addition, flagellar protein-encoding genes such as motC, motD, and flgQ, which are found absent in the above species but present in other model organisms such as *Sinorhizobium meliloti* and *Campylobacter jejuni*, were retrieved from KEGG and added to query datasets for homolog identification.

Homologue identification

A total of 308 representative sequences of 54 key flagellar protein-encoding genes assembled from the above species were used to identify homologs in complete genomes (Supplementary Table S1). Homologues were identified using iterative searches with Jackhmmer with criteria evalue 1E–10 (38). To further remove any false positives, we then utilised additional steps, including length filtering and structural information (Fig. 1A). We defined the length range for each flagellar protein-encoding gene based on the representative query sequences and excluded sequences significantly longer than this range.

Homologous sequences for each protein were clustered at 90% identity using CD-HIT (39), and representative sequences were aligned using MAFFT (40) with default parameters. These alignments were then used to generate a distance matrix and perform PCA for dimensionality reduction, enhancing data visualization and analysis by uncovering patterns and relationships. We selected the ten most distinctive points in the PCA plot, representing the most unique sequence homologs, for structural prediction with AlphaFold (41). This approach assumes that if the most variable sequences are structural homologues, then those preceding them may also be structural homologues. Foldseek (42) was then used to compare the predicted structures to the target query structures, and a probability cut-off of 1 was further used to filter the structural homologs with high confidence.

To explore if alternative dimensionality reduction methods could improve clustering patterns, we employed UMAP along with DBSCAN (43) clustering on individual protein sequence alignments. This approach improved clustering for proteins like fliG and fliI, which previously lacked structural homologues at distant sequences (Supplementary Fig. S7). We then compared the available predicted AlphaFold structures from these clusters using Foldseek to verify if they represented true fliG/fliI clusters. Sequences identified as non-structural homologues through PCA and UMAP-DBSCAN clustering were re-evaluated in CD-HIT clusters. Sequences associated with these false positives were removed using custom

shell scripts, resulting in a refined presence-absence list of flagellar protein-encoding genes for each bacterial strain.

Motility classifier

To identify genetic signatures and patterns of flagellar motility, we used an input file detailing each bacterial species and the number of flagellar proteins obtained from sequence homologue analysis. Histogram plots were generated using the ggplot package to examine distribution patterns. The bimodal distribution revealed two distinct peaks, which were classified into motile and non-motile clusters. The number of flagellar proteins (x-axis) at which these clusters separate was noted. Additionally, UMAP visualization was employed to analyze the presence-absence matrix of flagellar proteins across the genomes to identify any clear clusters separating motile from non-motile genomes.

Bayestraits analysis for traits evolution

To overlay the predicted motility onto bacterial evolutionary history, we first downloaded reference bacterial trees from ITOL (19) and GTDB (44). We then pruned taxa that were not in our species list to construct a reference bacterial tree tailored to our data (Supplementary phylogeny S1, Supplementary phylogeny S2). We then modelled the motility states obtained for each genome to estimate transition rates with BayesTraitsV4. BayesTraits is a Bayesian framework designed for model-based analyses of evolutionary selection of phenotypic or molecular traits(20). In this approach, we analysed the trait data (predicted motility) alongside the rooted phylogenetic tree (in Nexus format) for each taxon using different models such as multistate by setting transition rates to be equal and unequal for traits.

To find the best-fitting model, we used the MCMC (Markov Chain Monte Carlo); chains ran for 5 million iterations with an additional burn-in of 300,000 iterations and a sampling interval of 1,000 iterations implemented in BayesTraitsV4 (20) to estimate the log marginal likelihood for each model. The best-fitting model was determined using Log Bayes Factor (BF) to find the root probability and transition rates of different states. Ancestral character states of all nodes were estimated to identify the most likely state at each internal node in the phylogenetic tree. Histograms were generated for the distribution of transition rates using ggplot2, and posterior probabilities for each state of select nodes were tabulated using the BioGeoBEARS package in R (45). We then tested the correlation of the filament gene (fliC) and motility using Bayestraits. Discrete independent and dependent models were tested with the 4 states, such as 00 : non-motile-no fliC, 01 : non-motile fliC, 10 : motile - no fliC, 11 : motile-fliC. We then calculated Bayes Factors of models for each trait-fliC relationship to test whether the independent or dependent model of evolution was supported.

Data availability

All the data used is presented as tables in the Supplementary Information

Code availability

The code used is available at https://github.com/BakerLabBABS/classifier

References

1. Nakamura S, Minamino T. Flagella-Driven Motility of Bacteria. Biomolecules. 2019 Jul 14;9(7):279.

2. Sowa Y, Berry RM. Bacterial flagellar motor. Q Rev Biophys. 2008 May;41(2):103–32.

3. Minamino T, Imada K. The bacterial flagellar motor and its structural diversity. Trends Microbiol. 2015 May 1;23(5):267–74.

4. Terashima H, Kawamoto A, Morimoto YV, Imada K, Minamino T. Structural differences in the bacterial flagellar motor among bacterial species. Biophys Physicobiology. 2017 Dec 19;14:191–8.

5. Beeby M, Ribardo DA, Brennan CA, Ruby EG, Jensen GJ, Hendrixson DR. Diverse high-torque bacterial flagellar motors assemble wider stator rings using a conserved protein scaffold. Proc Natl Acad Sci U S A. 2016 Mar 29;113(13):E1917–26.

6. Chen S, Beeby M, Murphy GE, Leadbetter JR, Hendrixson DR, Briegel A, et al. Structural diversity of bacterial flagellar motors. EMBO J. 2011 Jun 14;30(14):2972–81.

7. Carroll BL, Liu J. Structural Conservation and Adaptation of the Bacterial Flagella Motor. Biomolecules. 2020 Oct 29;10(11):1492.

8. Liu R, Ochman H. Stepwise formation of the bacterial flagellar system. Proc Natl Acad Sci. 2007 Apr 24;104(17):7116–21.

9. Hilmen M, Silverman M, Simon M. The regulation of flagellar formation and function. J Supramol Struct. 1974;2(2–4):360–71.

10. Iino T, Komeda Y, Kutsukake K, Macnab RM, Matsumura P, Parkinson JS, et al. New unified nomenclature for the flagellar genes of Escherichia coli and Salmonella typhimurium. Microbiol Rev. 1988 Dec;52(4):533–5.

11. McCarter LL. Dual Flagellar Systems Enable Motility under Different Circumstances. Microb Physiol. 2004;7(1–2):18–29.

12. Coleman GA, Davín AA, Mahendrarajah TA, Szánthó LL, Spang A, Hugenholtz P, et al. A rooted phylogeny resolves early bacterial evolution. Science. 2021 May 7;372(6542):eabe0511.

13. Jambrich MA, Tusnady GE, Dobson L. How AlphaFold2 shaped the structural coverage of the human transmembrane proteome. Sci Rep. 2023 Nov 20;13(1):20283.

14. Zhu S, Sun X, Li Y, Feng X, Gao B. The common origin and degenerative evolution of flagella in Actinobacteria. mBio. 14(6):e02526-23.

15. Kanehisa M, Furumichi M, Sato Y, Kawashima M, Ishiguro-Watanabe M. KEGG for taxonomy-based analysis of pathways and genomes. Nucleic Acids Res. 2023 Jan 6;51(D1):D587–92.

16. El Baidouri F, Venditti C, Humphries S. Independent evolution of shape and motility allows evolutionary flexibility in Firmicutes bacteria. Nat Ecol Evol. 2016 Nov 21;1(1):1–6.

17. Madin JS, Nielsen DA, Brbic M, Corkrey R, Danko D, Edwards K, et al. A synthesis of bacterial and archaeal phenotypic trait data. Sci Data. 2020 Jun 5;7(1):170.

18. Whitman WB, Rainey F, Kämpfer P, Trujillo M, Chun J, DeVos P, et al. Bergey's manual of systematics of archaea and bacteria. Vol. 410. Wiley Online Library; 2015.

19. Letunic I, Bork P. Interactive Tree Of Life (iTOL) v5: an online tool for phylogenetic tree display and annotation. Nucleic Acids Res. 2021 Jul 2;49(W1):W293–6.

20. Pagel M, Meade A. Bayesian Analysis of Correlated Evolution of Discrete Characters by Reversible Jump Markov Chain Monte Carlo. Am Nat. 2006 Jun;167(6):808–25.

21. Coleman GA, Davín AA, Mahendrarajah TA, Szánthó LL, Spang A, Hugenholtz P, et al. A rooted phylogeny resolves early bacterial evolution. Science. 2021 May 7;372(6542):eabe0511.

22. Dial KP, Randall RJ, Dial TR. What Use Is Half a Wing in the Ecology and Evolution of Birds? BioScience. 2006 May 1;56(5):437–45.

23. Mivart G. On the Genesis of Species. BoD – Books on Demand; 2023. 358 p.

24. Islam MI, Ridone P, Lin A, Michie KA, Matzke NJ, Hochberg G, et al. Ancestral reconstruction of the MotA stator subunit reveals that conserved residues far from the pore are required to drive flagellar motility. microLife. 2023 Jan 1;4:uqad011.

25. Islam MI, Lin A, Lai YW, Matzke NJ, Baker MAB. Ancestral Sequence Reconstructions of MotB Are Proton-Motile and Require MotA for Motility. Front Microbiol. 2020 Dec 23;11:625837.

26. Pallen MJ, Matzke NJ. From The Origin of Species to the origin of bacterial flagella. Nat Rev Microbiol. 2006 Oct;4(10):784–90.

27. Ren CP, Beatson SA, Parkhill J, Pallen MJ. The Flag-2 Locus, an Ancestral Gene Cluster, Is Potentially Associated with a Novel Flagellar System from Escherichia coli. J Bacteriol. 2005 Feb;187(4):1430.

28. Schavemaker PE, Lynch M. Flagellar energy costs across the tree of life. eLife. 2022 Jul 26;11:e77266.

29. Poggio S, Abreu-Goodger C, Fabela S, Osorio A, Dreyfus G, Vinuesa P, et al. A complete set of flagellar genes acquired by horizontal transfer coexists with the endogenous flagellar system in Rhodobacter sphaeroides. J Bacteriol. 2007 Apr;189(8):3208–16.

30. Ferreira JL, Coleman I, Addison ML, Zachs T, Quigley BL, Wuichet K, et al. The "Jack-of-all-Trades" Flagellum From Salmonella and E. coli Was Horizontally Acquired From an Ancestral β-Proteobacterium. Front Microbiol. 2021 Mar 30;12:643180.

31. Switzer Blum J, Burns Bindi A, Buzzelli J, Stolz JF, Oremland RS. Bacillus arsenicoselenatis, sp. nov., and Bacillus selenitireducens, sp. nov.: two haloalkaliphiles from Mono Lake, California that respire oxyanions of selenium and arsenic. Arch Microbiol. 1998 Dec;171(1):19–30.

32. Ikeda T, Shinagawa T, Ito T, Ohno Y, Kubo A, Nishi J, et al. Hypoosmotic stress induces flagellar biosynthesis and swimming motility in Escherichia albertii. Commun Biol. 2020 Feb 28;3(1):1–7.

33. Parte AC, Sardà Carbasse J, Meier-Kolthoff JP, Reimer LC, Göker M. List of Prokaryotic names with Standing in Nomenclature (LPSN) moves to the DSMZ. Int J Syst Evol Microbiol. 2020 Nov 1;70(11):5607–12.

34. Wolfgang MC, Jyot J, Goodman AL, Ramphal R, Lory S. Pseudomonas aeruginosa regulates flagellin expression as part of a global response to airway fluid from cystic fibrosis patients. Proc Natl Acad Sci. 2004 Apr 27;101(17):6664–8.

35. Hajam IA, Dar PA, Shahnawaz I, Jaume JC, Lee JH. Bacterial flagellin—a potent immunomodulatory agent. Exp Mol Med. 2017 Sep;49(9):e373–e373.

36. Chaban B, Coleman I, Beeby M. Evolution of higher torque in Campylobacter-type bacterial flagellar motors. Sci Rep. 2018 Jan 8;8(1):97.

37. Gao B, Lara-Tejero M, Lefebre M, Goodman AL, Galán JE. Novel Components of the Flagellar System in Epsilonproteobacteria. mBio. 2014 Jun 24;5(3):10.1128/mbio.01349-14.

38. Finn RD, Clements J, Eddy SR. HMMER web server: interactive sequence similarity searching. Nucleic Acids Res. 2011 Jul 1;39(suppl_2):W29–37.

39. Fu L, Niu B, Zhu Z, Wu S, Li W. CD-HIT: accelerated for clustering the next-generation sequencing data. Bioinforma Oxf Engl. 2012 Dec 1;28(23):3150–2.

40. Katoh K, Standley DM. MAFFT Multiple Sequence Alignment Software Version 7: Improvements in Performance and Usability. Mol Biol Evol. 2013 Apr 1;30(4):772–80.

41. Jumper J, Evans R, Pritzel A, Green T, Figurnov M, Ronneberger O, et al. Highly accurate protein structure prediction with AlphaFold. Nature. 2021 Aug;596(7873):583–9.

42. van Kempen M, Kim SS, Tumescheit C, Mirdita M, Lee J, Gilchrist CLM, et al. Fast and accurate protein structure search with Foldseek. Nat Biotechnol. 2024 Feb;42(2):243–6.

43. Rehman SU, Asghar S, Fong S, Sarasvady S. DBSCAN: Past, present and future. Fifth Int Conf Appl Digit Inf Web Technol ICADIWT 2014. 2014 Feb;232–8.

44. Parks DH, Chuvochina M, Rinke C, Mussig AJ, Chaumeil PA, Hugenholtz P. GTDB: an ongoing census of bacterial and archaeal diversity through a phylogenetically consistent, rank normalized and complete genome-based taxonomy. Nucleic Acids Res. 2022 Jan 7;50(D1):D785–94.

45. Matzke NJ. BioGeoBEARS: BioGeography with Bayesian (and likelihood) evolutionary analysis in R Scripts. R Package Version 02. 2013;1:2013.

Author information

Authors and Affiliations

1. School of Biotechnology and Biomolecular Science, UNSW Sydney, Kensington 2052, Australia

   Jamiema Sara Philip, Sehhaj Grewal, Jacob Scadden & Matthew Arthur Barrington Baker
2. School of Biological Sciences, University of Auckland, Auckland, New Zealand

   Caroline Puente-Lelievre and Nicholas Matzke
3. School of Biological Sciences, University of Edinburgh, Edinburgh, United Kingdom

   Luke McNally

Contributions

JSP and MABB designed the analysis. JSP wrote original code, executed all final analysis including search, identification, classification and trait analysis. SG performed initial survey of genes in PATRIC database. JS examined phenotypes and provided validation of classifier. CPL assisted with structural homology. NJM and LM assisted with trait analysis and tree visualisation. MAAB. supervised the design, execution, and writing of the project. All authors contributed to writing and revision of the manuscript.

Corresponding author

Correspondence to Matthew AB Baker

matthew.baker@unsw.edu.au

Ethics declarations

Competing interests

The authors declare no competing interests.

Acknowledgements

Funding: MABB is supported by US Navy Office of Naval Research, Research grant no. N62909-22-1-2051. MABB, LM and NJM were supported by HFSP Project Grant RGY0072/21. MABB and NJM are supported by Australian Research Project Discovery Project DP240100462.

Supplementary Information

Supplementary Fig. S1. Confusion Matrix of Motility Classifications. Matrix comparing the motility classifications of 169 species from our classifier dataset with those from the Firmicutes motility dataset by Fouad El Baidouri et al. (2016).

Supplementary Fig. S2. Confusion Matrix of Motility Classifications. Matrix comparing the motility classifications of 1937 species in our classifier dataset with those from the curated dataset by Madin et al. (2020).

Supplementary Fig. S3. Transition rates of motility. Histograms showing the posterior distribution of transition rate estimates between motile (qMN) and non-motile (qNM). The red dotted line indicates the mean, and the blue line indicates the 95% CI.

Supplementary Fig. S4. Confusion Matrix of Filament Classifications. Matrix comparing the classification system based on filament genes and the whole gene motility classifier for 10960 species present in our dataset.

Supplementary Fig. S5. Confusion Matrix of fliC Gene Classification. Matrix comparing the classification system based on the fliC gene and the whole gene motility classifier for 10960 species present in our dataset.

Supplementary Fig. S6. Heatmap showing the abundance of different filament (a) and stator (b) proteins at the genus level. Presence-absence counts were aggregated by genera and normalised based on the number of species within each genus.

Supplementary Fig. S7. UMAP visualisation of DBSCAN clusters for FliI (A) and FliG (B) sequence alignments. Pairwise distance matrices from multiple sequence alignments were reduced to low-dimensional embeddings using UMAP and clustered with DBSCAN. Cluster annotations were assigned based on information retrieved from UniProt using ID mapping between PATRIC genome IDs and UniProt IDs. Each point represents a protein sequence, with colours indicating different clusters.

Supplementary Table S1: The complete list of species and flagellar pathway genes, together with the total number of genes present.

Supplementary Table S2: Log Marginal Likelihood (log ML) estimates using the BayesTraits Multistate algorithms for equal and unequal rates for qMN and qNM with MCMC and a stepping stone sampler for motility traits, as well as dependent and independent models for correlated motility and fliC presence.

Supplementary Table S3: The list of species with mismatched motility classification as compared with the phenotypic data curated from published literature [Fouad El Baidouri et al. (2016); Madin et al. (2020)].

Supplementary Phylogeny S1: Phylogenetic tree of 75 species pruned from reference ITOL tree in Newick format.

Supplementary Phylogeny S2: Phylogenetic tree of 4125 species pruned from reference GTDB tree in Nexus format.