nature microbiology

Supplementary information

https://doi.org/10.1038/s41564-026-02273-y

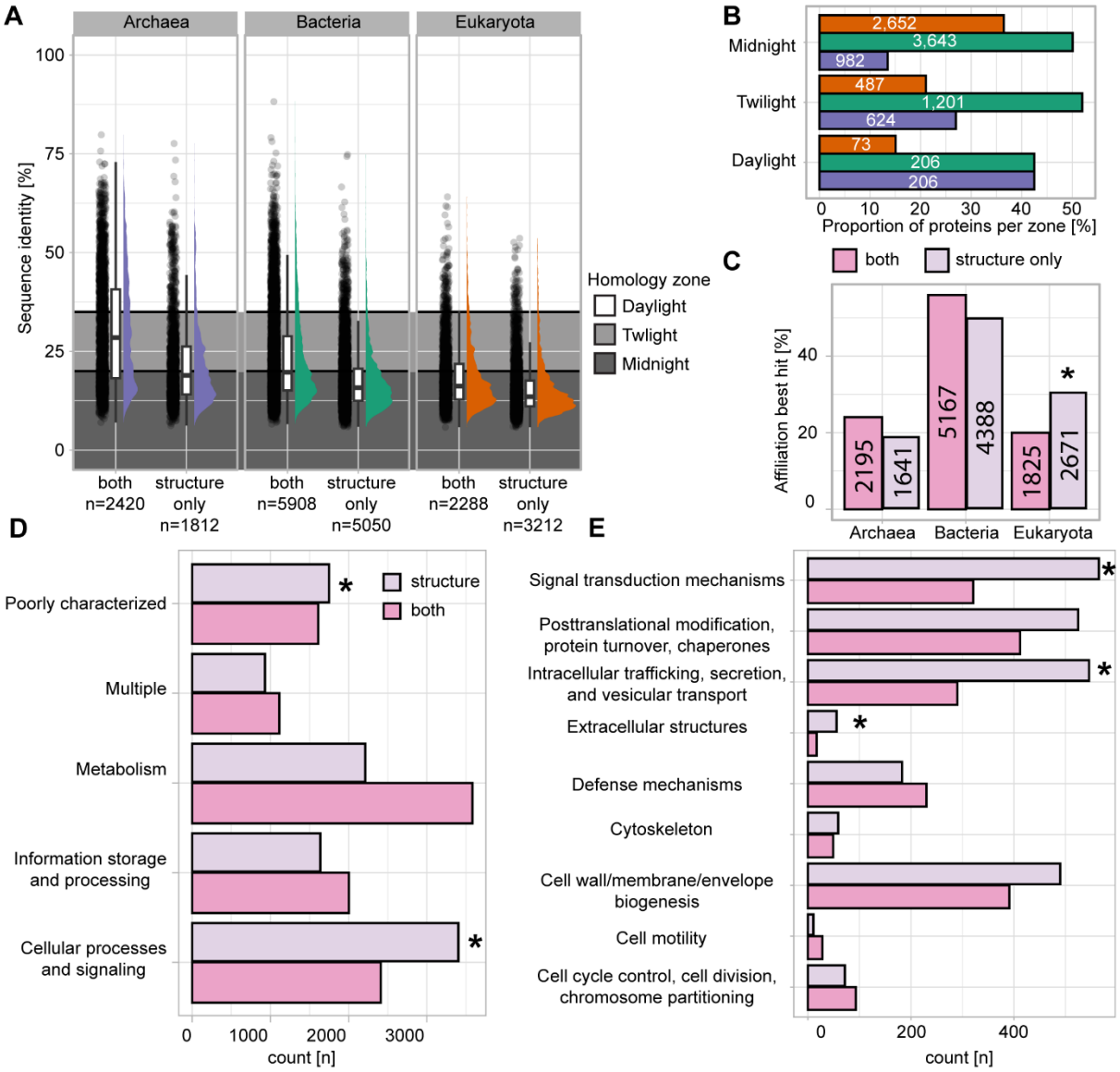
Prediction of eukaryotic cellular complexity in Asgard archaea using structural modelling

In the format provided by the authors and unedited

Comparison of sequence and structure-based annotations

We could annotate 8,891 proteins with both sequence and structure, finding agreement of the COG assignments for 96% of representative proteins and their respective best structural hits in SwissProt. Of note, almost half of the protein representatives with both a highly confident (sequence-based) COG and structural hit displayed less than 20% sequence identity to their best structure hit (n=4,263; median of 18.6%; interquartile range (IQR) =14.2-28.0%), falling below the 'twilight zone' of sequence identity (Supplementary Fig. 1A, B). This zone has been proposed for the range between 20–35% sequence identity where homology becomes challenging to predict with regular algorithms [22]. This demonstrates the high sensitivity of MSA-based searches. We found that all protein representatives that could only be annotated with MSA-based searches did have a structure hit in UniProt50 but belonged to protein families not annotated in SwissProt.

Subsequently, we tested whether differences in taxonomic assignment existed for proteins that could only be annotated based on structure, as these likely exhibit stronger sequence divergence. The 8,418 proteins exclusively annotated using structures showed significantly lower sequence identities to their best structure hits (medians=15.1% vs 18.6%; IQR=12.0-20.2% vs 14.2-28.0%, Wilcoxon signed-rank test p-value: $5 \cdot E^{-16}$; Extended Data Fig. 2). Interestingly, their annotated domains were enriched in best hits against eukaryotic protein structures (Supplementary Fig. 1C). This could indicate that, for Asgard archaeal proteins, their eukaryotic homologs diverged more extensively compared to their prokaryotic homologs. We further observed that these proteins are enriched in functional categories related to cellular processes and signaling (Bonferroni-corrected one-tailed Fisher's exact test p-value: $9.3 \cdot E^{-77}$), and more specifically in "intracellular trafficking, secretion, vesicular transport", "signal transduction", and "extracellular structures" (Bonferroni-corrected one-tailed Fisher's exact test p-values: $5 \cdot E^{-6}$, $3 \cdot E^{-4}$, $6.3 \cdot E^{-4}$, respectively) (Supplementary Fig. 1D,E).



Suppl. Fig. 1. Comparative assessment of Asgard archaeal structural clusters. **(A)** Sequence identity of representative Asgard archaeal proteins to SwissProt best-hit protein, segregated by the domain of life. Box plots: centre line = median; box bounds = 25th and 75th percentiles; whiskers = 1.5× IQR; outliers represented by points outside of whisker range. **(B)** Proportion of proteins per domain (colors follow panel A) per sequence homology zone. Absolute number of proteins indicated as white text within the respective bars. (**C**) Bar plot depicting the proportion of the domain-rank affiliation of best SwissProt structure hit based on annotation with both structural and sequence assignment (pink), or just structure (violet). The structure-only annotation is significantly enriched in eukaryotic best hits as indicated by the asterisk (one-tailed Fisher's exact test p-value: $8.8 \cdot E^{-22}$). **(D)** Bar plot indicating the number of hits per functional category of the best Swiss-Prot hit per Asgard archaeal representative protein. Asterisks indicate a functional enrichment based on a one-tailed Fisher's exact test with Bonferroni correction of "Cellular processes and signaling" and "Poorly characterized" with $9.3 \cdot E^{-77}$ and $1.4 \cdot E^{-7}$, respectively. **(E)** Bar plot indicating the number of hits per functional subcategory within "Cellular processes and signaling". Asterisks indicate a functional enrichment based on a one-tailed Fisher's exact test

with Bonferroni correction of “Intracellular trafficking, secretion, vesicular transport”, “signal transduction”, and “extracellular structures” with the respective adjusted p-values: $5 \cdot 10^{-6}$, $3 \cdot 10^{-4}$, $6.3 \cdot 10^{-4}$.

Supplementary Data

Data S1. (separate file)
Spread sheet with genome information of the outgroup genomes used for Fig. S1A and the dataset of 936 Asgard archaeal draft genomes.

Data S2. (separate file)
Spread sheet of UniProt protein IDs and annotations of sampled *Candidatus* Prometheoarchaeum syntrophicum.

Data S3. (separate file)
Spread sheet including the annotation of structures in the of ESPs and iESP structural clusters, and ESP and iESP proteins in *Candidatus* Prometheoarchaeum syntrophicum.