PLOS COMPUTATIONAL BIOLOGY

RESEARCH ARTICLE

The very early evolution of protein translocation across membranes

AJ Harris[1,2]*, **Aaron David Goldman**[2,3]*

1 Key Laboratory of Plant Resources Conservation and Sustainable Utilization, South China Botanical Garden, Chinese Academy of Sciences, Guangzhou, China, **2** Department of Biology, Oberlin College and Conservatory, K123 Science Center, Oberlin, Ohio, United States of America, **3** Blue Marble Space Institute of Science, Seattle, Washington, United States of America

* aj.harris@inbox.com (AJ-H); agoldman@oberlin.edu (ADG)

Citation: Harris AJ, Goldman AD (2021) The very early evolution of protein translocation across membranes. PLoS Comput Biol 17(3): e1008623. https://doi.org/10.1371/journal.pcbi.1008623

Editor: Dan S. Tawfik, Weizmann Institute of Science, ISRAEL

Received: June 7, 2020

Accepted: December 10, 2020

Published: March 8, 2021

Data Availability Statement: All relevant data are within the manuscript and its Supporting Information files.

Funding: This work benefited greatly from a start-up allocation of time on Comet (TG-BCS190003) made to AJ-H and was funded by grants from the National Aeronautics and Space Administration (80NSSC19M0069), the National Science Foundation (MRI1427949), and the Joint NASA-NSF Ideas Lab on the "Origins of Life" (NSF Solicitation 16-570) made to ADG. The funders had no role in study design, data collection and

Abstract

In this study, we used a computational approach to investigate the early evolutionary history of a system of proteins that, together, embed and translocate other proteins across cell membranes. Cell membranes comprise the basis for cellularity, which is an ancient, fundamental organizing principle shared by all organisms and a key innovation in the evolution of life on Earth. Two related requirements for cellularity are that organisms are able to both embed proteins into membranes and translocate proteins across membranes. One system that accomplishes these tasks is the signal recognition particle (SRP) system, in which the core protein components are the paralogs, FtsY and Ffh. Complementary to the SRP system is the Sec translocation channel, in which the primary channel-forming protein is SecY. We performed phylogenetic analyses that strongly supported prior inferences that FtsY, Ffh, and SecY were all present by the time of the last universal common ancestor of life, the LUCA, and that the ancestor of FtsY and Ffh existed before the LUCA. Further, we combined ancestral sequence reconstruction and protein structure and function prediction to show that the LUCA had an SRP system and Sec translocation channel that were similar to those of extant organisms. We also show that the ancestor of Ffh and FtsY that predated the LUCA was more similar to FtsY than Ffh but could still have comprised a rudimentary protein translocation system on its own. Duplication of the ancestor of FtsY and Ffh facilitated the specialization of FtsY as a membrane bound receptor and Ffh as a cytoplasmic protein that could bind nascent proteins with specific membrane-targeting signal sequences. Finally, we analyzed amino acid frequencies in our ancestral sequence reconstructions to infer that the ancestral Ffh/FtsY protein likely arose prior to or just after the completion of the canonical genetic code. Taken together, our results offer a window into the very early evolutionary history of cellularity.

Author summary

Cellularity is an ancient, fundamental organizing principle of life. Central to cellularity is the cell membrane, which separates a cell from the outside environment. Cell membranes

analysis, decision to publish, or preparation of the manuscript.

Competing interests: The authors have declared that no competing interests exist.

contain proteins that perform a range of functions including transport of compounds across the membrane barrier, sensing the external environment, and performing certain metabolic activities that must occur in proximity to the membrane. Therefore, embedding proteins into membranes and secreting proteins across membranes is an essential aspect of cellularity, not to mention an essential aspect of life itself. One cellular system that accomplishes embedding proteins into membranes and secreting proteins across membranes is the signal recognition particle (SRP) system. The SRP system has a core consisting of the proteins, FtsY and Ffh, which derive from a single FtsY/Ffh ancestral protein. The system is also associated with a protein-based passageway, the Sec channel, for embedding proteins within the membrane or allowing them to pass through it. To study the SRP system and the central protein of the Sec channel, SecY, in early life, we reconstructed evolutionary trees from protein sequences. Based on these trees, we infer that the last universal common ancestor (LUCA) of life had an SRP system and SecY channel that were similar to those in extant organisms, while an earlier ancestor of the LUCA possessed a more rudimentary system for embedding and secreting proteins. Moreover, the ancestral Ffh/FtsY protein probably arose prior to or soon after the final amino acids were added to the standard genetic code.

Introduction

The emergence of cellular organisms from non-cellular replicators is considered one of the major transitions, or key innovations, in evolutionary history [1] that may have been the prerequisite for the earliest speciation events [2] and for subsequent colonization of all the habitable environments on Earth [3,4]. Cellular organisms must build and maintain a cell membrane as well as both populate it with integral membrane proteins and secrete proteins out of it. In very early cellular life forms, the ability to embed proteins within the cell membrane would have been essential for a number of important biological processes, such as controlled cell division and any metabolism that required the import of large metabolites and that was reliant on ATP synthesis by proton motive force.

While several different systems have evolved that facilitate protein translocation into and across membranes, the signal recognition particle (SRP) system and its associated Sec secretion channel (illustrated in Fig 1) are particularly ubiquitous across the tree of life. The SRP system relies on two distinct, paralogous proteins. One of these proteins is a cytosolic protein known as Ffh in Bacteria and SRP54 in Archaea and Eukarya (hereafter Ffh) that binds a signal sequence in a nascent protein and guides the ribosome synthesizing the protein to the membrane [5–7]. The other protein is a membrane bound receptor for the Ffh-ribosome complex that is known as FtsY in Bacteria and SRα or SR receptor in Archaea and Eukarya (hereafter FtsY) [8–10]. In many Bacteria, the SRP system comprises only these two proteins and an RNA called SRP RNA, which facilitates the assembly and disassembly of the protein-synthesizing complex at the membrane [11,12]. In Archaea and Eukarya, the system includes these three components as well as other accessory proteins that are bound to Ffh and FtsY [13].

Ffh and FtsY belong to the SIMIBI class of proteins, which consist of dimerizing pairs that form GTPase domains, or G domains, at their interfaces [14–16]. Ffh and FtsY also each possess a helical bundle domain, which is situated in close proximity to the GTPase domain and is sometimes referred to as the N domain because it is closer to the N-terminus of the protein than the GTPase domain. Together, the helical bundle and GTPase domains comprise a conserved region shared by both FtsY and Ffh and are often collectively referred to as the NG

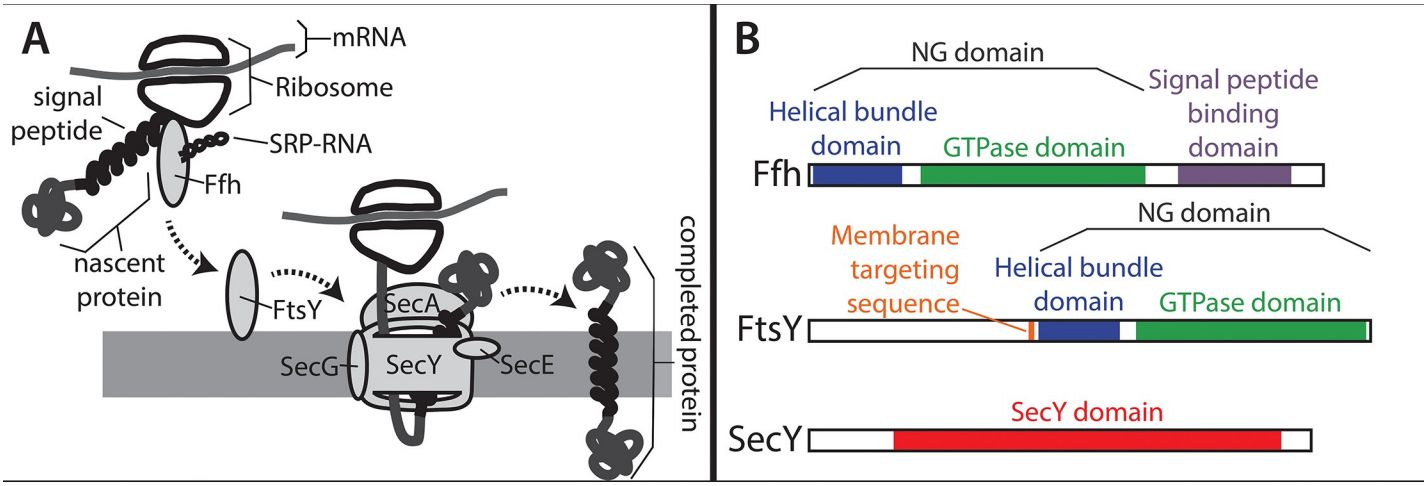

Fig 1. An overview of the canonical SRP/Sec membrane translocation system. SRP is responsible for binding signal peptides in nascent proteins and delivering the nascent protein and ribosome to the membrane surface. A) A signal peptide within the nascent protein is bound by Ffh. Ffh forms a complex on the membrane surface with FtsY and SRP-RNA along with the nascent protein and the ribosome that is synthesizing it. After the SRP complex is formed, the signal peptide on the nascent protein enters the Sec channel and newly added protein is translocated across the membrane while it is being synthesized. Upon release of the protein into the membrane, the signal peptide becomes a transmembrane domain. The Sec channel is a multiprotein complex composed of three proteins, SecYEG. Many auxiliary proteins promote this process, most notably SecA, which associates with SecYEG and drives translocation through the channel via ATP hydrolysis. B) Domain architectures of the three proteins analyzed in this study are shown based on their E. coli homologs. Both Ffh (Uniprot ID = P0AGD7) and FtsY (Uniprot ID = P10121) contain a a helical bundle domain (blue) and a GTPase domain (green), which together are called the NG domain. Ffh also contains a region called the M domain (purple), which binds the signal sequence of a nascent protein. FtsY also contains a membrane targeting sequence, or MTS, domain which binds to the surface of the membrane (orange). SecY (Uniprot ID P0AGA2) contains up to ten predicted transmembrane domains, but only one domain annotated by the pfam database, called "SecY" (red).

https://doi.org/10.1371/journal.pcbi.1008623.g001

domain. Ffh also possesses a C-terminal, methionine-rich M domain [17,18], while FtsY has a membrane targeting sequence (MTS) domain within or upstream of the helical bundle domain [19,20].

A protein that is targeted to the membrane by the SRP system will contain at least one signal peptide sequence. Once the ribosome synthesizes the signal sequence, Ffh will bind to it through the M domain [17,18], and this typically pauses the translation process. Subsequently, the ribosome-bound Ffh forms a heterodimer with FtsY on the membrane surface. The interaction between Ffh and FtsY is likely assisted by the SRP RNA in response to signal peptide binding in the M domain of Ffh [21]. In Bacteria and Archaea, the complex comprising FtsY, Ffh, SRP RNA, and the ribosome and nascent protein, forms on the interior surface of the plasma membrane (or the inner plasma membrane in Gram-negative bacteria), while in eukaryotes, it occurs on the exterior surface of the endoplasmic reticulum. Though FtsY is located on the surface of the membrane, it is not bound to the membrane through a typical hydrophobic transmembrane domain. Instead, the MTS domain is composed of positively charged amino acids, which bind to negative charges in phospholipid head groups on the surface of the membrane [22].

The interaction between Ffh and FtsY facilitates the transfer of the signal peptide sequence from Ffh to a translocation channel called SecYEG in Bacteria and Archaea, and Sec61 in eukaryotes (for review, see [23]). Within the translocation system, SecY is the largest protein subunit and forms the actual transmembrane channel. It is the only Sec subunit known to be present throughout the tree of life [24]. In addition to the channel, the Sec system as it has been described in *E. coli* also includes a protein that drives protein translocation by way of ATP hydrolysis, SecA, and a handful of other auxiliary proteins that promote the function of the channel or are used in special cases [25]. Via the translocation system, the signal sequence

is transferred to the interior of the membrane and translation resumes, with the new sequence being fed through the translocation channel. A protein that is destined to be secreted through the membrane usually contains an N-terminal signal sequence that is removed once it has been synthesized by a signal peptidase protein. Proteins that are destined to be embedded in the membrane contain one or more internal signal sequences that become transmembrane domains in the finished protein. After transfer of the signal peptide to the translocation channel, the Ffh/FtsY/SRP RNA complex dissociates. This dissociation is triggered by the hydrolysis of GTP molecules [26,27] that are bound at the interface of the Ffh-FtsY heterodimer through the GTPase domains contained in both proteins [14–16].

Both the SRP system and the SecY protein family are regarded as ancient and their nearly ubiquitous presence in extant organisms suggests that they had originated at least by the time of the last universal common ancestor of life, the LUCA [28]. However, there is no comprehensive protein phylogeny of the SRP system with systematic, robust taxonomic sampling and there is no current phylogeny of SecY (but see [29]) that takes into account recent advances in biodiversity discovery, taxonomic classifications, and newly available genome sequences of archaeal and bacterial species [30–33]. Therefore, in this study, we performed phylogenetic analyses of the SRP system and SecY proteins to elucidate the very early evolutionary history of these families. Our reconstructed phylogeny indicates that ancestors of Ffh, FtsY, and SecY were all present by the time of the LUCA. Furthermore, within the phylogeny, the Ffh and FtsY subtrees are connected by a well-supported pre-LUCA branch, supporting the notion that this protein family predates the origins of the LUCA. Based on the phylogeny, we used ancestral sequence reconstructions combined with protein function prediction to show that the ancestral Ffh and FtsY protein that predated the LUCA (hereafter, ancestral Ffh/FtsY) potentially arose before the completion of the genetic code, but that it likely contained many functional components of the modern SRP system allowing it to facilitate a rudimentary form of protein translocation and secretion across cell membranes. Together, these results provide a framework for understanding the very early evolution of membrane translocation, a foundational system required by all cellular organisms.

Results and discussion

Survey of SRP and Sec proteins across Bacteria and Archaea

We identified homologs of FtsY, Ffh, and SecY according to protein BLAST searches (BLASTp; [34]) and selected 20 sequences or the maximum number available (whichever was smaller; see Materials and Methods) from each bacterial and archaeal phylum or super phylum for inclusion in downstream analyses. Our final protein dataset for the SRP system comprised 938 protein sequences representing 18 bacterial and four archaeal phyla and superphyla. Of these protein sequences, 348 were of FtsY and 330 were of Ffh. Additional accessions represented two other paralogous protein families identified using BLASTp and occurring primarily within a subset of Bacterial lineages. Of these additional accessions, 46 comprised type III secretion system proteins and 214 represented the flagellar protein, FlhA (see explanation in Methods). The alignment of all 938 sequences consisted of 1666 characters (Fig 2 and S3 File) and had 21.1% average overall pairwise identity and 36.2% average positive pairwise identity based on the BLOSUM62 scoring matrix. Almost all bacterial and archaeal phyla and superphyla that were well-represented in the Genbank protein database were also well-represented in our dataset. The only exception was the bacterial phylum, Elusimicrobia, for which we obtained only three accessions; one each for FtsY, Ffh, and the type III secretion system. Other phyla and superphyla that were less well-represented in Genbank generally yielded fewer than 20 sequences.

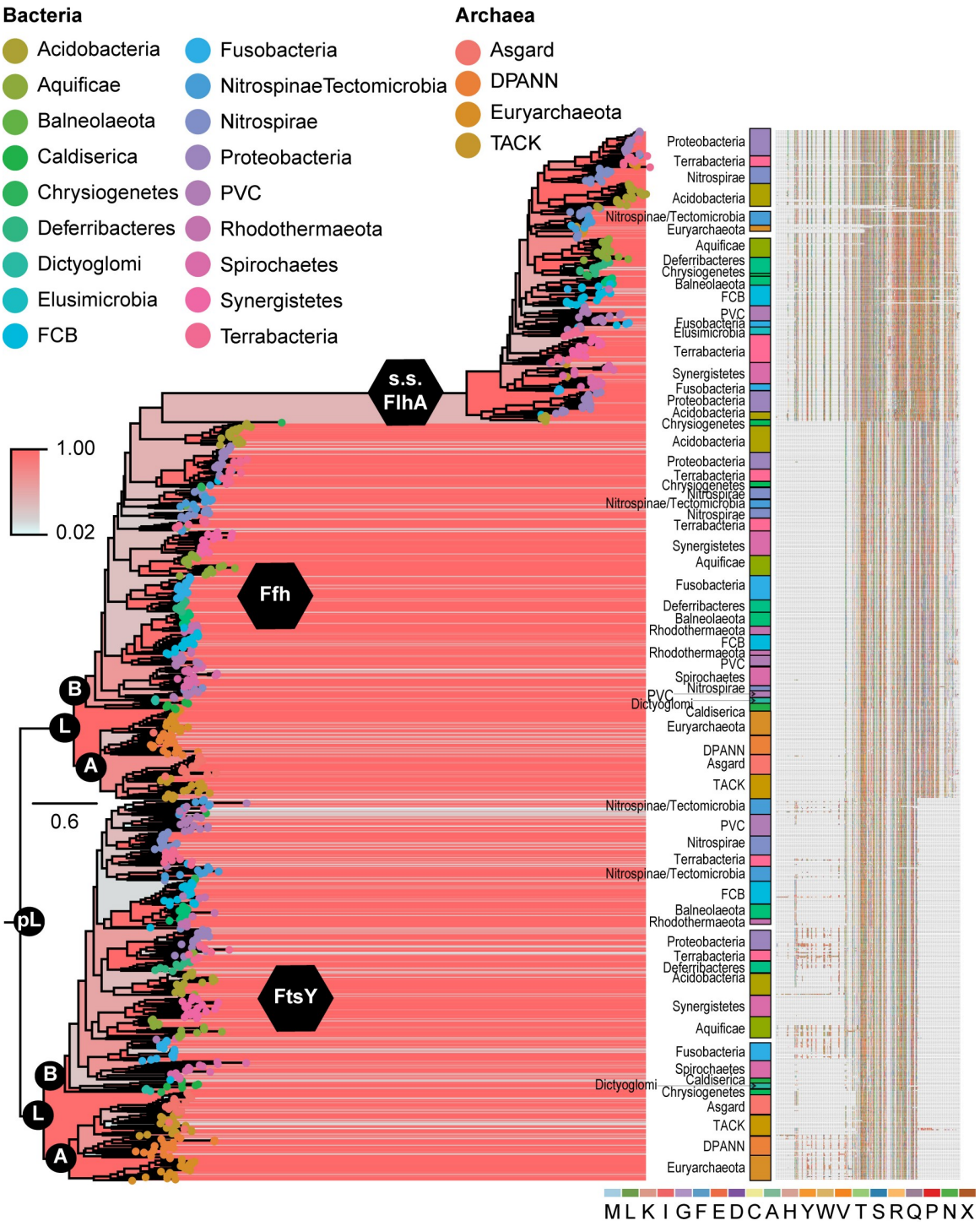

Fig 2. Maximum clade credibility tree resulting from Bayesian analysis of accessions of SRP system proteins, FtsY and Ffh, plus the flagellar protein, FlhA, and the type III secretion system proteins, EscV, YscV, and HrcV. Support values are indicated by color from red (higher) to blue (lower). Labeling of internal nodes indicates ancestral Ffh/FtsY (pre-LUCA; pL), the LUCA (L), and crown clades of Bacteria (B) and Archaea (A). Major clades comprising phyla or superphyla of Bacteria or Archaea are labeled to the right of terminals, notwithstanding one or a few nested accessions of other groups. The multiple sequence alignment from which we generated the phylogeny is shown to the far right. The complete phylogeny and multiple sequence alignment are available as supplementary files in newick and fasta formats, respectively.

https://doi.org/10.1371/journal.pcbi.1008623.g002

For SecY, we obtained 355 protein sequences representing the same taxonomic diversity of Bacteria and Archaea as for the SRP system. The sequences of SecY were available at roughly the rates expected based on taxonomic representation in Genbank, and Elusimicrobia was not under-represented. The alignment of these 355 sequences comprised 835 characters (Fig 2 and S4 File) and had an overall pairwise identity of 33.1% and a pairwise identity of 51.6% based on BLOSUM62 scoring.

We also attempted to use similar methods to obtain sequences of signal peptidase proteins, which cleave N-terminal signal sequences from translocated proteins. Previous literature has indicated that Signal Peptidase I, called LepB in Bacteria, was present in the LUCA [35]. One COG from the eggNOG database [36], COG0681, contains homologs of Signal Peptidase I and is found in both Bacteria and Archaea. However, a BLASTp search of *H. volcanii* proteins using LepB *E. coli* as a query yielded a poor match (query coverage = 19%, e value = 0.28) that failed a reciprocal best hit test. An additional search of the entire Uniprot database of archaeal proteins using LepB of *E. coli* as a query yielded only short hits, with the top hit consisting of only 30 amino acids.

We obtained a similar result for Signal Peptidase II, called LpsA in Bacteria. Signal Peptidase II is represented by eggNOG cluster COG0597, which contains both bacterial and archaeal proteins. A BLASTp search of *H. volcanii* proteins using a sequence of LpsA from *E. coli* as a query yielded no hits at all. However, a search of the entire Uniprot database of archaeal proteins using *E. coli* LpsA yielded a top hit to a putative protein from an archaean, *Halobellus sp. Atlit-31R* (NCBI Taxonomic ID = 2282130) that passed a reciprocal best hit test against the *E. coli* genome. When we applied this *Halobellus* protein as a BLAST query against the entire Uniprot database of archaeal proteins, we obtained only 82 significant hits under our e-value threshold. None of these protein sequences represented reviewed proteins, and all of them were from uncultured or unclassified taxa. Even if these results represent archaeal orthologs of LpsA, their scarcity is more consistent with a few horizontal gene transfer events than inheritance of Signal Peptidase II by descent from the LUCA. Therefore, neither Signal Peptidase I nor Signal Peptidase II warranted further phylogenetic analysis.

Additionally, we sought to determine if SecA of the Sec translocation channel may have been present in the LUCA. SecA drives the translocation of proteins through the membrane via ATP hydrolysis. SecA belongs to the eggNOG cluster, COG0653, which contains both bacterial sequences and archaeal sequences. The sequence of SecA protein from *E. coli* did not have a detectable homolog in *H. volcanii* based on a BLASTp search. A search of the entire Uniprot database of archaeal proteins using SecA of *E. coli* yielded a top hit to a putative protein from an unclassified archaean, Natrialbaceae archaeon XQ-INN 246 (NCBI Taxonomic ID = 2419781), that passed a reciprocal best hit test against the *E. coli* genome. However, this protein from Natrialbaceae yielded only 56 significant BLAST hits when used as a query against the entire Uniprot database of archaeal proteins, and many of the hits were very short compared to the query. None of these protein sequences represented reviewed proteins. As with Signal Peptidase II, if these results represent orthologs, their scarcity is more consistent with a few horizontal gene transfer events. Thus, we determined that the SecA family did not merit further phylogenetic analysis in this study.

Phylogenetic evidence for early ancestors of Ffh, FtsY, and SecY

The Bayesian phylogenetic reconstruction for the SRP system showed a generally well-resolved tree with two major lineages comprising FtsY and Ffh, each with clear bacterial and archaeal clades (Fig 2 and S5 File). This tree, therefore, provides very strong evidence that an ancestral version of Ffh and an ancestral version of FtsY were both present in the genome of the LUCA

and that a common ancestor of both proteins, an ancestral Ffh/FtsY, was present prior to the LUCA. Additionally, a mixed clade of FlhA and type III secretion system proteins arose within the bacterial Ffh clade on a very long evolutionary branch. Neither the type III secretion system proteins nor FlhA proteins were resolved as monophyletic groups, which is also consistent with mixed clusters of these proteins obtained in Clustal Omega (S1 File; e.g., cluster 2).

Within the phylogeny of the SRP system proteins, most protein accessions representing individual phyla and superphyla formed monophyletic groups. Additionally, many relationships between superphyla and phyla were consistent between subtrees of FtsY and Ffh, such as the relationship between the Asgard and TACK archaeal lineages, between the bacterial lineages of Aquificae and Synergistetes, as well as among Acidobacteria, Proteobacteria, and Terrabacteria. However, relationships within the clade of FlhA and type III secretion system proteins were less consistent with the other two subtrees. Across the entire tree, there was little phylogenetic evidence for horizontal gene transfers across taxonomic domains; namely, only one archaeal lineage was nested within the clade of bacterial FlhA and type III secretion system proteins.

The phylogenetic reconstruction of SecY resulted in a very highly supported tree (Fig 3 and S6 File). The tree consisted of two major clades corresponding to the bacterial and archaeal domains, therefore representing robust evidence for an ancestor of SecY being present in the genome of the LUCA. Within the tree, most bacterial and archaeal phyla and superphyla formed clades, and the longest branches were in the Elusimicrobia clade. Relationships among phyla and superphyla were somewhat inconsistent with those in the FtsY and Ffh subtrees. For example, accessions representing Aquificae and Synergistetes were only distantly related. However, similarities are also abundant, such as Caldiserica and Dictyoglomi being among the earliest diverging lineages in both the SecY and Ffh/FtsY trees, and accessions of Terrabacteria and Proteobacteria being closely related. We observed no phylogenetic evidence in the SecY tree of horizontal gene transfers across taxonomic domains.

Beyond the principal results that SecY was present in the LUCA and an ancestor of Ffh/FtsY existed prior to the LUCA, we observed an interesting trend within the Elusimicrobia phylum. Specifically, we found that few Elusimicrobia possess homologs of the SRP system and that several Elusimicrobia accessions of SecY were on long phylogenetic branches, which suggest considerable, possibly rapid, evolutionary change (Fig 3). It is possible that loss of FtsY and Ffh in some Elusimicrobia has driven the evolution of SecY to compensate, such as to enable SecY to interact with other proteins that have taken over the tasks of the SRP system. Our results indicate that more targeted studies of the SRP system and SecY in Elusimicrobia may be merited.

Sequence-based characterization of ancestral SRP system proteins and SecY

Our phylogenetic results clearly support the antiquity of both the Ffh/FtsY and the SecY protein families. Therefore, we sought to further examine the early evolution of these proteins by characterizing the ancestral sequences corresponding to ancient nodes within their respective phylogenetic trees. To accomplish this, we performed ancestral state reconstructions to infer the sequences of early Ffh, FtsY, and SecY proteins. Ancestral state reconstruction does not, on its own, provide a guide for the inclusion of indels, or gaps, in the reconstructed sequences. Therefore, we reconstructed ancestral sequences using a broad range of probability thresholds for inferring amino acids versus gaps.

The sequence reconstructions of the Ffh and FtsY proteins in the LUCA, as well as of the ancestral Ffh/FtsY protein, all predictably varied in length based on the threshold for including

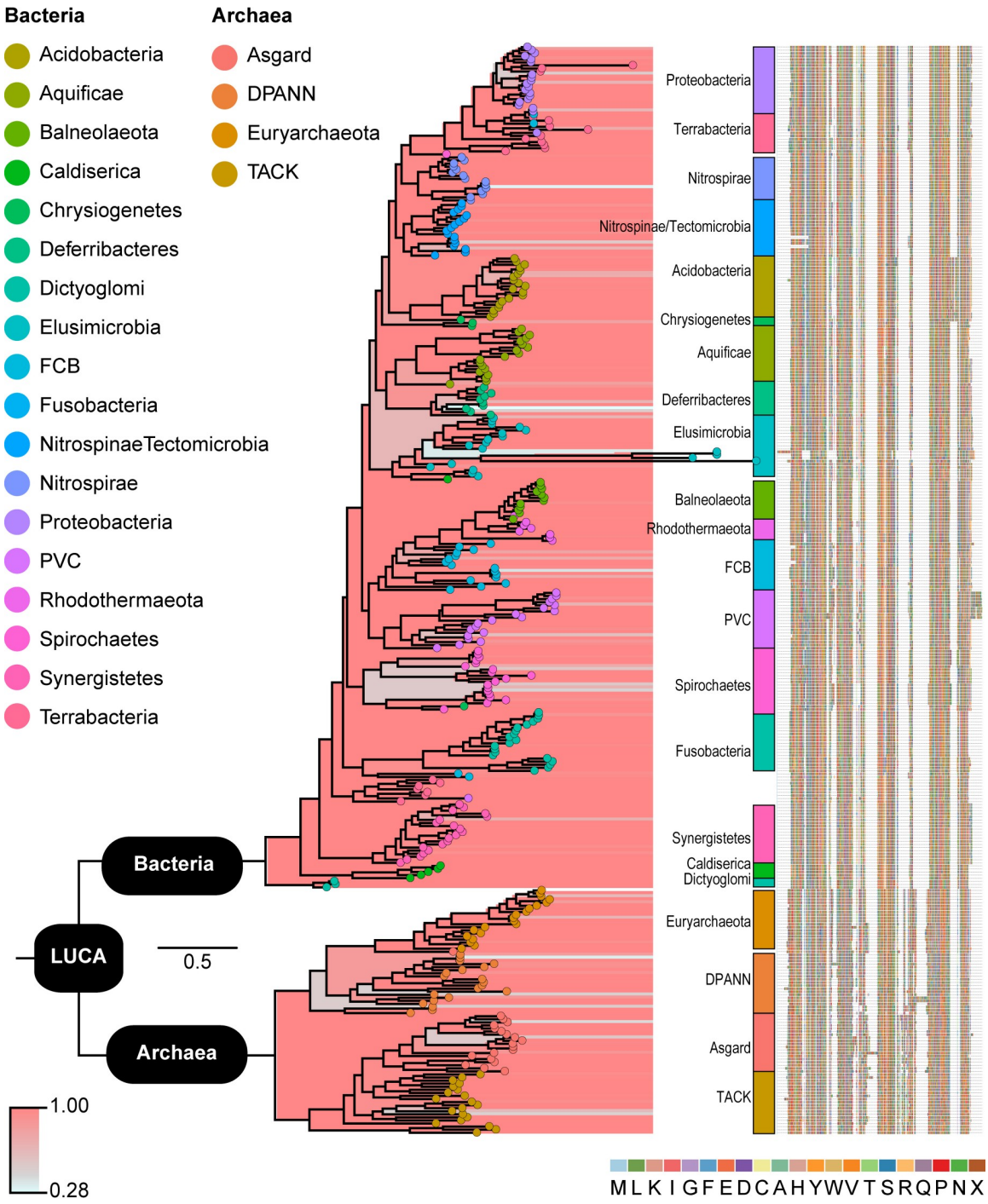

Fig 3. Maximum clade credibility tree resulting from Bayesian analysis of accessions of SecY. Support values are indicated by color from red (higher) to blue (lower). Major clades comprising phyla or superphyla of Bacteria or Archaea are labeled to the right of terminals, notwithstanding one or a few nested accessions of other groups. The multiple sequence alignment from which we generated the phylogeny is shown to the far right. The complete phylogeny and multiple sequence alignment are available as supplementary files in newick and fasta formats, respectively.

https://doi.org/10.1371/journal.pcbi.1008623.g003

gaps (S7 and S8 Files). With a 90% probability threshold for inferring gaps, the average lengths of 100 reconstructed sequences in the SRP system were 763 (SD: 33.80) for ancestral Ffh/FtsY and, in the LUCA, 681 (SD: 118.10) for FtsY and 665 (SD: 124.14) for Ffh. With only a 10%

Table 1. Sequence-based function prediction for ancestors of Ffh, FtsY, and SecY.

Protein	Inferred Origin	PFAM (alignment locus|e-value)	Protein localization based on SOSUI (hydrophobicity index)
Ancestral Ffh/FtsY	pre-LUCA	SRP54-type protein, helical bundle domain (6–81|1.4e-76) SRP54-type protein, GTPase domain (95–279|1.7e-15)	cytoplasmic (0.0792)
Ffh	LUCA	SRP54-type protein, helical bundle domain (5–80|8.0e-21) SRP54-type protein, GTPase domain (94–278|1.3e-77) Signal peptide binding domain (309–408|8.7e-43)	cytoplasmic (-0.3597)
FtsY	LUCA	SRP54-type protein, helical bundle domain (96–175|7.2e-11) SRP54-type protein, GTPase domain (189–373| 1.6e-76)	cytoplasmic (-0.4542)
SecY	LUCA	SecY translocase (58–377|4.9e-101)	trans-membrane(0.9908)

https://doi.org/10.1371/journal.pcbi.1008623.t001

threshold, the averages were 289 (SD: 28.98), 427 (SD: 125.57), and 435 (SD: 5.48) respectively. For the 50% threshold, the values were intermediate between these.

The sequences of ancestral Ffh/FtsY obtained using the 10%, 50%, and 90% thresholds for inferring gaps all possessed helical bundle, GTPase, and MTS domains, while the sequences reconstructed using the 50% and 90% cut-offs also included extended N- and C-termini. An extended C-terminal region is typical of Ffh proteins and includes the M domain that contains a signal peptide binding motif. We recovered this extended C-terminal region in the sequence of ancestral Ffh/FtsY protein reconstructed using a 90% threshold for indel inclusion. An extended N-terminal region occurs in some FtsY proteins, but it appears to have arisen several times independently in evolutionary history based on our phylogeny and is not conserved in either length or sequence (Fig 2, S3 File). Therefore, unsurprisingly, the N-terminal region of ancestral Ffh/FtsY obtained using the 90% threshold did not contain any homologs to domains or motifs within the PFAM database [37,38]. Overall, the 10% threshold for indel inclusion yielded a sequence for ancestral Ffh/FtsY that most closely resembled the NG domain shared among all modern Ffh and FtsY proteins, so we chose this sequence for downstream analyses except as noted.

Domains and motifs of the sequence reconstruction resulting from the 10% indel threshold show that the ancestral Ffh/FtsY protein very likely contained the helical bundle (e-value: 1.4e-76) and GTPase domains (e-value: 1.7e-15; Table 1 and Fig 4) that are found in modern Ffh and FtsY proteins. The helical bundle domain of ancestral Ffh/FtsY also appears to contain a

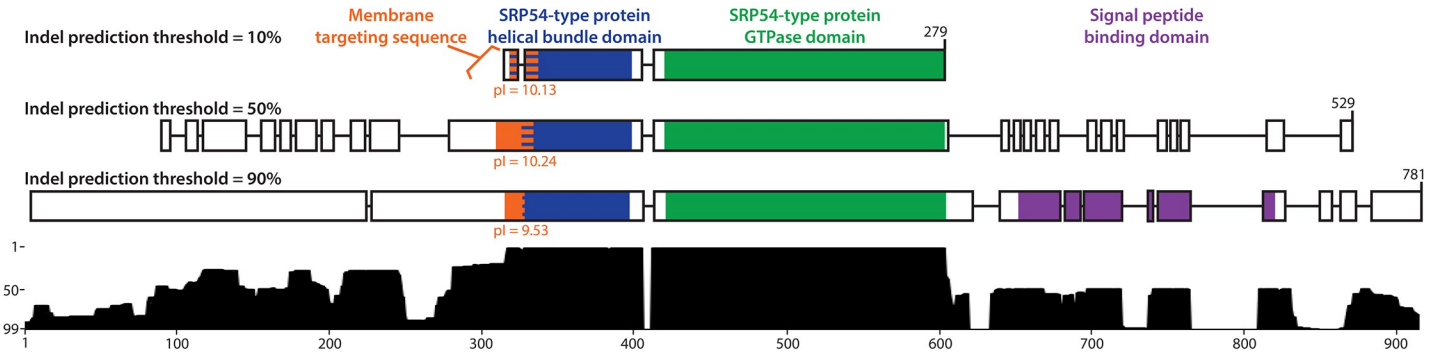

Fig 4. Functional annotation of ancestral Ffh/FtsY. Domains and motifs from the Pfam Database are mapped onto sequences of the reconstructions of ancestral Ffh/FtsY using three different thresholds for inference of gaps (top sequence = 10%, middle sequence = 50%, and bottom sequence = 90%). The histogram below represents the frequency that an amino acid (as opposed to a gap) appears in each position across all thresholds for inferences of gaps (1%-99%). There is agreement across all three sequence reconstructions that ancestral Ffh/FtsY contained the GTPase domain (green) and the four helical bundle domain (blue) that are typical of both Ffh and FtsY proteins. Sequence alignment to *E. coli* Ffh (see Methods) revealed an MTS domain (orange) in all three reconstructed ancestral sequences with characteristically basic isoelectric points. In addition to these domains, the ancestral sequence that was reconstructed using a 90% threshold for inference of gaps contains a C-terminal peptide binding domain (purple), but this domain is typical only of modern Ffh proteins and is not observed in modern FtsY proteins.

https://doi.org/10.1371/journal.pcbi.1008623.g004

membrane targeting sequence (MTS) motif in its N-terminal helix that has an isoelectric point similar to a recently characterized MTS motif in *E. coli* (10.13 and 12.01, respectively). Thus, this suggests that the ancestral Ffh/FtsY protein may have been able to bind to membranes through electrostatic attraction as in modern FtsY proteins. Similar to the ancestral Ffh/FtsY and modern proteins, the Ffh and FtsY of the LUCA both possessed a helical bundle domain (e-values: 8.0e-21, 7.2e-11, respectively) and a GTPase domain (e-values: 1.3e-77, 1.6e-76, respectively) (Table 1). The sequence reconstruction for Ffh was also inferred to possess a C-terminal M domain (e-value: 8.7e-43; Table 1).

In addition to characterizing the functional domains and motifs within these ancestral proteins, we also used SOSUI [39] to investigate whether they were most likely to be cytoplasmic (i.e., soluble in an aqueous solvent) or embedded within a membrane. SOSUI predicts whether a protein is located in the membrane or is cytoplasmic based on its chemical properties and constituent amino acids, especially focusing on hydrophobicity [39]. Analyses in SOSUI showed that the reconstructed sequence of ancestral Ffh/FtsY as well as FtsY and Ffh of the LUCA were cytoplasmic proteins with average hydrophilicity indices of 0.0792, 0.4541, and -0.3597, respectively (Table 1). Note that while FtsY is membrane bound, it is not an integral membrane protein, and is therefore expected to have a hydrophilicity index typical of cytoplasmic proteins.

For the ancestor of SecY within the LUCA, the average lengths of reconstructed sequences were 393 (SD 0.00), 598 (SD: 24.44), and 696 (SD: 23.45) for the 10%, 50%, and 90% thresholds for inferences of gaps. We used the 10% threshold for downstream analyses to avoid biases in our treatments of this protein family compared to the FtsY and Ffh ancestors and because the unaligned sequence lengths that were recovered using the 10% threshold showed no variation among 100 reconstructions, suggesting strong support for the placement of gaps among the results. The sole PFAM annotation for the reconstructed SecY in the LUCA shows that the majority of this ancestral sequence aligns to the SecY translocase domain family (e-value: 4.9e-101; Table 1). Moreover, the reconstructed sequence of SecY for the LUCA was predicted to be a membrane protein with an average hydrophobicity of 0.9908 (Table 1) based on SOSUI, and this is consistent with SecY in modern species.

Structure-based characterization of ancestral SRP system proteins

In order to further characterize the functions of the ancestral Ffh/FtsY protein as well as Ffh and FtsY of the LUCA, we performed protein structure prediction and structure-based function prediction on the reconstructed sequences using I-TASSER [40], a highly accurate method that performs both of these tasks. The predicted structure for the ancestral Ffh/FtsY protein in I-TASSER showed similarities to known structures of FtsY and Ffh from all three domains of extant life that have been solved by X-ray diffraction (e.g., PDB IDs 3ndb, 3dm5, 2og2, and 4ak9; [18,41,42]). Proteins from the SRP systems of extant organisms with the greatest predicted structural similarity to the inferred ancestral protein were Ffh proteins from the archaean, *Methanocaldococcus jannaschii* (Jones et al. 1984) Whitman 2002 strain DSM 2661 (RMSD = 1.16; PDB ID 3ndbB) and *Pyrococcus furiosus* Erauso *et al.* 1993 (RMSD = 1.37; PDB ID 3dm5A), and the bacterium, *Thermus aquaticus* Brock & Freeze, 1969 (RMSD = 1.94; PDB ID 1ng1). Structure-based functional characterization (Table 2) predicted that ancestral Ffh/FtsY could bind GTP (GO:0005525), SRP RNA (GO:0008312), and proteins (GO:0005515) and would be capable of nucleoside-triphosphatase activity (GO:0017111).

Ancestral Ffh/FtsY was also predicted to be localized to the plasma membrane (GO:0005886) just as FtsY of modern species is (S1 Table). However, in contrast to both FtsY and Ffh of modern species (e.g., *E. coli* and *H. volcanii*; S1 Table), I-TASSER did not predict

Table 2. Structure and structure-based function prediction for ancestors of Ffh, FtsY, and SecY.

Protein	Inferred Origin	Most Structurally Similar Protein (PDB accession)	RMSD score	Molecular Function (GO term| score)	Biological Process (GO term| score)	Cellular Component (GO term|score)
FtsY-Ffh Ancestor	pre-LUCA	Signal recognition 54 kDa protein (3NDBB)	1.16	GTP binding (GO:0005525| 1.00) nucleoside-triphosphatase activity (GO:0017111|1.00) 7S RNA binding (GO:0008312| 0.98) protein binding (GO:0005515|0.92)	SRP-dependent cotranslational protein targeting to membran (GO:0006614|1.00) cell division (GO:0051301|0.71) cell cycle (GO:0007049|0.71)	signal recognition particle, endoplasmic reticulum targeting (GO:0005786|0.98) plasma membrane (GO:0005886|0.71)
Ffh	LUCA	Signal recognition 54 kDa protein (3DM5A)	0.67	GTP binding (GO:0005525|0.91) nucleoside-triphosphatase activity (GO:0017111|0.86) 7S RNA binding (GO:0008312| 0.69) protein binding (GO:0005515|0.32)	SRP-dependent cotranslational protein targeting to membrane (GO:0006614|0.91)	signal recognition particle, endoplasmic reticulum targetin (GO:0005786|0.69)
FtsY	LUCA	SRP receptor alpha subunit (6FRKy)	2.24	7S RNA binding (GO:0008312| 0.92) GTPase activity (GO:0003924|0.61) signaling receptor activity (GO:0038023|0.41) protein binding (GO:0005515|0.40) GTP binding (GO:0005525|0.37)	SRP-dependent cotranslational protein targeting to membrane (GO:0006614|0.92) cell division (GO:0051301|0.63) cell cycle (GO:0007049|0.63)	intrinsic component of plasma membrane (GO:0031226|0.40) cytosol (GO:0005829|0.40) signal recognition particle, endoplasmic reticulum targeting (GO:0005786|0.37)
SecY	LUCA	Protein translocase subunit SecY (5NCOg)	1.07	P-P-bond-hydrolysis-driven protein transmembrane transporter activity (GO:0015450|0.95) protein binding (GO:0005515|0.43) signal sequence binding (GO:0005048|0.43)	transmembrane transport (GO:0055085|0.92) SRP-dependent cotranslational protein targeting to membrane (GO:0006614|0.43)	integral component of membrane (GO:0016021|0.92) plasma membrane (GO:0005886|0.73) Ssh1 translocon complex (GO:0071261|0.43)

1 Top scoring GO terms from among consensus terms

https://doi.org/10.1371/journal.pcbi.1008623.t002

any GO terms suggesting cytosolic or cytoplasmic localization of ancestral Ffh/FtsY. Additionally, the predicted structure of ancestral Ffh/FtsY showed unexpected predictions for the biological processes of cell division (GO:0051301) and the cell cycle (GO:0007049), which are not annotated for FtsY and Ffh sequences of representative modern species (S1 Table). Within the LUCA, structural predictions suggest that both FtsY and Ffh facilitated cotranslational protein targeting to the cell membrane (GO:0006614), and FtsY is predicted to be localized to both a cytosolic and membrane component of the cell (GO:0005829, GO:0031226, respectively). No cellular component predictions for Ffh were recovered among the consensus GO terms produced by I-TASSER. Similar to ancestral Ffh/FtsY, the FtsY ancestor in the LUCA is also predicted to have functions in cell division and the cell cycle.

The predictions that the pre-LUCA Ffh/FtsY ancestor and FtsY in the LUCA are associated with cell division and the cell cycle are probably spurious. They come from local structural similarities to well-characterized proteins. While there is a direct relationship between protein structure and molecular function, there is only an indirect relationship between protein structure and the biological processes that the protein is associated with. That said, while FtsY of modern species is generally not known to operate directly within the cell cycle (S1 Table; [43]), it occurs in an operon with two vital cell division proteins, FtsE and FtsX, in some bacterial species, such as *E. coli*. Notably, an evolutionary explanation for the inclusion of FtsY within the *ftsYEX* operon is lacking [43,44]. Thus, it is possible that FtsY had an ancient function in cell division in addition to its role in membrane translocation and embedding of proteins.

The predicted structure of ancestral Ffh/FtsY also showed close structural alignment to both the Ffh protein (RMSD = 2.3) and FtsY protein (RMSD = 1.5) from *E. coli* that were solved in complex with GDP and an SRP-RNA tetraloop (PDB ID 4c7o). This structural alignment produced a model of ancestral Ffh/FtsY as a homodimer in complex with SRP-RNA and two guanosine nucleotides (Fig 5). In this model, the MTS domains are present in the N-terminal helices of each subunit and are located at the region of the complex that would likely make contact with the membrane. The GTPase domains make contact to the SRP-RNA and form a common active site that accommodates both GDP molecules. Taken together, these sequence- and structure-based characterizations of ancestral Ffh/FtsY suggest that this protein was, at minimum, capable of forming a homodimer in complex with an SRP-RNA, binding and hydrolyzing GTP, and binding to the membrane through MTS domains.

The predicted structure of SecY in the LUCA depicts a helix-rich protein (S1 Fig) with the greatest structural similarity to SecY of *Geobacillus thermodenitrificans* (Manachini et al. 2000)

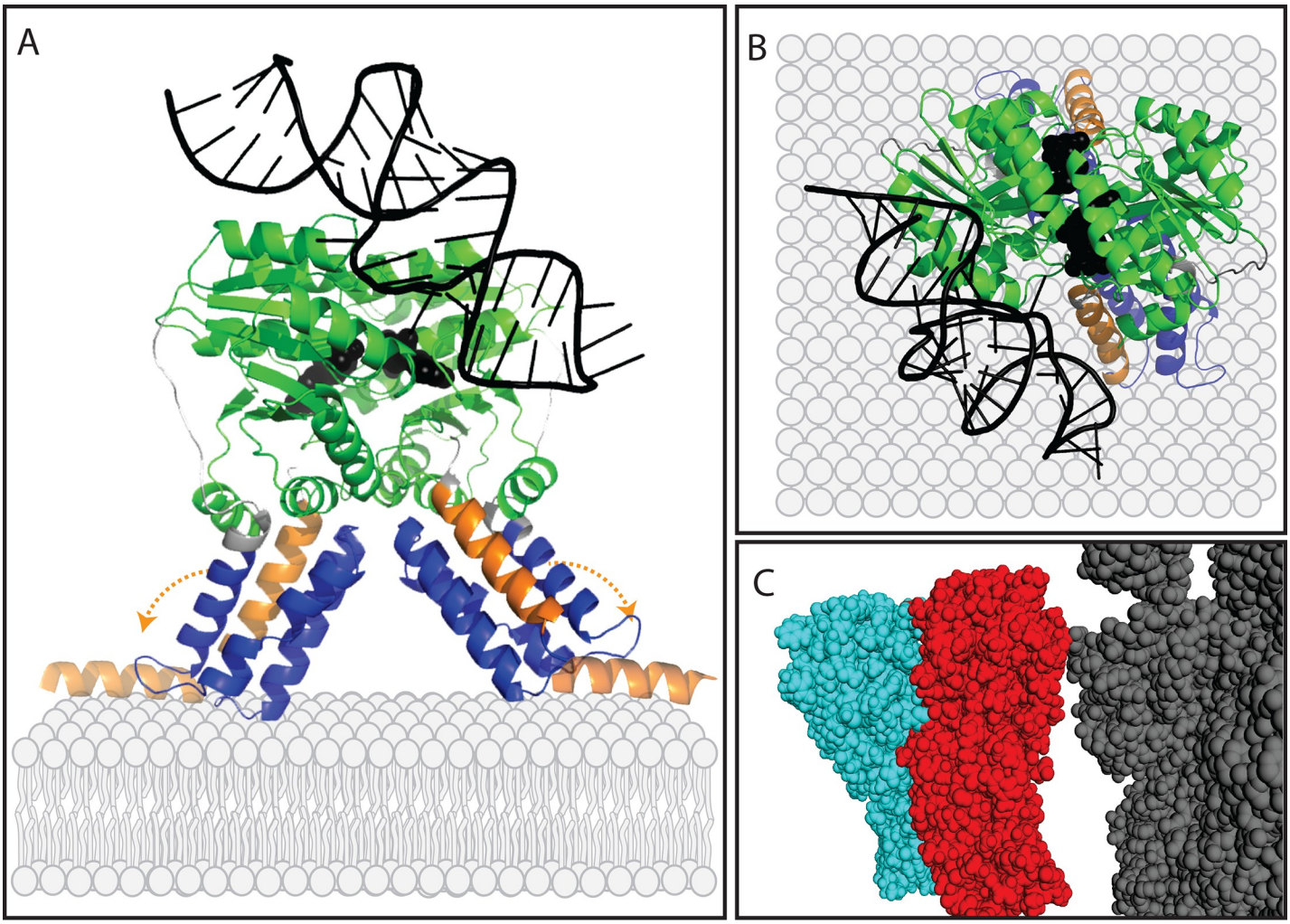

Fig 5. Predicted structural model of ancestral Ffh/FtsY. Predicted three dimensional structures of ancestral Ffh/FtsY are aligned to extant Ffh and FtsY proteins from x-ray diffraction structures (PDB IDs 4c7o and 5nco; [16,70]). In panels A and B, a potential homodimer of ancestral Ffh/FtsY is shown in complex with an SRP-RNA. The GTPase domains (green) form a GTP binding site that accommodates two GDP molecules (black), while the helical bundle domains (blue) and MTS domains (orange) are oriented toward the membrane. Orange arrows in panel A indicate the ability of the N-terminal MTS domains to reorient and attach to the interior surface of the membrane. Panel C depicts a potential homodimer of ancestral Ffh/FtsY (red and cyan) interacting with the 23S ribosomal subunit (gray).

https://doi.org/10.1371/journal.pcbi.1008623.g005

Nazina et al. 2001 emend. Coorevits et al. 2012 strain NG80-2 (RMSD = 0.49; PDB ID 6itcy; [45]). The structure-based function prediction from I-TASSER suggested that ancestral SecY in the LUCA was capable of membrane transport (GO: 0015450, GO:0055085) and was an integral membrane protein (GO:0016021; Table 2).

Amino acid compositions inferred for ancestral proteins in the SRP system

The evolution of the canonical genetic code represents a stage in evolutionary history that likely occurred well before the time of the LUCA [46–51]. In order to test the hypothesis that the ancestral Ffh/FtsY arose in life prior to completion of the canonical genetic code, we analyzed the frequencies of late evolving amino acids in the reconstructed protein. Though there is disagreement about the exact chronology of the addition of amino acids to the genetic code, consensus between forty different published studies suggested the following order from earliest to latest is Gly/Ala, Val/Asp, Pro, Ser, Glu/Leu, Thr, Arg, Asn, Lys, Gln, Ile, Cys, His, Phe, Met, Try, and Trp [52]. A slightly more recent chronology based on phylogeny [53] is largely in agreement, especially regarding the latest five amino acids added to the code, which are predicted to be Cys, Phe, Tyr, Met, and Trp. A previous study by Fournier and Alm [54] analyzed the enzymes tyrosine aminoacyl tRNA synthetase (TyrRS) and tryptophan aminoacyl tRNA synthetase (TrpRS), which, like Ffh and FtsY, are paralogs that arose in an ancestor predating the LUCA. The authors found that Trp was missing from most of their sequence reconstructions for a pre-LUCA ancestor but became more likely in later nodes of their protein family tree, such as in the LUCA and the last shared ancestors of the bacterial and archaeal domains. Fournier and Alm [54] also performed ancestral sequence reconstructions of simulated sequences based on the WAG substitution matrix to show that the lack of Trp in the ancestor of the LUCA was not an artifact of Trp being a rare amino acid. The WAG substitution matrix is an empirically determined set of amino acid substitution rates from which sequences can be simulated to provide a base rate (i.e., typical rate) of amino acid usage for comparison against amino acid usage rates of reconstructed sequences.

We performed a similar set of analyses to Fournier and Alm [54] in order to evaluate amino acid compositions during the evolutionary history of FtsY and Ffh. Our reconstructed sequence of ancestral Ffh/FtsY contained no Trp, while modern Ffh and FtsY proteins contain an average of 1.6 Trp, ranging from zero to five Trp in a single sequence. Reconstructed sequences of subsequent ancestors on the FtsY branch contained between one and two Trp, and ancestors on the Ffh branch contained no Trp (Fig 6). In contrast, ancestral sequence reconstructions based on the Ffh/FtsY tree (Fig 2) but containing terminal sequences simulated from the WAG substitution matrix (i.e., following Fournier and Alm [54]) yielded an average of 4.5 Trp at the node representing ancestral Ffh/FtsY. Therefore, the paucity of Trp in the ancestral Ffh/FtsY protein is not due to the rarity of Trp in modern proteins.

We also performed a finer resolution characterization of the evolutionary incorporation of Trp into the SRP system by examining the occurrences of this amino acid within protein domains through time at reconstructed nodes and in modern species. Interestingly, we found that Trp, if it was utilized at all in the SRP system of the LUCA, was most likely incorporated into the GTPase domain of FtsY, while, in Ffh, it could have been used within the signal peptide recognition domain, which is the only domain not shared between the two protein families (Fig 7). In the most recent common ancestor (MRCA) of Archaea, Trp may have been utilized within the signal peptide recognition domain of Ffh and in the helical bundle domain of FtsY. In contrast, the MRCA of Bacteria may not have used Trp at all within Ffh, but may have used it within the GTPase domain of FtsY. This is consistent with the modern usage among sampled species, in which Trp is more likely to be found in archaeal helical bundle

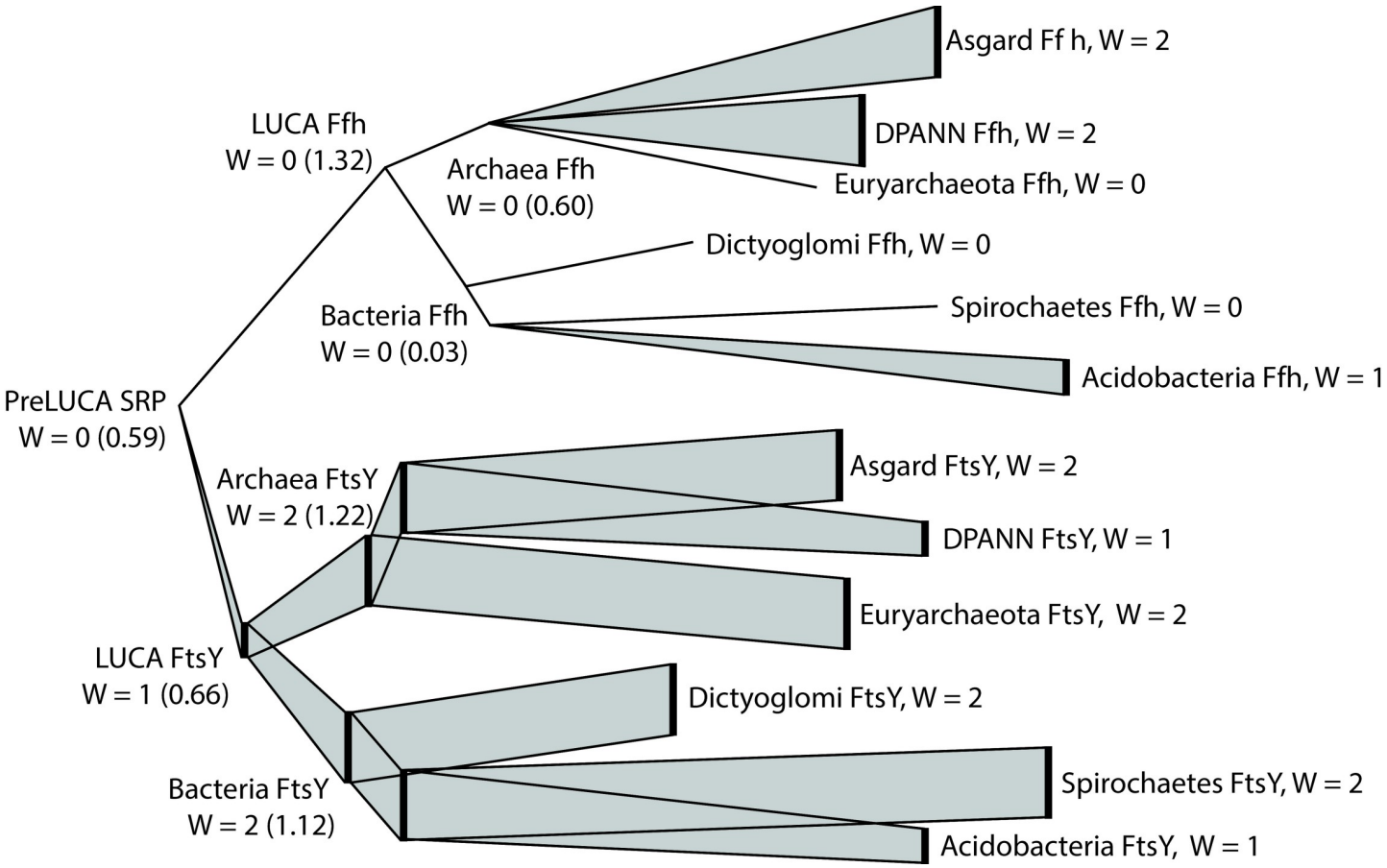

Fig 6. Inferred usage of tryptophan in the SRP system protein family from ancestral Ffh/FtsY to present. The tree topology and branch lengths are consistent with the SRP system tree shown in Fig 1. Branch widths correspond to the tryptophan usage at each node. For ancestral nodes, Trp usage is represented as the total Trp count in the consensus ancestral sequence and, in parentheses, the average Trp count across all sequence reconstructions at the corresponding node. Trp usage in extant Ffh and FtsY sequences is shown as median tryptophan usage in three phyla or superphyla of bacteria and three of archaea, which were chosen because their Ffh and FtsY sequences comprised clades without paraphyly or polyphyly (or nearly so) and represented the broadest sequence diversity within their respective taxonomic domains.

https://doi.org/10.1371/journal.pcbi.1008623.g006

domains and bacterial GTPase domains in FtsY. In Ffh, the archaeal lineage makes greater use of Trp in all three protein domains than Bacteria. It appears that the incorporation of Trp into the SRP system differed between the two protein families and between the two major prokaryotic evolutionary lineages.

The above results regarding Trp should be interpreted with caution because of its rarity within proteins of the modern SRP system and, thus, due to the difficulty of inferring its prevalence in the past. Notably, many Ffh and FtsY proteins of modern organisms contain no Trp at all. We therefore extended this analysis to other late amino acids as per Trifonov [52] and Jordan et al. [53] (i.e., the two studies mentioned above on the chronological sequence of amino acid additions to the genetic code). In addition to Trp, the reconstructed sequence of ancestral Ffh/FtsY showed lower usage of all of the latest amino acids (Fig 8). Specifically, usages of the latest evolving amino acids in the ancestral Ffh/FtsY are much lower than in modern Ffh and FtsY proteins and much lower than predicted for the ancestor according to sequence simulations based on the WAG model. Taken together, these results suggest that the ancestral Ffh/FtsY dates back to the final stages of genetic code evolution, or at least, that it evolved at a time when the last amino acids to be incorporated into the genetic code were much less common within proteins.

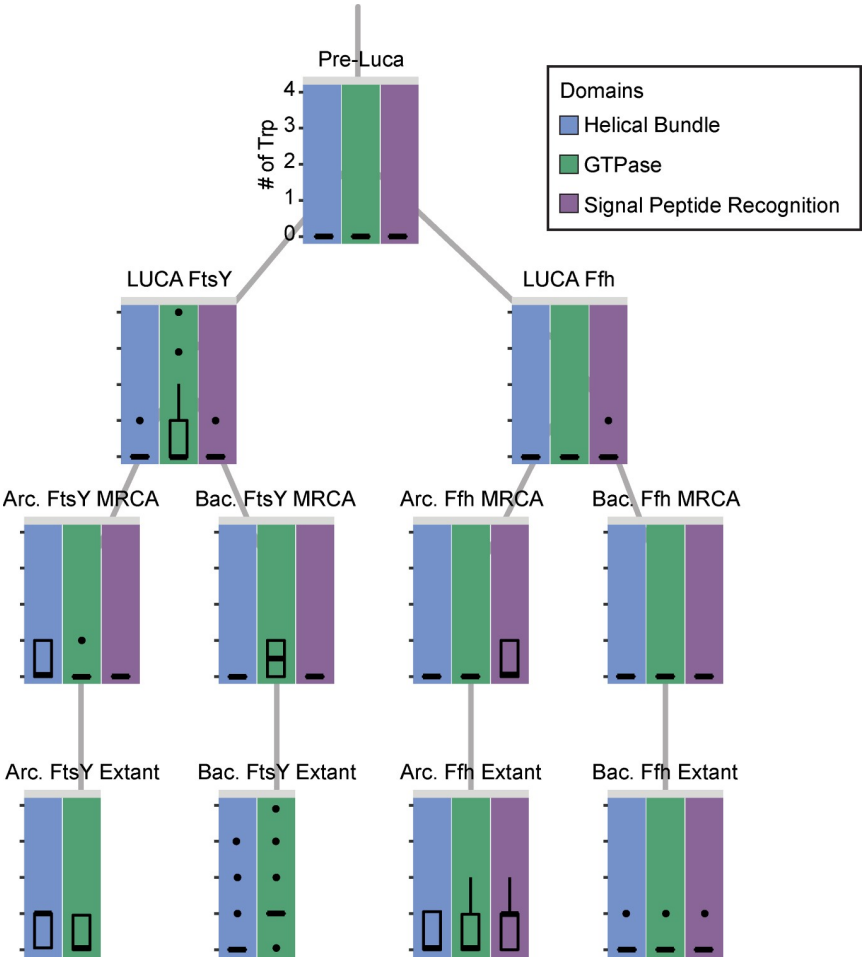

Fig 7. Incorporation of tryptophan into FtsY and Ffh protein domains through time based on ancestral sequence reconstructions and modern species. Box plots representing Pre-LUCA, LUCA, and bacterial (Bac) and archaeal (Arc) most recent common ancestors (MRCAs) are based on 100 ancestral sequence reconstructions using different subsets of the alignment of modern species. Subsets comprised one randomly selected sequence from each cluster identified in Clustal Omega (see Methods). For the modern species, tryptophan usage within extant proteins was determined only for Ffh and FtsY proteins, not homologs representing FlhA or the type III secretion system proteins. All box plots utilize the same scale for the y-axis presented adjacent to the set of plots for the pre-LUCA (top).

https://doi.org/10.1371/journal.pcbi.1008623.g007

Early evolution of SRP/Sec membrane translocation

The cellular organization of organisms appears to have evolved sometime between the origin of life and the LUCA, a process that occurred roughly 3.5–4 billion years ago [55–60]. Within origins of life settings, the presence of amphipathic molecules with membrane-forming potential is well-documented [61,62]. However, in addition to the formation of membranes, cellularity also requires the ability to embed proteins in the membrane and secrete proteins through it. The evidence we present here constitutes the first thorough account of an ancient system of protein translocation and its evolution prior to the time of the LUCA. Specifically, we show that three central protein components of membrane translocation, FtsY, Ffh, and SecY, were all in place by the time of the LUCA and that they had functional and structural characteristics that were similar to their modern counterparts.

At this time, the available data in public databases do not indicate support for the presence of signal peptidase proteins or the SecA protein in the LUCA based on lack of orthologs well-

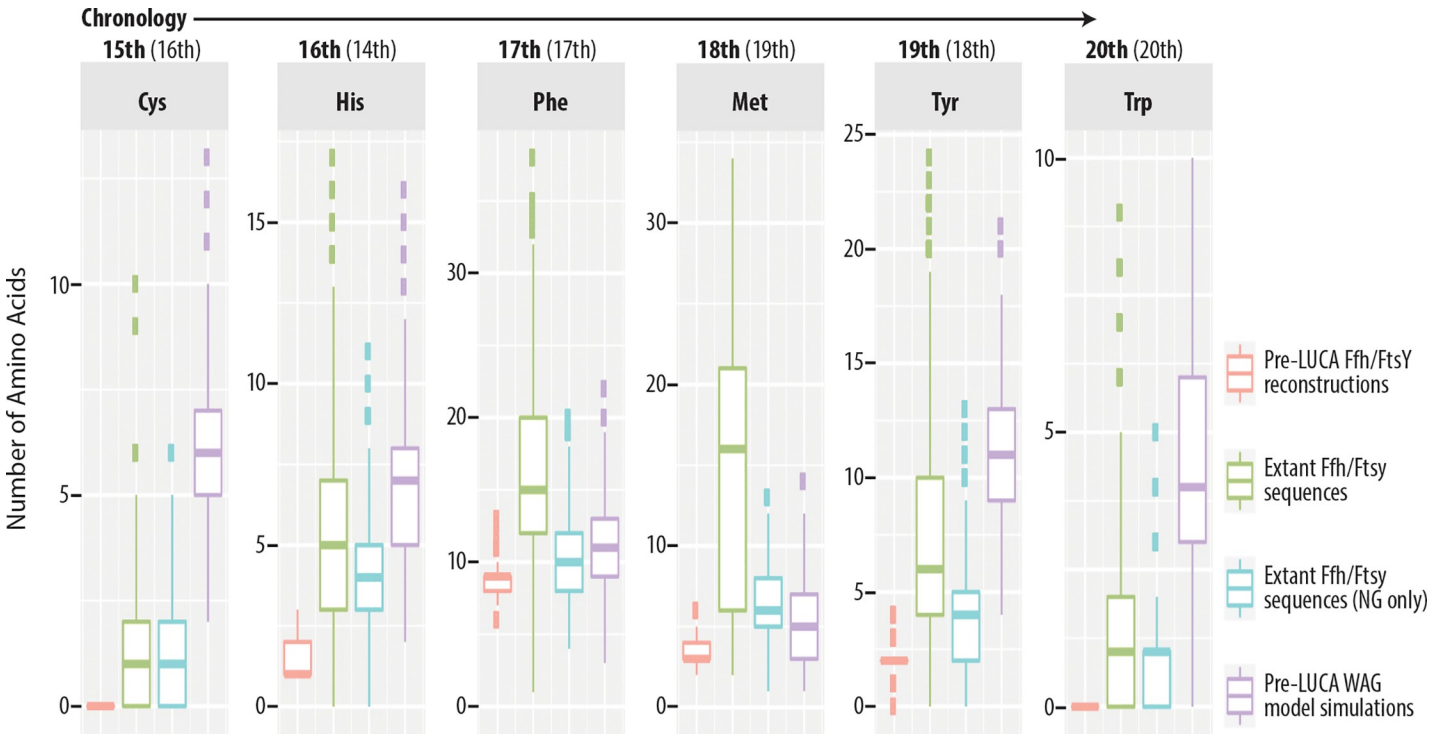

Fig 8. Paucity of late amino acids in ancestral Ffh/FtsY. The chronological order of the addition of late amino acids to the genetic code is taken from Trifonov [52] and Jordan *et al.* [53] (the latter shown within parentheses). Distributions of amino acid counts from the ancestral Ffh/FtsY sequence (red) represent 100 sequence reconstructions for the pre-LUCA node using the 10% gap threshold in FastML. Distributions of amino acid counts are shown for extant SRP system proteins (green) or only the region containing the helical bundle and GTPase domains that aligns to the ancestral Ffh/FtsY sequence (blue). When analyzing extant proteins, we excluded FlhA and type III secretion system proteins from the analysis. In all cases, the inferred ancestral Ffh/FtsY sequences contain fewer late amino acids than extant homologs. Reconstructions based on simulated sequences derived from the WAG model (purple) demonstrate that the paucity of late amino acids in ancestral Ffh/FtsY is not an artifact of the ancestral sequence reconstruction methods or the background rate of amino acid usage across proteins in general.

https://doi.org/10.1371/journal.pcbi.1008623.g008

represented among taxonomic domains and phyla or superphyla. These proteins perform seemingly essential functions; namely removing N-terminal signal sequences from secreted proteins and powering protein translocation through ATP hydrolysis. It may be the case that these proteins only enhance the SRP/Sec system and were not necessary to the system as it existed in the LUCA. Though we found that the Ffh sequence of the LUCA likely contained a signal recognition domain, this feature of the system may have been used only for internal signal sequences that need not be removed by a signal peptidase enzyme. However, more likely, these proteins may have undergone non-orthologous displacement [63] (or replacement) within the ancestral lineage of Archaea such that their presence in the LUCA can no longer be detected using phylogenetic methods. The process of non-orthologous gene displacement has been previously documented (e.g., [64,65]) and may explain other key differences in how the modern proteomes of Archaea and Bacteria accomplish fundamental processes such as DNA synthesis [66,67].

Beyond this account of the SRP/Sec system as it existed in the LUCA, we additionally show that the ancestral Ffh/FtsY protein that predated the LUCA had several functional characteristics found in modern Ffh and FtsY proteins and that this protein likely predates the completion of the canonical genetic code. Even though the ancestral Ffh/FtsY protein likely dates to this early stage in evolutionary history, it appears to have been capable of at least most of the functions performed by both proteins today. It had a membrane targeting sequence that could

have bound to the membrane surface through electrostatic interactions in a manner similar to FtsY. It also had an N-terminal helical bundle and a GTPase domain, which, upon dimerization, created an active site that could bind and hydrolyze GTP (Fig 5). The GTPase domain, itself, belongs to the P-loop containing nucleoside triphosphate hydrolase superfamily, which is responsible for the majority of ATP and GTP hydrolysis within proteomes and is considered one of the most ancient tertiary structures and functional domains within all proteins [14,51,68]. It is unlikely that ancestral Ffh/FtsY contained a signal peptide binding domain, such as in Ffh, given that this domain is only present in sequences reconstructed with highly permissive inclusion of gaps.

Even without the ability to bind the signal sequence of a nascent protein, it is possible that ancestral Ffh/FtsY was capable of binding a ribosome during active translation, given that recent structural studies of the SRP-ribosome complex show contact between the large ribosomal subunit and the GTPase domain of Ffh [69,70]. Furthermore, though the main binding site for SRP-RNA is located in the M domain of Ffh [21], structural characterization of SRP-RNA co-crystalized with the Ffh and FtsY GTPase domains [16] suggests that ancestral Ffh/FtsY could have formed a homodimer in complex with an SRP-RNA (Fig 5). The presence of an MTS domain in the ancestral SRP system protein further suggests that this complex could have bound to a membrane surface.

Given that the ancestral Ffh/FtsY was able to bind the membrane but not able to bind signal peptides in a nascent protein, our results suggest that this ancestral protein was more like FtsY than Ffh and that a duplication event prior to the LUCA ultimately yielded the Ffh component of the SRP system. Under this scenario, Ffh appears to represent a neofunctionalization that likely increased efficiency of the SRP system through targeted signal sequence binding and, therefore, facilitating specialization of FtsY as a membrane bound receptor for Ffh, despite that FtsY appears to also retain ancestral capabilities to homodimerize [16]. Neofunctionalizations are sometimes regarded as resulting most frequently from an unobserved subfunctionlization step within the pathway of protein evolution following a duplication event [71,72], which, in this case, could be the loss of the membrane binding function prior to the gain of the signal peptide binding function.

Of course, it is possible that ancestral Ffh/FtsY was capable of all of the functions that are currently performed by both Ffh and FtsY (e.g., as in the reconstructions with 90% indel threshold; Fig 4, S8 File) and that SRP RNA and the Sec translocation channel were present alongside ancestral Ffh/FtsY in pre-LUCA organisms. However, even though we found that SecY was present in the LUCA, it is not a universal paralog, and, therefore, we cannot use phylogenetic analysis to infer its presence in organisms older than the LUCA. Similarly, the SRP RNA was likely present in the LUCA [73] and may potentially be much older than the LUCA [74], but it is not directly traceable to organisms predating the LUCA using phylogenetic methods.

Nevertheless, even without SecY and SRP RNA, ancestral Ffh/FtsY could have provided an early mechanism for delivering proteins to a membrane (Fig 9). The ancestral SRP system protein could have bound ribosomes through their GTPase domains. Ribosomes would then be temporarily bound to the interior of the membrane while they were translating. If the nascent peptide exiting the ribosome had certain chemical properties, it could spontaneously translocate across the membrane as is observed in some viral coat proteins (e.g.,[75,76]) and which is an important delivery system of peptide-based drugs [77]. If the nascent protein did not contain sequences capable of spontaneous membrane translocation, it would be released into the cytoplasm of the cell. The GTPase domain of the ancestral homodimer would regulate the formation and dissociation of the complex as it does in the Ffh/FtsY heterodimer, today. This

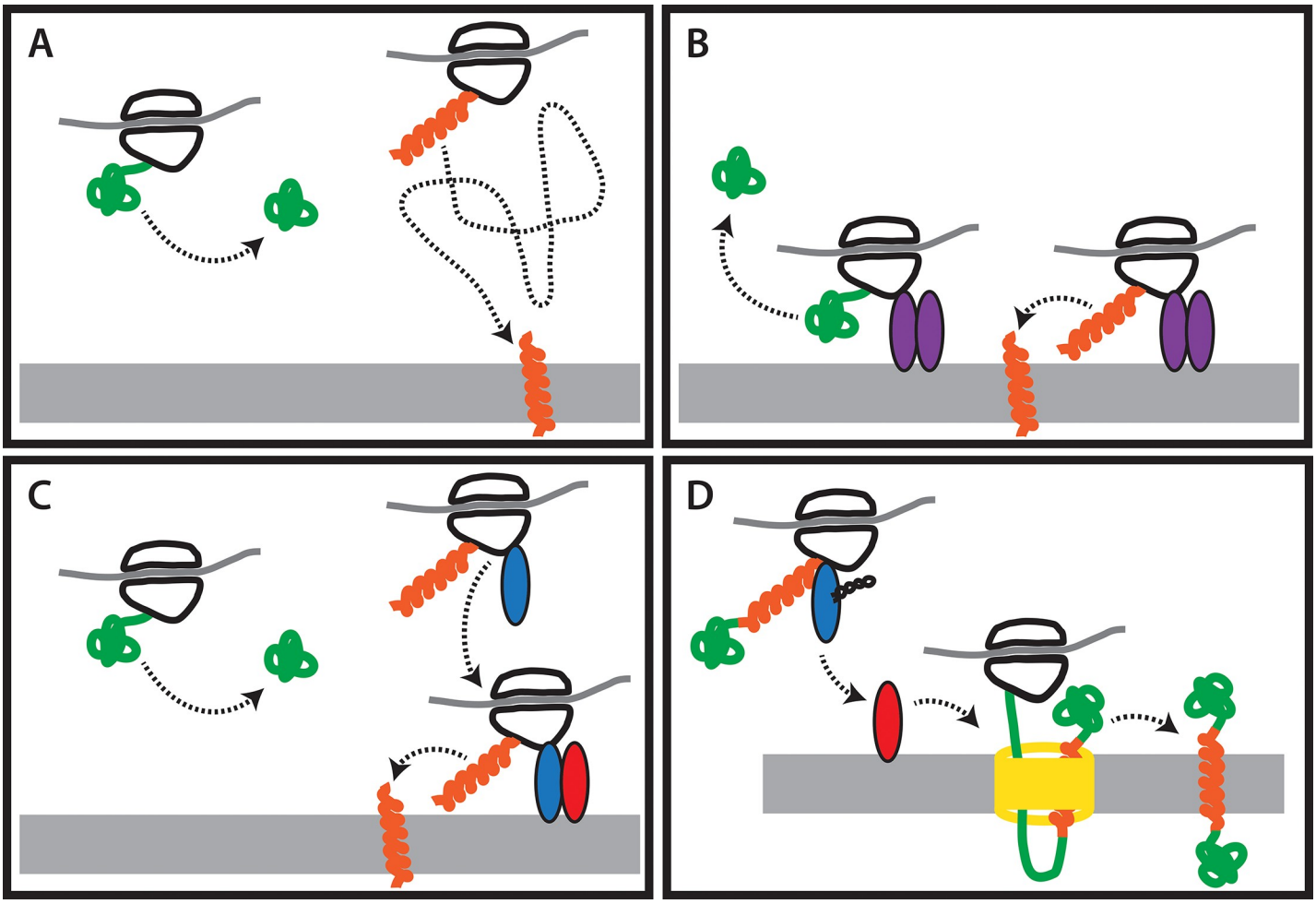

Fig 9. An incremental model for the evolution of the SRP membrane translocation system. A) At first, no components of the system exist. Ribosomes are present in the cytoplasm and some peptides are capable of spontaneous membrane translocation due to their chemical properties (orange), but only after they make their way to the membrane. B) The ancestral Ffh/FtsY protein (purple) evolves and acts to anchor ribosomes to the membrane, facilitating the spontaneous translocation of membrane proteins. C) The subsequent evolution of Ffh (blue) and the specialization of FtsY (red) as a receptor for Ffh allows one component of the system to bind ribosomes in the cytoplasm while they are synthesizing membrane proteins or secreted proteins. These ribosomes are then bound to the membrane surface through the interaction of Ffh and FtsY. D) The addition of the SecY membrane channel (yellow) to the system permits the translocation of large proteins that could not have done so spontaneously. Another equally likely order of events is that the SecY membrane channel preceded the divergence of the ancestral Ffh/FtsY protein into Ffh and FtsY.

https://doi.org/10.1371/journal.pcbi.1008623.g009

system may not have required SRP RNA, which appears to function in response to the binding of a signal sequence.

This model (Fig 9) supported by our results suggests that the ancestral Ffh/FtsY, which predated the LUCA, could have served a similar function to its modern counterparts despite lacking some components of the modern system. The later functional refinement of the Ffh and FtsY proteins, along with the addition of the SRP RNA and SecY translocation would have significantly enhanced the membrane translocation process such that a highly sophisticated system had evolved by the time of the LUCA. As our evidence suggests, the earliest component of this system had evolved even before the final stages of genetic code evolution (Fig 7). This indicates that even very early life forms actively maintained cellular organization. Future work examining the early evolution of membrane proteins and of the enzymes that synthesized membrane components will further enhance our understanding of the most ancient forms of cellular organization.

Materials and methods

Sequence collection of FtsY, Ffh, and SecY homologs

We conducted BLASTp searches to obtain homologous proteins in four archaeal and 18 bacterial superphyla or phyla that were represented by at least one complete genome in the NCBI database and well-characterized taxonomically (i.e., excluding *incertae sedis* despite the relevance of poorly characterized bacterial lineages to the tree of life [32]). We performed the BLASTp using query sequences of *Escherichia coli* (Migula 1895) Castellani and Chalmers 1919 (Proteobacteria; accessions P10121, P0AGD7, P0AGA2 of the SwissProt database [78], a model bacterium, and *Haloferax volcanii* (Mullakhanbhai and Larsen, 1975) Torreblanca et al., 1986 (Euryarchaeota; accessions Q977V3, D4GYW6, Q977V2), an emerging model archaean [79,80]. Prior to performing the BLAST searches, we verified that the representative sequences of FtsY, Ffh, and SecY in *E. coli* and *H. volcanii* were reciprocal top hits of each other within their respective genomes from the RefSeq database ([81]; accessions: ASM584v2, ASM2568v1) based on e-values. We conducted the BLASTp searches using systematic taxonomic sampling within superphyla or phyla according to the NCBI taxonomic database [82,83] to account for diversity across the tree of life. Preliminary BLAST search results and surveys of the literature suggested that the flagellar synthesis protein, FlhA, and the EscV/YscV/HrcV genes of the type III secretion system, evolved within the SRP system protein family [84]. Therefore, we also performed searches for FlhA or type III secretion system proteins using accessions P76298, A0A2J6N6Z4, O85633, and A0A2K3J983 as query sequences. After obtaining a dataset containing 20 (or as many as available) taxonomically diverse proteins of each type from each phylum or superphylum, we performed reciprocal BLAST searches to ensure their orthology with the target proteins.

For all BLAST searches, we accepted only those results with e-values of 0.0001 or lower and query and target coverage of at least 75%. Where possible, we selected different genera or, less ideally, different species from among the top hits for a total of 20 sequences or the maximum number available meeting our criteria. We performed reciprocal BLAST searches of the obtained proteins into the genomes of *E. coli* and *H. volcanii* from RefSeq ([81]; accessions: ASM58ASM584v2, ASM2568v1) to ensure their orthology with the target proteins and discarded sequences for which the top hit was not at the expected locus.

Phylogenetic analysis

For the final protein datasets of the SRP system and SecY, we aligned sequences in Geneious v. 11.1.5 [85] using MAFFT [86–88] with a gap open penalty of 3.0 and extension penalty of 1.0. We allowed Geneious to automatically select the best algorithm within MAFFT for our data. We adjusted the alignments using MAFFT on either side of highly conserved blocks as well as performing manual adjustments.

We conducted Bayesian phylogenetic inference using ExaBayes [89] on the Comet supercomputing cluster of XSEDE for the dataset of sequences representing SRP system, including FlhA and type III secretion system proteins, and separately for the dataset of sequences representing SecY. The analyses for each dataset comprised two independent runs of 20 million generations with sampling every 5000 generations. We applied the WAG amino acid substitution matrix [90] based on preliminary analyses under mixed models in ExaBayes and MrBayes 3.2 [91–93]. Following the analyses, we assessed convergence of independent runs and stationarity according to ESS values > 200 in Tracer v1.6 [94]. We performed 10% burnins for each run and combined the remaining trees in Log Combiner of the BEAST v2.0 package [95,96]. From among the combined trees, we obtained the maximum clade credibility tree using Tree

Annotator, also part of the BEAST package, with median branch lengths and no posterior probability cut-off.

The MAFFT sequence alignment for the SRP system comprised 1666 amino acids, and showed large gaps, especially in the N- and C- terminal regions. Within the N-terminal regions, the gaps largely occurred because some FtsY proteins have extended, non-conserved sequence in this region, while in the C-terminal regions, Ffh proteins contain the signal peptide binding motif that is lacking in FtsY. The SecY alignment, having a length of 835 amino acids, also contained large gaps, such as may be anticipated when aligning sequences across domains of life. To ensure that retaining gap-rich regions did not significantly impact tree topologies, we also performed phylogenetic analyses on matrices, in which we removed highly gapped regions. To remove these regions, we applied Gblocks [97] to both the full SRP and SecY alignments.

Within Gblocks, we set the minimum number of sequences for a conserved or flanking position to the minimum allowable (ca. just over half the number of total sequences in the alignment; 470 of 938 sequences for SRP and 178 of 355 for SecY). We set Gblocks to accept up to 100 contiguous non-conserved positions, a minimum length block of two, and to allow any number of positions within the conserved block to contain gaps. This represents a relaxed set of parameters [98] for selecting conserved blocks of protein sequences, and more conservative settings yielded no blocks in preliminary analyses. The resulting alignment of the SRP system proteins (S9 File) comprised 706 sites consisting of 49.0% gaps (compared to 68.3% in the complete alignment). Through comparison with the complete alignment for the SRP system, we confirmed that the use of Gblocks resulted primarily in the trimming of the N- and C-terminal regions that are present only in some FtsY and most Ffh, respectively. For SecY, the alignment based on Gblocks (S10 File) contained 667 sites and 38.6% gaps (compared to 43.0% in the complete alignment).

We performed phylogenetic analyses for the trimmed alignments using ExaBayes and summarized the resulting trees as described above for the complete alignments. We visually compared the trees resulting from the Gblocks and complete alignments using the cophylo function of the Phytools library [99] for R (S2 and S3 Figs). We also investigated the posterior probabilities for nodes between the Gblocks and complete alignment trees using cumulative fraction plots (S4 and S5 Figs) and by performing one-tailed t-tests (assuming equal variance based on prior F-tests). These results show that, in the SecY tree, only a few terminal taxa occur at different positions in the trimmed, Gblocks alignment compared to the complete one (S3 Fig). For SRP, several superphyla or phyla are in different positions (S2 Fig). However, in both cases, the use of Gblocks did not change major findings, such as low rates of inter-domain HGT, overall tree shape, or, within the SRP system, the relationships among proteins. Moreover, the use of Gblocks significantly ($\alpha = 0.05$) lowered support overall across the tree topologies (S4 and S5 Fig), as speculated in prior studies (e.g., [100]). Thus, we utilized the complete alignments for downstream analyses.

Ancestral sequence reconstruction

We performed ancestral state reconstructions using FastML [101]. Preliminary analyses of sequence reconstructions in FastML revealed that $\geq$300 input sequences yielded an error, which appeared related to very small values rounded to zero (i.e., a precision error). Therefore, we reduced the size of our datasets for ancestral sequence reconstruction using Clustal Omega [102] with soft bounds on the number of sequences per cluster; five for the SRP system and three for SecY. We expected these soft bounds to yield between 200–300 clusters for each dataset, and we obtained 286 clusters for the SRP system (S1 File) and 166 for SecY (S2 File). To

avoid biases or anomalies in sequence selection within the clusters, we assembled 100 datasets for the SRP system and SecY by randomly selecting an accession from each cluster. We performed sequence reconstruction in FastML on each of the 100 datasets and assembled the results for each ancestral node of interest in the phylogeny into a consensus sequence for downstream analyses.

In FastML, we conducted marginal sequence reconstructions over the SRP and SecY phylogenies with maximum likelihood inference of gaps, and we used our maximum clade credibility trees of the SRP system and SecY as guides rather than allowing FastML to reconstruct internal guide trees. For each analysis, we pruned the input tree to represent the 100 selected sequences. For the ancestral sequence reconstructions, our nodes of interest were the LUCA node in SecY that is also the root of the tree and, in the SRP system tree, the LUCA nodes of Ffh and FtsY, and the ancestral Ffh/FtsY representing the root of the tree. For each of these nodes, we obtained a consensus sequence from the 100 results using the simple majority method in the seqinr library for R [103].

Protein function prediction of ancestral sequences

In addition to identifying conserved domains and motifs of the reconstructed ancestral sequences in the PFAM database, we also identified potential MTS domains, which are not represented in PFAM, in ancestral Ffh/FtsY by alignment to the *E. coli* sequence in which this domain was first characterized [Protein Data Bank (PDB; [104,105]) ID: 2yhs; [22]] using the ClustalW2 webserver (https://embnet.vital-it.ch/software/ClustalW.html; [106]) on default settings. We calculated the isoelectric point of the MTS domain of *E. coli* and predicted MTS domains in our reconstructed ancestral sequence using the ProtParam webserver [107].

To complement the sequence-based approaches, we also predicted the three-dimensional protein structures of the ancestral proteins in I-TASSER [40]. In addition to predicting three-dimensional structures, I-TASSER also finds the most structurally similar proteins in the PDB according to the root mean square deviation (RMSD) of the distance between atoms in structural alignments and maps proteins to their most probable gene ontology (GO) terms [108,109] using the structure-based function prediction database, BioLiP [110]. To better assess interactions of ancestral Ffh/FtsY with GTP and SRP RNA, we aligned the top-ranked structural model from I-TASSER to an x-ray diffraction structure of the conserved NG domains of both Ffh and FtsY, bound as a heterodimer and co-crystallized with GDP and the SRP-RNA tetraloop (PDB ID: 4c7o; [16]) using the "align" function in MacPyMOL (Schrödinger LLC.; [111]).

Amino acid composition analysis of ancestral sequences

We compared the frequencies of the most recently added amino acids, Cys, His, Phe, Met, Tyr, and Trp in the reconstruction of ancestral Ffh/FtsY to the frequencies of their modern descendants represented by our dataset of protein sequences for the SRP system. These latest additions to the canonical genetic code were based on consensus between the influential chronology published by Trifonov [52], which itself is based on consensus of 40 prior studies, and a more recent chronology published by Jordan et al. [53]. We performed two comparisons; one of the complete alignment, including FlhA and the type III secretion system proteins, and another in which we removed FlhA and type III secretion system protein sequences and trimmed the remaining alignment to contain only the conserved NG domains of FtsY and Ffh. Additionally, for Trp, which is widely believed to be the latest amino acid added to the genetic code [52,53], we also analyzed its frequency in reconstructed sequences of the archaeal and

bacterial crown clades within the FtsY and Ffh clades of the SRP system (i.e., based on the full alignment; see above).

We sought to ensure that the frequency of late arising amino acids in ancestral Ffh/FtsY was not an artifact of the method of ancestral sequence reconstruction in FastML. Therefore, we simulated sequences across the phylogenetic tree of the SRP system in the R package phangorn [112] under the WAG model. We used the WAG model as a standard against which to compare because it was the model selected to represent sequence evolution of the SRP system in the RMC analysis and because we based our analyses of ancestral amino acid composition on Fournier and Alm [54], who also used the WAG model. We performed 100 simulations, in which we generated new datasets with the same total number of sequences (938; including FlhA and the secretion system) of the same length as the existing alignment (1666 amino acids). Subsequently, we placed gaps in the simulated sequences for each terminal taxon at the loci where they occurred in the existing alignment, thereby using the empirical data to determine the evolution of indels rather than modeling them. We used the newly simulated sequences with indels as alignments to infer ancestral sequences in Fast ML. We performed these analyses as above by generating 100 smaller datasets using Clustal Omega, generating only one alignment of reduced size for each dataset of simulated sequences. For each reduced alignment, we pruned the SRP system tree accordingly so that it could be applied as a guide. In FastML, we estimated gaps at a 10% threshold, and we regarded the consensus sequences at internal nodes and the root node (i.e., representing an ancestor of the LUCA) to be the simulated ancestral sequence reconstructions.

Supporting information

S1 File. Output of Clustal Omega showing 938 proteins representing the SRP system organized into 286 groups.
(TXT)

S2 File. Output of Clustal Omega showing 355 proteins representing SecY organized into 166 groups.
(TXT)

S3 File. Complete gapped alignment of 938 sequences representing the SRP system used in this study. Sequence names follow the format gene_domain_phylum_Accession#, where gene refers to either FtsY, Ffh, FlhA or the secretion system (SS), domains and phyla or superphyla are according to NCBI taxonomy, and accession numbers are from the NCBI nr protein database.
(TXT)

S4 File. Complete gapped alignment of 938 sequences representing the SRP system used in this study. Sequence names follow the format gene_domain_phylum_Accession#, where gene is always SecY, domains and phyla or superphyla are according to NCBI taxonomy, and accession numbers are from the NCBI nr protein database.
(TXT)

S5 File. Maximum clade credibility tree resulting from Bayesian analysis of SRP system proteins (see S3 File) in ExaBayes showing branch lengths and posterior probability support values.
(TXT)

S6 File. Maximum clade credibility tree resulting from Bayesian analysis of SecY proteins (see S4 File) in ExaBayes showing branch lengths and posterior probability support values.
(TXT)

S7 File. Ancestral sequences reconstructions for SecY and the SRP system. For the SRP system, we show results for three different cutoffs for inferring gaps: 10%, 50%, and 90%. For SecY, we used only the cut-off of 50%. In each case, there are 100 sequences resulting from analyses of reduced datasets derived from the clusters predicted in Clustal Omega (see S1 and S2 Files).
(TXT)

S8 File. Consensus sequences of 100 ancestral reconstructions of reduced datasets of SecY and the SRP system. For the SRP system, we show consensus sequences for 10%, 50%, and 90% thresholds for inferring gaps, while for SecY, we show only the 50% threshold.
(TXT)

S9 File. Alignment of 937 sequences representing the SRP system used in this study processed in Gblocks. Sequence names follow the format gene_domain_phylum_Accession#, where gene refers to either FtsY, Ffh, FlhA or the secretion system (SS), domains and phyla or superphyla are according to NCBI taxonomy, and accession numbers are from the NCBI nr protein database. Note that the sequence SS_Archaea_TACK_TBR20608 is present in the ungapped alignment (S3 File and Fig 2) but not the alignment processed in Gblocks, which led to the removal of all sites containing this sequence.
(TXT)

S10 File. Alignment of 938 sequences representing the SRP system used in this study processed in Gblocks. Sequence names follow the format gene_domain_phylum_Accession#, where gene is always SecY, domains and phyla or superphyla are according to NCBI taxonomy, and accession numbers are from the NCBI nr protein database.
(TXT)

S1 Table. Comparison representative sequences of FtsY (light colors) and Ffh (dark colors) among *E. coli* (blue) and *H. volcanii* (yellow) annotated for GO terms representing Biological Function (Green), Cellular Process (Purple), and Cellular Component (Orange). Annotations are according to Uniprot for accessions P10121, P0AGD7, Q977V2, and D4GYW6. Adjacent sequences match between organisms and bold highlighting indicates that the ancestral Ffh/FtsY was also annotated with the term.
(XLSX)

S1 Fig. Predicted structure of SecY in the LUCA according to I-TASSER based on ancestral sequence reconstruction. The image represents the most highly supported of two models based on a C-score of 1.57.
(PNG)

S2 Fig. Maximum clade credibility trees of the SRP system resulting from analysis of the complete, gapped alignment (right) and Gblocks-processed alignment (right) in ExaBayes. Trees are rotated to minimize crossings of lines that connect identical accessions between trees. This provides a visual representation of the similarity between the trees resulting from the two analyses. The visual was generated using Phytools in R.
(PDF)

S3 Fig. Maximum clade credibility trees of SecY resulting from analysis of the complete, gapped alignment (right) and Gblocks-processed alignment (right) in ExaBayes. Trees are rotated to minimize crossings of lines that connect identical accessions between trees. This provides a visual representation of the similarity between the trees resulting from the two

analyses. The visual was generated using Phytools in R.
(PDF)

S4 Fig. Cumulative distribution plot and t-tests comparing posterior probabilities of nodes based on ExaBayes analyses of complete, gapped alignments and alignments with Gblocks processing of SRP system proteins.
(PDF)

S5 Fig. Cumulative distribution plot and t-tests comparing posterior probabilities of nodes based on ExaBayes analyses of complete, gapped alignments and alignments with Gblocks processing of SecY.
(PDF)

Acknowledgments

We gratefully acknowledge the vital roles of the Comet (San Diego Supercomputer Center of the University of California, San Diego, a component of XSEDE) and Sciurus (Oberlin College) supercomputing resources for accomplishing the work herein and the managers of these resources, Mark Miller and Chris Mohler, respectively.

Author Contributions

Conceptualization: AJ Harris, Aaron David Goldman.

Formal analysis: AJ Harris, Aaron David Goldman.

Funding acquisition: AJ Harris, Aaron David Goldman.

Investigation: AJ Harris, Aaron David Goldman.

Methodology: AJ Harris, Aaron David Goldman.

Project administration: AJ Harris, Aaron David Goldman.

Resources: Aaron David Goldman.

Software: AJ Harris.

Supervision: Aaron David Goldman.

Validation: AJ Harris, Aaron David Goldman.

Visualization: AJ Harris, Aaron David Goldman.

Writing – original draft: AJ Harris, Aaron David Goldman.

Writing – review & editing: AJ Harris, Aaron David Goldman.

References

1. Szathmáry E, Smith JM. The major evolutionary transitions. Nature. 1995; 374:227–232. https://doi.org/10.1038/374227a0 PMID: 7885442
2. Woese CThe universal ancestor. Proceedings of the National Academy of Sciences of the United States of America. 1998; 95:6854–6859. https://doi.org/10.1073/pnas.95.12.6854 PMID: 9618502
3. Cantine MD, Fournier GP. Environmental adaptation from the origin of life to the last universal common ancestor. Origins of Life and Evolution of Biospheres. 2018; 48. https://doi.org/10.1007/s11084-017-9542-5 PMID: 28685374
4. Takagi YA, Nguyen DH, Wexler TB, Goldman AD. The coevolution of cellularity and metabolism following the origin of life. Journal of Molecular Evolution. 2020; 88:598–617. https://doi.org/10.1007/s00239-020-09961-1 PMID: 32809045

5. Kurzchalia T, Wiedmann M, Girshovich A, Bochkareva E, Bielka H, Rapoport T. The signal sequence of nascent preprolactin interacts with the 54K polypeptide of the signal recognition particle. Nature. 1986; 320:634. https://doi.org/10.1038/320634a0 PMID: 3010127
6. Poritz M, Bernstein H, Strub K, Zopf D, Wilhelm H, Walter P. An E. coli ribonucleoprotein containing 4.5S RNA resembles mammalian signal recognition particle. Science. 1990; 250:1111–1117. https://doi.org/10.1126/science.1701272 PMID: 1701272
7. Bhuiyan SH, Gowda K, Hotokezaka H, Zwieb C. Assembly of archaeal signal recognition particle from recombinant components. Nucleic acids research. 2000; 28:1365–1373. https://doi.org/10.1093/nar/28.6.1365 PMID: 10684931
8. Bernstein HD, Poritz MA, Strub K, Hoben PJ, Brenner S, Walter P. Model for signal sequence recognition from amino-acid sequence of 54K subunit of signal recognition particle. Nature. 1989; 340:482–486. https://doi.org/10.1038/340482a0 PMID: 2502718
9. Young JC, Ursini J, Legate KR, Miller JD, Walter P, Andrews DW. An amino-terminal domain containing hydrophobic and hydrophilic sequences binds the signal recognition particle receptor α subunit to the β subunit on the endoplasmic reticulum membrane. Journal of Biological Chemistry. 1995; 270:15650–15657. https://doi.org/10.1074/jbc.270.26.15650 PMID: 7797564
10. Ronimus RS, Musgrave DR. Identification of a gene in the euryarchaeal *Thermococcus* species AN1 encoding a protein homologous to the alpha subunit of the eukaryal signal recognition particle (SRP) receptor. Biochimica et Biophysica Acta (BBA)-Gene Structure and Expression. 1997; 1351:1–8. https://doi.org/10.1016/s0167-4781(96)00236-9 PMID: 9116022
11. Peluso P, Herschlag D, Nock S, Freymann DM, Johnson AE, Walter P. Role of 4.5S RNA in assembly of the bacterial signal recognition particle with its receptor. Science. 2000; 288:1640–1643. https://doi.org/10.1126/science.288.5471.1640 PMID: 10834842
12. Chandrasekar S, Sweredoski MJ, Sohn CH, Hess S, Shan S. Co-evolution of Two GTPases enables efficient protein targeting in an RNA-less chloroplast signal recognition particle pathway. Journal of Biological Chemistry. 2017; 292:386–396.
13. Nagai K, Oubridge C, Kuglstatter A, Menichelli E, Isel C, Jovine L. Structure, function and evolution of the signal recognition particle. The EMBO journal. 2003; 22:3479–3485. https://doi.org/10.1093/emboj/cdg337 PMID: 12853463
14. Leipe DD, Wolf YI, Koonin EV, Aravind L. Classification and evolution of P-loop GTPases and related ATPases11Edited by J. Thornton. Journal of Molecular Biology. 2002; 317:41–72. https://doi.org/10.1006/jmbi.2001.5378 PMID: 11916378
15. Gasper R, Meyer S, Gotthardt K, Sirajuddin M, Wittinghofer A. It takes two to tango: regulation of G proteins by dimerization. Nature Reviews Molecular Cell Biology. 2009; 10:423–429. https://doi.org/10.1038/nrm2689 PMID: 19424291
16. Voigts-Hoffmann F, Schmitz N, Shen K, Shan S, Ataide Sandro F, Ban N. The Structural Basis of FtsY Recruitment and GTPase Activation by SRP RNA. Molecular Cell. 2013; 52:643–654. https://doi.org/10.1016/j.molcel.2013.10.005 PMID: 24211265
17. Batey RT, Rambo RP, Lucast L, Rha B, Doudna JA. Crystal Structure of the Ribonucleoprotein Core of the Signal Recognition Particle. Science. 2000; 287:1232–1239. https://doi.org/10.1126/science.287.5456.1232 PMID: 10678824
18. Hainzl T, Huang S, Meriläinen G, Brännström K, Sauer-Eriksson AE. Structural basis of signal-sequence recognition by the signal recognition particle. Nature Structural & Molecular Biology. 2011; 18:389–391. https://doi.org/10.1038/nsmb.1994 PMID: 21336278
19. Parlitz R, Eitan A, Stjepanovic G, Bahari L, Bange G, Bibi E, et al. *Escherichia coli* signal recognition particle receptor FtsY contains an essential and autonomous membrane-binding amphipathic helix. Journal of Biological Chemistry. 2007; 282:32176–32184. https://doi.org/10.1074/jbc.M705430200 PMID: 17726012
20. Weiche B, Bürk J, Angelini S, Schiltz E, Thumfart JO, Koch H-G. A cleavable N-terminal membrane anchor is involved in membrane binding of the Escherichia coli SRP receptor. Journal of Molecular Biology. 2008; 377:761–773. https://doi.org/10.1016/j.jmb.2008.01.040 PMID: 18281057
21. Bradshaw N, Walter P. The signal recognition particle (SRP) RNA links conformational changes in the SRP to protein targeting. Molecular biology of the cell. 2007; 18:2728–2734. https://doi.org/10.1091/mbc.e07-02-0117 PMID: 17507650
22. Stjepanovic G, Kapp K, Bange G, Graf C, Parlitz R, Wild K, et al. Lipids trigger a conformational switch that regulates signal recognition particle (SRP)-mediated protein targeting. Journal of Biological Chemistry. 2011; 286:23489–23497. https://doi.org/10.1074/jbc.M110.212340 PMID: 21543314
23. Gold VAM, Duong F, Collinson I. Structure and function of the bacterial Sec translocon (Review). Molecular Membrane Biology. 2007; 24:387–394. https://doi.org/10.1080/09687680701416570 PMID: 17710643

24. Hand NJ, Klein R, Laskewitz A, Pohlschröder M. Archaeal and bacterial SecD and SecF homologs exhibit striking structural and functional conservation. Journal of Bacteriology. 2006; 188:1251–1259. https://doi.org/10.1128/JB.188.4.1251-1259.2006 PMID: 16452406

25. Veenendaal AKJ, van der Does C, Driessen AJM The protein-conducting channel SecYEG. Biochimica et Biophysica Acta (BBA)—Molecular Cell Research. 2004; 1694: 81–95.

26. Römisch K, Webb J, Herz J, Prehn S, Frank R, Vingron M, et al. Homology of 54K protein of signal-recognition particle, docking protein and two E. coli proteins with putative GTP–binding domains. Nature. 1989; 340:478. https://doi.org/10.1038/340478a0 PMID: 2502717

27. Egea PF, Stroud RM, Walter P. Targeting proteins to membranes: structure of the signal recognition particle. Current Opinion in Structural Biology. 2005; 15:213–220. https://doi.org/10.1016/j.sbi.2005.03.007 PMID: 15837181

28. Gribaldo S, Cammarano P. The root of the universal tree of life inferred from anciently duplicated genes encoding components of the protein-targeting machinery. Journal of Molecular Evolution. 1998; 47:508–516. https://doi.org/10.1007/pl00006407 PMID: 9797401

29. Cao TB, Saier MH. The general protein secretory pathway: phylogenetic analyses leading to evolutionary conclusions. Biochimica et Biophysica Acta (BBA)–Biomembranes. 2003; 1609:115–125. https://doi.org/10.1016/s0005-2736(02)00662-4 PMID: 12507766

30. Rinke C, Schwientek P, Sczyrba A, Ivanova NN, Anderson IJ, Cheng J-F, et al. Insights into the phylogeny and coding potential of microbial dark matter. Nature. 2013; 499: 431. https://doi.org/10.1038/nature12352 PMID: 23851394

31. Garrity GM, Holt JG. Thermodesulfobacteria phy. nov. In: Whitman WB, Rainey F, Kämpfer P, Trujillo M, Chun J, DeVos P, Hedlund B, Dedysh S, editors. Bergey's Manual of Systematics of Archaea and Bacteria. Hoboken, NJ: Wiley; 2015.

32. Hug LA, Baker BJ, Anantharaman K, Brown CT, Probst AJ, Castelle CJ, et al. A new view of the tree of life. Nature Microbiology. 2016; 1:16048. https://doi.org/10.1038/nmicrobiol.2016.48 PMID: 27572647

33. Zaremba-Niedzwiedzka K, Caceres EF, Saw JH, Bäckström D, Juzokaite L, Vancaester E, et al. Asgard archaea illuminate the origin of eukaryotic cellular complexity. Nature. 2017; 541:353. https://doi.org/10.1038/nature21031 PMID: 28077874

34. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. Journal of Molecular Biology. 1990; 215:403–410. https://doi.org/10.1016/S0022-2836(05)80360-2 PMID: 2231712

35. Iwabe N, Kuma K, Hasegawa M, Osawa S, Miyata T. Evolutionary relationship of archaebacteria, eubacteria, and eukaryotes inferred from phylogenetic trees of duplicated genes. Proceedings of the National Academy of Sciences of the United States of America. 1989; 86:9355–9359. https://doi.org/10.1073/pnas.86.23.9355 PMID: 2531898

36. Huerta-Cepas J, Szklarczyk D, Heller D, Hernández-Plaza A, Forslund SK, Cook H, et al. eggNOG 5.0: a hierarchical, functionally and phylogenetically annotated orthology resource based on 5090 organisms and 2502 viruses. Nucleic acids research. 2019; 47:D309–D314. https://doi.org/10.1093/nar/gky1085 PMID: 30418610

37. Sonnhammer EL, Eddy SR, Durbin R. Pfam: a comprehensive database of protein domain families based on seed alignments. Proteins: Structure, Function, and Bioinformatics. 1997; 28:405–420. https://doi.org/10.1002/(sici)1097-0134(199707)28:3<405::aid-prot10>3.0.co;2-l PMID: 9223186

38. El-Gebali S, Mistry J, Bateman A, Eddy SR, Luciani A, Potter SC, et al. The Pfam protein families database in 2019. Nucleic acids research. 2018; 47:D427–D432.

39. Hirokawa T, Boon-Chieng S, Mitaku S. SOSUI: classification and secondary structure prediction system for membrane proteins. Bioinformatics (Oxford, England). 1998; 14: 378–379. https://doi.org/10.1093/bioinformatics/14.4.378 PMID: 9632836

40. Roy A, Kucukural A, Zhang Y. I-TASSER: a unified platform for automated protein structure and function prediction. Nature Protocols. 2010; 5:725. https://doi.org/10.1038/nprot.2010.5 PMID: 20360767

41. Chandrasekar S, Chartron J, Jaru-Ampornpan P, Shan S. Structure of the chloroplast signal recognition particle (SRP) receptor: domain arrangement modulates SRP–receptor interaction. Journal of Molecular Biology. 2008; 375:425–436. https://doi.org/10.1016/j.jmb.2007.09.061 PMID: 18035371

42. Träger C, Rosenblad MA, Ziehe D, Garcia-Petit C, Schrader L, Kock K, et al. Evolution from the Prokaryotic to the Higher Plant Chloroplast Signal Recognition Particle: The Signal Recognition Particle RNA Is Conserved in Plastids of a Wide Range of Photosynthetic Organisms. The Plant Cell. 2012; 24:4819–4836. https://doi.org/10.1105/tpc.112.102996 PMID: 23275580

43. Graham B. Characterization of the ftsYEX operon of *Escherichia coli* [Dissertation]. Ames, Iowa: Iowa State University; 2002. 150 p.

44. Itoh T, Takemoto K, Mori H, Gojobori T. Evolutionary instability of operon structures disclosed by sequence comparisons of complete microbial genomes. Molecular Biology and Evolution. 1999; 16:332–346. https://doi.org/10.1093/oxfordjournals.molbev.a026114 PMID: 10331260

45. Ma C, Wu X, Sun D, Park E, Catipovic MA, Rapoport TA, et al. Structure of the substrate-engaged SecA-SecY protein translocation machine. Nature communications. 2019; 10:2872. https://doi.org/10.1038/s41467-019-10918-2 PMID: 31253804

46. Woese CR. On the evolution of the genetic code. Proceedings of the National Academy of Sciences of the United States of America. 1965; 54:1546–1552. https://doi.org/10.1073/pnas.54.6.1546 PMID: 5218910

47. Crick FHC. The origin of the genetic code. Journal of Molecular Biology. 1968; 38:367–379. https://doi.org/10.1016/0022-2836(68)90392-6 PMID: 4887876

48. Knight RD, Freeland SJ, Landweber LF. Selection, history and chemistry: the three faces of the genetic code. Trends in Biochemical Sciences. 1999; 24:241–247. https://doi.org/10.1016/s0968-0004(99)01392-4 PMID: 10366854

49. Szathmáry E. The origin of the genetic code: amino acids as cofactors in an RNA world. Trends in Genetics. 1999; 15:223–229. https://doi.org/10.1016/s0168-9525(99)01730-8 PMID: 10354582

50. Freeland SJ, Knight RD, Landweber LF, Hurst LD. Early fixation of an optimal genetic code. Molecular Biology and Evolution. 2000; 17:511–518. https://doi.org/10.1093/oxfordjournals.molbev.a026331 PMID: 10742043

51. Goldman AD, Samudrala R, Baross JA. The evolution and functional repertoire of translation proteins following the origin of life. Biology Direct. 2010; 5:15. https://doi.org/10.1186/1745-6150-5-15 PMID: 20377891

52. Trifonov EN. Consensus temporal order of amino acids and evolution of the triplet code. Gene. 2000; 261:139–151. https://doi.org/10.1016/s0378-1119(00)00476-5 PMID: 11164045

53. Jordan IK, Kondrashov FA, Adzhubei IA, Wolf YI, Koonin EV, Kondrashov AS, et al. A universal trend of amino acid gain and loss in protein evolution. Nature. 2005; 433:633–638. https://doi.org/10.1038/nature03306 PMID: 15660107

54. Fournier GP, Alm EJ. Ancestral reconstruction of a pre-LUCA aminoacyl-tRNA synthetase ancestor supports the late addition of trp to the genetic code. Journal of Molecular Evolution. 2015; 80:171–185. https://doi.org/10.1007/s00239-015-9672-1 PMID: 25791872

55. Abramov O, Mojzsis SJ. Microbial habitability of the Hadean Earth during the late heavy bombardment. Nature. 2009; 459:419–422. https://doi.org/10.1038/nature08015 PMID: 19458721

56. Djokic T, Van Kranendonk MJ, Campbell KA, Walter MR, Ward CR. Earliest signs of life on land preserved in ca. 3.5 Ga hot spring deposits. Nature Communications. 2017; 8:1–9. https://doi.org/10.1038/s41467-016-0009-6 PMID: 28232747

57. Dodd MS, Papineau D, Grenne T, Slack JF, Rittner M, Pirajno F, et al. Evidence for early life in Earth's oldest hydrothermal vent precipitates. Nature. 2017; 543:60–64. https://doi.org/10.1038/nature21377 PMID: 28252057

58. Tashiro T, Ishida A, Hori M, Igisu M, Koike M, Méjean P, et al. Early trace of life from 3.95 Ga sedimentary rocks in Labrador, Canada. Nature. 2017; 549:516–518. https://doi.org/10.1038/nature24019 PMID: 28959955

59. Pearce BK, Tupper AS, Pudritz RE, Higgs PG. Constraining the time interval for the origin of life on Earth. Astrobiology. 2018; 18:343–364. https://doi.org/10.1089/ast.2017.1674 PMID: 29570409

60. Betts HC, Puttick MN, Clark JW, Williams TA, Donoghue PCJ, Pisani D. Integrated genomic and fossil evidence illuminates life's early evolution and eukaryote origin. Nature ecology & evolution. 2018; 2:1556–1562. https://doi.org/10.1038/s41559-018-0644-x PMID: 30127539

61. Deamer DW, Pashley R. Amphiphilic components of the Murchison carbonaceous chondrite: surface properties and membrane formation. Origins of Life and Evolution of the Biosphere. 1989; 19:21–38. https://doi.org/10.1007/BF01808285 PMID: 2748144

62. Namani T, Walde P. From decanoate micelles to decanoic acid/dodecylbenzenesulfonate vesicles. Langmuir. 2005; 21:6210–6219. https://doi.org/10.1021/la047028z PMID: 15982022

63. Koonin EV, Mushegian AR, Bork P. Non-orthologous gene displacement. Trends in genetics: TIG. 1996; 12:334–336. PMID: 8855656

64. van Hooff JJE, Snel B, Kops GJPL. Unique phylogenetic distributions of the Ska and Dam1 complexes support functional analogy and suggest multiple parallel displacements of Ska by Dam1. Genome Biology and Evolution. 2017; 9:1295–1303. https://doi.org/10.1093/gbe/evx088 PMID: 28472331

65. de Crécy-Lagard V, Ross RL, Jaroch M, Marchand V, Eisenhart C, Brégeon D, et al. Survey and validation of tRNA modifications and their corresponding genes in *Bacillus subtilis* sp subtilis strain 168. Biomolecules. 2020; 10:977. https://doi.org/10.3390/biom10070977 PMID: 32629984

66. Forterre P. Displacement of cellular proteins by functional analogues from plasmids or viruses could explain puzzling phylogenies of many DNA informational proteins. Molecular Microbiology. 1999; 33:457–465. https://doi.org/10.1046/j.1365-2958.1999.01497.x PMID: 10417637

67. Goldman AD, Landweber LF. *Oxytricha* as a modern analog of ancient genome evolution. Trends in Genetics. 2012; 28:382–388. https://doi.org/10.1016/j.tig.2012.03.010 PMID: 22622227

68. Wang M, Yafremava LS, Caetano-Anollés D, Mittenthal JE, Caetano-Anollés G; Reductive evolution of architectural repertoires in proteomes and the birth of the tripartite world. Genome Research. 2007; 17: 1572–1585. https://doi.org/10.1101/gr.6454307 PMID: 17908824

69. Jomaa A, Boehringer D, Leibundgut M, Ban N. Structures of the *E. coli* translating ribosome with SRP and its receptor and with the translocon. Nature Communications. 2016; 7:1–9.

70. Jomaa A, Fu Y-HH, Boehringer D, Leibundgut M, Shan S, Ban N. Structure of the quaternary complex between SRP, SR, and translocon bound to the translating ribosome. Nature Communications. 2017; 8:15470. https://doi.org/10.1038/ncomms15470 PMID: 28524878

71. He X, Zhang J. Rapid subfunctionalization accompanied by prolonged and substantial neofunctionalization in duplicate gene evolution. Genetics. 2005; 169:1157–1164. https://doi.org/10.1534/genetics.104.037051 PMID: 15654095

72. Rastogi S, Liberles DA. Subfunctionalization of duplicated genes as a transition state to neofunctionalization. BMC Evolutionary Biology. 2005; 5:28. https://doi.org/10.1186/1471-2148-5-28 PMID: 15831095

73. Hoeppner MP, Gardner PP, Poole AM. Comparative analysis of RNA families reveals distinct repertoires for each domain of life. PLoS Computational Biology. 2012; 8:e1002752–e1002752. https://doi.org/10.1371/journal.pcbi.1002752 PMID: 23133357

74. Jeffares DC, Poole AM, Penny D. Relics from the RNA world. Journal of Molecular Evolution. 1998; 46:18–36. https://doi.org/10.1007/pl00006280 PMID: 9419222

75. Engelman D, Steitz T. The spontaneous insertion of proteins into and across membranes: the helical hairpin hypothesis. Cell. 1981; 23:411–422. https://doi.org/10.1016/0092-8674(81)90136-7 PMID: 7471207

76. Kuhn A. Major coat proteins of bacteriophage Pf3 and M13 as model systems for Sec-independent protein transport. FEMS Microbiology Reviews. 1995; 17:185–190. https://doi.org/10.1111/j.1574-6976.1995.tb00201.x PMID: 7669345

77. Macchi S, Signore G, Boccardi C, Di Rienzo C, Beltram F, Cardarelli F. Spontaneous membrane-translocating peptides: Influence of peptide self-aggregation and cargo polarity. Scientific Reports. 2015; 5:16914. https://doi.org/10.1038/srep16914 PMID: 26567719

78. The UniProt Consortium. UniProt: the universal protein knowledgebase. Nucleic acids research. 2017; 45:D158–D169. https://doi.org/10.1093/nar/gkw1099 PMID: 27899622

79. Hartman AL, Norais C, Badger JH, Delmas S, Haldenby S, Madupu R. The complete genome sequence of *Haloferax volcanii* DS2, a model archaeon. PLOS One. 2010; 5:e9605. https://doi.org/10.1371/journal.pone.0009605 PMID: 20333302

80. Pohlschroder M, Schulze S. *Haloferax volcanii*. Trends in Microbiology. 2019; 27:86–87. https://doi.org/10.1016/j.tim.2018.10.004 PMID: 30459094

81. O'Leary NA, Wright MW, Brister JR, Ciufo S, Haddad D, O'Leary, et al. Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. Nucleic acids research. 2015; 44: D733–D745. https://doi.org/10.1093/nar/gkv1189 PMID: 26553804

82. Sayers EW, Barrett T, Benson DA, Bryant SH, Canese K, Chetvernin V, et al. Database resources of the National Center for Biotechnology Information. Nucleic acids research. 2009; 37:D5–D15. https://doi.org/10.1093/nar/gkn741 PMID: 18940862

83. Benson DA, Karsch-Mizrachi I, Lipman DJ, Ostell J, Sayers EW. GenBank. Nucleic acids research. 2009; 37:D26–D31. https://doi.org/10.1093/nar/gkn723 PMID: 18940867

84. Bange G, Kümmerer N, Grudnik P, Lindner R, Petzold G, Kressler D, et al. Structural basis for the molecular evolution of SRP-GTPase activation by protein. Nature Structural & Molecular Biology. 2011; 18:1376.

85. Kearse M, Moir R, Wilson A, Stones-Havas S, Cheung M, Sturrock S, et al. Geneious Basic: an integrated and extendable desktop software platform for the organization and analysis of sequence data. Bioinformatics. 2012; 28:1647–1649. https://doi.org/10.1093/bioinformatics/bts199 PMID: 22543367

86. Katoh K, Misawa K, Kuma K-i, Miyata T. MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. Nucleic acids research. 2002; 30:3059–3066. https://doi.org/10.1093/nar/gkf436 PMID: 12136088

87. Katoh K, Kuma K, Toh H, Miyata T. MAFFT version 5: improvement in accuracy of multiple sequence alignment. Nucleic acids research. 2005; 33:511–518. https://doi.org/10.1093/nar/gki198 PMID: 15661851

88. Katoh K, Standley DM. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. Molecular Biology and Evolution. 2013; 30:772–780. https://doi.org/10.1093/molbev/mst010 PMID: 23329690

89. Aberer AJ, Kobert K, Stamatakis A. ExaBayes: massively parallel Bayesian tree inference for the whole-genome era. Molecular Biology and Evolution. 2014; 31:2553–2556. https://doi.org/10.1093/molbev/msu236 PMID: 25135941

90. Goldman N, Whelan S. A general empirical model of protein evolution derived from multiple protein families using a maximum-likelihood approach. Molecular Biology and Evolution. 2001; 18:691–699. https://doi.org/10.1093/oxfordjournals.molbev.a003851 PMID: 11319253

91. Huelsenbeck JP, Ronquist F. MRBAYES: Bayesian inference of phylogenetic trees. Bioinformatics. 2001; 17:754–755. https://doi.org/10.1093/bioinformatics/17.8.754 PMID: 11524383

92. Ronquist F, Huelsenbeck JP. MrBayes 3: Bayesian phylogenetic inference under mixed models. Bioinformatics. 2003; 19:1572–1574. https://doi.org/10.1093/bioinformatics/btg180 PMID: 12912839

93. Ronquist F, Teslenko M, van der Mark P, Ayres DL, Darling A, Höhna S, et al. MrBayes 3.2: Efficient Bayesian phylogenetic inference and model choice across a large model space. Systematic Biology. 2012; 61:539–542. https://doi.org/10.1093/sysbio/sys029 PMID: 22357727

94. Rambaut A, Drummond A. Tracer. 1.4 ed. 2007

95. Drummond A, Rambaut A. BEAST: Bayesian evolutionary analysis by sampling trees. BMC Evolutionary Biology. 2007; 7:214. https://doi.org/10.1186/1471-2148-7-214 PMID: 17996036

96. Bouckaert R, Heled J, Kuehnert D, Vaughan T, Wu C-H, Xie D. BEAST 2: a software platform for Bayesian evolutionary analysis. PLoS Computational Biology. 2014; 10:e1003537. https://doi.org/10.1371/journal.pcbi.1003537 PMID: 24722319

97. Castresana J. Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis. Molecular Biology and Evolution. 2000; 17:540–552. https://doi.org/10.1093/oxfordjournals.molbev.a026334 PMID: 10742046

98. Talavera G, Castresana J. Improvement of phylogenies after removing divergent and ambiguously aligned blocks from protein sequence alignments. Systematic Biology. 2007; 56:564–577. https://doi.org/10.1080/10635150701472164 PMID: 17654362

99. Revell LJ. phytools: an R package for phylogenetic comparative biology (and other things). Methods in Ecology and Evolution. 2012; 3:217–223.

100. Harris AJ, Chen Y, Olsen RT, Lutz S, Wen J. On merging Acer sections Rubra and Hyptiocarpa: Molecular and morphological evidence. PhytoKeys. 2017; 9. https://doi.org/10.3897/phytokeys.86.13532 PMID: 29033667

101. Ashkenazy H, Penn O, Doron-Faigenboim A, Cohen O, Cannarozzi G, Zomer O, et al. FastML: a web server for probabilistic reconstruction of ancestral sequences. Nucleic acids research. 2012; 40: W580–W584. https://doi.org/10.1093/nar/gks498 PMID: 22661579

102. Sievers F, Wilm A, Dineen D, Gibson TJ, Karplus K, Li W, et al. Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. Molecular systems biology. 2011; 7. https://doi.org/10.1038/msb.2011.75 PMID: 21988835

103. Charif D, Lobry JR. SeqinR 1.0–2: a contributed package to the R project for statistical computing devoted to biological sequences retrieval and analysis. In: Bastolla U, Porto M, Roman HE, Vendruscolo M, editors. Structural Approaches to Sequence Evolution. Berlin: Springer; 2007. pp. 207–232.

104. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, et al. The Protein Data Bank. Nucleic acids research. 2000; 28:235–242. https://doi.org/10.1093/nar/28.1.235 PMID: 10592235

105. Rose P, Beran B, Bi C, Bluhm W, Dimitropoulos D, Goodsell D, et al. The RCSB Protein Data Bank: redesigned web site and web services. Nucleic Acids Res. 2010;39. https://doi.org/10.1093/nar/gkp877 PMID: 19854952

106. Thompson JD, Higgins DG, Gibson TJ. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Res. 1994; 22:4673–4680. https://doi.org/10.1093/nar/22.22.4673 PMID: 7984417

107. Gasteiger E, Hoogland C, Gattiker A, Wilkins MR, Appel RD, Bairoch A. Protein identification and analysis tools on the ExPASy server. In: Walker JM, editor. The Proteomics Protocols Handbook. Totowa, NJ: Humana Press Inc. 2005. pp. 571–607.

108. Gene Ontology Consortium, Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, et al. Gene Ontology: tool for the unification of biology. Nature Genetics. 2000; 25:25–29. https://doi.org/10.1038/75556 PMID: 10802651

109. Gene Ontology Consortium. The Gene Ontology (GO) database and informatics resource. Nucleic acids research. 2004; 32:D258–D261. https://doi.org/10.1093/nar/gkh036 PMID: 14681407

110. Yang J, Roy A, Zhang Y. BioLiP: a semi-manually curated database for biologically relevant ligand–protein interactions. Nucleic Acids Research. 2012; 41:D1096–D1103. https://doi.org/10.1093/nar/gks966 PMID: 23087378

111. DeLano WL. Pymol: An open-source molecular graphics tool. CCP4 Newsletter on Protein Crystallography. 2002; 40:82–92.

112. Schliep KP. phangorn: phylogenetic analysis in R. Bioinformatics. 2010; 27:592–593. https://doi.org/10.1093/bioinformatics/btq706 PMID: 21169378