An origin for a eukaryotic lipid transfer protein fold in Asgard archaea

Nicolas-Frédéric Lipp[1] and Itay Budin[1*]

[1] Department of Chemistry and Biochemistry, University of California San Diego, La Jolla, CA 92093, USA

NFL ORCID: 0000-0001-7963-7415
IB ORCID: 0000-0001-9706-4294

*Corresponding author: ibudin@ucsd.edu

Abstract

The evolution of eukaryotic cells necessitated the advent of machinery to transport molecular building blocks for organelles to proliferate. Extant eukaryotes share several classes of highly conserved lipid transfer proteins (LTP) that associate with donor membranes, bind individual lipid molecules, and shuttle them to acceptor membranes. Because cells lacking organelles do not require extensive lipid transport networks, it is not known if this machinery pre-dated eukaryotic organelles or had to evolve alongside them. Here we describe a class of putative lipid transporters in the Asgard archaeota superphylum that share a common ancestry with eukaryotic LTPs in the START domain superfamily. We identified three classes of Asgard START proteins, StarAsg1-3, which are conserved across most Asgard phyla. Of these, StarAsg1 family proteins contain the predicted structural features necessary for lipid transfer: large, hydrophobic binding pockets lined with amphipathic motifs for membrane docking. In contrast, StarAsg2 and StarAsg3 family proteins contain smaller binding cavities and minimal predicted membrane interactions. We found that StarAsg1 from Lokiarcheota interacts with anionic membranes both in vitro and in yeast cells and can transfer sterols between liposomes. Phylogenetic analysis of START domains across the tree of life indicates that eukaryotic LTPs share a common ancestry with StarAsg1 homologs, while StarAsg2 and StarAsg3 form a monophyletic group with eukaryotic heat shock protein cochaperones. We propose that the emergence of proteins for inter-membrane lipid transporters in the ancestors of eukaryotic cells could have facilitated the subsequent development of intracellular organelles.

Introduction

The architecture of eukaryotic cells relies on their capacity to synthesize and transport insoluble lipid molecules as structural components of membranes. Most organelles are bound by membranes that are composed of distinct lipid components, which tune their physicochemical properties and interactions with protein machinery. However, lipid metabolism itself is not replicated across organelles, with the endoplasmic reticulum (ER) being the primary production site for most major lipid classes. Thus, cells rely on post-metabolic transport processes to bring lipid molecules – which by their hydrophobic nature cannot diffuse through the cytosol – from the ER to all other organelles. Lipid transfer between membranes is much faster than protein trafficking[1,2], and this robust transport is essential due the high lipid fluxes needed for organelles to grow and replicate. In extant eukaryotes, these processes also act to generate compositional heterogeneities between organelles that are specialized for their function[3–7].

Lipid transport was originally thought to occur via vesicular trafficking, but in recent years it has been discovered that many distinct lipid transfer proteins instead accomplish this task. Some LTPs are channel-shaped proteins that form bridges across contact sites, where organelles sit within 10s of nm from each other[8,9]. However, most LTPs are small, soluble proteins that transiently interact with membranes and extract individual lipid molecules and transport them between membranes[10,11]. These shuttle-like LTPs may contain domains that tether them to membranes or contact site components, yet they still rely on soluble domains with amphipathic membrane-binding sites, regulated by dynamic bound-to-soluble switching mechanisms that support second-to-minute-scale lipid transport rates[12,13]. The exact roles and specificities of individual LTPs within cells are often changing to parse[14–17], but as a whole their capacity to shuttle lipids between opposing membranes is required for the existence of organelles that define eukaryotic cells. Despite their importance to the emergence of eukaryotic cells, our understanding of the evolution of LTPs remains limited. Lipid transfer mechanisms have been identified across the periplasmic space of gram-negative bacteria involving LolA/B and MlaC homologs[18], but no direct eukaryotic counterparts to these systems have been identified. Currently, the origin of any eukaryotic LTP class remains unknown.

Eukaryogenesis occurred via endosymbiotic event(s) between archaea and bacteria 1.5 and 2 billion years ago[19–22]. Until recently, it was accepted that eukaryotes form a sister group of archaea, with mitochondria descended from an engulfed proto-symbiont sister to alphaproteobacteria[23–25]. The discovery of Asgardarchaeota (hereafter called Asgards) as an archaeal superphylum that forms a monophyletic group with Eukarya changed our understanding of eukaryogenesis[26] and recent analyses suggest eukaryotes belong to Asgards *sensu stricto*, arising from the order Hodarchaeales[27,28]. Such a relationship is supported by the presence of Eukaryotic Signature Proteins (ESPs) in Asgards, including homologs of key components of eukaryotic membrane machinery: components of the Endosomal Sorting Complexes Required for Transport (ESCRT) system[29,30], GTPases, Roadblock/longin families of protein, and TRAnsport Protein Particle (TRAPP) complex[31,32]. Cultured Asgard, like the Lokiarchaea *Promethearchaeum syntrophicum*[33,34], also show complex membrane topologies including tubular protrusions and surface bound vesicles. These features, yet to be molecularly defined, suggest that the evolution of complex membrane biology seen in eukaryotes might have roots predating eukaryogenesis.

We hypothesized that lipid transport represented an inherent challenge during eukaryogenesis: organelle-containing cells absolutely require it, yet it is unclear what could have driven the evolution of specific LTPs in the absence of their organellar substrates. We therefore asked if any eukaryotic LTPs could have been

present in Asgard ancestors, thus facilitating the subsequent evolution of organelles, analogous to other recently identified ESPs. We identified a set of conserved three Asgard proteins in the START domain family, which also include several classes of extant eukaryotic LTPs. One of these families is specialized for interaction with anionic membranes and extraction of lipid ligands, suggesting a possible last ancestor for many eukaryotic LTPs.

Results

Identification of putative membrane binding START domain in Asgardarchaeota

We first examined the distribution of 14 annotated classes of shuttle-like LTPs in bacteria, archaea, viruses and eukaryotes (**Fig. 1**). Each of these LTP classes, often defined by a family or a superfamily of proteins, contains a globular domain hosting one or two lipidic ligands in a buried binding cavity. Nine of them are exclusive to eukaryotes with very rare occurrences in other groups, suggesting that the diversity of LTPs expanded alongside eukaryogenesis. In contrast, proteins containing START and Sterol Carrier Protein 2 (SCP-2) domains are distributed among eukaryotes and archaea, but also found in bacteria. From this observation, we identified potential LTPs coding sequences in the genome of the isolated Lokiarchaeota *Prometheoarchaeum syntrophicum* MK-D1. MK-D1 codes for one SCP-2 homolog and three START-like domains, which we named StarLok1, StarLok2 and StarLok3. We focused further attention on these candidates because the structural features and ability to bind hydrophobic ligands of START domains are better defined than for SCP-2[35,36].

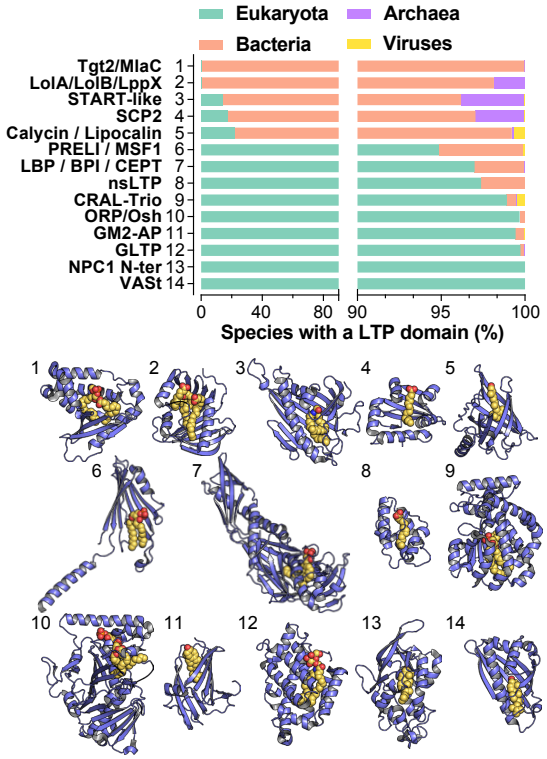


Fig. 1: Distribution and structural diversity of shuttle-like LTPs across biological systems. *Top panel:* The proportion of taxonomic species expressing each shuttle-like LTP class. For each class, the InterPro accession number is given in the method section. Most known LTP classes are exclusive to eukaryotes; archaea contain examples of three (2-4). *Bottom panel:* Structures of example proteins from each shuttle-like LTP class show in (B) containing a bound lipid ligand (alkanes, yellow; oxygen, red; nitrogen, blue; phosphate, orange). 1. MlaC (lipid: phosphatidylethanolamine, pdb: 7VR6). 2. Lppx (α-linolenic acid and docosahexaenoic acid, 2BY0), 3. START domain of CERT (ceramide, 2E3P), 4. SCP-2 (palmitate, 4JGX), 5. RBPA (retinol, 1RBP), 6. Ups1 (phosphatidic acid, 4YTX), 7. BPI (phosphatidylcholine (PC), 1BP1), 8. nsLTP (palmitoleate, 1FK3), 9. Sec14 (PC, 3B7Z), 10. Osh4 (phosphatidylinositol 4-phosphate, 3SPW), 11. NPC2 (ergosterol, 6R4N), 12. GLTP domain of 4-phosphate adaptor protein 2 (FAPP2) (*N*-oleoyl-galactosylceramide, 5KDI), 13. N-terminal domain of NPC1 (cholesterol, 3GKI), 14. Lam2 (ergosterol, 6CAY).

A combination of structural prediction and phylogenetics allowed us to explore the hypothesis that Asgard START-domains are capable of transporting lipids from membranes. Most of the 54 Asgard START-domain proteins in the AlphaFold 2 database displayed the specific helix-grip fold, with a seven-strand core sheet that is consistent with >390 solved structures in this protein superfamily. To characterize the

phylogeny of START domains within Asgards, we first analyzed the relationships of 55 homologous sequences of this class across Asgard phyla. The resulting tree indicates that StarLok1 is part of a monophyletic clade comprising 21 sequences, with strong branch support (aLRT = 0.8939) **(Fig. 2A)**. Proteins of this clade ('StarAsg1') featured predicted binding cavities large enough (MK-D1 1595 Å^3; Heimdall AB_125, 1579 Å^3) to accommodate a wide range of lipids, including sterols (~ 600 Å^3) and phospholipids (> 1000 Å^3) **(Fig. 2B)**. In contrast, the clades containing StarLok2 (StarAsg2) and StarLok3 (StarAsg3) had smaller binding cavities, with StarAsg3 homologues possessing residual pockets under 300 Å^3. StarLok2 and StarLok3 proteins, but not StarAsg1, were identified as homologs of the activators of the heat shock protein 90 (Hsp90) ATPase (Aha1) by Protein Natural Language Model (ProtNLM). Thus, Asgards contain three classes of START domain proteins, represented by orthologues of StarAsg1 and StarAsg2, and StarAsg3.

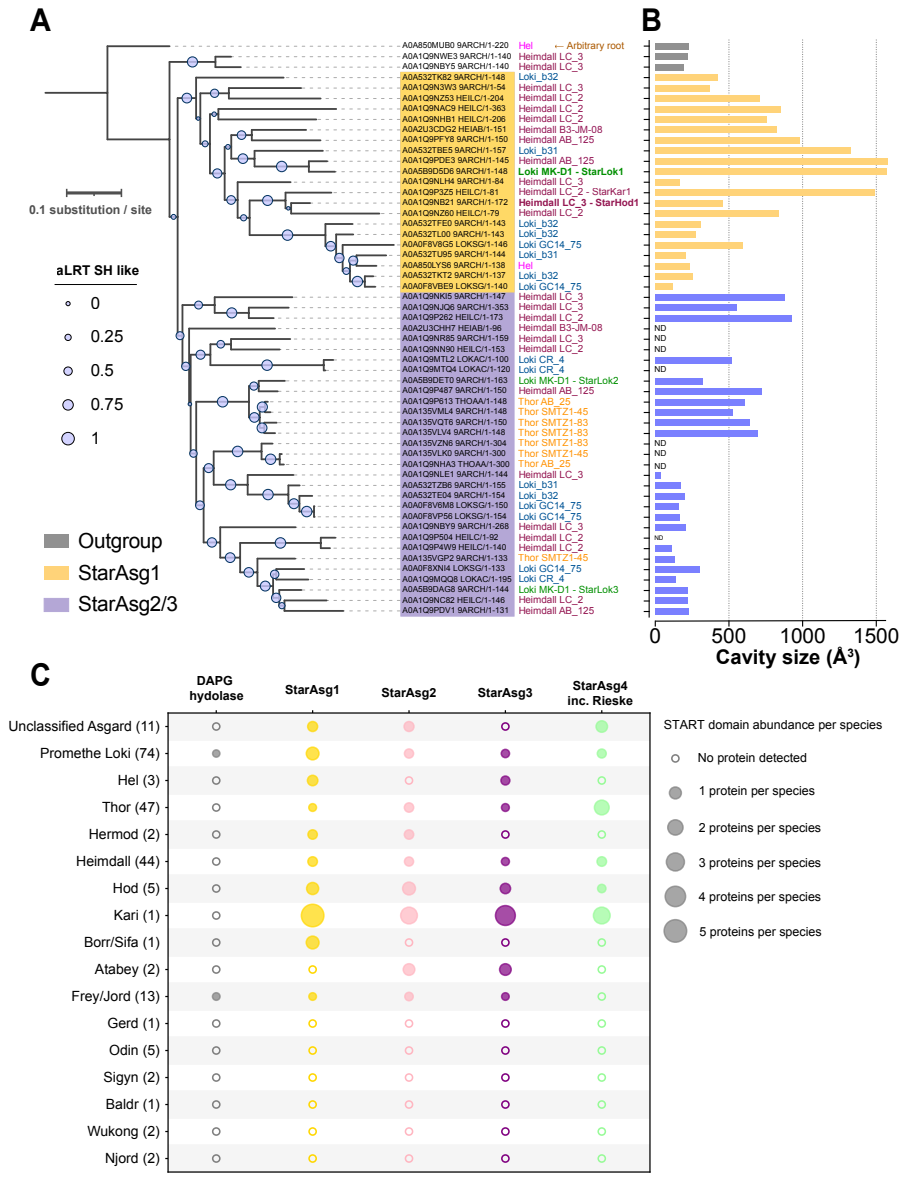


Fig. 2: Distinct clades of START domain proteins across Asgards. A. Maximum likelihood phylogenetic tree (PhyML Q.Pfam + R4 + F) of the START proteins identified in Asgards. Branch support is indicated by aLRT SH-like values, with the circle size corresponding to the support value. The tree is arbitrarily rooted with a single sequence identified in the Helarchaeota (strain CR_Bin_291). Species names are color-coded according to their taxonomic groups: Helarchaeota (magenta), Heimdallarchaeota (wine), Thorarchaeota (orange), Lokiarchaeota (blue), except for MK-D1 (green). Sequence accessions and lengths corresponding to the monophyletic clade of StarAsg1 are highlighted in yellow, while the monophyletic clade of StarAsg2 and StarAsg3 is highlighted in purple. **B.** Comparative analysis of cavity sizes in StarAsg1 and its homologs in Asgards. The Uniprot accession numbers are listed along the y-axis, with the corresponding cavity sizes represented by horizontal bars. Yellow bars indicate proteins belonging to the StarAsg1 clade, and purple bars indicate proteins from the StarAsg2 and StarAsg3 clades. **C.** Abundance of the START domain classes in different Asgard phyla. The dot size is proportional to the number of identified START domains per species; an empty circle indicates the absence of START domains in that phylum. The average number of determined species in the dataset is given next to each phyla's name in parenthesis.

In a holistic analysis of Asgard proteomes, we identified 411 START domain sequences distributed across 203 current species. We did not observe any START homolog in several Asgard phyla, including Odin and Wukong. However, the three START domains clades (StarAsg1-3) are well represented in Asgardarchaea that are considered closest to the last common ancestors of eukaryotes: Kari, Hod and Heimdall phyla all display the four classes of START domain, including StarAsg1 **(Fig. 2C)**. We also identified three occurrences of DAPG hydrolase (PFAM 18089) as well as twenty-nine Rieske [2Fe-2S] type-protein diverging from a group of START domains, which we named StarAsg4 **(Fig. 2C** and **Supplementary Fig. 1).** Overall, this analysis suggests a clear distinction between StarAsg1 (value of bootstrap and aLRT of 83.3 and 83, respectively), StarAsg3 (73, 81) and the phylum including StarAsg4 and StarAsg3 (86.1, 71) (**Supplementary Fig. 1**).

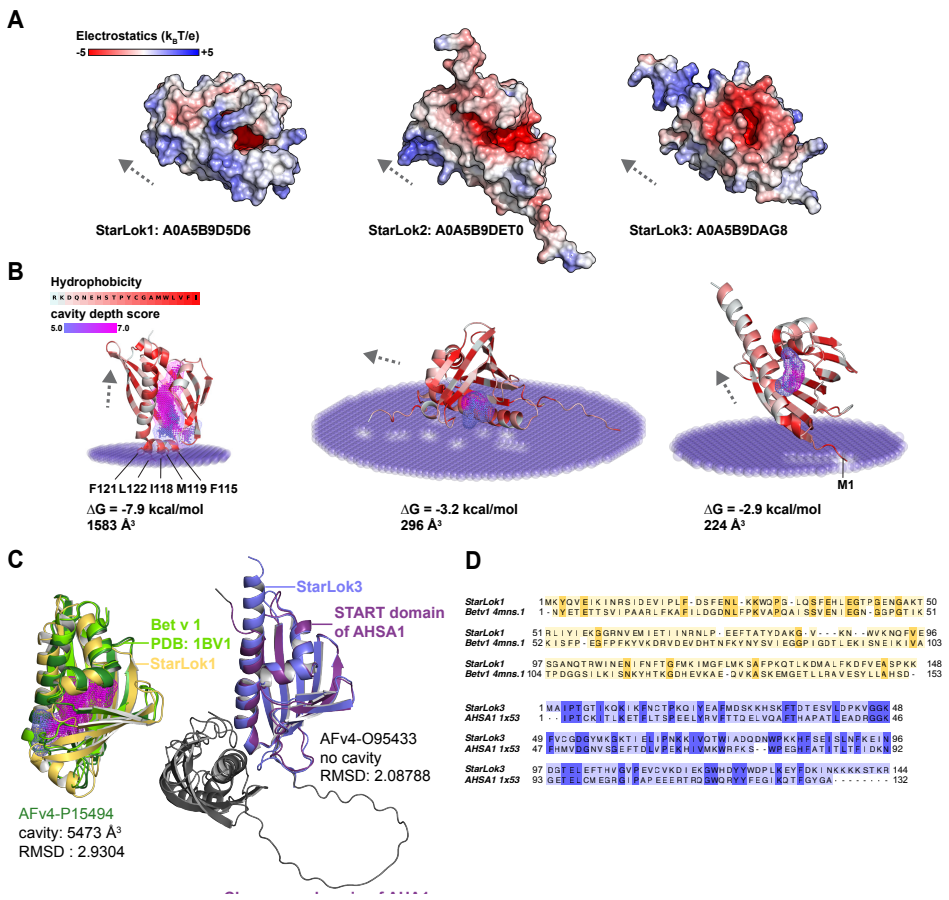


Fig. 3: StarAsg proteins from *Promethearcheum syntrophicum* have divergent structural features. A. Structural prediction of StarLok1, StarLok2 and StarLok3, represented as electrostatic surface potential ranging from -5 to +5 k_BT/e and shown as a color gradient from red (negative) to white (neutral) to blue (positive). The three proteins were aligned and oriented to show the cavity entrance of StarLok1. Arrows indicate the orientation of the main α-helix of each START domain, from N- to C-terminus. **B.** PPM prediction of StarLok1, 2 and 3 (from left to right) docked to a model archaeal lipid membrane. The structure is colored according to the hydrophobicity of each residue from pale cyan (low hydrophobicity) to red (high hydrophobicity). Computed cavities are visualized as a 3D point cloud, color-coded by LigSite depth scores, from periwinkle blue (lower scores) to magenta (higher scores). Protein-membrane interaction energies, computed cavity sizes are provided for each protein. Key hydrophobic residues of StarLok1 embedded in the lipid monolayer are indicated with connecting lines: phenylalanine (F), leucine (L), isoleucine (I) and methionine (M). **C.** *Left panel:* Structural similarities between the AF structure of StarLok1 (yellow) and that of Bet v 1 (dark green). The crystal structure of Bet v 1 (pdb: 1BV1, light green) was also aligned. The cavity of Bet v 1 is displayed as in panel A using LigSite depth scoreq. *Right panel:* Structural similarities between the Alphafold structure of StarLok3 (blue) and that of the START domain of Aha1 (Purple). The co-chaperone binding domain of Aha1 is indicated in black. RMSD of AFv4-P15494 and AFv4-O95433 aligned to StarLok1 and StarLok3, respectively, are indicated. As well, the cavity of AFv4-P15494 is reported. **D.** SWISS-MODEL hit template alignment of StarLok1 and Bet v 1 (yellow) and StarLok3 and Aha1 (blue) sequences.

Using computational approaches, we further characterized the predicted structural features in the three Asgard START domains from the Lokiarchaeon *P. syntrophicum* MK-D1. We observed a charge distribution of StarLok1 that is strongly polarized, with one side featuring a large cationic patch surrounding the mouth of the protein cavity **(Fig. 3A).** This configuration suggested that StarLok1 may interact with membranes, as peripheral membrane proteins are often electrostatically polarized to interact with

phospholipids[37–40]**.** A network of hydrophobic residues (F20, L26, F23, W29, F77, V90 W104, F110, F138), two opposite tyrosines (Y54, Y81), and two distant anionic residues (D21, E66), characterize the cavity of StarLok1, supporting the binding of hydrophobic ligand. Notably, the cavity lacks key polar residues generally required for enzymatic activity, such as ketone cyclization[41]. The Positioning of Protein in Membrane (PPM) algorithm suggested that StarLok1 could dock to membranes through the insertion of five hydrophobic residues, reminiscent of the binding mode of the sterol transporter STARD4[42]. **(Fig. 3B).** In contrast, StarLok2 and StarLok3 have lower electrostatic polarization, residual cavities and did not show a clear mode of membrane interaction by PPM. Overall, structure-sequence analyses StarLok1 suggest closest homology to the plant allergen Bet v 1, while StarLok3 resembled the START domain of Aha1 (**Fig. 3C**).

<u>StarLok1 is specialized for interactions with anionic membranes</u>

To biochemically characterize Asgard START proteins, we expressed and purified recombinant StarLok1, StarLok2 and StarLok3 from *E. coli*. StarLok1 was fused to a C-terminal 6xHis tag that to prevent interference with membrane binding region or β-sheet stability (**Supplementary Fig. 2**), while StarLok2 and 3 have naturally extended disordered N-termini that were suitable for tag placement. Folding of the StarLok1 was confirmed by Circular Dichroism (CD) spectroscopy, which showed a secondary structure profile consistent with the AlphaFold structures (**Supplementary Fig. 2G**) and previous CD spectra measured for similar proteins[43,44]. StarLok2 and StarLok3 were readily purified as soluble proteins, but StarLok1 readily aggregated when concentrated, likely due to its extreme polarity in both hydrophobicity and electrostatics. Extensive screening identified specific detergents or buffers that could support higher StarLok1 concentrations or refolding from inclusion bodies (**Supplementary Fig. 3**). In general, detergent-free StarLok1 aggregates unless kept at acidic pH, which promotes increased cationic electrostatic repulsion along the amphipathic protein face. However, StarLok1 was additionally stabilized by acetate, as it still showed aggregation in citrate buffer at the same pH (**Supplementary Fig. 3E**). We also sought to express and purify StarHod1, a close ortholog of StarLok1 present in Hodarchaeale **(Fig. 2A)**. StarHod1 has similar predicted structural feature **(Fig. 2A, Supplementary Fig. 4A)** and can be stabilized at low pH in the presence of acetate **(Supplementary Fig. 4B)**. Asgards, including Lokiarcheatoa, features an acetogenic metabolism, in which carbon dioxide is fixed into acetyl-CoA via the Wood-Ljungdahl carbon fixation pathway[45]. Such organisms then generate ATP from acetyl-CoA, producing acetate as a byproduct[46]. Thus, it is possible that intracellular acetate controlled by central metabolism might alter oligomerization or other interactions of StarLok1 in Asgard cells.

We next assayed the interaction of each Lokiarcheatoa START protein with lipid membranes. Using a dilute solution of each protein, we measured changes in light scattering upon incubation with liposomes containing the neutral lipid di-oleoyl-phosphatidylcholine (DOPC) and the anionic lipid phosphatidylserine (PS). Addition of StarLok1 triggered clustering of liposomes containing 5% or higher PS concentrations, which were minimized in the presence of neutral DOPC liposomes (**Fig. 4A**). In contrast, StarLok2 and StarLok3 showed minimal interactions with PS-containing liposomes (**Fig. 4B**). We could also detect gravity separation of StarLok1 in solutions with PS-containing liposomes, but not DOPC liposomes **(Supplementary Fig. 5G**). StarLok1 also showed features consistent with its hypothesized hydrophobic nature, including aggregation in the absence of liposomes (**Supplementary Fig. 3**), high levels of staining with hydrophobic dyes on SDS-PAGE (**Supplementary Fig. 5C**), and accumulation at air-water interfaces in density-based separation experiments **(Supplementary Fig. 5G**).

We also asked if StarAsg1 proteins can interact with cell membranes *in vivo*. For this, we expressed StarLok1, Lok2, Lok3 and Hod1 fused to a C-terminal GFP in *Saccharomyces cerevisiae*. Visualization of StarLok1-GFP logarithmic phase yeast cells distributions revealed a strong plasma membrane (PM) association of with additional enrichment at regions surrounding the bud neck of cells. More specifically, StarHod1 localized exclusively in the PM in a polarized manner (**Fig. 4B, Supplementary Fig. 6**). This distribution echoed previously observed for a PS-binding sensor, the C2 domain of lactadherin ($C2_{LACT}$) in yeast cells[47]. Expression the mRFP-C2Lact senor in StarLok1-GFP expressing cells confirmed colocalization of StarLok1 with membranes enriched in PS. These experiments suggest that StarAsg1 proteins have specialized for binding to anionic membranes in cells. In contrast, StarLok2 and StarLok3 remained soluble in the yeast cytosol **(Fig. 4C)**

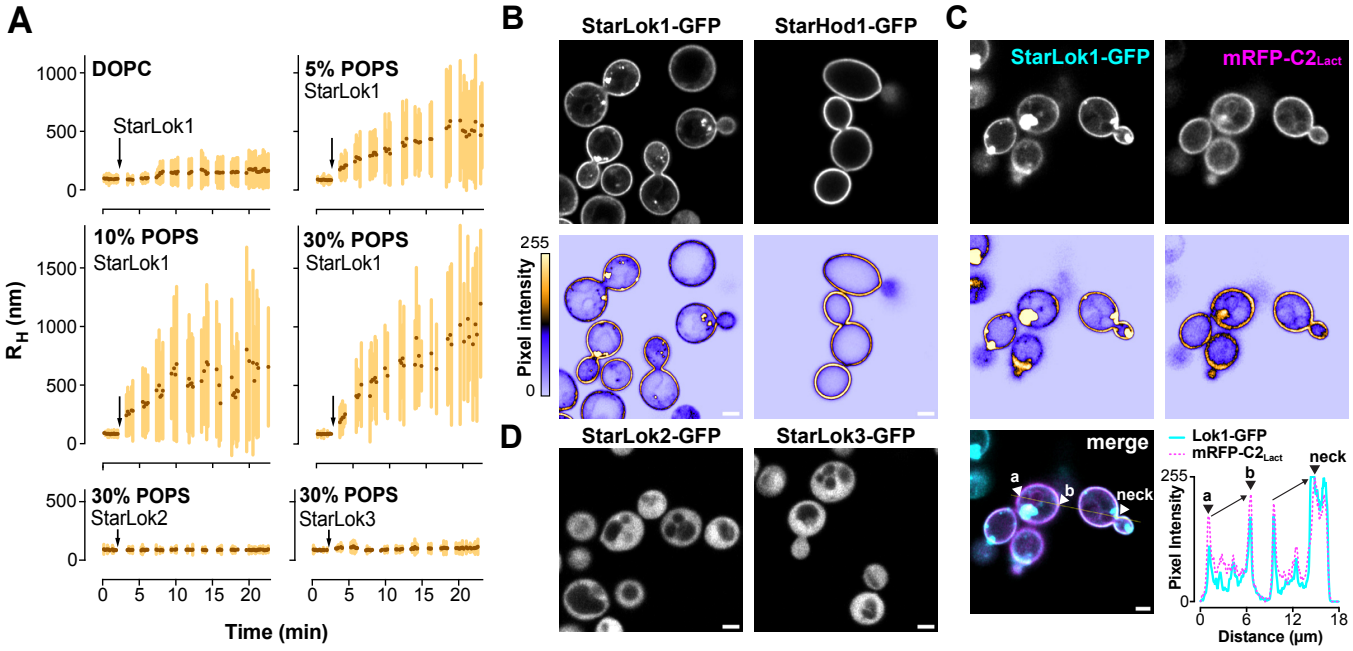


Fig. 4: StarAsg1 proteins are specialized for interaction with anionic membranes. A. DLS of StarAsg1 with liposomes of different lipid compositions. Light scattering was recorded for 2 minutes after the addition of 50 μM liposomes composed of pure DOPC, or DOPC doped with 5 mol% POPS, 10 mol% POPS, or 30 mol% POPS. At t = 2 min, 1 μM StarAsg1 was added to the mixture, and measurements were acquired for an additional 20 minutes. Increasing POPS causes protein-mediated aggregation of the liposomes. Addition of StarLok2 and StarLok3 addition to DOPC liposomes with 30 mol% POPC does not induce aggregation. **B.** StarLok1-GFP and StarHod1-GFP expressed in budding yeast binds the PM. The lower panels are intensity heat maps of the corresponding confocal image to highlight the polarized distribution of StarAsg1 in membranes surrounding the yeast bud neck and budding daughter cells. C. StarLok1 localization and polarization correlates with PS abundance in the OM. PS is detected with the protein sensor mRFP-$C2_{LACT}$ as previously descibed[47]. Corresponding intensity heatmaps is reported as in B. The lower left composite image of StarLok1-GFP (cyan) and mRFP-$C2_{LACT}$ (magenta) was used to determine colocalization shown on the left curve. An increase of signal from "a" to "b" across the PM is indicated by an arrow. Scale bar, 2 μm.

<u>StarLok1 transfers sterols between liposomes</u>

Given its similarity to Bet v 1, a known steroid binder[48], we investigated the potential for StarAsg1 proteins to bind and transport hydrophobic ligands in its cavity. Asgards, including Lokarchaea, have been recently shown to contain terpene cyclases[49], which are related to cyclases used to generate sterol backbones in

all eukaryotes. We thus tested the ability of Asgard START domains to transport sterols as a model substrate. The movement of the fluorescent phytosterol dehydroergosterol (DHE) across liposomes can be quantified using Förster resonance energy transfer (FRET) with the membrane-anchored fluorophore dansyl-phosphatidylethanolamine (dansyl-PE) (**Fig. 5A**). Transport of DHE out of dansyl-PE-containing liposomes is quantified by its dequenching, an assay that has been used to measure activity of several sterol LTPs, including other START-domain proteins[50].

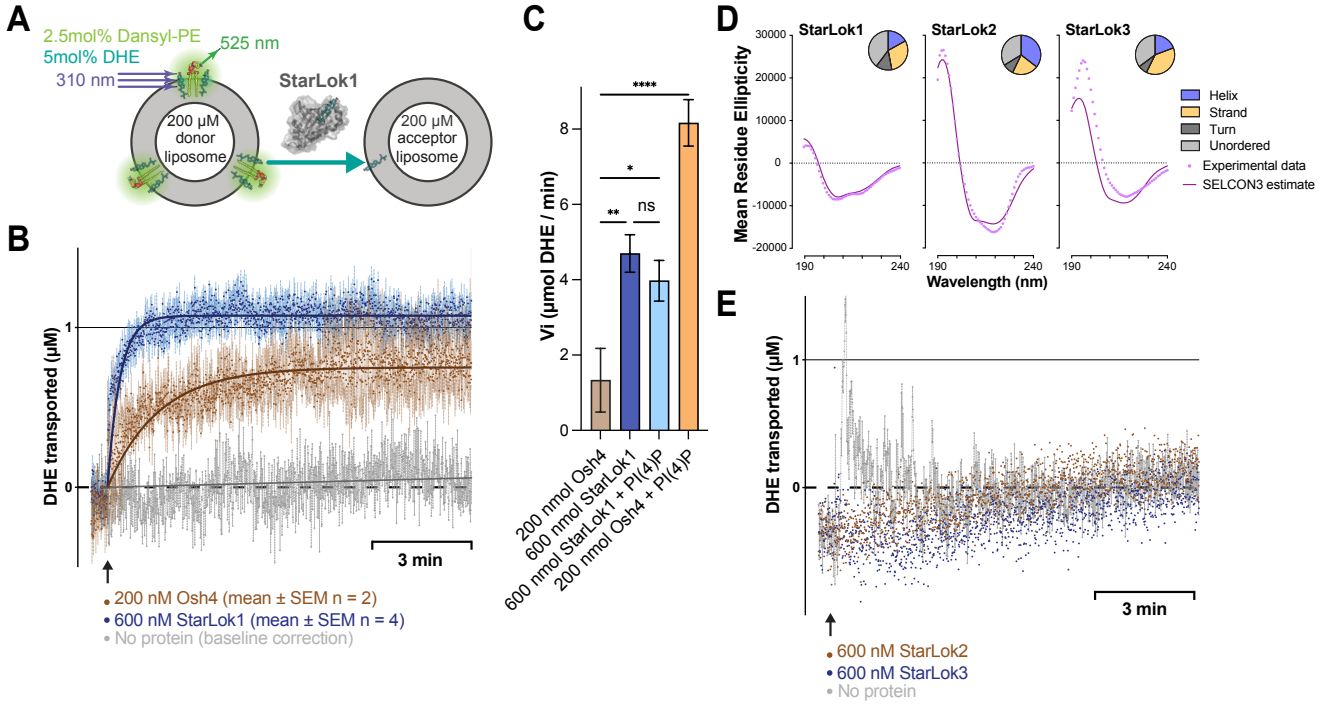


Fig. 5: StarLok1 mediates the transfer of sterols. A. Schematic representation of the FRET-based DHE (dehydroergosterol) transfer assay using StarAsg1. Donor liposomes containing 2.5% Dansyl-PE and 5% DHE are incubated with acceptor liposomes composed of pure DOPC. Upon excitation of DHE, Dansyl-PE fluoresces due to FRET. StarLok1 facilitates the transfer of DHE from donor to acceptor liposomes, which is monitored by a FRET decrease. **B.** Kinetics of DHE transport measured with StarLok1 (blue) or Osh4 proteins (brown). DOPC donor liposomes (200 μM lipids) containing 2.5% Dansyl-PE and 5% DHE were mixed with DOPC acceptor liposomes (200 μM) at 30°C. At t = 1 minute post-stabilization, StarLok1 (600 nM, n = 4) or Osh4 (200 nM, n = 2) were added. The scatter plots show the mean amount (± SEM) of DHE (in μM) transferred from donor to acceptor liposomes. The curves represent non-linear fitting of the measured signal. The gray trace corresponds to a representative condition with no protein, used for baseline correction. **C.** Initial rates (V_i) of DHE transport by StarAsg1 (600 nmol) and Osh4 (200 nmol) with and without 5 mol% PI(4)P in acceptor liposomes. The slopes of each experiment were analyzed for statistical significance. After validating homogeneity of variance, an F-test with post-hoc Tukey indicated p-values as *p<0.05, **p<0.01, ****p<0.0001, ns: not significant. **D.** Circular dichroism spectra of StarLok1, 2 and 3 reveal their secondary-structure content. Experimental mean residue ellipticity (MRE in deg.cm^2.dmol^{-1}) was recorded from 190 to 240 nm (dotted lines) and analyzed by SELCON3 deconvolution (solid lines). Pie charts summarize the estimated proportion of α-helix, β-strand, turn and unordered regions for each protein. **E.** Similar experiments as in B for StarLok2 and StarLok3 (600 nM) do not show any DHE transfer capacity.

Upon StarLok1 addition to DOPC liposomes containing DHE and dansyl-PE, we observed a decrease in FRET corresponding to DHE transport (**Fig. 5B**). The measured initial transfer rate of 7.83 ± 0.83 DHE min^{-1} protein^{-1} (**Fig. 5C**) of StarLok1 was similar to those previously measured for the sterol transporters GramD1 and STARD4, and exceeded those for Ysp2[39,50,51]. In comparison, DHE transfer by the native yeast LTP Osh4, which is responsible for the bulk transfer of ergosterol out of the ER, was similar to that of StarLok1 (6.67 ± 4.23 DHE min^{-1} protein^{-1}) in the absence of its agonist phosphatidylinositol 4-phosphate

(PI(4)P), but faster in the presence of PI(4)P (40.83 ± 2.09 DHE min^{-1} $protein^{-1}$). We compared the folding of StarLok1, 2 and 3 purified in native buffer condition using circular dichroism (**Fig. 5D**). Though StarLok1 specific ellipticity intensity was reduced, it retains a secondary structure like that of StarLok2 and StarLok3. Consistent with their predicted structures, StarLok2 has an increase helical signal and reduced β-strands proportion, while StarLok3 has a reduced number of turns. These measurements also suggest that StarLok1 is less structured than StarLok2/3, likely due to the absence of stabilizing ligand in its large binding cavity. In contrast to StarLok1, addition of StarLok2 and StarLok3 to DOPC liposomes containing DHE and dansyl-PE did not cause any FRET change, indicating that these proteins do not transport DHE (**Fig. 5E**). Even though it is unlikely to be the protein's natural ligand in *P. syntrophicum* cells, StarLok1 shows robust transfer of DHE *in vitro*, suggesting an underlying capacity for lipid transport that is unique among the START domain proteins in Lokiarcheatoa.

Phylogeny of eukaryotic START domains reveals archaeal origins

Given similarities between StarAsg1 and specific eukaryotic lipid binding proteins and StarAsg2/3 with eukaryotic Hsp90 co-chaperones, we sought to understand the broader phylogenetic relationship between START domains across the tree of life. Because START domains often exist in tandem repeats in eukaryotic genomes or as parts of multi-domain proteins, we developed a tool (ProtDomRetriever) to isolate and align individual domains from 161 representative taxa containing eukaryotes, archaea, and bacteria (**Fig. 6A**). This yielded a taxonomic dataset containing 748 START domains containing sequences for which we retrieved 805 AlphaFold structures. We then applied both sequence-based maximum likelihood approaches and newer, structure-based analysis using Foldtree[52], which generates trees based on structural divergences measured by Foldseek[53]. Although we found agreement between the two approaches across several key facets of the START domain phylogeny, as described below, Foldtree generally provided more robust trees for the START domains. This is likely due to the high sequence plasticity in this protein class – less than 25% of START domains passed the Chi2 composition test – that resulted in high entropy and gap rich sequence alignments. In contrast, Foldtree accurately identified the evolutionary history of known taxa and our domain extraction method further reduced long branch attractions artifacts.

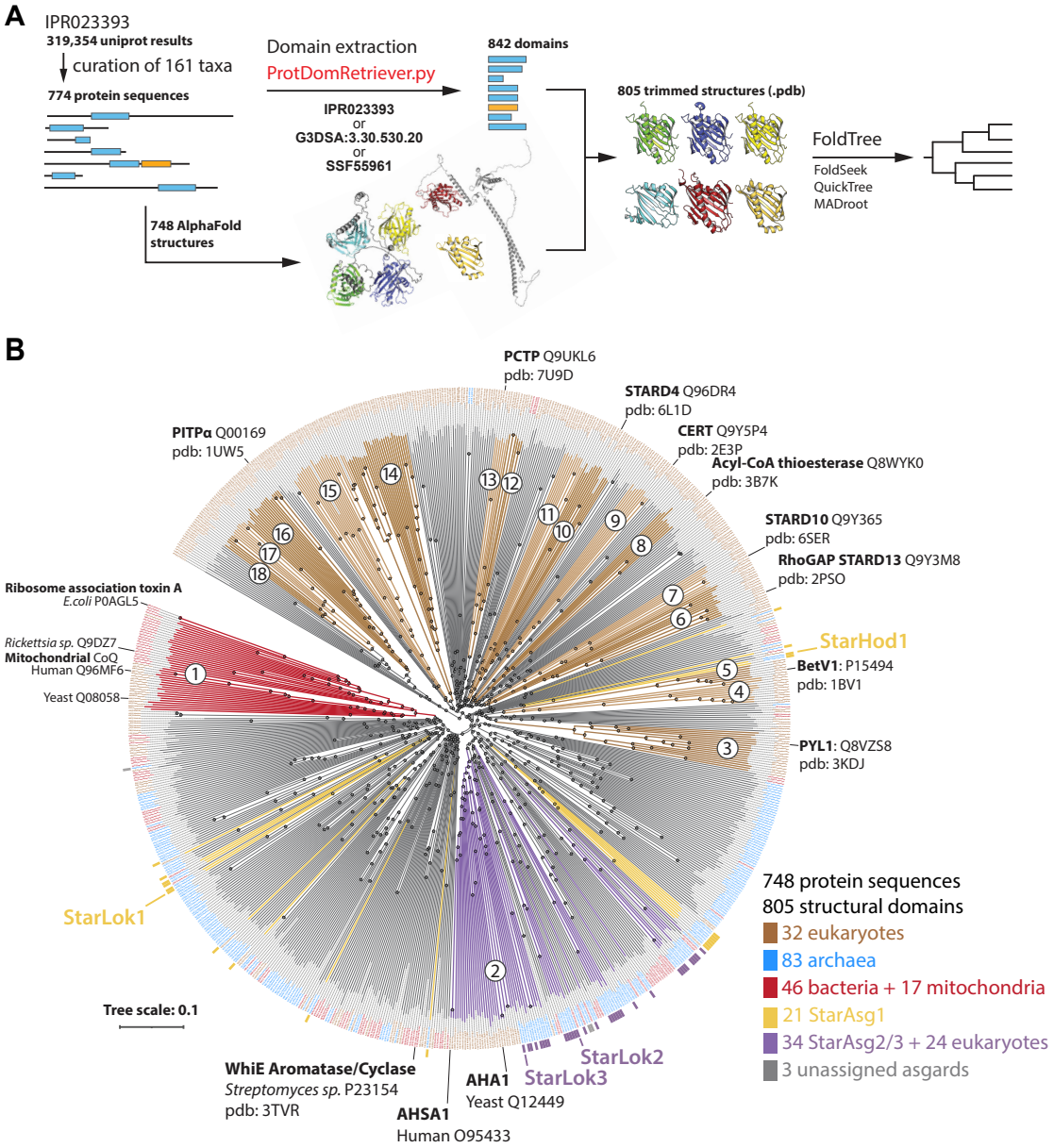


Fig. 6: Large structural domain phylogeny supports orthology between StarAsg1-like proteins and most eukaryotic START LTPs. A. Workflow diagram for large scale structural domain phylogeny. From a collection of species spanning the tree of life, ProtDomRetriever was used to collect the position of the identified domain in each sequence. For the START domain, the InterPro superfamily accession, CATH/gene3D annotation and SUPERFAMILY HMM library accession was used. From the position of each START domain, the corresponding structural domain was extracted from the available AlphaFold structures and phylogenetic analysis was performed using the FoldTree pipeline. **B.** Rooted structural phylogenetic tree of 805 START domains from 161 species. Labels at branch tips are colored in brown, blue and red respectively for eukaryotes, archaeal and bacterial proteins. Yellow, purple or gray stripes near branch tips labels indicate Asgard sequences belonging respectively to the StarAsg1, StarAsg2/3 or undefined clades in Fig. 1B. Branches in red and purple highlight monophyly of the mitochondrial CoQ binding protein 10 homologs with bacterial RatA. Branches related to StarAsg2/3 and AHA1 are colored in purple. Branches in yellow and brown indicates StarAsg1 members and Eukaryotic START LTP respectively. Each monophyletic class of START domain are numbered and their name is reported on a simplified cladogram in Fig.7A. Nodes are represented as black dots. The tree scale indicates the number of substitutions per site. When available, the indicated PDB accession is provided for major clades.

All our trees estimated an ancient origin to the START-domain that preceded eukaryogenesis. Sequence alignment-based maximum likelihood phylogeny (**Supplementary Fig. 7**) and structure-based trees (**Fig. 6B**) showed a distinct branching group consisting of mitochondrial Coenzyme Q-binding proteins (COQ10-like class) and Ribosome association toxin A (RatA) in bacteria, including alphaproteobacteria such as *Rickettsia sp*. that share a common ancestry with mitochondria. As previously noted[54], bacterial COQ10-like proteins are of the Ribosome association toxin A (RatA) class. This START domain was present in 26 out of 36 bacteria in the dataset, including *Rickettsia sp*. and *E. coli*. The divergence of COQ10 from alphaproteobacteria RatA homologs was supported by strong branch support (aLRT = 0.9225) in sequence trees. These analyses reflect a likely bacterial origin of the nucleus-encoded but mitochondrial-localized COQ10 binding protein.

We also confirmed the archaeal origins of co-chaperones of Hsp-90 (Aha1, ASHA1, ASHA2), as was suggested by ProtNLM analysis of StarAsg2/3. Trees showed Aha1-like proteins directly stem their origin in Asgards and are part of a monophyletic group that include nine StarAsg3 homologs, which diverged from StarAsg2 clade in Asgards. The node Aha1 and StarAsg3 is well supported (aLRT = 0.9133) and diverged from StarAsg2-like proteins (aLRT = 0.8727), confirming our maximum likelihood results within Asgards (Fig. 2b). We did note several bacterial START domains with structural similarity to StarAsg2 (but not StarAsg3) and we did not detect a clear phylogenetic signal for these proteins and are not aware of experimental data supporting interaction with HSP-90 in bacteria. Thus, Aha1-like proteins are likely to have been inherited from an ancestral form of StarAsg3 in Asgards to the first eukaryotic common ancestor. In eukaryotes, START domains are the basis for multiple classes of LTPs that shuttle lipids intracellularly. Foldtree analysis supported previous models for the classifications of START domain LTPs within eukaryotes[36,55,56], but provided several additional details (**Fig. 7A**). It showed orthology between STARD4/5/6, as well as STARD1/3; both of which radiate independently in the Holozoa clade, as they are found in animals and their closest unicellular relatives, like *Capsaspora owczarzaki*. Similarly, the two START domains containing acyl-CoA thioesterases (STARD14 and STARD15) also show orthology and arose prior to the divergence of animals and *C. owczarzaki* in Holozoa. More distantly, CERT (STARD11) is grouped as an independent class presumably emerging further back in Holozoa, before the divergence of choanoflagellates. We also observed orthology between STARD8, STARD12, and STARD13, all of which contain a START domain fused to a small GTPase activating protein Rho domain, which are also common to choanoflagellates. STARD2 (PCTP) and STARD7 are grouped together in animals and choanoflagellates but are also paraphyletic to few fungal and planktonic proteins. In contrast to previous analyses, we observed a separation of STARD10 orthologs from an ancestral EDR2-like START domain earlier, with occurrence in a broader group of unicellular eukaryotes (Apusozoa, Excavata, and Cryptophyceae). Finally, our tree suggests a new model in which PCTP and STARD7 form a paralogous sister group from other members of the STARD group.

We observed only a scattering of misgrouped bacterial sequences across the STARD-like sequences, suggesting that the 296 eukaryotic proteins of this class in our dataset stem from a protein scaffold already present in the last eukaryotic ancestor. This includes the parent of the START domain of HD-ZIP proteins strongly conserved and redundant in plants, but also PITP isoforms, PITM and PITC, which are known to transport phosphatidylinositol and phosphatidic acid between organelles in mammals[57]. Foldtree analysis support that all the LTP START domains in our dataset find their roots with a fragmental peptide in Hod LC_3 (A0A1Q9NLH4), which belongs to the StarAsg1 clade. Although we did not observe a phylogenetic signal strong enough to confirm an Asgard origin for STARD-like LTPs, it was clear that few StarAsg1-like sequences are conserved in Hod and Heimdall archaea separating STARD-like LTPs from Bet v 1-like

proteins, the fourth major eukaryotic clade of START domain containing protein that radiated in land plants. These Hod sequences includes A0A1Q9NZ53 in LC_2, A0A1Q9P3Z5 in LC_2, and A0A1Q9NB21 in LC_3, which corresponds to StarHod1. They all belong to the StarAsg group of proteins that feature an enlarged binding cavity, as also seen in eukaryotic LTPs of the START superfamily (**Supplementary Fig. 8**). Based on these analyses and the established relationship of eukaryotes within Asgard archaea, the most parsimonious grouping is of eukaryotic START domain LTPs diversifying from an ancestral protein within the StarAsg1 clade (**Fig. 7B**).

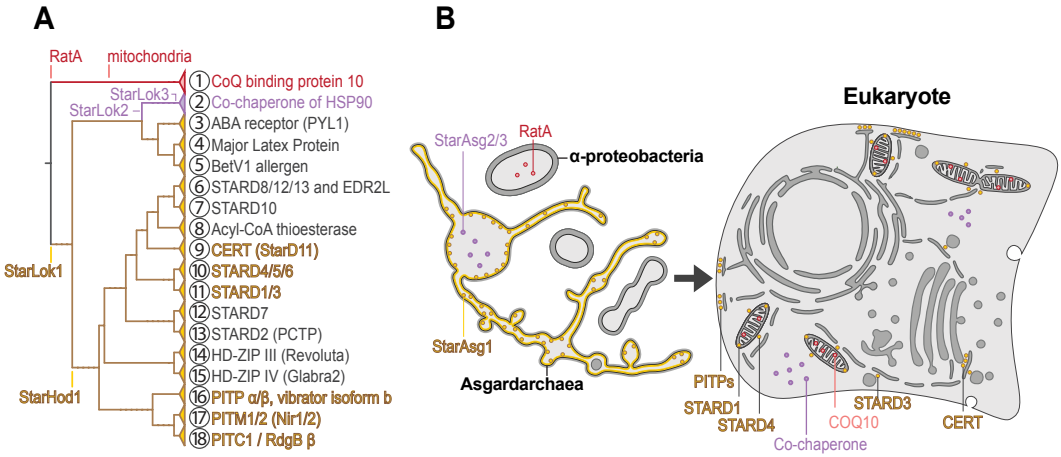


Fig. 7: Model for the evolution of START domain proteins in eukaryotes. A. Simplified cladogram of the tree shown in Figure 6, highlighting the major eukaryotic START domain protein families and their phylogeny. Proposed origins for Aha1/Co-chaperone of HSP90 (StarAsg3), Coq10/CoQ binding protein (alphaproteobacterial RatA), and START domains LTPs (StarAsg1), are highlighted. **B.** Schematic with proposed relationships between START proteins in the prokaryotic ancestors of eukaryotic cells (left) and their localization in eukaryotes. StarAsg1 is a putative membrane-bound lipid binding protein in Asgardarchaea, which shares a common origin with several START domain proteins acting as LTPs. These are predominantly localized to membrane contact sites between organelles. StarAsg2 and StarAsg3, proteins of unknown functions in Asgards, share an origin with the conserved HSP90 co-chaperone in the cytosol of eukaryotes. RatA, a START domain protein in alpha-proteobacteria, shares ancestry with the mitochondrial protein COQ10.

Discussion

In this study, we asked whether the ancestral Asgard superphylum from which eukaryotes originated contain proteins capable of transporting lipids. Lipid trafficking is a hallmark of eukaryotic cells and is a prerequisite for the stable coexistence of multiple membrane organelles. Identification of LTPs within Asgard would support a model in which eukaryotic ancestors already contained proteins supporting lipid transport or ones that could have quickly adapted to serve such a purpose. To this end, we described three conserved clades of Asgard proteins containing START domains, the same fold that forms the basis for a large and diversified group of eukaryotic LTPs. One, StarAsg1, is composed of proteins containing predicted structural features commonly found in LTPs: a large, hydrophobic binding pocket lined with an amphipathic and basic region for interaction with anionic lipid bilayers. Heterologously expressed StarLok1 behaved in a manner consistent with structural predictions, it strongly interacted with anionic membranes and displayed a hydrophobicity that made the protein insoluble even at moderate concentrations, unless hydrophobic interactions are compensated by low pH using volatile short chain fatty acid such as acetate or methanoate (data not shown). In yeast cells, the protein localizes to the anionic PM. These findings suggest StarAsg1 recognizes membranes rich in anionic phospholipids, such as phosphatidylserine and phosphatidylinositol-phosphates (PIPs), which are the hallmarks the PM in eukaryotic cells[5]. Notably,

cultured Asgard express archaetidylserine synthase and archaeatidylinositolphosphate synthase[33], the archaeal analogues of eukaryotic PS and PIPs. StarLok1 facilitates the transport of a fluorescent sterol analogue, DHE, similarly to several eukaryotic START LTPs. In contrast, StarLok2 and StarLok3 did now show these features, indicating that StarAsg1 proteins specialized for membrane interactions and potentially lipid transport.

Phylogenetic and structural homology analysis suggests a broader relationship between Asgard START proteins and those in eukaryotes. We developed a domain trimming method to compare a large dataset of isolated START domains across archaea, bacteria, and eukaryotes. Analysis with recently developed methodologies based on Foldseek structural alignment was able to generate a putative tree of START domains across kingdoms that was taxonomically consistent and was supported by purely sequence-based analyses in key respects. It suggests that STARD-like proteins and similar eukaryotic LTPs emerged from an ancient START domain fold, presumably present in Asgards as StarAsg1 homologues. A second class of Asgard START domains, StarAsg3, includes ancestral proteins for Activator of Heat Shock Protein 90 (Hsp90) ATPases (AHAs), an essential co-chaperone for protein stability and maturation in eukaryotes. StarAsg3 itself diverged from the Asgard StarAsg2 group, which may harbor a cryptic ligand binding site. Another class of START domain proteins conserved in eukaryotes, the Coenzyme-Q binding protein 10, shares a common ancestry with bacterial Ribosome association toxin A proteins and was thus likely inherited from the alphaproteobacterial ancestor of mitochondria, consistent with previous reports [54,58]. This analysis demonstrates how robust structural homology-based evolutionary trees applied to specific protein domains can be especially valuable for lipid binding proteins, which are defined by their overall fold and shape of ligand pockets.

Given that our data is derived from genome-encoded protein features, we cannot yet make definitive claims about the physiological functions of Asgard START proteins. We observed robust sterol transport activity for StarAsg1 *in vitro*, but sterol metabolism itself has not yet been identified in Asgardarchaeota. However, archaea, including Lokiarchaea[33], synthesize abundant terpenes, the base building blocks of sterols and related molecules. Even well-studied lipid binding START domains, like Bet v 1 and its homologues, accommodate a diversity of ligands[59], and so their exact biological functions remain elusive based on purely biochemical analyses. LTPs rarely show exacting specificity for similar ligands in isolation[60]. Instead, their activity is often understood through their cellular contexts and mode of interactions with specific membranes[61] and other proteins, like contact site tethers[62]. Thus, it is likely that StarAsg1 proteins act on yet-unidentified lipid molecules in Asgard archaea. In addition to transport functions, lipid-binding domains can also function as sensors[63], enzymes[41], and regulatory molecules[64]. Similar to putative LTPs, the exact native activities of the ancestral protein for Aha1 (Asgard StarAsg3) remains unknown, but could participate in binding of unfolded, peptides as Aha1[65] and other co-chaperones like DnaJ do[66]. To specialize for function as an LTP, existing protein families would need to adopt enlarged hydrophobic binding cavities lined by a membrane-interacting motif, and that is the feature we observe in StarAsg1 proteins compared to their sistergroups StarAsg2/3. More broadly, we observe that an increase in binding cavity size is a general feature for the specialization of START domains for lipid binding across both Asgards and eukaryotes (**Supplementary Fig. 7**).

The presence of START domains in Asgards could have primed the evolution and diversification of lipid transfer systems during the origin of eukaryotic cells. Pre-existing LTPs or other protein scaffolds with the capacity to quickly adapt into ones could have been a necessary component for early eukaryotic cells to transition from endosymbiotic systems to those with integrated organellar systems. As a process, eukaryogenesis required a massive proliferation of lipid transfer mechanisms and other forms of inter-

organelle communication to support endosymbiont and, eventually, endomembrane systems. These mechanisms could have arisen from existing machinery that supported symbiosis between prokaryotes, yet few examples of secreted transporters have been identified. Alternatively, either of the endosymbiotic ancestors of eukaryotic cells could have brought an ancestral capacity to transfer lipids. Given recent 'inside-out' models suggesting the expansion of the archaeal cytosol to that of the proto-eukaryote[67,68], it is logical for the Archaeal ancestor to potentially bring this capacity. Consistent with these models, we identified a potential origin of a class of eukaryotic LTPs in Asgard archaea, which may have facilitated the emergence of detached endomembranes and a centralized lipid metabolism located in the ER.

It is likely that the first eukaryotic LTPs emerged from other proteins that binding hydrophobic ligands, like chaperones, and acted as metabolic facilitators by fortuitously transferring newly synthesized lipids between membranes down their concentration gradient to acceptor membrane. This passive mechanism may have enabled the proliferation of multiple compartments, thereby favoring the evolution of more complex cellular architectures. The START proteins identified here could have fulfilled this function with their large, hydrophobic binding cavity and capacity to bind to membranes. Later, the emergence of more sophisticated LTPs that couple transfer of multiple lipids, like the Osh/ORP proteins that counter exchange ligands like sterols for PI(4)P[69] would have allowed for the generation of heterotopic organelle membranes with specialized lipid compositions. Bridge-like LTPs, like those in the VPS13 class, could have later emerged due to the need for rapid membrane proliferation in newer aspects of eukaryotic cell biology, like mitochondrial network extension[70], the generation of autophagosomes[71] or spores[72], and then later adapted to supporting routine lipid transfer in larger and more metabolically active cells[73,74]. In this way, the increasing complexity of eukaryotic cell biology could have been enabled by biochemical advances in the ability of proteins to transport lipids.

Acknowledgments

Yongxuan Su and the UCSD Molecular Mass Spectrometry Facility aided in analyses. Deshmukh, Devaraj, Ghosh and Herzik labs provided instrumentation support. Guillaume Drin provided Osh4 expression plasmid. The work was supported by the National Institutes of Health (GM142960), the Moore–Simons Project on the Origin of the Eukaryotic Cell (GBMF9734, grant DOI 10.37807/GBMF9734), and the Paul G. Allen Family Foundation.

Data Availability

Alignments and trees generated are available at iTol (https://itol.embl.de/shared/BudinLab), scripts used at GitHub (https://github.com/NicoFrL/ProtDomRetriever). Data files used to generate figures will be made available on FigShare.

Competing interests

The authors have no competing interests to declare.

Materials and methods

Nomenclature

START: StAR-related transfer, it refers to START-like domain superfamily *stricto sensu (IPR:* IPR023393*).* In opposition to other "Starkin" domains such as members of VAD1 Analog of StAR-related lipid transfer family and PRELI/MSF1, both display a helix-grip fold analog to START domains. StAR: Steroidogenic

acute regulatory protein. SRPBCC: NCBI Conserved Protein Domain superfamily associated to the START domain (NCBI: cl14643), START/RHO_alpha_C/PITP/Bet_v1/CoxG/CalC ligand-binding domain superfamily. STARD: (StAR)-related lipid transfer domain.

Lipid transfer protein classification analysis

The 14 classes of LTP we identified were previously described[51,75–78]. The Interpro accession number of each of this class was used to determine the number of taxonomic species expressing at least one of these proteins: 1) Tgt2/MlaC (Superfamily IPR042245), 2) LolA/LolB/Lppx (Superfamily IPR029046), 3) START-like domain (Superfamily IPR023393), 4) SCP2 sterol-binding domain (superfamily IPR036527), 5) Calycin-Lipocalin (superfamily IPR012674),PRELI/MSF1 (Domain IPR006797), LBP/BPI/CEPT (i.e. Bactericidal permeability-increasing protein, alpha/beta domain; Superfamily IPR017943), nsLTP (i.e. Bifunctional inhibitor/plant lipid transfer protein/seed storage helical domain; Superfamily IPR036312); CRAL/Trio (Superfamily IPR036865), ORP/Osh (Superfamily IPR037239), GM2-AP (Superfamily IPR036846), GLTP (Superfamily IPR036497), NPC1 N-ter (i.e. Niemann-Pick C1, N-terminal; Domain IPR032190), VASt (i.e. VAD1 Analog of StAR-related lipid transfer; Domain IPR031968).

Expression and purification of StarLok1

A codon optimized 492 bp cDNA coding for StarLok1 (UniprotKB: A0A5B9D5D6) C-terminally fused to a thrombin cleavage site (LVPRGS) followed by a 6xHis tag was synthesized and cloned into a pET-21a(+) vector (GeneScript) and expressed in BL21-Gold(DE3) (Agilent). StarAsg1 was purified either from native soluble fraction or refolded from an insoluble fraction. For native purification, a 1L culture LB Lennox with 50 *µ*g/mL ampicillin was grown to an OD_{600} of 0.5 at 37°C then induced (0.2 mM IPTG) before overnight growth at 16°C. Cell pellets were resuspended in cold TN buffer (50 mM Tris, 500 mM NaCl, pH 8.0, at 4°C) supplemented with protease inhibitor cocktail (Roche cOmplete EDTA-free, 10 µM bestatin, 1 µg/mL pepstatin A, 10 µM phosphoramidon), 10 mM imidazole, 1 mM DTT, 20 µg / mL of DNase I (Roche), 5 mM $MgCl_2$ and 1 mg/mL of lysozyme. Cells were then disrupted by sonication and the lysate was clarified by ultracentrifugation (186,000 x *g,* 1h, 4°C). Fifty volumes of supernatant were applied to one volume of Ni-NTA resin slurry (cOmplete His-Tag Purification Resin) and incubated overnight at 4°C. The resin suspension was applied to an Econo-Pac chromatography column (Bio-Rad) successively washed with TN containing 10 mM and 20 mM imidazole. The protein was eluted with TN containing 250 mM imidazole and the eluate concentrated at 2500 *x g* using Amicon Ultra centrifugal unit (MWCO 3000). The protein was further purified by SEC on an Akta pure 25 FPLC with a Sephacryl S200 HR column equilibrated with TN buffer at a flow rate of 1mL/min and fractionation volume of 2.5 mL. Glycerol was added to the buffer (10% v/v) and protein-containing fractions were pooled and concentrated up to 20 *µ*M. Concentration was determined by both A_{280} (DeNovix DS-11 spectrophotometer) and Sypro Orange (ThermoFisher) staining of SDS-PAGE using BSA standards and samples were frozen in liquid N_2 and stored at -80°C.

For refolding of insoluble protein, StarLok1 expression was performed as described in the previous section with minor modification. When OD_{600} reached 0.6, 1mM IPTG was added into the culture that was then incubated three more hours at 37°C before collecting the cells and preparing the pellet for freezing. Pellets were resuspended in 50 mM HEPES-KOH pH 7.00 with a protease inhibitor cocktail (Roche cOmplete EDTA-free, 10 µM bestatin, 1 µg/mL pepstatin A, 10 µM phosphoramidon) and passed 4 times through a high-pressure homogenizer at 1 bar (EmusiFlex-c3, Avestin). 20 µg/mL of DNase I (Roche), 5 mM $MgCl_2$ were added, and the inclusion bodies were then pulled down by ultracentrifugation (186,000 x *g*, 1 hour, 4°C). The supernatant was discarded, and the white pellet was solubilized in HU buffer (50 mM HEPES-KOH pH 7.0 with 6M urea) by mild vortexing and pipetting and one passage through high-pressure homogenizer (1000 bar). After centrifugation (186,000 x *g,* 1 hour at 4°C), 10 mM imidazole was added to

the clear supernatant, which was applied to 50:1 to Ni-NTA resin slurry (cOmplete His-Tag Purification Resin) and incubated 3 hours at 21°C. The suspension was applied to an Econo-Pac chromatography column (Biorad), washed twice with 10 bed volumes of HU with 20 mM imidazole, and eluted with eight bed volumes of HU with 250 mM imidazole. Refolding was performed by dialysis (3.5K MWCO, Slide-A-Lyzer™ MINI Dialysis Devices) twice against 150 volumes of acetate buffer (164 mM acetic acid (OHAc), 36 mM sodium acetate (NaOAc), pH of 4.04) over 14 hours. Screening for refolding conditions was initially done by light microscopy (ThermoFisher EVOS) after low-volume dialysis. Refolded samples were further purified by FPLC on a HiPrep 16/60 Sephacryl S200 HR column (Cytiva) equilibrated with acetate buffer at a flow rate of 1 mL/min and fractionation volume of 2.5 mL. Purification and refolding were monitored by SDS-PAGE. This protocol yielded 5 to 7 mg of protein per 1 L of bacterial culture, as determined by A_{280}.

Expression and purification of StarLok2 and StarLok3

A codon optimization 489 bp cDNA coding for StarLok2 (UniprotKB: A0A5B9DET0) and 432 bp coding for StarLok3 (UniprotKB: A0A5B9DAG8), each N-terminally fused to a 6xHis tag followed by a thrombin cleavage site (LVPRGS), were synthesized and cloned into a pET-28a(+) vector (GeneScript) and expressed in BL21-Gold(DE3) (Agilent). Each protein was purified from 1L culture in its native form as described for StarLok1 in TN buffer supplemented with 1mM DTT. In this standard condition, the protocol yielded more than 10 mg of protein per 1 L of bacterial culture. Protein concentration and purity were controlled by SDS-PAGE stained with InstantBlue (Abcam).

Expression and purification of Osh4

Yeast Osh4 was purified from BL21-Gold(DE3) as previously described[79], with the following modifications: cells pellets were resuspended in Lysis Buffer (50 mM Tris, 150 mM NaCl, pH 7.4, at 4°C) supplemented with a protease inhibitor cocktail (one tablet of Roche cOmplete EDTA-free per 50 mL, 10 µM bestatin, 1 µg/mL pepstatin A, 10 µM phosphoramidon) and 2 mM DTT. The cells were disrupted using a high-pressure homogenizer at 1000 bar (EmusiFlex-c3, Avestin). DNAse I was added to the lysed cells and the solution ultracentrifuged at 186 000 x *g*. SEC was performed on prepacked HiPrep 16/60 Sephacryl S200 HR on an Akta Pure FPLC system (Cytiva). Protein concentration was determined by A_{280}.

Circular Dichroïsm

CD was performed on a Jasco J-1500 spectrometer using a 350 µL synthetic quartz glass cell with a 1 mm path length (CV1Q035AE2, ThorLabs). Measurements were taken of StarAsg1 in either 20 mM Tris, 120 mM NaF, pH 7.4 (native), or in a solution prepared by diluting the stock refolded protein (HOAc/NaOAc, pH 4.0) in milli-Q water. Each spectrum represents the average of five continuous scans recorded from 185 to 260 nm with the following parameters: bandwidth of 1 nm, step size of 0.1 nm, scan speed of 50 nm·min−1, CD scale of 200 mdeg/0.1 dOD, and digital integration time (D.I.T) of 4 seconds. A control spectrum of buffer samples without protein was subtracted from each protein spectrum. For temperature interval scans, an equilibration time of 20 seconds was set before each scan at 10°C intervals. The measured ellipticity was converted to molar ellipticity, considering the cuvette path length and the mean residue concentration of proteins. For secondary structure determination, the spectra were analyzed in the 185–240 nm range using the SELCON3 algorithm and the SP175t dataset on the DICHROWEB server[80] and compared with the algorithm available on the BeStSel web server[81]. Thermal denaturation results were displayed using matplotlib.

Protein structure analysis

Structures were rendered using Pymol (v2.5.8, Schrödinger, LLC). Structure alignment comparisons were conducted using the cealign algorithm and the RMSD values were reported. Model-template alignments

were obtained using SwissModel[82]. Cavities point clouds were calculated using the CavitOmiX plugin (v. 1.0, 2022, Innophore GmbH)[83] and cavities were calculated using a modified LIGSITE algorithm[84]. To compute the cavity size, a script calling CavFind and LigSite modules from CavitOmiX was used to extract the number of cavities, their respective volumes and ligsite score. Protein electrostatic surface potential was determined using APBS (pdb2pqr)[85] with a grid spacing of 0.26. Hydrophobicity was determined using the Eisenberg scale[86]. The Position of Protein on Membrane (PPM) was determined using PPM 3.0 web server[87], using a flat theoretical archaeal plasma membrane composed of Phosphatidylglycerol, Cardiolipin, Archaeol and Menaquinone 8 at a ratio 64:7:8:21. Changing the membrane composition did not significantly change the obtained results. The AlphaFold2 structure of StarLok1 with a thrombin cleavage linked to 6x(His)-tag was computed using ColabFold v1.5.5 notebook. The secondary structure composition was determined using either STRIDE[88] and DSSP algorithm[89]. The theoretical CD spectra of StarLok1 was determined using either SESCA[90] and PDB2CD[91] for comparison with experimental measurement.

Mass Spectrometry

StarLok1 sample purified in native condition was dialyzed three times (one hour, overnight and one hour) l against 1L of TN buffer (10 mM Tris-HCl, 50 mM NaCl, pH 7.4) using a 0.5 mL Slide-A-Lyzer cassette, to ensure complete glycerol removal. For StarLok1 purified in acetate buffer, no dialysis was performed. An additional one-hour dialysis in a new TN buffer was performed. Protein samples were analyzed for purity and mass identification by liquid chromatography with an Agilent 6230 time-of-flight mass spectrometer (TOFMS) with JetStream electrospray ionization source ESI (LC-ESI-TOFMS)

Liposome preparation

Stock solutions of 1,2-dioleoyl-sn-glycero-3-phosphocholine (DOPC), ergosta-5, 7, 9(11), 22-tetraen-3β-ol (DHE), 1, 2-dioleoyl-sn-glycero-3 -phosphoethanolamine-N -(5-dimethylamino- 1-naphthalenesulfonyl), (18:1 Dansyl-PE), 1-palmitoyl -2-oleoyl-sn-glycero-3 -phospho-L-serine (POPS), and L-α-phosphatidylinositol-4-phosphate (PI(4)P) from porcine brain (Avanti Polar Lipids) were prepared in chloroform ($CHCl_3$) or in a 2:1 CHCl3:MeOH and mixed at the indicated molar ratios in pear-shaped flasks. The solvent was evaporated under vacuum (54.9 mbar) for 30 minutes on a rotary evaporator at 37°C. Residual solvent was eliminated by further drying in a vacuum chamber for at least one hour. The resulting lipid films were hydrated with 2 mL HK buffer (50 mM HEPES-KOH, 120 mM K-acetate, pH 7.4) and vigorously vortexed with glass beads. The resulting multilamellar vesicles suspension was subjected to five freeze-thaw cycles (liquid nitrogen/37°C) followed by extrusion through a 0.2 µm polycarbonate filter using a mini-extruder (Avanti Polar Lipids). Freshly extruded liposomes that were prepared on the same day and shielded from light were used for all experiments.

Dynamic Light Scattering

Measurements were performed using a DynaPro instrument with a Flex-99 correlator (ProteinSolutions) and quartz cuvettes. The instrument sensitivity was set to 30% with a maximum acquisition time of 10 seconds and a signal-to-noise threshold of 2.5. The cuvette holder temperature was maintained at 25°C. The buffer refractive index was set to 1.333. For kinetics experiments, 20 min autocorrelation curves were acquired from 200 nm extruded liposomes (50 µM lipid concentration) in degassed HK buffer (50 mM HEPES-KOH, 120 mM K-acetate, pH 7.4) with 1 µM of StarLok1 in acetate buffer added after 1 minute. Alternatively, StarLok2 and StarLok3 were tested. Data filtering and analysis were performed using DYNAMICS software (v.5.25.44) with a maximum Sum Of Square (SOS) limit at 2100, under baseline limit at 0.995, over baseline limit at 1.02, under amplitude limit of 0.05. For the correlation function, the first 2

coefficients were ignored, and the analysis was truncated at channel 120. The autocorrelation functions were then analyzed using the regularization algorithm.

DHE transfer assay

The DHE transfer assay was conducted as previously described [79], with minor modifications. Fluorescence of 500 µL samples was read on a Cary Eclipse Fluorescence Spectrophotometer (Agilent) with a quartz SUPRASIL rectangular cuvette (Perkin Elmer) stirred with a magnetic bar and maintained at 30°C. Samples were injected by pipetting. Experiments began with a suspension (475 µL) of acceptor liposomes (L_B) (DOPC, or DOPC with 5 mol% PI(4)P) in HK buffer (200 µM lipid concentration). At 1 min, 25 µL of donor liposomes (L_A) containing 2.5 mol% of DNS-PE and 5 mol% of DHE was added (200 µM total lipid, final concentration). Under these conditions, all DHE (10 µM) is accessible to proteins due to its rapid equilibrium between the inner and outer leaflets ($t_{1/2}$ < 1 min) [92]. After 3 min, StarLok1/2/3 (600 nM) or Osh4 (200 nM) was injected. During this time-course experiment, the dansyl fluorescence signal was recorded at 525 nm upon DHE excitation at 310 nm. The amount of DHE (in µM) transferred from L_A to L_B liposomes was determined by normalizing the loss of FRET as:

$$\text{Conc} = 10 \times (1 - ((F - F_0) / (F_{max} - F_0)))$$

Where F is the recorded fluorescence, F_{max} is the average fluorescence value recorded over 1 min before protein injection, and F_0 is the average fluorescence signal measured over 10 seconds at t = 6 min in an independent experiment using L_A liposomes containing 2.5 mol% dansyl-PE, but devoid of DHE. The amount of DHE transported was baseline-corrected by subtracting the normalized FRET signal of L_A and L_B without protein from the concentration value (Conc) to eliminate the effects of photobleaching and any potential spontaneous transfer of DHE. This baseline correction involved fitting a second-order polynomial (quadratic) model to the normalized FRET signal of L_A and L_B measured without protein. The coefficients obtained from this regression were used to calculate baseline values, which were then subtracted from the concentration values (Conc) to obtain the corrected amount of DHE effectively transported. Conc curves were fitted using a nonlinear curve fitting method (robust regression) with a one-phase association exponential model. The initial rate (Vi ± SEM) of DHE transport was determined as the slope of a linear fit (least squares regression) of the corrected concentration values over the initial 9.8 s. Statistical analysis of the initial rates was performed using an F-test to confirm homoscedasticity, followed by ordinary one-way ANOVA with Tukey's multiple comparisons test for post-hoc analysis. Normality of residuals was assumed, despite the small dataset. Statistical analyses were performed using GraphPad Prism (10.2.3).

Yeast expression

E. coli codon optimized sequences for StarLok1, StarLok2, StarLok3 and StarHod1 were cloned into the yeast integration vector pCD256 under a TEF1 promoter and fused with a C-terminal yeGFP. The expression construct was integrated into the *URA3* locus of background strain W303a after double digestion of pCD256 with Pme1/Pac1 and chemical transformation. Clones were isolated by selective plating with G418 and grown in CSM medium with 2.0% w/v glucose. The C2 domain of Bovin Lactadherin N-terminally fused to mRFP was a gifted from Sergio Grinstein (Addgene plasmid #74061) was cloned into a pRS415 vector under the control of a PGK1 promoter and terminator sequences. Yeast cells expressing StarLok1 were transformed and selected on CSM-Leu with 250 µg/mL G418, then grown in CSM-Leu medium with 2.0% w/v glucose. For imaging, stationary phase cells were added to concanavalin-1A-coated 8-well coverslip dishes and imaged with a Zeiss LSM 880 confocal laser scanning microscope equipped with an AxioObserver stand and a Plan-Apochromat 63×/1.4 NA Oil DIC M27 objective. Images were acquired with the following settings: GFP was excited using a 488 nm argon laser at 2.3% power, and fluorescence emission was collected in the 494.95-565.07 nm range using a photomultiplier tube (PMT) detector set at 750 V. Brightfield (BF) images were simultaneously captured using a non-descanned

transmission PMT detector. Imaging parameters included a pixel dwell time of 4.1 µs, 8× line averaging, and a confocal pinhole of 0.997 Airy units (50.4 µm) to optimize the signal-to-noise ratio while maintaining optical sectioning. Images were digitized as 16-bit single optical sections with a frame size of 1024×1024 pixels and subsequently cropped to 512×512 pixels for further processing. The physical pixel size was 65.9 nm, yielding a field of view of 33.7×33.7 µm. For super-resolution imaging, the same microscope was used with its Airyscan detector. Z-stacks were collected with a step size of 0.18 µm, covering a total depth of 11.88 µm (66 slices). For optimal resolution, the pixel size was set to 0.04×0.04 µm in the x-y dimension, yielding voxel dimensions of 0.04×0.04×0.18 µm. For dual-channel imaging (FFP/mRFP), Z-stacks were collected with 0.185 µm steps (75 slices, 13.87 µm total depth) at 0.0488×0.0488 µm pixel size, yielding voxel dimensions of 0.0488×0.0488×0.185 µm. Excitation was performed using 488 nm (argon laser) and 561 nm (DPSS laser) at 2% power each. Raw Airyscan images were processed using the ZEISS ZEN software with the standard Airyscan processing algorithm in 3D mode. In FIJI (imageJ), the "Sharpen" filter was applied once to enhance structural details. No additional processing was performed on individual slices to maintain the integrity of the original signal distribution. For visualization purposes, all images were cropped to 512×512 pixels to focus on the regions of interest.

Interpro Sequence alignments and phylogeny of START proteins in Asgards

The initial dataset consists of the 56 identified sequences corresponding to the InterPro accession number IPR023393 in annotated Asgard strains including Lokiarchaeota (MK-D1, CR_4, b31, b32 and GC14_75), Thorarchaeota (SMTZ1_45, SMTZ1_83 and AB25), Heimdallarchaeota (including LC_3 and B3_JM_08 both reclassified as Hodarchaeales, LC_2 as Kariarchaeaceae, and AB_125 as Heimallarchaeia[28]) and Helarchaeota. One sequence (A0A135VUP7) from Thor was removed because it was identical to A0A135VGP2. Full length protein sequences were aligned using MAFFT L-INS-i with default parameters and Blosum62 matrix[93]. The alignment results was used to compute a maximum likelihood phylogeny using PhyML server (PhyML 3.3.2), with Smart Model Selection based on Bayesian information Criteria (SMS BIC), using a neighbor joining starting tree (BioNJ) and aLRT-SH like branch support[94]. The tree was then displayed and annotated using the iTOL online tool[95] and a singular sequence from Helarchaeota was selected as a root.

NCBI sequences alignment and phylogeny of START proteins in Asgards

Asgard proteome sequences were collected from the NCBI database (GenBank Release 264), comprising 711,493 Identical Protein Groups (IPGs). Protein domain analysis was performed using InterProScan version 5.72-103.0[96] deployed via Docker Desktop 4.37.2 (179585) on macOS. Analysis was restricted to Gene3D, Pfam, and SUPERFAMILY search programs. A custom bash script divided protein sequences into smaller batches and ran analyses in parallel (2 concurrent processes, 2 CPU cores each). The pre-calculated match lookup service was utilized with default configuration. Output files were organized by batch and the 508,901 newly annotated sequences were concatenated for downstream analysis. After assigning taxonomic ranks to each sequence using NCBI annotation, 411 START domains were collected using InterPro annotations. (SSF55961; G3DSA:3.30.530.20; PF10604; PF11485; PF08982; PF06240; PF19569; PF02121; PF08327; PF03364; PF01852; PF00407; IPR023393). Each sequence was trimmed according to the longest non-overlapping domain position given by the InterPro Analysis. A subset of START domain-containing proteins in the Thorarchaea clade contained two domains; these were retained as distinct protein sequences. The trimmed sequences were aligned using Muscle5[97], and the maximum likelihood phylogenetic tree was reconstructed after model selection testing, bootstrap calculation (1000 replicates), and approximate likelihood ratio test based on the Shimodaira-Hasegawa test (aLRT SH-like, 1000 replicates) using IQTree[98]. Five monophyletic domain classes were identified and counted. For domain abundance representation, we determined the average number of species in each Asgard archaeal

taxon using seven universal single-copy proteins (SecY, uS13, uS3, uS4, aIF2, Arginine and Serine RNA ligase) among the 508,901 annotated sequences. The average number of proteins in each taxon was used to estimate species abundance. START domain abundance in each taxon was normalized by the estimated species abundance.

Large sequence dataset construction

A set of 83 species of archaea including 14 Asgards, 4 Stygia, 4 Acherontia, 18 Stenoarchaea, 5 Methamonada, 4 Diaforarchaea, 31 TACK group members and 2 DPANN group members, was collected based on recent phylogeny[99]. Similarly, 32 eukaryotes were selected including Obazoa, Amoebozoa, Excavata, Cryptista, Archaeplastida and TSAR reflecting previous molecular-clock analysis[100]. 46 Bacterial species were also selected a posteriori, to increase the depth of the tree by including members of Acidobacteriota, Bdellovibrionota, Campylobacterota, Chrysiogenota, Coprothermobacterota, Deferribacterota, FCB group, Fusobacteriota, Myxococcota, Pseudomonadota (Alpha, Beta and Gammaproteobacteria), PVC group, Terrabacteria group, (Actinomycetota, Deinococcota), Thermodesulfobacteriota and Spirochaetota. For each species, START-containing protein sequences are collected using IPR023393 accession as a request to Uniprot and then used CD-HIT (-c 0.9 -n 5) to identify putative in-paralog groups[101]. For each of these groups, spurious redundant sequences were manually removed depending on their annotation in genomic databases such as genbank, Gramene, Flybase, EmsemblMetazoa and RefSeq. When locus annotations were lacking between every in-paralog, the longest transcript was selected. From this new protein dataset, as a list of accession file, ProtDomRetriever.py was executed using as an input the InterPro entries IPR023393, CATH G3DSA:3.30.530.20, and SUPERFAMILY SSF55961.

Reduced sequence-based maximum likelihood phylogeny of START domains

Trimmed sequences corresponding to 842 START domains from 774 protein sequences were aligned using Muscle5[97]. IQ-tree[98] was then used to perform a composition chi-square test for every sequence in the alignment. Sequences with compositional bias were rejected. The 242 obtained sequences were aligned with Muscle5 and then the tree was built using PhyML, as described above. The tree was displayed and annotated with iTol[102] and was rooted at midpoint for clarity.

Large structural phylogeny of START domains

AlphaFold structures were retrieved from Alphafold Protein Structure database. A trimming algorithm was applied to the PDB structures using the position file generated by ProtDomRetriever, to generate new .pdb files excluding any atoms that are not part of the START domain. When multiple START domains are present in a structure, each of these domains are extracted and copied in an independent new file. This set of trimmed structures was submitted to the FoldTree (build d0f0239) pipeline[52], using Google Colab Notebook on a Python3 runtime accelerated by a A100 GPU. The rooted tree was then displayed and annotated with iTol without modifying the topology. From this tree, a smaller collapsed cladogram was constructed by deleting all sequences not belonging to a known clade of functional protein.

References

1. Kaplan, M. R. & Simoni, R. D. Transport of cholesterol from the endoplasmic reticulum to the plasma membrane. *J. Cell Biol.* **101**, 446–453 (1985).
2. Iglesias-Artola, J. M. *et al.* Quantitative imaging of species-specific lipid transport in mammalian cells. *Cell Biology* (2024).
3. Bigay, J. & Antonny, B. Curvature, lipid packing, and electrostatics of membrane organelles: defining cellular territories in determining specificity. *Dev. Cell* **23**, 886–895 (2012).
4. Menon, A. K. Sterol gradients in cells. *Curr. Opin. Cell Biol.* **53**, 37–43 (2018).
5. Platre, M. P. *et al.* A Combinatorial Lipid Code Shapes the Electrostatic Landscape of Plant Endomembranes. *Dev. Cell* **45**, 465-480.e11 (2018).
6. Banerjee, T. *et al.* Spatiotemporal dynamics of membrane surface charge regulates cell polarity and migration. *Nat. Cell Biol.* **24**, 1499–1515 (2022).
7. Santinho, A., Carpentier, M., Lopes Sampaio, J., Omrane, M. & Thiam, A. R. Giant organelle vesicles to uncover intracellular membrane mechanics and plasticity. *Nat. Commun.* **15**, 3767 (2024).
8. Khaddaj, R. & Kukulski, W. Piecing together the structural organisation of lipid exchange at membrane contact sites. *Curr. Opin. Cell Biol.* **83**, 102212 (2023).
9. Voeltz, G. K., Sawyer, E. M., Hajnóczky, G. & Prinz, W. A. Making the connection: How membrane contact sites have changed our view of organelle biology. *Cell* **187**, 257–270 (2024).
10. Wong, L. H., Gatta, A. T. & Levine, T. P. Lipid transfer proteins: the lipid commute via shuttles, bridges and tubes. *Nat. Rev. Mol. Cell Biol.* **20**, 85–101 (2019).
11. Reinisch, K. M. & Prinz, W. A. Mechanisms of nonvesicular lipid transport. *J. Cell Biol.* **220**, (2021).
12. Wong, L. H., Čopič, A. & Levine, T. P. Advances on the Transfer of Lipids by Lipid Transfer Proteins. *Trends Biochem. Sci.* **42**, 516–530 (2017).
13. Zhang, Y., Ge, J., Bian, X. & Kumar, A. Quantitative Models of Lipid Transfer and Membrane Contact Formation. *Contact* **5**, 1–21 (2022).
14. Grabon, A., Bankaitis, V. A. & McDermott, M. I. The interface between phosphatidylinositol transfer protein function and phosphoinositide signaling in higher eukaryotes. *J. Lipid Res.* **60**, 242–268 (2019).
15. Lipp, N.-F., Ikhlef, S., Milanini, J. & Drin, G. Lipid Exchangers: Cellular Functions and Mechanistic Links With Phosphoinositide Metabolism. *Front Cell Dev Biol* **8**, 663 (2020).
16. Drin, G. Creating and sensing asymmetric lipid distributions throughout the cell. *Emerg Top Life Sci* **7**, 7–19 (2023).
17. Hanada, K. Metabolic channeling of lipids via the contact zones between different organelles. *Bioessays* **46**, e2400045 (2024).
18. Yeow, J. & Chng, S.-S. Of zones, bridges and chaperones - phospholipid transport in bacterial outer membrane assembly and homeostasis. *Microbiology* **168**, (2022).
19. Knoll, A. H., Javaux, E. J., Hewitt, D. & Cohen, P. Eukaryotic organisms in Proterozoic oceans. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* **361**, 1023–1038 (2006).
20. Eme, L., Spang, A., Lombard, J., Stairs, C. W. & Ettema, T. J. G. Archaea and the origin of eukaryotes. *Nat. Rev. Microbiol.* **16**, 120 (2018).
21. Betts, H. C. *et al.* Integrated genomic and fossil evidence illuminates life's early evolution and eukaryote origin. *Nat Ecol Evol* **2**, 1556–1562 (2018).
22. López-García, P. & Moreira, D. The Syntrophy hypothesis for the origin of eukaryotes revisited. *Nat Microbiol* **5**, 655–667 (2020).
23. Williams, T. A., Foster, P. G., Cox, C. J. & Embley, T. M. An archaeal origin of eukaryotes supports only two primary domains of life. *Nature* **504**, 231–236 (2013).
24. Williams, T. A., Cox, C. J., Foster, P. G., Szöllősi, G. J. & Embley, T. M. Phylogenomics provides robust support for a two-domains tree of life. *Nat Ecol Evol* **4**, 138–147 (2020).
25. Muñoz-Gómez, S. A. *et al.* Site-and-branch-heterogeneous analyses of an expanded dataset favour mitochondria as sister to known Alphaproteobacteria. *Nat Ecol Evol* **6**, 253–262 (2022).

26. Spang, A. *et al.* Complex archaea that bridge the gap between prokaryotes and eukaryotes. *Nature* **521**, 173–179 (2015).
27. Zaremba-Niedzwiedzka, K. *et al.* Asgard archaea illuminate the origin of eukaryotic cellular complexity. *Nature* **541**, 353–358 (2017).
28. Eme, L. *et al.* Inference and reconstruction of the heimdallarchaeial ancestry of eukaryotes. *Nature* **618**, 992–999 (2023).
29. Hatano, T. *et al.* Asgard archaea shed light on the evolutionary origins of the eukaryotic ubiquitin-ESCRT machinery. *Nat. Commun.* **13**, 3398 (2022).
30. Souza, D. P. *et al.* Evolutionarily conserved principles of ESCRT-III-mediated membrane remodelling revealed by a two-subunit Asgard archaeal system. *bioRxiv* 2024.07.01.601590 (2024) doi:10.1101/2024.07.01.601590.
31. Tran, L. T., Akıl, C., Senju, Y. & Robinson, R. C. The eukaryotic-like characteristics of small GTPase, roadblock and TRAPPC3 proteins from Asgard archaea. *Commun Biol* **7**, 273 (2024).
32. Vargová, R. *et al.* Arf family GTPases are present in Asgard archaea. *bioRxiv* 2024.02.28.582541 (2024) doi:10.1101/2024.02.28.582541.
33. Imachi, H. *et al.* Isolation of an archaeon at the prokaryote-eukaryote interface. *Nature* **577**, 519–525 (2020).
34. Imachi, H. *et al.* Promethearchaeum syntrophicum gen. nov., sp. nov., an anaerobic, obligately syntrophic archaeon, the first isolate of the lineage “Asgard” archaea, and proposal of the new archaeal phylum Promethearchaeota phyl. nov. and kingdom Promethearchaeati regn. nov. *Int. J. Syst. Evol. Microbiol.* **74**, (2024).
35. Iyer, L. M., Koonin, E. V. & Aravind, L. Adaptations of the helix-grip fold for ligand binding and catalysis in the START domain superfamily. *Proteins* **43**, 134–144 (2001).
36. Radauer, C., Lackner, P. & Breiteneder, H. The Bet v 1 fold: an ancient, versatile scaffold for binding of large, hydrophobic ligands. *BMC Evol. Biol.* **8**, 286 (2008).
37. Ferguson, K. M., Lemmon, M. A., Sigler, P. B. & Schlessinger, J. Scratching the surface with the PH domain. *Nat. Struct. Biol.* **2**, 715–718 (1995).
38. Murray, D., Arbuzova, A., Honig, B. & McLaughlint, S. The role of electrostatic and nonpolar interactions in the association of peripheral proteins with membranes. in *Current Topics in Membranes* vol. 52 277–307 (Academic Press, 2002).
39. Iaea, D. B., Dikiy, I., Kiburu, I., Eliezer, D. & Maxfield, F. R. STARD4 membrane interactions and sterol binding. *Biochemistry* **54**, 4623–4636 (2015).
40. Lipp, N.-F. *et al.* An electrostatic switching mechanism to control the lipid transfer activity of Osh6p. *Nat. Commun.* **10**, 3926 (2019).
41. Caldara-Festin, G. *et al.* Structural and functional analysis of two di-domain aromatase/cyclases from type II polyketide synthases. *Proc. Natl. Acad. Sci. U. S. A.* **112**, E6844-51 (2015).
42. Talandashti, R. *et al.* Membrane specificity of the human cholesterol transfer protein STARD4. *J. Mol. Biol.* **436**, 168572 (2024).
43. Husslik, F. *et al.* Folded or Not? Tracking Bet v 1 Conformation in Recombinant Allergen Preparations. *PLoS One* **10**, e0132956 (2015).
44. McBride, J. K., Cheng, H., Maleki, S. J. & Hurlburt, B. K. Purification and characterization of pathogenesis related class 10 panallergens. *Foods* **8**, 609 (2019).
45. Orsi, W. D. *et al.* Metabolic activity analyses demonstrate that Lokiarchaeon exhibits homoacetogenesis in sulfidic marine sediments. *Nat. Microbiol.* **5**, 248–255 (2020).
46. Ragsdale, S. W. & Pierce, E. Acetogenesis and the Wood-Ljungdahl pathway of CO(2) fixation. *Biochim. Biophys. Acta* **1784**, 1873–1898 (2008).
47. Fairn, G. D., Hermansson, M., Somerharju, P. & Grinstein, S. Phosphatidylserine is polarized and required for proper Cdc42 localization and for development of cell polarity. *Nat. Cell Biol.* **13**, 1424–1430 (2011).
48. Kofler, S. *et al.* Crystallographically mapped ligand binding differs in high and low IgE binding isoforms of birch pollen allergen bet v 1. *J. Mol. Biol.* **422**, 109–123 (2012).
49. Baker, B., McShea, H., De Anda, V., Brocks, J. & Welander, P. Archaeal lineages related to eukaryotes encode functional diterpenoid cyclases. *Research Square* (2025) doi:10.21203/rs.3.rs-6009237/v1.

50. Mesmin, B. *et al.* STARD4 abundance regulates sterol transport and sensing. *Mol. Biol. Cell* **22**, 4004–4015 (2011).
51. Horenkamp, F. A., Valverde, D. P., Nunnari, J. & Reinisch, K. M. Molecular basis for sterol transport by StART-like lipid transfer domains. *EMBO J.* **37**, (2018).
52. Moi, D. *et al.* Structural phylogenetics unravels the evolutionary diversification of communication systems in gram-positive bacteria and their viruses. *Bioinformatics* (2023).
53. van Kempen, M. *et al.* Fast and accurate protein structure search with Foldseek. *Nat. Biotechnol.* **42**, 243–246 (2024).
54. Tsui, H. S. *et al.* Human COQ10A and COQ10B are distinct lipid-binding START domain proteins required for coenzyme Q function. *J. Lipid Res.* **60**, 1293–1310 (2019).
55. Alpy, F. & Tomasetto, C. Give lipids a START: the StAR-related lipid transfer (START) domain in mammals. *J. Cell Sci.* **118**, 2791–2801 (2005).
56. Wyckoff, G. J., Solidar, A. & Yoden, M. D. Phosphatidylinositol transfer proteins: sequence motifs in structural and evolutionary analyses. *J. Biomed. Sci. Eng.* **3**, 65–77 (2010).
57. Shadan, S. *et al.* Dynamics of lipid transfer by phosphatidylinositol transfer proteins in cells. *Traffic* **9**, 1743–1756 (2008).
58. Fino, C. *et al.* PasT of Escherichia coli sustains antibiotic tolerance and aerobic respiration as a bacterial homolog of mitochondrial Coq10. *Microbiologyopen* **9**, e1064 (2020).
59. Aglas, L. *et al.* Ligand binding of PR-10 proteins with a particular focus on the Bet v 1 allergen family. *Curr. Allergy Asthma Rep.* **20**, 25 (2020).
60. Hamaï, A. & Drin, G. Specificity of lipid transfer proteins: An in vitro story. *Biochimie* (2024) doi:10.1016/j.biochi.2024.09.007.
61. Srinivasan, S. *et al.* The conformational plasticity of structurally unrelated lipid transport proteins correlates with their mode of action. *PLoS Biol.* **22**, e3002737 (2024).
62. Hanada, K. Lipid transfer proteins rectify inter-organelle flux and accurately deliver lipids at membrane contact sites. *J. Lipid Res.* **59**, 1341–1366 (2018).
63. Little, M. & Ortlund, E. A. Structure, function, and lipid sensing activity in the thioesterase superfamily. *Biochem. Soc. Trans.* **52**, 1565–1577 (2024).
64. Dittrich, M. *et al.* The role of Arabidopsis ABA receptors from the PYR/PYL/RCAR family in stomatal acclimation and closure signal integration. *Nat. Plants* **5**, 1002–1011 (2019).
65. Tripathi, V., Darnauer, S., Hartwig, N. R. & Obermann, W. M. J. Aha1 can act as an autonomous chaperone to prevent aggregation of stressed proteins. *J. Biol. Chem.* **289**, 36220–36228 (2014).
66. Schröder, H., Langer, T., Hartl, F. U. & Bukau, B. DnaK, DnaJ and GrpE form a cellular chaperone machinery capable of repairing heat-induced protein damage. *EMBO J.* **12**, 4137–4144 (1993).
67. Baum, B. & Spang, A. On the origin of the nucleus: a hypothesis. *Microbiol. Mol. Biol. Rev.* **87**, e0018621 (2023).
68. Baum, D. A. & Baum, B. An inside-out origin for the eukaryotic cell. *BMC Biol.* **12**, 76 (2014).
69. Antonny, B., Bigay, J. & Mesmin, B. The oxysterol-binding protein cycle: Burning off PI(4)P to transport cholesterol. *Annu. Rev. Biochem.* **87**, 809–837 (2018).
70. Guillén-Samander, A. *et al.* VPS13D bridges the ER to mitochondria and peroxisomes via Miro. *J. Cell Biol.* **220**, (2021).
71. Valverde, D. P. *et al.* ATG2 transports lipids to promote autophagosome biogenesis. *J. Cell Biol.* **218**, 1787–1798 (2019).
72. Park, J.-S. & Neiman, A. M. VPS13 regulates membrane morphogenesis during sporulation in Saccharomyces cerevisiae. *J. Cell Sci.* **125**, 3004–3011 (2012).
73. Kumar, N. *et al.* VPS13A and VPS13C are lipid transport proteins differentially localized at ER contact sites. *J. Cell Biol.* **217**, 3625–3639 (2018).
74. Park, J.-S. *et al.* Yeast Vps13 promotes mitochondrial function and is localized at membrane contact sites. *Mol. Biol. Cell* **27**, 2435–2449 (2016).
75. Khafif, M., Cottret, L., Balagué, C. & Raffaele, S. Identification and phylogenetic analyses of VASt, an uncharacterized protein domain associated with lipid-binding domains in Eukaryotes. *BMC Bioinformatics* **15**, 222 (2014).
76. Chiapparino, A., Maeda, K., Turei, D., Saez-Rodriguez, J. & Gavin, A.-C. The orchestra of lipid-transfer proteins at the crossroads between metabolism and signaling. *Prog. Lipid Res.* **61**, 30–39

(2016).
77. Missaoui, K. *et al.* Plant non-specific lipid transfer proteins: An overview. *Plant Physiol. Biochem.* **171**, 115–127 (2022).
78. Miliara, X. *et al.* An intermolecular hydrogen bonded network in the PRELID-TRIAP protein family plays a role in lipid sensing. *Biochim. Biophys. Acta: Proteins Proteomics* **1871**, 140867 (2023).
79. Lipp, N.-F. & Drin, G. In Vitro Strategy to Measure Sterol/Phosphatidylinositol-4-Phosphate Exchange Between Membranes. *Methods Mol. Biol.* **1949**, 269–292 (2019).
80. Miles, A. J., Ramalli, S. G. & Wallace, B. A. DichroWeb, a website for calculating protein secondary structure from circular dichroism spectroscopic data. *Protein Sci.* **31**, 37–46 (2022).
81. Micsonai, A. *et al.* BeStSel: webserver for secondary structure and fold prediction for protein CD spectroscopy. *Nucleic Acids Res.* **50**, W90–W98 (2022).
82. Waterhouse, A. *et al.* SWISS-MODEL: homology modelling of protein structures and complexes. *Nucleic Acids Res.* **46**, W296–W303 (2018).
83. Gruber, K., Steinkellner, G. & Gruber, C. Determining novel enzymatic functionalities using three-dimensional point clouds representing physico chemical properties of protein cavities. *US Patent* (2020).
84. Hendlich, M., Rippmann, F. & Barnickel, G. LIGSITE: automatic and efficient detection of potential small molecule-binding sites in proteins. *J. Mol. Graph. Model.* **15**, 359–63, 389 (1997).
85. Dolinsky, T. J., Nielsen, J. E., McCammon, J. A. & Baker, N. A. PDB2PQR: an automated pipeline for the setup of Poisson-Boltzmann electrostatics calculations. *Nucleic Acids Res.* **32**, W665-7 (2004).
86. Eisenberg, D., Schwarz, E., Komaromy, M. & Wall, R. Analysis of membrane and surface protein sequences with the hydrophobic moment plot. *J. Mol. Biol.* **179**, 125–142 (1984).
87. Lomize, A. L., Todd, S. C. & Pogozheva, I. D. Spatial arrangement of proteins in planar and curved membranes by PPM 3.0. *Protein Sci.* **31**, 209–220 (2022).
88. Frishman, D. & Argos, P. Knowledge-based protein secondary structure assignment. *Proteins* **23**, 566–579 (1995).
89. Kabsch, W. & Sander, C. Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers* **22**, 2577–2637 (1983).
90. Nagy, G., Hoffmann, S. V., Jones, N. C. & Grubmüller, H. Reference data set for circular dichroism spectroscopy comprised of validated intrinsically disordered protein models. *Appl. Spectrosc.* **78**, 897–911 (2024).
91. Mavridis, L. & Janes, R. W. PDB2CD: a web-based application for the generation of circular dichroism spectra from protein atomic coordinates. *Bioinformatics* **33**, 56–63 (2017).
92. John, K., Kubelt, J., Müller, P., Wüstner, D. & Herrmann, A. Rapid transbilayer movement of the fluorescent sterol dehydroergosterol in lipid membranes. *Biophys. J.* **83**, 1525–1534 (2002).
93. Katoh, K., Rozewicki, J. & Yamada, K. D. MAFFT online service: multiple sequence alignment, interactive sequence choice and visualization. *Brief. Bioinform.* **20**, 1160–1166 (2017).
94. Guindon, S. *et al.* New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. *Syst. Biol.* **59**, 307–321 (2010).
95. Letunic, I. & Bork, P. Interactive Tree of Life (iTOL) v6: recent updates to the phylogenetic tree display and annotation tool. *Nucleic Acids Res.* **52**, W78–W82 (2024).
96. Jones, P. *et al.* InterProScan 5: genome-scale protein function classification. *Bioinformatics* **30**, 1236–1240 (2014).
97. Edgar, R. C. Muscle5: High-accuracy alignment ensembles enable unbiased assessments of sequence homology and phylogeny. *Nat. Commun.* **13**, 6968 (2022).
98. Minh, B. Q. *et al.* IQ-TREE 2: New models and efficient methods for phylogenetic inference in the genomic era. *Mol. Biol. Evol.* **37**, 1530–1534 (2020).
99. Aouad, M. *et al.* A divide-and-conquer phylogenomic approach based on character supermatrices resolves early steps in the evolution of the Archaea. *BMC Ecol. Evol.* **22**, 1 (2022).
100. Strassert, J. F. H., Irisarri, I., Williams, T. A. & Burki, F. A molecular timescale for eukaryote evolution with implications for the origin of red algal-derived plastids. *Nat. Commun.* **12**, 1879 (2021).
101. Li, W. & Godzik, A. Cd-hit: a fast program for clustering and comparing large sets of protein or

nucleotide sequences. *Bioinformatics* **22**, 1658–1659 (2006).
102. Letunic, I. & Bork, P. Interactive Tree Of Life (iTOL) v5: an online tool for phylogenetic tree display and annotation. *Nucleic Acids Res.* **49**, W293–W296 (2021).

Supplementary Information

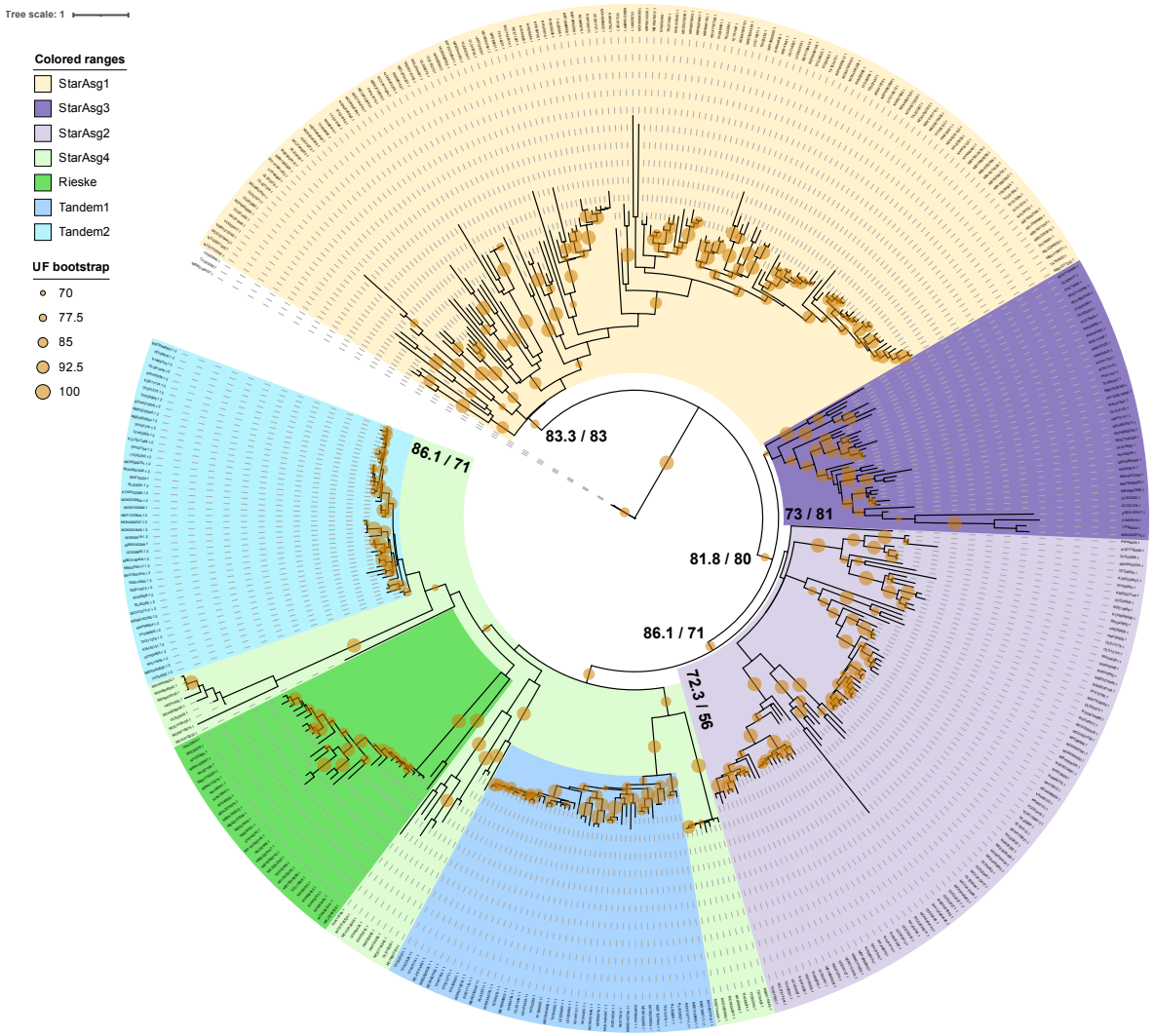
For

An origin for a eukaryotic lipid transfer protein fold in Asgard archaea

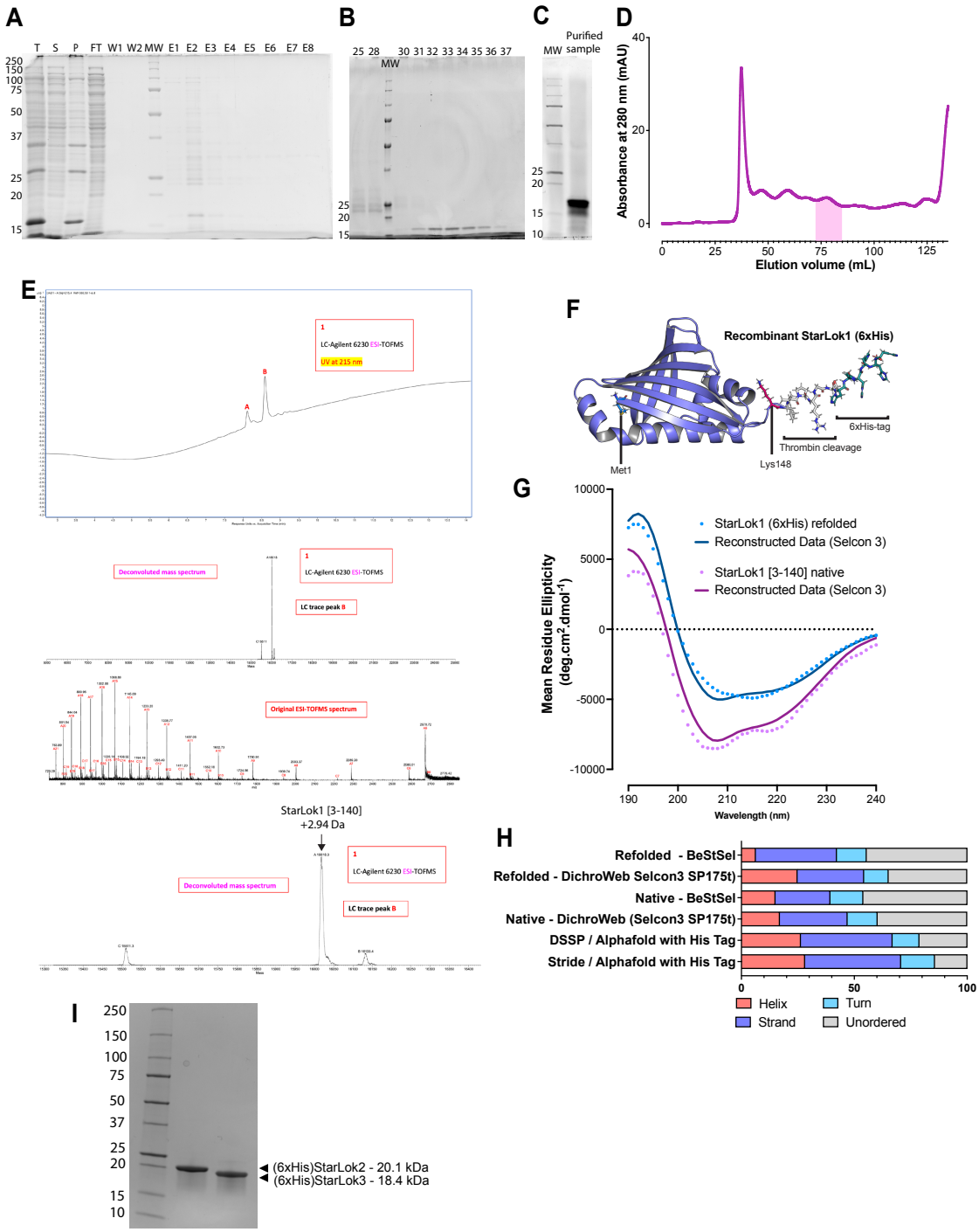
Nicolas-Frédéric Lipp and Itay Budin

This file includes:

Figures S1-S8

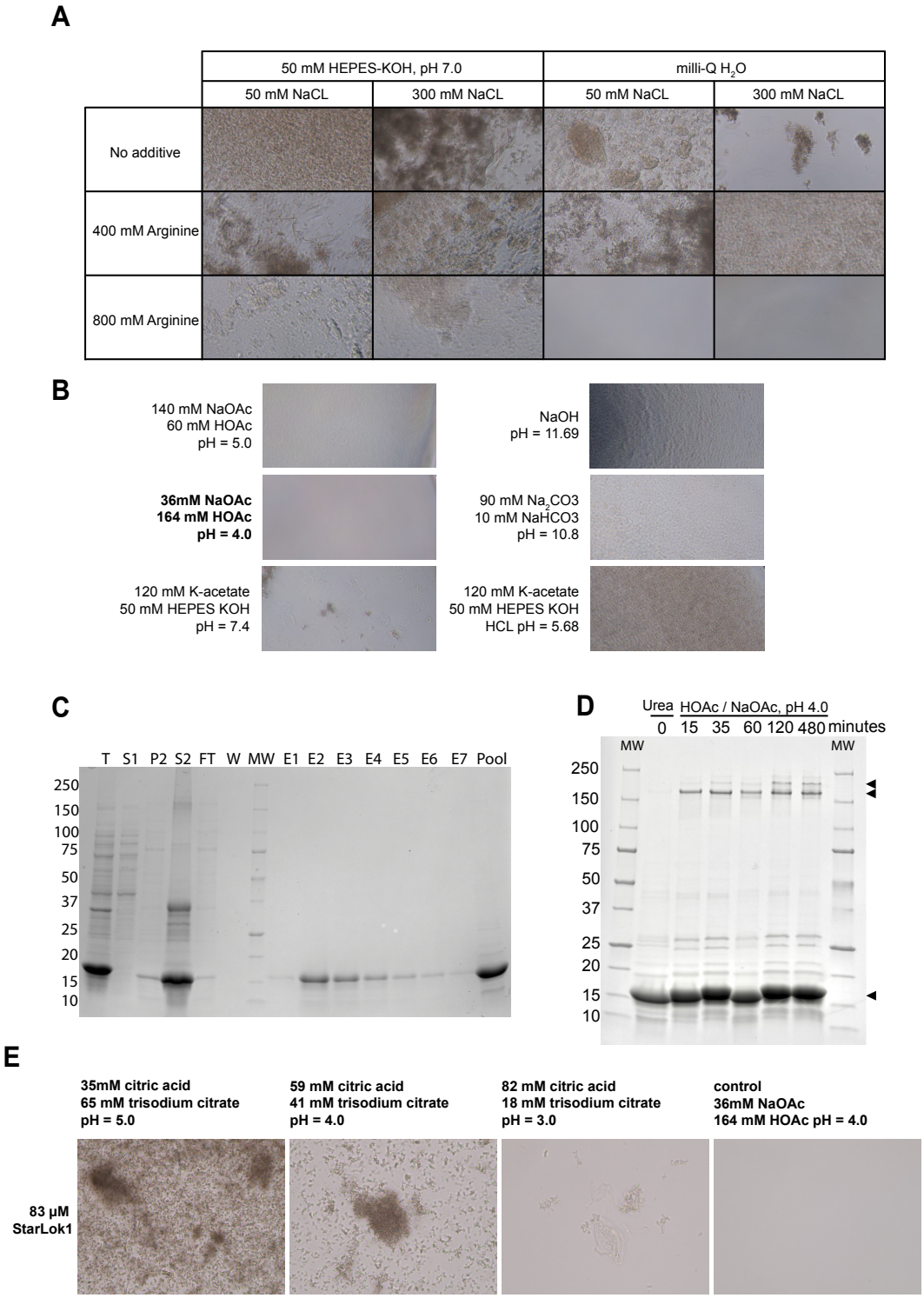


Supplementary Fig. 1: Phylogenetic tree of START domains identified in Asgards. Maximum-likelihood tree (IQ-Tree, WAG + F + I + R5) of trimmed START domains in Asgard proteomes. Each color indicates a START domain subclass. Tandem1 and Tandem2 belong to the same protein, and together with Rieske define the class StarAsg4. Branch support is shown as brown circles when both bootstrap and aLRT SH-like values are ≥ 70. Relevant support values are reported as bootstrap/aLRT SH-like values respectively.



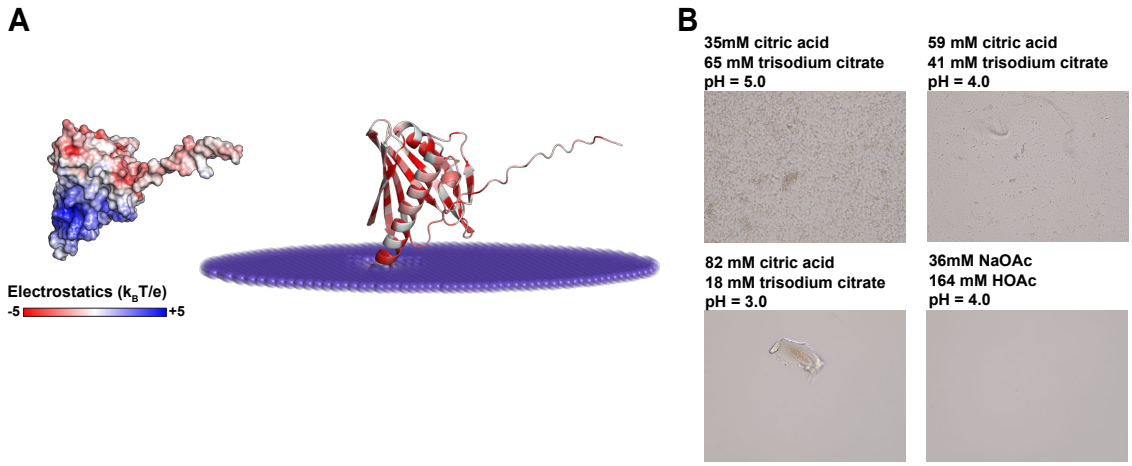
Supplementary Fig. 2: Purification and folding of StarLok proteins expressed in *E. coli*. A. SDS-PAGE results of recombinant StarLok1 expression and purification using immobilized metal affinity chromatography (IMAC). Letter indicates purification steps as follows: T: total, S: supernatant, P: pellet, FT: Flow through, W1 and W2: Washes, E1 to E8 corresponds to elution fractions. MW: Molecular weight marker. Numbering on the left indicates the apparent molecular weight in kDa. **B.** SDS-PAGE results from samples recovered from each fractions after size exclusion chromatography. MW: Molecular weight marker. Numbering on the left indicates the apparent molecular weight in kDa. **C.** SDS-PAGE showing 1.5

µL equivalent sample of stock of purified StarLok1. **D.** SEC chromatogram, the plain pink area under the curve indicates the concentrated fraction (31 to 36) containing StarLok1. **E.** LC-ESI-TOFMS analysis of recombinant StarLok1 (6xHis) purified from *E. coli*. The chromatograms depicted the deconvoluted mass spectrum of the main detected peak. **F.** Alphafold2 model of the recombinant StarLok1 (6xHis). **G.** Circular dichroism spectra of StarLok1 purified natively (blue) and purified at pH 4.0 (pink). Reconstructed data using Selcon 3 algorithms are displayed. **H.** Comparison between the secondary structure determination of StarLok1 obtained from CD spectra with secondary structure distribution of AlphaFold structure. **I.** SDS-PAGE showing StarLok2 and StarLok3 (6xHis) purified from *E. coli*.

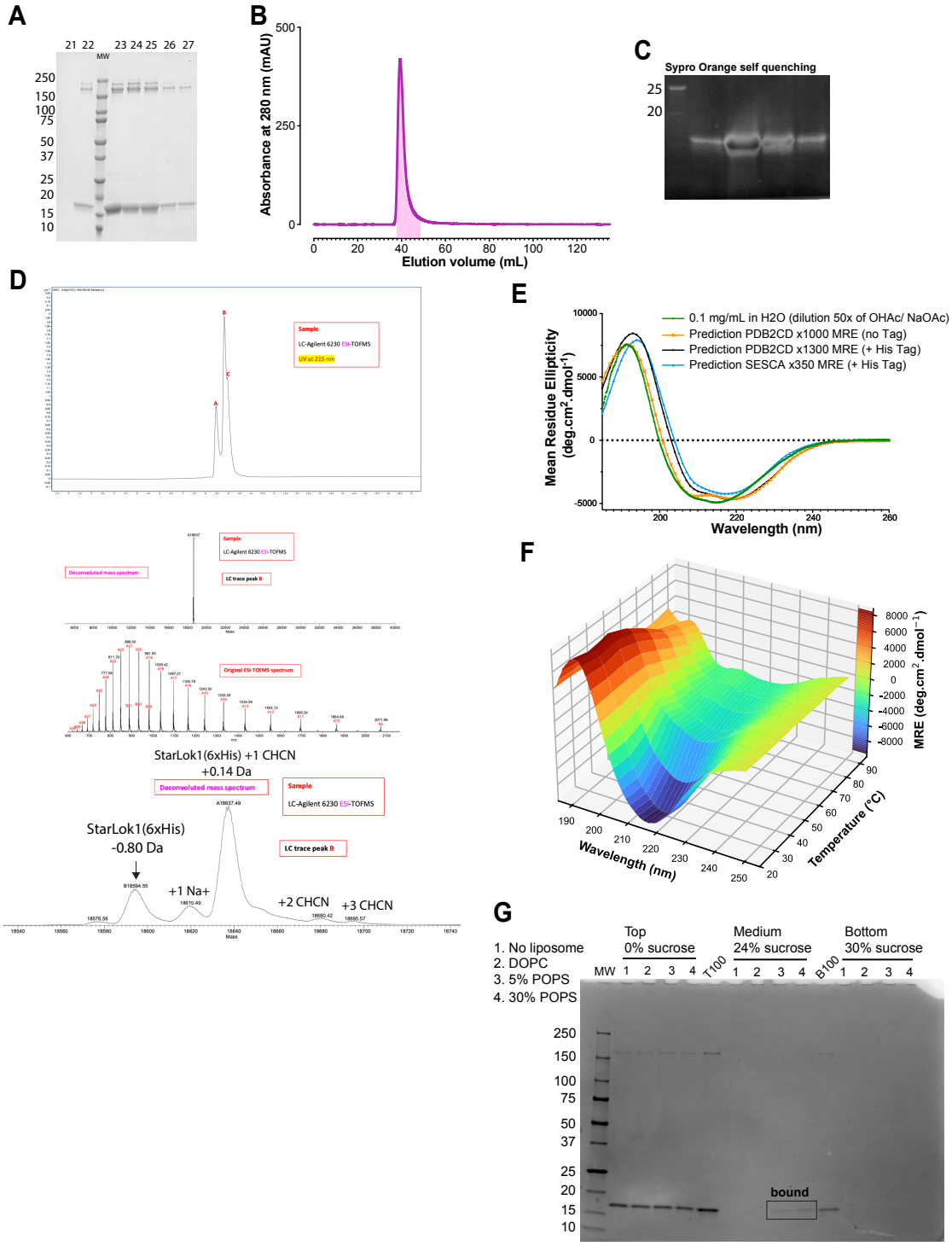


Supplementary Fig. 3: Aggregation of concentrated StarLok1 is pH and solute dependent. A. Light microscopy analysis of the StarLok1-containing sample after dialysis in buffers with different arginine and NaCl concentrations. Different levels of protein aggregation are observed in most conditions, but not in unbuffered solution of 800mM arginine. **B.** Light microscopy analysis of the StarLok1-containing sample after dialysis in buffers with different pH values and buffer compositions. At pH 4.0 maintained by a sodium acetate solution, StarLok1 does not show signs of aggregation unlike at higher pH. **C.** SDS-PAGE results

of StarLok1 IMAC purification under denaturing condition with 6M urea. T: total, S1: supernatant before Urea addition, P2: pellet after urea denaturation, S2, supernatant after urea denaturation, FT: Flow through, W1 and W2: Washes, E1 to E7 corresponds to elution fractions. The pool fraction corresponds to the concentrated protein before refolding. MW: Molecular weight marker. Numbering on the left indicates the apparent molecular weight in kDa. **D.** SDS-PAGE analysis of the kinetics of refolding of StarLok1. Arrow indicated the presence of StarLok1 as a monomer at the expected molecular weight of 18.6 kDa and the formation of multimers at an apparent molecular weight of ~ 180 kDa within the first 15 minutes of dialysis. **E.** Comparison of StarLok1 dialyzed in sodium citrate vs. sodium acetate. In acetate buffered solution, the protein is soluble at pH 4.0, but in citrate aggregates are still observed at pH 3.

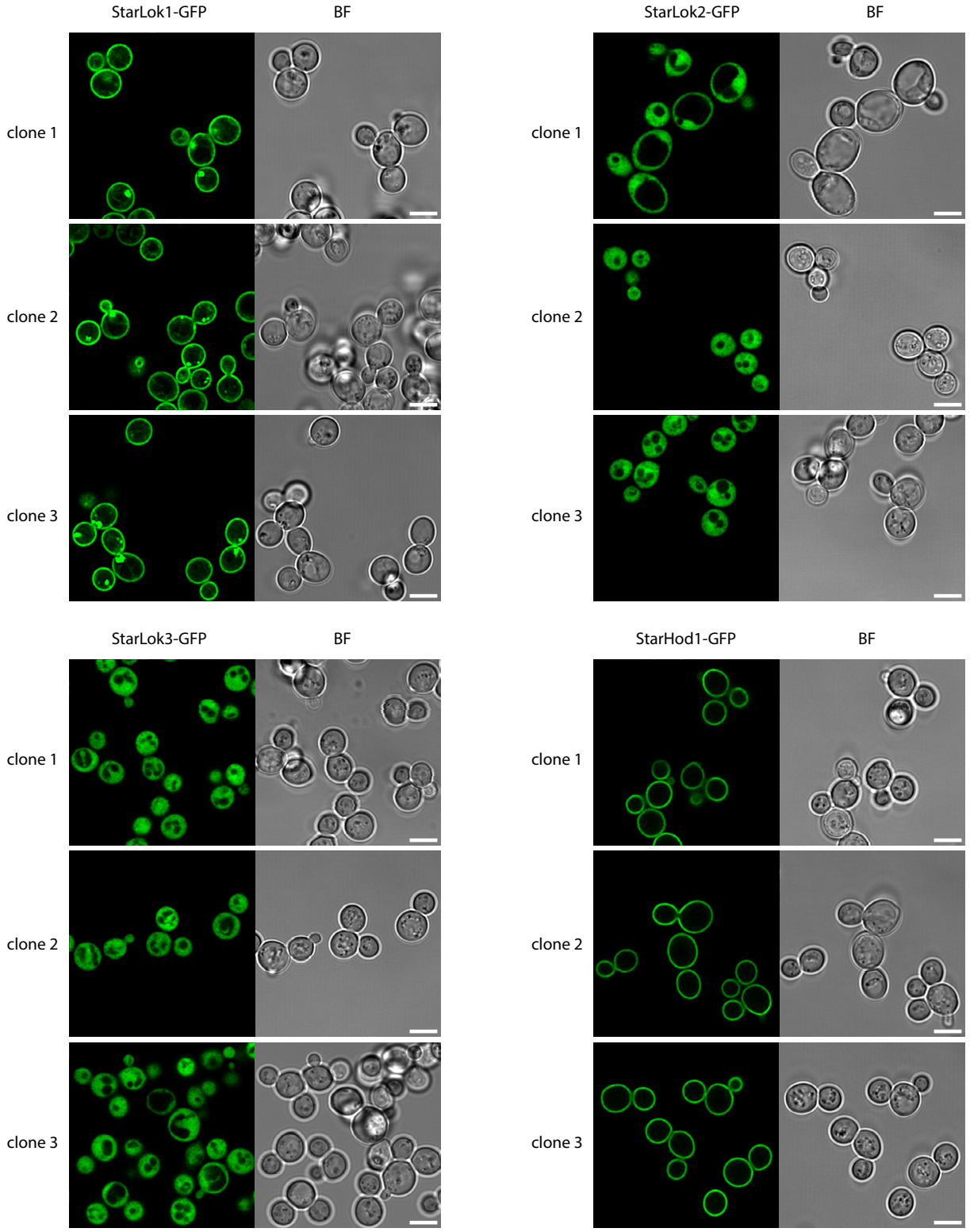


Supplementary Fig. 4: Similarities between StarHod1 and StarLok1. A. Structural prediction of StarHod1 as electrostatic surface potential ranging from -5 to +5 k_BT/e and shown as a color gradient from red (negative) to white (neutral) to blue (positive) and corresponding membrane docking mode calculation on a model archaeal lipid membrane. The structure is colored according to the hydrophobicity of each residue from pale cyan (low hydrophobicity) to red (high hydrophobicity). **B.** Refolding of StarHod1 in different acidic buffers. Similarly to StarLok1, a combination of acetate and low pH is required for StarHod1 solubility.



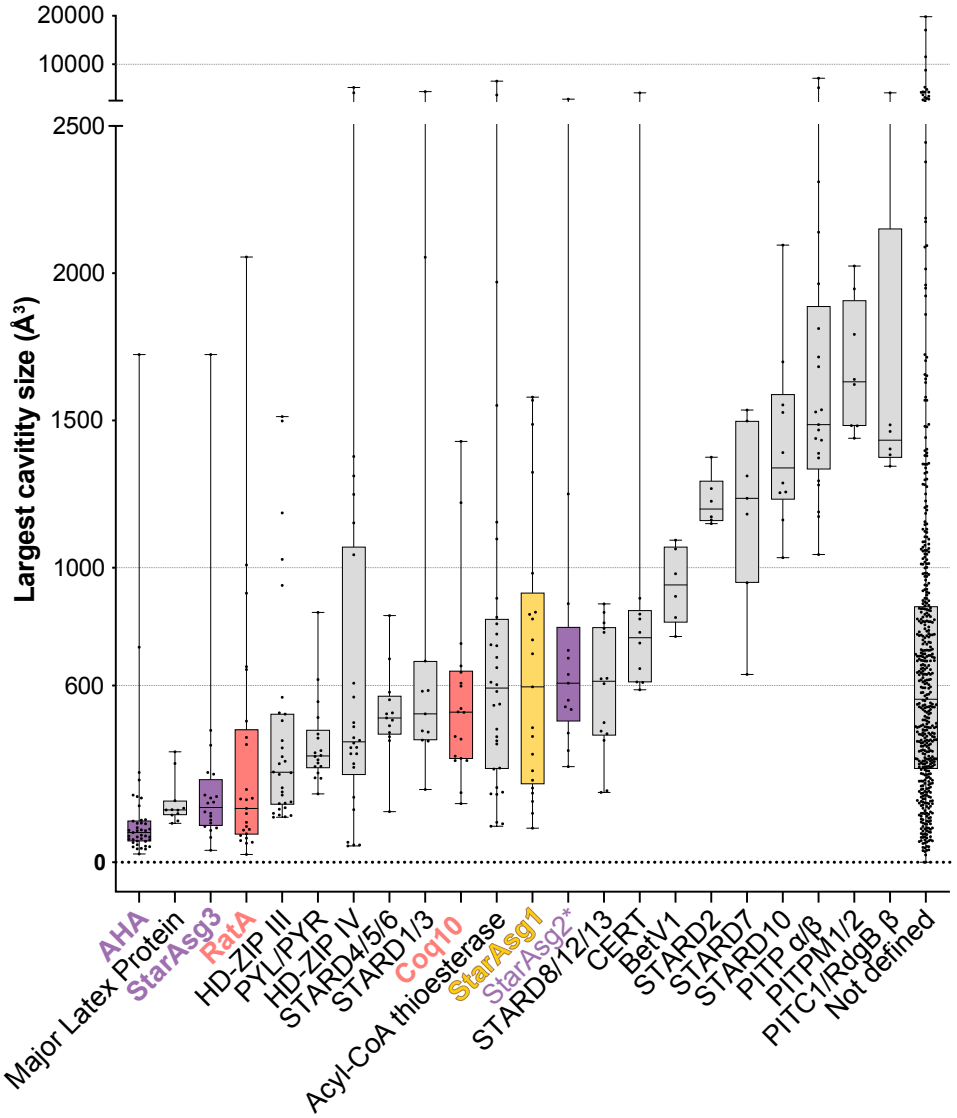
Supplementary Fig. 5: Denaturation and refolding of StarLok1. A. SDS-PAGE analysis from samples recovered from each fraction after size exclusion chromatography (fraction numbers are indicated at the top). MW: Molecular weight marker. Numbering on the left indicates the apparent molecular weight in kDa. **B.** SEC chromatogram, the plain pink area under the curve indicates the concentrated fraction (22 to 27) containing StarLok1. **C.** SDS-PAGE of the refolded StarLok1 in acetate buffer from fractions after size

exclusion chromatography and stained Sypro orange. The dark spot in the second lane indicates oversaturation of the stain. **D.** LC-ESI-TOFMS analysis of recombinant StarLok1 (6xHis) purified from *E. coli i*n denatured condition and refolder in acetate buffer at pH 4.0. The chromatograms depicted the deconvoluted mass spectrum of the main detected peak. **E.** The CD spectrum of StarLok1 refolded in sodium acetate buffer is compared to theoretical CD spectra determined from Alphafold structures. **F.** Thermal denaturation of StarLok1 refolded in sodium acetate buffer and diluted in water as in panel E. The ellipticity of StarLok1 was monitored between 190 and 250 nm with increasing temperature from 20 to 90°C . The color gradient indicates MRE values and corresponding values are indicated on the right scale. As anticipated, a substantial decrease in β-sheet content at temperatures above 60°C is evidenced by the diminished characteristic negative ellipticity at 218 nm. In contrast, the α-helical structures exhibit greater thermal stability, maintaining their distinctive negative ellipticity peaks at 208 nm and 222 nm up to 90°C. This indicates that α-helical regions retain their secondary structure at elevated temperatures where β-sheet structures have predominantly denatured. **G.** Flotation assay of StarLok1 protein across sucrose density gradients following ultracentrifugation at 240,000 × *g* for 1 hour in the presence or absence of liposomes. SDS-PAGE analysis was used to determine the protein content in low-density (0% sucrose), intermediate-density (24% sucrose), and high-density (30% sucrose) fractions. StarLok1 was incubated with liposomes of varying compositions (samples 1–4) and homogenized in the high-density sucrose layer (30% w/v) before centrifugation. In the absence of anionic membranes, StarLok1 predominantly accumulates in the low-density fraction. However, when incubated with liposomes containing anionic membranes, the migration of StarLok1 to the low-density fraction is delayed, as evidenced by detectable amounts of the protein in the intermediate-density fraction (24% sucrose).



Supplementary Fig. 6: Localization of StarLok1-GFP to the yeast PM. Data shows confocal micrographs of three independent clones of StarLok1-GFP expression in *S. cerevisiae* W303a expressing the indicated fluorescent fusion protein. StarLok2-GFP and StarLok3-GFP are cytosolic. Corresponding bright field images are shown to the right. Scale bars, 5 μm.

Supplementary Fig. 7: Sequence tree of the START domain superfamily. Maximum-likelihood phylogenetic tree (PhyML Q.pfam + R4 + F) of 242 START proteins with compositional homogeneity from 22 archaea (including 8 Asgards), 29 eukaryotes and 22 bacteria indicated at the tip labels colored in blue, green and magenta respectively. Branch support is indicated by aLRT SH-like values, with the circle size corresponding to the support value from 0.9 to 1. The tree was rooted at midroot for clarity. Tree scale is indicated as the number of substitutions per site. Yellow, purple or gray stripes near branch tips labels indicate Asgard sequences belonging respectively to the StarAsg1, StarAsg2/3 or undefined clades in the



Supplementary Fig. 8: Cavity size distribution of each orthologous group of 805 START domain spanning the tree of life. Each dot represents the largest cavity size identified in each START domain. Data are presented as box plots, showing the minimum, maximum, median, 25th percentile, and 75th percentile values for cavity sizes within each phylogenetic class.