Integrating Phylogenetics With Intron Positions Illuminates the Origin of the Complex Spliceosome

Julian Vosseberg [1,2] Daan Stolker,[1] Samuel H.A. von der Dunk [1] and Berend Snel *[,1]

[1]Theoretical Biology and Bioinformatics, Department of Biology, Faculty of Science, Utrecht University, 3584 CH Utrecht, the Netherlands

[2]Laboratory of Microbiology, Wageningen University & Research, 6700 EH Wageningen, the Netherlands

***Corresponding author:** E-mail: b.snel@uu.nl.

Associate editor: Irina Arkhipova

Abstract

Eukaryotic genes are characterized by the presence of introns that are removed from pre-mRNA by a spliceosome. This ribonucleoprotein complex is comprised of multiple RNA molecules and over a hundred proteins, which makes it one of the most complex molecular machines that originated during the prokaryote-to-eukaryote transition. Previous works have established that these introns and the spliceosomal core originated from self-splicing introns in prokaryotes. Yet, how the spliceosomal core expanded by recruiting many additional proteins remains largely elusive. In this study, we use phylogenetic analyses to infer the evolutionary history of 145 proteins that we could trace back to the spliceosome in the last eukaryotic common ancestor. We found that an overabundance of proteins derived from ribosome-related processes was added to the prokaryote-derived core. Extensive duplications of these proteins substantially increased the complexity of the emerging spliceosome. By comparing the intron positions between spliceosomal paralogs, we infer that most spliceosomal complexity postdates the spread of introns through the proto-eukaryotic genome. The reconstruction of early spliceosomal evolution provides insight into the driving forces behind the emergence of complexes with many proteins during eukaryogenesis.

Key words: **spliceosome, eukaryogenesis, introns.**

Introduction

The spliceosome is a dynamic ribonucleoprotein (RNP) complex that assembles on the pre-mRNA to remove introns, intervening sequences between the exons. The exons are spliced together to form a mature mRNA. Like the complex, the exon-intron structure of protein-coding genes is characteristic of eukaryotes. Transcription and splicing occur in the nucleus, which physically separates these processes from protein translation. Failure of correct splicing generally results in non-functional proteins.

The composition of the spliceosome changes during the splicing cycle (Wilkinson et al. 2020). It consists of five small nuclear RNAs (snRNAs) U1, U2, U4, U5, and U6, which are bound by multiple proteins to form small nuclear RNPs (snRNPs), and several additional subcomplexes and factors. In the splicing reaction, the 5′ splice site first reacts with the adenosine branch point, forming a lariat structure. Subsequently, the exons are ligated and the lariat intron is released. The components of the spliceosome orchestrate different activities in a precisely ordered manner: they recognize the splice sites and the branch point sequences, prevent a premature reaction, perform the splicing reaction, and assemble, remodel, or disassemble the complex. The spliceosome is one of the most complex molecular machines in eukaryotic cells and a complex spliceosome was present in the last eukaryotic common ancestor (LECA) (Collins and Penny 2005).

Eukaryotes have two types of introns that are recognized by different spliceosome complexes. A vast majority of introns are of the U2-type and are recognized by the major spliceosome; the U12-type introns comprise a small minority (Moyer et al. 2020). The minor spliceosome specifically recognizes the U12-type introns and most proteins of the major spliceosome as well as U5 snRNA are also part of the minor spliceosome (Turunen et al. 2013; Bai et al. 2021). The other snRNAs have a minor-spliceosome equivalent (U11, U12, U4atac, and U6atac) and a few minor-spliceosome-specific proteins have been identified, especially in the U11/U12 di-snRNP (Turunen et al. 2013). The minor spliceosome and U12-type introns were also present in LECA (Russell et al. 2006).

In sharp contrast to a probably intron-rich LECA (Csuros et al. 2011; Vosseberg et al. 2022) with a complex spliceosome, prokaryotic genes lack spliceosomal introns, which must have emerged at some time during eukaryogenesis. Spliceosomal introns and the key spliceosomal protein PRPF8 are thought to derive from self-splicing group II introns in prokaryotic genomes. This is based on similarities in the splicing reaction, function, and structure of the RNAs involved, as well as the homology inferred between the spliceosomal protein PRPF8 and the single

© The Author(s) 2023. Published by Oxford University Press on behalf of Society for Molecular Biology and Evolution.
This is an Open Access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted reuse, distribution, and reproduction in any medium, provided the original work is properly cited.

Open Access

Downloaded from https://academic.oup.com/mbe/article/40/1/msad011/6985000 by Wageningen University and Research – Library user on 27 May 2023

Article

protein encoded by group II introns, the intron-encoded protein (IEP) (Zimmerly and Semper 2015). Recent work has suggested that the emergence of intragenic introns might have been an early event during eukaryogenesis (Vosseberg et al. 2022). The evolutionary histories of a few gene families in the spliceosome have been described (Anantharaman et al. 2002; Veretnik et al. 2009; Califice et al. 2012) and they suggest that gene duplications played a pivotal role in the emergence of the complex spliceosome. Yet, a detailed picture of the origins of the full spliceosome, one of the most complex machines to emerge during eukaryogenesis, is lacking.

This paper details in-depth phylogenetic analyses to reconstruct the spliceosome in LECA and the evolutionary histories of these LECA proteins in the prokaryote-to-eukaryote transition. Subsequent integration of the phylogenetic trees with the positions of introns allows to investigate the relation between the origin of the spliceosome and the emergence of intragenic introns. Our findings underline the role of gene duplications in establishing the complex LECA spliceosome and we detected a strong evolutionary link with the ribosome. The intron analyses suggest that the emergence of a complex spliceosome occurred late, relative to the spread of introns.

Results

Complex Composition of the LECA Spliceosome

To infer the evolutionary origin of the LECA spliceosome, it is first necessary to establish which proteins were likely present in the spliceosome in LECA. The most recent systematic inventory of the composition of LECA's spliceosome stems from 2005 (Collins and Penny 2005), and since then, multiple additional proteins, such as the minor-spliceosome-specific proteins, have been traced back to the eukaryotic ancestor. In conjunction with the enormous increase in genomic data, this provides ample reasons to update the reconstruction of the composition of the LECA spliceosome. We carried out this reconstruction by performing homology searches with spliceosomal proteins of humans (supplementary Table S1, Supplementary Material online) and baker's yeast (supplementary Table S2, Supplementary Material online), two species whose spliceosomes are well-studied. We used a strict definition of the spliceosome, which excludes proteins that function in related processes such as the coupling of splicing with transcription and the regulation of splicing. If we identified orthologs in multiple Opimoda and Diphoda species (Derelle et al. 2015) (see Materials and Methods), we inferred that a spliceosomal protein was ancestral. With these criteria, 145 spliceosomal orthogroups (OGs) could be traced to LECA (fig. 1, supplementary Table S3, Supplementary Material online). This number is nearly twice as large as the previously estimated 78 spliceosomal proteins in LECA (Collins and Penny 2005), a consequence of the expanded genomic sampling of eukaryotic biodiversity and increased knowledge on eukaryotic spliceosomes. The inferred number of spliceosomal LECA OGs is slightly lower than the number of spliceosomal proteins in humans (164, only one LECA OG missing) and substantially larger than the number of proteins in the yeast spliceosome (99, 86 LECA OGs present). In addition to these proteins, five major spliceosomal snRNAs and four minor-spliceosome-specific snRNAs were also present in LECA.

Unresolved Origin of PRPF8 From Intron-encoded Protein and Additional Group II Introns in Asgard Archaea

As described above, the U5-snRNP protein PRPF8 is a remnant of self-splicing group II introns. The prokaryotic origins of this system could, in principle, be inferred from the phylogenetic affinity of IEP and the spliceosomal PRPF8 protein, as the reverse transcriptase (RT)-like domain in PRPF8 is homologous to the RT domain in IEP (Dlakić and Mushegian 2011; Qu et al. 2016; Zhao and Pyle 2016). However, phylogenetic analysis of this domain is hindered by the high sequence divergence of PRPF8, and to a lesser extent its paralog telomerase, relative to prokaryotic RT domains. In our analyses, the nuclear homologs of IEP are not clearly associated with a particular IEP type and their exact phylogenetic position in the IEP tree is unresolved (supplementary fig. S1*a*, Supplementary Material online).

Group II introns occur predominantly in bacteria. A recent study showed that most complete archaeal genomes do not contain group II introns, with the exception of Methanomicrobia (Miura et al. 2022). We detected group II introns in several Asgard archaeal genomes, which are from multiple different IEP types (supplementary fig. S1*b*, Supplementary Material online). This finding expands the set of observed IEP types in archaea to also include ML, D, E, CL2A, and a separate CL type. The presence of these "bacterial" mobile elements in Asgard archaea is in good agreement with the diverse mobile elements that were recently found in circular *Heimdallarchaeum* genomes and the proposed continuous influx of bacterial genes in Asgard archaea (Wu et al. 2022). This so far unappreciated wide diversity of self-splicing group II introns in Asgard archaea might indicate the presence of such elements in the archaeal ancestor of eukaryotes.

Expansion of the Emerging Spliceosome Through Extensive Gene Duplication

All other 144 spliceosomal OGs do not have a homolog in group II introns. We performed phylogenetic analyses to infer their respective evolutionary origins (supplementary Table S4, Supplementary Material online). A few OGs had a complex evolutionary history since they contain multiple domains with a separate history and resulted from a fusion event (Supplementary Information). Among them, 56 OGs are most closely related to another spliceosomal OG (fig. 2*a*), and therefore, their preduplication ancestor was, probably, already part of the spliceosome. By collapsing such close paralogous clades of spliceosomal OGs, we identified 102 ancestral spliceosomal units (supplementary fig. S2, Supplementary Material online). Duplications of

Downloaded from https://academic.oup.com/mbe/article/40/1/msad011/6985000 by Wageningen University and Research – Library user on 27 May 2023

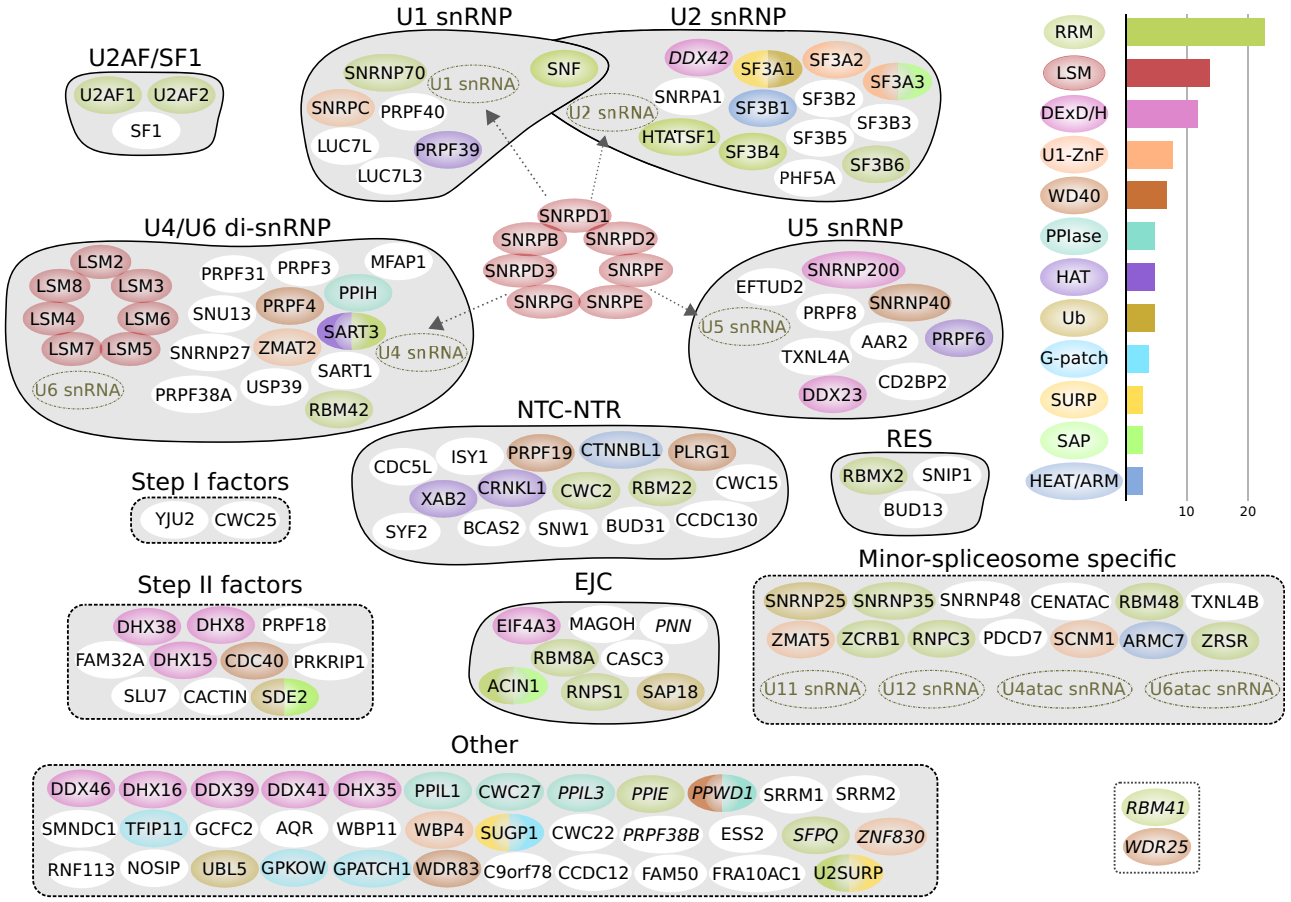


Fig. 1. The spliceosome inferred in LECA. The names of OGs with a lower confidence score are in italics (possibly spliceosomal in LECA, see Materials and Methods). The OGs are grouped based on the subcomplex they are in or another collection (dashed line), and they are colored based on their domain composition. Only domains that are present in at least three OGs are shown. The bar plot shows the number of OGs per domain. OGs that are only present in the minor spliceosome are displayed as minor-spliceosome-specific. The main differences between the major and minor spliceosomes are the presence of a U11/U12 di-snRNP instead of U1 and U2 snRNPs and the replacement of U4 and U6 snRNA with U4atac and U6atac snRNA. Two candidate minor-spliceosome-specific proteins that we identified in this study are shown in the dotted box. snRNP, small nuclear ribonucleoprotein; snRNA, small nuclear RNA; NTC, Prp19-associated complex; NTR, Prp19-related complex; RES, retention and splicing complex; EJC, exon-junction complex; RRM, RNA recognition motif; ZnF, zinc finger; PPIase, peptidylprolyl isomerase; HAT, half-a-tetratricopeptide repeat; Ub, ubiquitin; HEAT/ARM, HEAT or armadillo repeats.

spliceosomal genes increased the number of spliceosomal proteins with a factor of 1.4. The ancestral spliceosomal units themselves also originated in most cases from duplication, but then, from a gene with another function in the proto-eukaryotic cell. For 33 ancestral units, we could not detect other homologs and these were, therefore, classified as proto-eukaryotic inventions. One single spliceosomal OG, AAR2, was surprisingly found to be one-on-one orthologous to a gene in a limited number of prokaryotes, including Loki- and Gerdarchaeota (Supplementary Information). Over a hundred proteins seemed to have been recruited to the emerging spliceosome at different points during eukaryogenesis. Subsequent duplications of these proteins resulted in an even more complex spliceosome in LECA.

Eukaryotic genomes are chimeric in nature, with genes originating from the Asgard archaea-related host, the alphaproteobacteria-related protomitochondrion or other prokaryotes by means of horizontal gene transfer. The eukaryotic spliceosome mirrors this general trend. It contains considerable numbers of genes from archaeal and bacterial origin, making it a chimeric complex in phylogenetic origin (fig. 2*b*). The largest group, however, is comprised of genes for which we could not detect ancient homologs in prokaryotes and possibly originated de novo. This suggests that novel eukaryote-specific folds played a major role in shaping the emerging spliceosome. It is noteworthy that none of the acquisitions from bacteria could be traced back to alphaproteobacteria. This argues against a direct contribution of the mitochondrial endosymbiont to the spliceosome.

Spliceosomal Proteins Originated Predominantly From Ribosomal Biogenesis, Translation, and RNA Processing Proteins

A relatively large number of spliceosomal OGs were acquired from genes that functioned in ribosome biogenesis and translation (fig. 2*c*), especially OGs from archaeal

Downloaded from https://academic.oup.com/mbe/article/40/1/msad011/6985000 by Wageningen University and Research – Library user on 27 May 2023

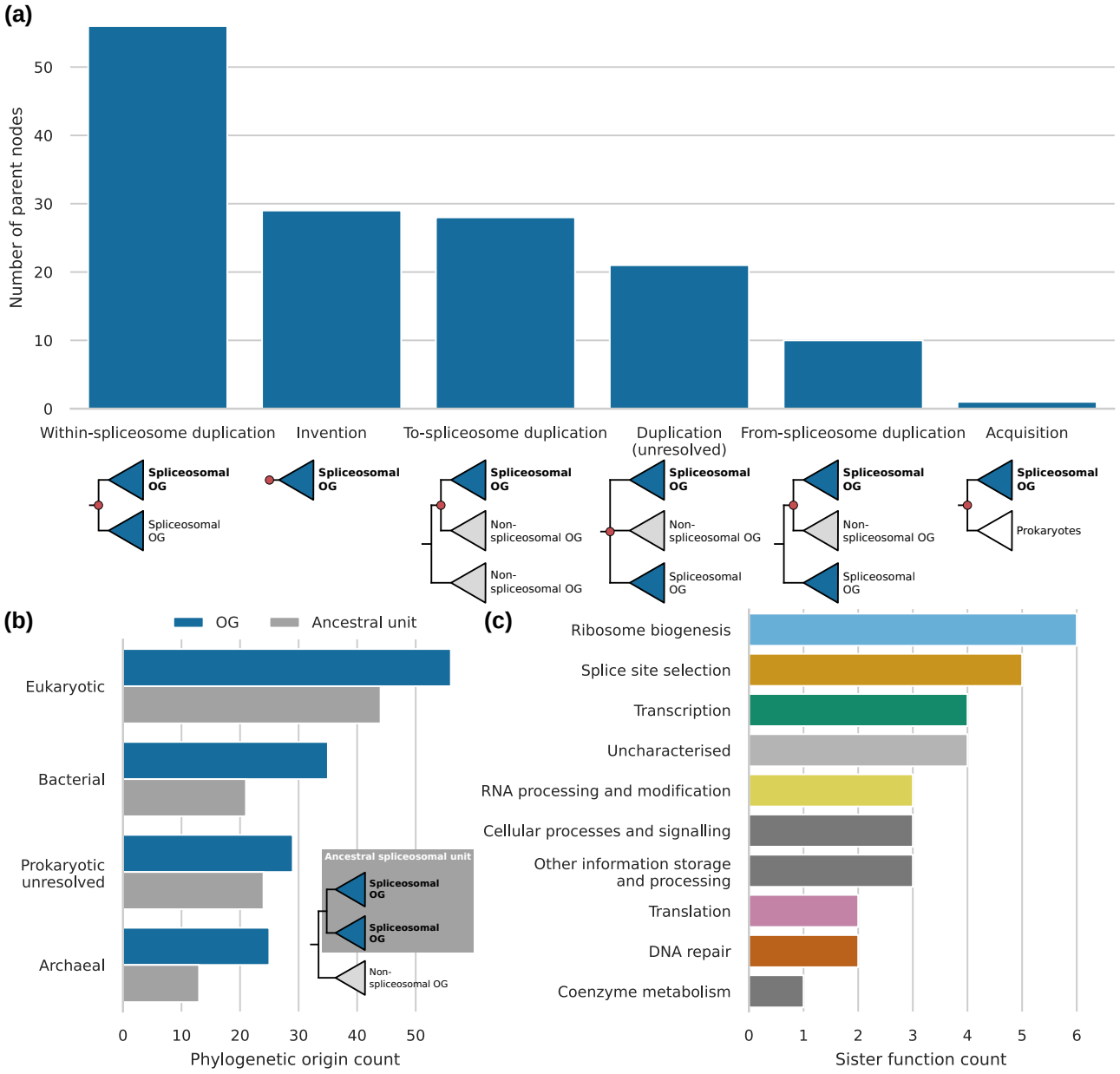


Fig. 2. Evolutionary history of spliceosomal proteins before LECA. (*a*) Annotations of the parent nodes of spliceosomal OGs. These parent nodes are shown in red in the example trees below. If the sister OG of a spliceosomal OG is also spliceosomal, the parent node was classified as a within-spliceosome duplication. If no other homologs outside the OG could be detected, the parent node was classified as invention. In case of a non-spliceosomal sister OG, to- and from-spliceosome duplications were distinguished based on the function of other homologs. If the sister group of the OG consisted of prokaryotic sequences, it was classified as an acquisition. (*b*) Bar plot showing the phylogenetic origins of spliceosomal OGs and ancestral spliceosomal units, in which within-spliceosome duplications had been collapsed. (*c*) Functions of the sister OGs of ancestral spliceosomal units.

origin. The U5 snRNP protein EFTUD2 is a paralog of elongation factor 2 (fig. 3*a*), which catalyzes ribosomal translocation during translation elongation. The archaeal ortholog performs the same translocation function, yet, also probably plays a role in ribosome biogenesis that is similar to the other proto-eukaryotic paralog EFL1 (Lo Gullo et al. 2021). The U4/U4atac-binding proteins SNU13 and PRPF31 (Nottrott et al. 2002) can be linked to the C/D-box snoRNP (fig. 3*b* and *c*). SNU13 is also part of this snoRNP (Watkins et al. 2000) and PRPF31 originated from a C/D-box snoRNP protein. The archaeal orthologs NOP5 and RPL7Ae are part of the functionally equivalent C/D box sRNP (fig. 3*d*), which is involved in ribosome biogenesis by modifying rRNA (Aittaleb et al. 2003; Breuer et al. 2021). The eukaryotic DDX helicases, of which six are part of LECA's spliceosome, evolved from prokaryotic DEAD and RHLE proteins, which also function in ribosome assembly (Charollais et al. 2004; Jain 2008) (fig. 3*e* and *f*). A large group of related RNA helicases is the DHX helicases. The ancestral function of DHX

Downloaded from https://academic.oup.com/mbe/article/40/1/msad011/6985000 by Wageningen University and Research – Library user on 27 May 2023

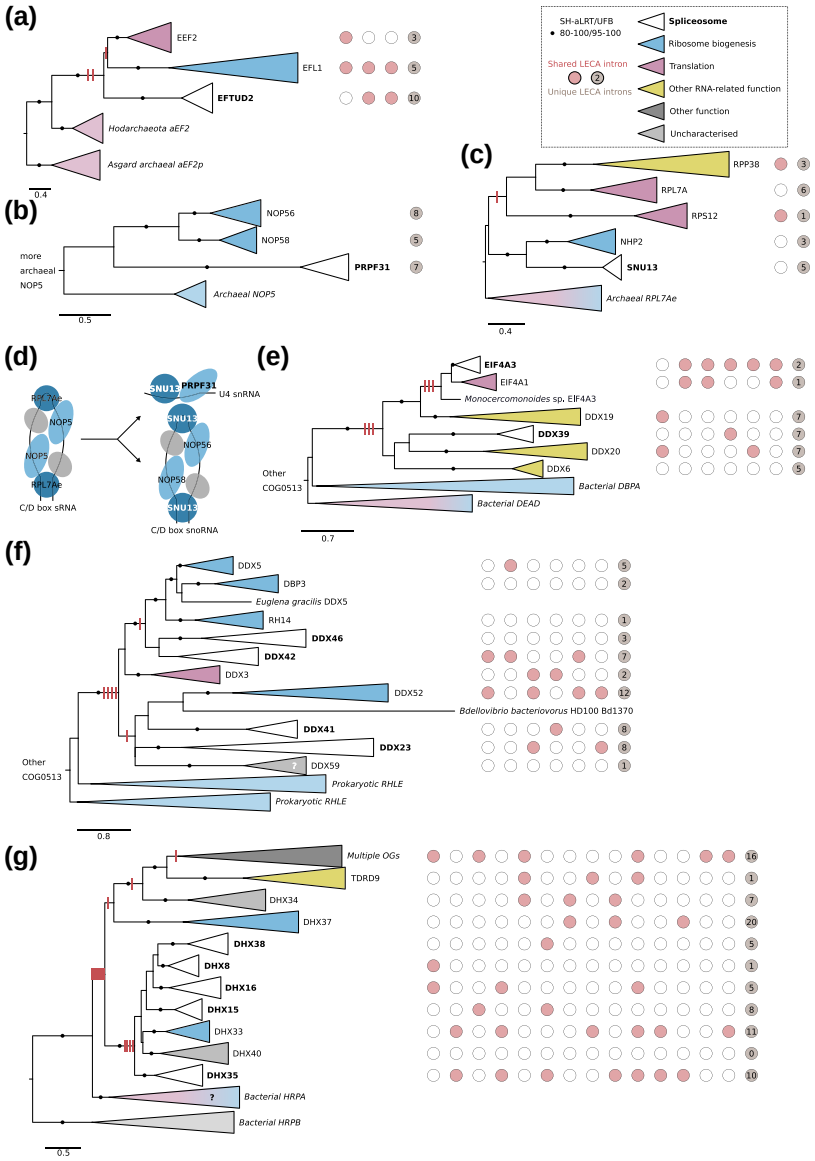


Fig. 3. Spliceosomal proteins that originated from ribosome-related proteins. (*a*) Phylogenetic tree of the EF2 family. (*b*) Phylogenetic tree of the NOP family. (*c*) Phylogenetic tree of the RPL7A family. (*d*) Evolution of the C/D box snoRNP and U4 snRNP proteins SNU13 and PRPF31 in LECA from the C/D box sRNP in the archaeal ancestor of eukaryotes. Homologous proteins are shown in the same color. SNU13 was present in both complexes in LECA. The gray protein corresponds with fibrillarin. (*e*, *f*) Phylogenetic tree of the DDX helicase family, displaying two separate acquisitions during eukaryogenesis in two separate panels. The function of DDX59 has not been characterized, but its phylogenetic profile is similar to minor-spliceosome-specific proteins (de Wolf et al. 2021). (*g*) Phylogenetic tree of the DHX helicase family. (*a*–*c*, *e*–*g*) Eukaryotic LECA OGs are collapsed and colored based on their function, as are the prokaryotic clades. Introns inferred in LECA are depicted; columns with red/white circles correspond with the presence of introns at homologous positions. The gain of introns before duplications as reconstructed using Dollo parsimony is shown with red stripes on the branches. Scale bars correspond with the number of substitutions per site. Clades with significant support as assessed with the SH-like approximate likelihood ratio (SH-aLRT) and ultrafast bootstrap (UFB) values are indicated with filled circles.

helicases is probably related to ribosome biogenesis (fig. 3*g*). Recruitment into the spliceosome and duplications resulted in five spliceosomal DHX helicases.

The Lsm and Sm heptamer rings that bind U6 or U6atac snRNA and other snRNAs, respectively, are also of archaeal origin. The archaeal homologs, called Sm-like archaeal proteins (SmAPs), are poorly characterized RNA-binding proteins that might function in tRNA processing and RNA degradation (Lekontseva et al. 2021). The SmAP genes are located directly adjacent to ribosomal protein RPL37e (Mura et al. 2013), emphasizing the potential link with translation. The eukaryotic Lsm ring is involved in different forms of RNA processing besides splicing (Mura et al. 2013), including rRNA maturation (Kufel et al. 2003). During eukaryogenesis, the Lsm ring gained a U6(atac) binding function and was recruited into the spliceosome. Subsequent gene duplications resulted in two types of heteromeric rings of Lsm/Sm proteins in the spliceosome (supplementary fig. S3, Supplementary Material online, Supplementary Information).

Downloaded from https://academic.oup.com/mbe/article/40/1/msad011/6985000 by Wageningen University and Research – Library user on 27 May 2023

A substantial fraction of the LECA spliceosome OGs contains an RNA recognition motif (RRM) (fig. 1). The proteins in this family perform diverse functions, as this domain can not only bind RNA, but is also involved in protein–protein interactions (Maris et al. 2005). RRM proteins were likely acquired from a bacterium during eukaryogenesis, as proteins with this domain are present in some bacteria. Although the tree is largely unresolved due to the short length of the motif, multiple recruitments into the spliceosome can be observed, some followed by intraspliceosome duplications (supplementary fig. S4*a*, Supplementary Material online). Functions of other RRM proteins that are closely related to the spliceosome OGs include transcription, splice site selection, and mRNA degradation. Some OGs contain multiple RRMs, pointing at a rich history of domain and gene duplications before LECA in this family.

Many Minor-spliceosome-Specific Proteins are Closely Related to a Major Spliceosome Protein

The major and minor spliceosomes share many subunits (Turunen et al. 2013; Bai et al. 2021) and this was very likely also the case in LECA. We inferred 13 minor-spliceosome-specific proteins in LECA (fig. 1). Six of these are closely related to a major-spliceosome-specific protein. The RRM proteins SNRNP35 and ZRSR have a major spliceosome equivalent as their sister paralog (supplementary fig. S4*b* and *c*, Supplementary Material online). RNPC3 is closely related to SNF but probably not as its sister paralog (supplementary fig. S4*d*, Supplementary Material online). The sister paralog of RNPC3 is the poorly characterized RBM41. Its phylogenetic profile, however, corresponds with minor spliceosome OGs (de Wolf et al. 2021). If RBM41 is part of the minor spliceosome, the RNPC3-RBM41 duplication would represent the only identified duplication within the minor-spliceosome-specific OGs. The phylogenetic position of the other minor spliceosome OGs with an RRM is unresolved (supplementary fig. S4*a*, Supplementary Material online). ZMAT5 and SCNM1 are members of the U1-type zinc finger family. The equivalent of ZMAT5 in the major spliceosome is SNRPC (Will et al. 2004) and SCNM1 functions as a combination of SF3A2 and SF3A3 (Bai et al. 2021). Although the phylogenetic tree of this family is unresolved (supplementary fig. S5*a*, Supplementary Material online), it is likely that these major and minor spliceosome equivalents are sister paralogs. In contrast, TXNL4B is a clear sister paralog of the major spliceosome OG TXNL4A (supplementary fig. S5*c*, Supplementary Material online). The sister paralog of the WD40-repeat protein CDC40, called WDR25 (supplementary fig. S5*b*, Supplementary Material online), has a presence pattern across eukaryotes that is typical of minor spliceosome OGs (de Wolf et al. 2021), like RBM41. This protein has not been characterized either, yet its phylogenetic profile strongly suggests a function in the minor spliceosome.

A peculiar observation that we made for all major/minor pairs mentioned above is that the branch in the phylogenetic tree leading from duplication to the minor-spliceosome-specific OG is considerably shorter than the one leading to the major-spliceosome-specific OG (supplementary fig. S5*d*, Supplementary Material online). This means that these major-spliceosome-specific OGs have diverged more from the ancestral preduplication state and suggests that the function of the minor-spliceosome-specific SNRNP35, ZRSR, RNPC3, ZMAT5, SCNM1, TXNL4B, and possibly WDR25, better reflect the ancestral state.

Substantial Intron Spread Predating Spliceosomal Duplications

In a previous study, we investigated the spread of introns in proto-eukaryotic paralogs (Vosseberg et al. 2022). Intron positions that are shared between genes that duplicated during eukaryogenesis are likely shared because they were present in the gene before it duplicated. By analyzing intron positions in spliceosomal OGs, we can relate duplications in the primordial spliceosome to the spread of the elements that they function on, the introns. Therefore, we applied the same approach as in our previous study to the paralogs in the spliceosome. 45% of duplications that probably resulted in a novel spliceosomal gene had at least one intron traced back to the preduplication state (13 of the 29 to-spliceosome duplications). For 46% of the within-spliceosome duplications, we detected shared introns between paralogs in the spliceosome (18 out of 39).

The presence of introns in ancestral genes that themselves likely did not function in the spliceosome is strikingly illustrated by the DDX and DHX helicases, with three to seven introns traced back to before the first duplication after the acquisition from prokaryotes (fig. 3). Introns shared between spliceosomal paralogs are also found in the LSM, PPIase, and WD40 families (supplementary fig. S3*a*, Supplementary Material online, supplementary Table S5, Supplementary Material online). The U5 snRNP proteins SNRNP200 and EFTUD2, which interact with PRPF8, shared multiple introns with paralogs outside the spliceosome and likely contained introns before they became part of the spliceosome (fig. 3, supplementary Table S5, Supplementary Material online). These numbers and cases suggest that introns were already present in a substantial number of ancestral genes before the corresponding proteins were recruited into the spliceosome and, subsequently, duplicated within the spliceosome.

Duplication and Subfunctionalization Completed Multiple Times After Eukaryogenesis: U1A/U2B″

A notable difference between the LECA spliceosome and the human and yeast spliceosomes is the presence of two proteins in both human and yeast stemming from a single SNF protein in LECA. In early studies, the single SNF protein in *Drosophila melanogaster* was seen as the derived state and two separate proteins, U1A and U2B″, were proposed to represent the ancestral state (Polycarpou-Schwarz et al. 1996; Williams and Hall 2010).

Downloaded from https://academic.oup.com/mbe/article/40/1/msad011/6985000 by Wageningen University and Research – Library user on 27 May 2023

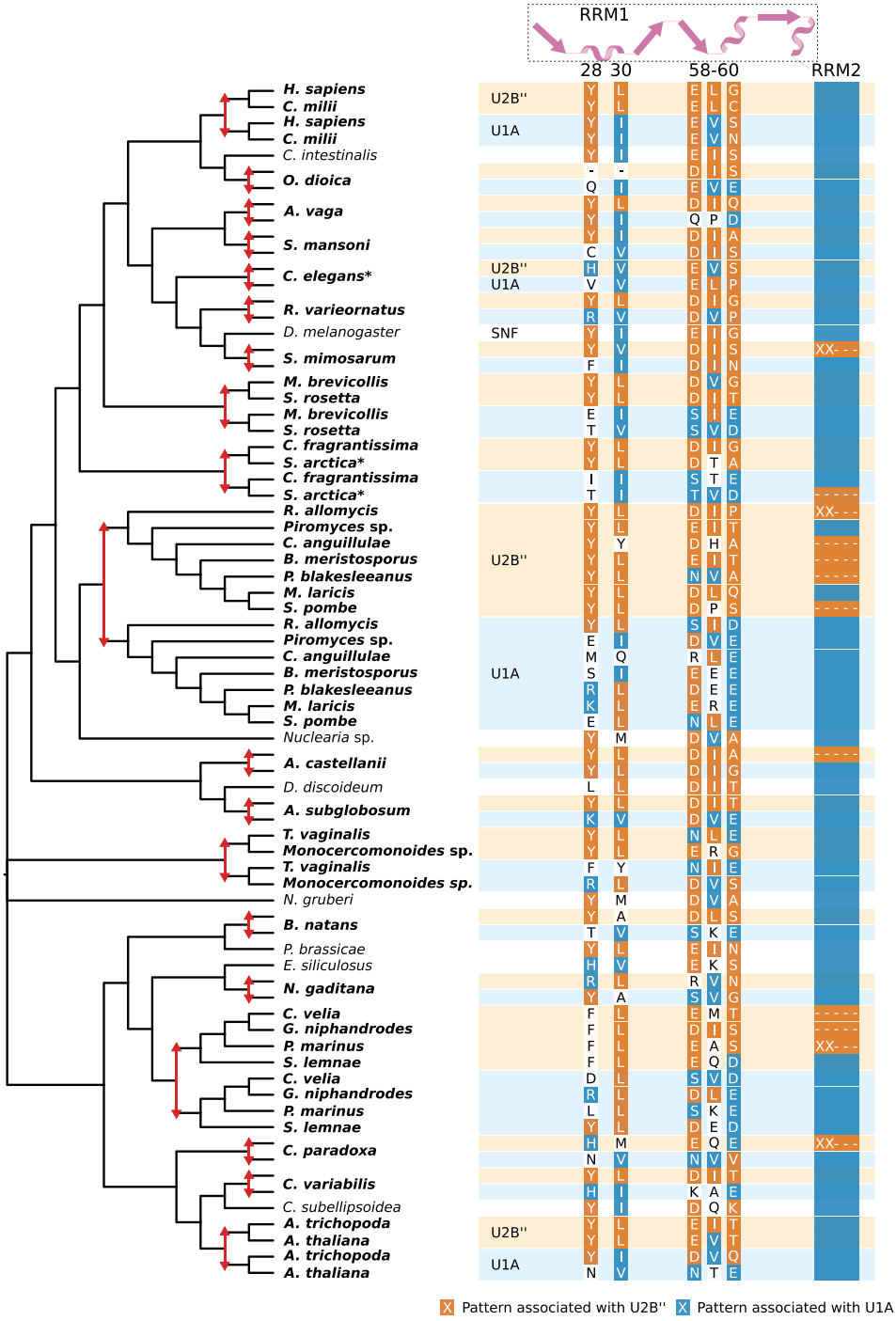


FIG. 4. Independent gene duplications and recurrent sequence evolution in the SNF family. The reconciled tree (see supplementary fig. S6a, Supplementary Material online for the full tree) shows the positions of gene duplications (red arrows) and the species names with duplicates are in bold. The colored rectangles next to the species names correspond with the predicted fate of the duplicates. The most prominent recurrent patterns are depicted with colors corresponding with the fate this pattern is associated with. For the second RRM (RRM2), the pattern is the presence (blue bar), absence (dashes), or partial presence (XX---) of this domain. The secondary structure of the first RRM (RRM1) and the position of the patterns in the *D. melanogaster* sequence are shown at the top. The duplications in *Sphagnum fallax* and *Emiliania huxleyi* are not shown because the duplicates are identical for the positions that are displayed.

However, with the availability of more genomes, the human and yeast proteins were shown with high confidence as the result of separate gene duplications (Williams et al. 2013). Additional SNF duplications were identified in other animal lineages (Williams et al. 2013). We observed even more independent SNF duplications, 22 in total using our set of eukaryotic genomes (supplementary fig. S6a, Supplementary Material online). *Guillardia theta* even had an additional third one, probably from the secondary endosymbiont (Supplementary Information).

Downloaded from https://academic.oup.com/mbe/article/40/1/msad011/6985000 by Wageningen University and Research – Library user on 27 May 2023

Drosophila SNF has a dual role in the spliceosome. It is part of the U1 snRNP, where it binds U1 snRNA, and part of the U2 snRNP, where it binds U2 snRNA and U2A' (Weber et al. 2018). In humans and yeast, U1A and U2B″ have subfunctionalized and perform the respective functions as indicated by the snRNP in their name. To assess whether a similar subfunctionalization has occurred in other lineages where SNF had duplicated, we looked for patterns of recurrent sequence evolution in the different paralogs with our previously published pipeline (von der Dunk and Snel 2020). Two fates could be distinguished, which we refer to as U1A and U2B″ based on the fates in model organisms. This distinction was based on a diffuse, mainly U1A-specific signal. Upon inspection of the two fate clusters and comparison with single SNF orthologs, the fate separation seemed to be predominantly based on recurrent substitutions in the first RRM of U1A and the recurrent loss of the second RRM in U2B″ (fig. 4). We inferred 16 RRM loss events in U2B″-fate proteins (supplementary fig. S6*b*, Supplementary Material online). These recurrent sequence changes allow us to predict which inparalog is likely to have a U1A function and which one has a U2B" function in organisms where detailed biochemical studies are lacking. Besides these remarkable findings on recurrent sequence evolution, the repeated post-LECA duplications suggest that the complexification of the spliceosome by duplication during eukaryogenesis could, in part, have been driven by the same process as happened multiple times after LECA.

Discussion

A Chimeric Complex Spliceosome that Postdates the Initial Proliferation of Introns Through the Genome

The spliceosome is one of the most complex molecular machines in present-day eukaryotes. In this study, we reconstructed the composition of the spliceosome in LECA and traced the sometimes byzantine evolutionary histories of these 145 inferred spliceosomal proteins prior to LECA. Previous research has established that the core of the spliceosome—the U2, U5, and U6 snRNAs and PRPF8—as well as the spliceosomal introns themselves evolved from self-splicing group II introns (Zimmerly and Semper 2015). Proteins of archaeal and bacterial origin were added to this core, especially proteins that performed a function in ribosome biogenesis or translation. For many proteins, we could not detect other homologous proteins, suggesting that the primordial spliceosome expanded with spliceosome-specific folds. Subsequent expansions resulted from the numerous gene duplications that we observed. These duplications enabled us to assess the extent of intron positions that were shared between paralogs and likely predated the duplication event (Vosseberg et al. 2022). Our ancestral intron position reconstructions support the presence of introns in almost half of the proteins before their recruitment into the spliceosome. This suggests that introns were already widespread through the genome when most components of the complex spliceosome emerged. The increase in spliceosomal complexity did not coincide with the initial widespread increase in intron numbers, but followed it instead. Additional introns were probably inserted in spliceosomal genes after duplication. We propose a scenario in which intragenic introns emerged early in eukaryogenesis and the complex spliceosome emerged relatively late.

The group II introns that gave rise to the spliceosomal introns are commonly proposed to have come from the protomitochondrion (Cavalier-Smith 1991; Martin and Koonin 2006). Notwithstanding the extent of horizontal gene transfer of organellar group II introns among eukaryotes (Zimmerly et al. 2001), group II introns were probably present in the mitochondria in LECA (Kim et al. 2022). Our analysis did not yield sufficient phylogenetic signal to confidently position PRPF8 in the IEP tree. However, the identification of multiple intron types in Asgard archaea makes an alternative scenario plausible, in which group II introns were present in the archaeal genome before mitochondrial endosymbiosis (Vosseberg and Snel 2017; Vosseberg et al. 2022).

Proteins involved in the assembly and functioning of another large RNP in the cell, the ribosome, became part of the primordial spliceosome, supplemented with other RNA-binding proteins. The evolutionary link with the ribosome emphasizes the comparable composition as a RNP with catalyzing RNA molecules (ribozymes). In contrast with the other spliceosomal snRNAs, the U1/U11 and U4/U4atac snRNAs did probably not originate from the introns themselves. However, an evolutionary link with translation and rRNA processing is present for these snRNAs too. U1/U11 snRNA likely evolved from a tRNA (Hogeweg and Konings 1985). The evolutionary histories of SNU13 and PRPF31 and similarities between U4 and C/D-box RNAs suggest that the U4(atac) snRNP evolved from a C/D-box snoRNP (Watkins et al. 2000).

The contribution of gene duplications in shaping the LECA spliceosome is in line with the central role of duplications in establishing eukaryotic features during eukaryogenesis (Makarova et al. 2005; Vosseberg et al. 2021). Gene duplications were key for the emergence of spliceosome-specific proteins from proteins that were part of other complexes as well as for expanding proteins that were already part of the spliceosome. This pattern has also been observed for the kinetochore (Tromer et al. 2019). These kinetochore proteins, however, came from a wider variety of cellular processes compared with the spliceosome. The origin of another eukaryote-specific complex, the nuclear pore, compares well with the spliceosome regarding the chimeric prokaryotic ancestry of its components (Mans et al. 2004). This is unlike complexes and processes that predated eukaryogenesis, such as transcription and translation, which have a more consistent phylogenetic signal (Pittis and Gabaldón 2016; Vosseberg et al. 2021).

Downloaded from https://academic.oup.com/mbe/article/40/1/msad011/6985000 by Wageningen University and Research – Library user on 27 May 2023

Origin of Two Types of Introns and Two Types of Spliceosomes

Two types of introns were present in the LECA genome, U2 and U12, which were removed from the primary transcripts by the LECA major and minor spliceosome, respectively (Russell et al. 2006). The far majority of introns were probably of the U2-type (Vosseberg et al. 2022). Different scenarios have been postulated for the emergence of two types of introns (Burge et al. 1998). In some scenarios, different intron types diverged from an ancestral set of introns, either in the same proto-eukaryotic lineage or two separate lineages that later fused. An alternative scenario proposes that the two types of introns originated from two separate introductions of group II introns in the genome. Previously, we called the separate introductions scenario unlikely based on the observed U12-type introns that are shared between proto-eukaryotic paralogs (Vosseberg et al. 2022). The enormous overlap in composition between the major and minor spliceosomes (Turunen et al. 2013; Bai et al. 2021) refutes separate origins of these complexes from different group II introns. Many minor-spliceosome-specific proteins have a close homolog in the major spliceosome and the minor-spliceosome-specific snRNAs have equivalents in the other spliceosome type. This suggests that the divergence between the major and minor spliceosomes occurred relatively late in pre-LECA spliceosome evolution, after the addition of U1 and U4 snRNA and U1 and U2 snRNP proteins. The minor-spliceosome-specific proteins were estimated to have accumulated fewer substitutions after the duplications that separated major- and minor-spliceosome-specific OGs. This suggests that the latter better reflect the ancestral situation. The U12-type introns and the minor spliceosome might, therefore, have originated earlier than the abundant U2-type introns and the major spliceosome.

Evolution of Spliceosomal Complexity

During eukaryogenesis, the recruitment of proteins and gene duplications resulted in an increase in spliceosomal complexity. Spliceosomal evolution after LECA is, in most eukaryotic lineages, dominated by simplification. A clear example is the minor spliceosome, which was lost recurrently at least 23 times (Supplementary Information). Certain lineages have experienced a substantial loss of spliceosomal genes that were part of the LECA spliceosome (supplementary fig. S7, Supplementary Material online). Only 59% of the LECA OGs are present in *Saccharomyces cerevisiae*, for example. Reduced spliceosomes have also been described in red algae and diplomonads (Hudson et al. 2019; Wong et al. 2022).

The most prominent example of a more complex spliceosome after LECA is the duplication of SNF in at least 22 lineages. To the best of our knowledge, this is the highest number of independent gene duplications in eukaryotes reported so far. It is slightly more than the 16 MadBub duplications (Tromer et al. 2016) and the 20 EF1β/δ duplications that were described before (von der Dunk and Snel 2020). We detected patterns of recurrent sequence evolution in the resulting paralogs, pointing at similar fates of these paralogs across eukaryotes. Given the described fates of the SNF paralogs in vertebrates, fungi, plants, and *Caenorhabditis elegans*, a similar subfunctionalization into dedicated U1 and U2 snRNP proteins in other lineages with duplications is likely.

The recurrent loss of the second RRM in proteins with a predicted U2B″ fate suggests that the function of this RRM is mainly restricted to the U1A role. Whereas the function of the first RRM has been described as binding to U1 and U2 snRNA, the function of the second RRM has remained elusive (Williams et al. 2013). The observation of recurrent loss of this RRM in specifically U2B″ proteins provides possible directions for further molecular research.

The dual-function SNF protein seems to be poised for duplication and subsequent subdivision of the roles in the U1 and U2 snRNP. It is tempting to speculate that the recurrent duplication of SNF indicates that this specific gene duplication and subsequent subfunctionalization could, in principle, have occurred during eukaryogenesis instead. Because it did not happen to be duplicated then, it could be seen as "unfinished business" during eukaryogenesis. The cases of independent gene duplications after LECA might be used as a model for proto-eukaryotic gene duplications. Because these duplications happened relatively recently, experiments based on ancestral protein reconstructions can be performed more reliably, as has been done for the SNF family in deuterostomes (Williams et al. 2013; Delaney et al. 2014). These experiments can provide insight into the role of adaptive or neutral evolution (Finnigan et al. 2012) in creating the complex spliceosome (Vosseberg and Snel 2017).

Investigating the Emergence of the Complex Eukaryotic Cell

Our study provides a comprehensive view on the origin of the numerous proteins in this complex molecular machine, also in relation to the spread of the introns it functions on. Further studies on the spliceosome composition in diverse eukaryotes have the potential to identify more spliceosomal proteins in LECA. New developments in detecting deep homologies (Jumper et al. 2021; Monzon et al. 2022) could reveal additional links for the spliceosomal proteins that we classified as inventions in this study. Phylogenetic analyses combined with intron analyses on the numerous other complexes that emerged during eukaryogenesis could further illuminate their origin and, thereby, the major transition from prokaryotes to eukaryotes.

Materials and Methods

Data

We used a diverse set of 209 eukaryotic and 3,466 prokaryotic (predicted) proteomes, as compiled for a previous study (Vosseberg et al. 2021) from different sources

Downloaded from https://academic.oup.com/mbe/article/40/1/msad011/6985000 by Wageningen University and Research – Library user on 27 May 2023

(Huerta-Cepas et al. 2016; Zaremba-Niedzwiedzka et al. 2017; Deutekom et al. 2019). Proteins from 167 of the eukaryotic species had been grouped in OGs using different approaches (Deutekom et al. 2021). To illuminate the evolutionary history of some protein families (see below), we made use of the widely expanded set of Asgard archaeal genomes that has come available since. By including genomes from numerous studies (Liu et al. 2018, 2021; Tully et al. 2018; Huang et al. 2019; Seitz et al. 2019; Imachi et al. 2020; Farag et al. 2021; Sun et al. 2021; Zhao and Biddle 2021; Wu et al. 2022), the number of Asgard archaeal proteomes in our expanded set amounted to 133 in total. If no predicted proteome was available, the genomes were annotated with Prokka v1.13 (Seemann 2014) for the genomes from (Liu et al. 2018; Seitz et al. 2019) or v1.14.6 with the metagenome option for the genomes from (Farag et al. 2021; Liu et al. 2021).

Reconstructing LECA's Spliceosome

To infer the composition of the spliceosome in LECA, we searched for orthologs of proteins in the well-studied *Homo sapiens* and *Saccharomyces cerevisiae* spliceosome complexes in other eukaryotic proteomes. A list of human and budding yeast spliceosomal proteins was obtained from the UniProt database (The UniProt Consortium 2019) on February 26, 2020, only including manually reviewed proteins (supplementary Table S1, 2, Supplementary Material online). Proteins that are involved in other processes (such as transcription and polyadenylation) and splice site selection and splicing regulation were removed. The list was supplemented with human spliceosomal proteins from recent literature (Bai et al. 2021; de Wolf et al. 2021; Sales-Lee et al. 2021). Initial evolutionary scenarios of these proteins were inferred based on the approach of Van Hooff et al. (2019). In short, the human and yeast protein sequences were searched against our in-house eukaryotic proteome database (Deutekom et al. 2019) with blastp (Altschul et al. 1990). Significant hits (E-value 0.001 or lower) in *H. sapiens*, *Xenopus tropicalis*, *D. melanogaster*, *Salpingoeca rosetta*, *S. cerevisiae*, *Schizosaccharomyces pombe*, *Spizellomyces punctatus*, *Thecamonas trahens*, *Acanthamoeba castellanii*, *Dictyostelium discoideum*, *Arabidopsis thaliana*, *Chlamydomonas reinhardtii*, *Cyanidioschyzon merolae*, *Ectocarpus siliculosus*, *Plasmodium falciparum*, *Plasmodiophora brassicae*, *Naegleria gruberi*, *Leishmania major*, *Giardia intestinalis*, and *Monocercomonoides* sp. were aligned with MAFFT v7.310 (E-INS-i option) (Katoh and Standley 2013). These alignments were trimmed with trimAl v1.4.rev15 (gappyout option) (Capella-Gutiérrez et al. 2009) and a phylogenetic tree was inferred with IQ-TREE v1.6.4 (Nguyen et al. 2015) using the LG + G4 model to establish the initial scenario (van Hooff et al. 2019): (i) easy, in case of orthologs in a diverse set of eukaryotes including at least two Opimoda and two Diphoda species (Derelle et al. 2015); (ii) ancient (pre-LECA) duplication, when the set of homologs also includes clades of more distantly related homologs across eukaryotes; (iii) lineage-specific (post-LECA) duplication, when the spliceosomal function likely originated after LECA; (iv) taxonomically limited, with homologs in a limited set of eukaryotes. The latter cases were further studied by checking hits in the complete set of eukaryotes. For SNRNP27, CASC3, and WBP11, hits to the more sensitive Pfam models PF08648, PF09405, and PF09429 (Finn et al. 2016) detected before (Vosseberg et al. 2021) were used instead of the BLAST-based homologs.

In case of an easy or ancient duplication scenario, a LECA OG was defined. If members of this LECA OG were present in both the human and yeast spliceosomes, it was classified as a LECA spliceosome OG. Yeast LIN1 (CD2BP2 ortholog) and PRP24 (SART3 ortholog) and human LUC7L and LUC7L2 (LUC7 orthologs) were not in the initial set, but their ortholog was. These were included in the original list because these were also clearly described as spliceosomal in the literature. If an ortholog was not present in yeast, spliceosomal annotations for orthologs in *S. pombe*, *A. thaliana* (both in the UniProt database) or *Cryptococcus neoformans* (Sales-Lee et al. 2021) were checked. If an ortholog was not present in humans, the function of the *A. thaliana* ortholog was investigated. If these orthologs were not characterized, they were classified as spliceosomal in LECA if their close paralog was also in the spliceosome, or if they only had an annotated spliceosomal function. If their main function was in the spliceosome or if they were not well-characterized, they were classified as possibly spliceosomal. In case of multiple functions, the OG was discarded. The reconstruction of spliceosome OGs in LECA is summarized in supplementary Table S3, Supplementary Material online.

Inferring pre-LECA Evolutionary Histories

To trace the pre-LECA histories of the inferred spliceosomal LECA proteins, we performed phylogenetic analyses of these proteins with other eukaryotic OGs and with prokaryotic proteins that are homologous to the spliceosomal proteins. We started by analyzing the domain composition of the proteins and looking for these domains or full-length proteins in trees that we created for a previous study (Vosseberg et al. 2021). Additional phylogenetic analyses were performed for the families described below. Multiple sequence alignments were made with MAFFT v7.310 (Katoh and Standley 2013) and, subsequently, trimmed to remove parts of the alignment of low quality with trimAl v1.4.rev15 (Capella-Gutiérrez et al. 2009) or Divvier v1.0 (Ali et al. 2019) (maximum of 50% gaps per position). The chosen options per family are shown in supplementary Table S6, Supplementary Material online. Phylogenetic trees were inferred using IQ-TREE v2.1.3 (Minh et al. 2020) with the best substitution model among nuclear models including LG + C{10,20,30,40,50,60} mixture models identified by ModelFinder (Kalyaanamoorthy et al. 2017). Mixtures models with an F-class were not considered, as recently recommended (Baños et al. 2022). Branch supports were calculated with 1,000 ultrafast bootstraps

Downloaded from https://academic.oup.com/mbe/article/40/1/msad011/6985000 by Wageningen University and Research – Library user on 27 May 2023

(Hoang et al. 2018) and the SH-like approximate likelihood ratio test (Guindon et al. 2010). Topologies were compared using the approximately unbiased test (Shimodaira 2002) with 10,000 replicates.

Duplications that resulted in spliceosomal OGs were functionally annotated based on the function of other homologous sequences and the tree topology using Dollo parsimony. In two cases for which we could only detect one paralogous OG with a non-spliceosomal function, the preduplication ancestor was annotated with the non-spliceosomal function. Clades containing only within-spliceosome duplications were collapsed to obtain ancestral spliceosomal units.

IEP-PRPF8

Representative sequences of prokaryotic and organellar IEP sequences and other prokaryotic RT-containing sequences were chosen from two datasets (Candales et al. 2012; Toro and Nisa-Martínez 2014) and supplemented with four Asgard archaeal IEP sequences (Zaremba-Niedzwiedzka et al. 2017). We also selected slowly evolving representatives for PRPF8 and TERT. For the tree that included PRPF8 and TERT, separate alignments were made for the prokaryotic and organellar IEP (E-INS-i algorithm), PRPF8 and TERT sequences (both with L-INS-i). We extracted the RT fingers-palm and thumb domains from these alignments based on a published structural alignment (Qu et al. 2016). The extracted domains were aligned and a tree was inferred. A constrained tree search with a monophyletic PRPF8 and TERT clade was additionally performed.

We used eggNOG 4.5 (Huerta-Cepas et al. 2016) annotations to identify additional Asgard archaeal IEPs by executing emapper-1.0.3 (Huerta-Cepas et al. 2017) with DIAMOND v0.8.22.84 (Buchfink et al. 2015) searches on the expanded Asgard set. Proteins assigned to COG3344 were combined with the selection of IEP sequences; non-IEP COG3344 hits were discarded based on a preliminary phylogenetic tree.

AAR2

Only three prokaryotic AAR2 homologs were detected in the initial dataset based on hits to the PF05282 model (Vosseberg et al. 2021), one in *Limnospira maxima* and two in *Lokiarchaeum*. We used the same approach to detect additional hits in the expanded set of Asgard archaea by running hmmsearch (HMMER v3.3.2 (Eddy 2011)) with the Pfam 31.0 hidden Markov models (Finn et al. 2016) using the gathering thresholds. Additionally, hmmsearch with the PF05282.14 model was performed on the EBI server (https://www.ebi.ac.uk/Tools/hmmer/search/hmmsearch) against the UniProtKB database on April 21, 2022.

EFTUD2

The EF2 family has undergone multiple duplications in archaeal and eukaryotic evolution resulting in two orthologs in the last Asgard archaeal common ancestor and three in LECA (Narrowe et al. 2018). The latter are represented in eukaryotic eggNOG families (euNOGs) KOG0467, KOG0468, and KOG0469. To increase the phylogenetic resolution, we used a ScrollSaw-inspired approach (Elias et al. 2012; Vosseberg et al. 2021; van Wijk and Snel 2022) to select slowly evolving sequences from four main eukaryotic clades (Amorphea, Diaphoretickes, Discoba, and Metamonada). Asgard archaeal sequences assigned to COG0480 were aligned with E-INS-i. The alignment was trimmed with trimAl (-gt 0.5) and a tree was inferred using the LG + G4 model. Hodarchaeal representatives and other Asgard sequences from the same Asgard archaeal OG (see Supplementary Information) were combined with the eukaryotic sequences.

PRPF31 and SNU13

For PRPF31, the sequences in the PF01798 tree were replaced with the corresponding full-length sequences to increase the phylogenetic signal. Based on the PF01248 tree, which includes SNU13, we chose two slowly evolving Opimoda and two Diphoda sequences (Derelle et al. 2015) per OG, supplemented with the archaeal RPL7Ae sequences. Full-length sequences were used for subsequent phylogenetic inference.

DDX Helicases

Slowly evolving eukaryotic DDX helicase sequences were selected using the ScrollSaw-based approach on the sequences that were assigned to the euNOGs that are part of the COG0513 cluster (Makarova et al. 2005). An alignment of these sequences was created (E-INS-i, trimAl -gt 0.5) and a phylogenetic tree was inferred with FastTree v2.1.10 (LG model) (Price et al. 2010). From this tree, we selected per OG the sequence on the shortest branch for each of the four eukaryotic clades (if present and not on a deviating long branch). The selected sequences were split into the two inferred acquisitions and combined with prokaryotic COG0513 representatives.

DHX Helicases

A similar approach as for the DDX helicases was applied to the COG1643 cluster (Makarova et al. 2005). The initial tree was based on an alignment created with E-INS-i and trimAl (gappyout option) and made using the LG + F + R8 model in IQ-TREE. An unclear clade with multiple OGs was reduced and sequences from the missing DHX40 OG were added.

LSM

To elucidate the pre-LECA history of the Lsm/Sm proteins, we initially made a tree combining the eukaryotic sequences from LECA OGs in the Sm-like Pfam clan (PF01423, PF12701, and PF14438). We selected slowly evolving sequences as described for the DDX and DHX helicases from the resulting tree (alignment with FFT-NS-I, trimming with trimAl (-gt 0.1), tree with the LG + G4 model). LSM14 and ATXN2 were not included in the selection because of their divergent nature. The full-length sequences in the expanded

Downloaded from https://academic.oup.com/mbe/article/40/1/msad011/6985000 by Wageningen University and Research – Library user on 27 May 2023

set of Asgard archaea that were PF01423 hits were used for the SmAP tree. We selected representatives from the different clades and combined these with the full-length versions of the previously selected eukaryotic sequences. We also performed a constrained tree search with one monophyletic eukaryotic clade.

RRM and TXNL4

We identified LECA OGs in the PF00076 (RRM) tree based on automatic annotation and manual assessment (i.e., a high support value and substantial pre-LECA branch length). Per OG, the Opimoda and Diphoda sequences on the shortest branch were selected. For the different subtrees, we selected full-length sequences in the OGs from *H. sapiens*, *A. castellanii*, *A. thaliana*, *Aphanomyces astaci*, *Monocercomonoides* sp., and *N. gruberi*. For RBM41, the *Selaginella moellendorffii* sequence was included to replace the missing *A. thaliana* ortholog. To illustrate the relationship between TXNL4A and TXNL4B in the larger thioredoxin family, we used orthologs from the same species as chosen for the RRM subtrees.

U1-type Zinc Finger

Slowly evolving sequences from the euNOGs in the smart00451 cluster (Makarova et al. 2005), supplemented with the SCNM1 euNOG ENOG410IW6J, were selected with the aforementioned ScrollSaw-based approach. These sequences were aligned with the E-INS-i algorithm and the resulting alignment was trimmed with trimAl (-gt 0.25). Based on the inferred tree with the VT + R4 model, we selected the shortest branching sequences per OG from each of the four eukaryotic groups.

WD40

The ScrollSaw-based approach was also applied to the euNOGs in the COG2319 cluster (Makarova et al. 2005), using bidirectional best hits between the Opimoda and Diphoda species instead, because of the size of this protein family. An alignment of the selected sequences was made (E-INS-i, trimAl gappyout) and a tree was inferred (LG + R4 model). Per OG, the shortest branching Opimoda and Diphoda sequences were chosen. PPWD1 and some potential sister OGs based on the BLAST trees were not in the COG2319 cluster. We followed a similar approach to identify slowly evolving sequences for these euNOGs (KOG0882, ENOG410IQTX, −0KD7K, and −0IF90), using a different gap threshold (50%) and substitution model (LG + R3). Based on the BLAST trees and the COG2319 cluster tree, we identified potential sister OGs and inferred a tree with these OGs and the spliceosomal OGs.

Ancestral Intron Position Reconstructions

We performed ancestral intron position reconstructions for the identified pre-LECA paralogs in the entire clade or only for the spliceosomal OGs and sister OGs (supplementary Table S5, Supplementary Material online), depending on the number of OGs in an acquisition or invention. To establish the content of the OGs, we started with the euNOG assignments. If the taxonomic distribution of the euNOG was limited, we continued with the Broccoli (Derelle et al. 2020) OG assignments (Deutekom et al., 2021). A phylogenetic tree of the OG was inferred to check for the presence of non-orthologous or dubious sequences and remove these (E-INS-i, trimAl -gt 0.5 or -gappyout, FastTree -lg). After cleaning up the OGs, a final E-INS-i alignment was made. Except for the alignment with PRPF8 and TERT, which was based on the RT domain (see "IEP-PRPF8' above), full-length sequences were used for this alignment. Intron positions were mapped onto the alignment using the method described before (Vosseberg et al. 2022). LECA introns were inferred with Malin (Csűrös 2008) using the intron gain and loss rates that we previously estimated for the KOG clusters (Vosseberg et al. 2022). Pre-duplication introns were inferred using Dollo parsimony.

Recurrent Duplication and Subfunctionalization of SNF

To identify post-LECA duplications, SNF sequences were aligned with E-INS-i and this alignment was trimmed with Divvier. The SNF tree was inferred with the LG + C50 + R6 model and manually reconciled with the species tree to annotate gene duplication events. We looked at potential duplications in more detail by remaking trees of specific parts of the tree, including additional species from our original set (Deutekom et al. 2019). Prior to making the final alignment, we removed additional in-paralogs, probable fission events or partial annotations and the sequences from *Guillardia theta*, which had likely acquired a third copy from its endosymbiont. The final alignment was made with the E-INS-i algorithm. This alignment and the annotated duplication events were used as input for our previously published pipeline to identify patterns of recurrent sequence evolution after independent gene duplications (von der Dunk and Snel 2020).

Statistical Analysis

Statistical analyses were performed in Python using NumPy v1.21.141 (Harris et al. 2020) and pandas v1.3.142 (McKinney 2010). Figures were created with Matplotlib v3.4.245 (Hunter 2007), seaborn v0.11.146 (Waskom 2021), and FigTree v1.4.3 (https://github.com/rambaut/figtree).

Supplementary material

Supplementary data are available at *Molecular Biology and Evolution* online.

Data Availability

Fasta files, phylogenetic trees, and mapped intron files are available in figshare (https://doi.org/10.6084/m9.figshare.20653575).

Downloaded from https://academic.oup.com/mbe/article/40/1/msad011/6985000 by Wageningen University and Research – Library user on 27 May 2023

Acknowledgments

We thank the members of the Theoretical Biology & Bioinformatics group for useful discussions. This work is part of the research program VICI with project number 016.160.638, which is financed by the Netherlands Organisation for Scientific Research (NWO).

Author Contributions

J.V. and B.S. conceived the study. J.V. and D.S. performed the research. S.H.A.v.d.D. aided with the recurrent sequence evolution analysis of the SNF family. J.V., D.S., and B.S. analyzed and interpreted the results. J.V. wrote the manuscript, which was edited and approved by the other authors.

References

Aittaleb M, Rashid R, Chen Q, Palmer JR, Daniels CJ, Li H. 2003. Structure and function of archaeal box C/D sRNP core proteins. *Nat Struct Mol Biol*. **10**:256–263.

Ali RH, Bogusz M, Whelan S. 2019. Identifying clusters of high confidence homologies in multiple sequence alignments. *Mol Biol Evol*. **36**:2340–2351.

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. 1990. Basic local alignment search tool. *J Mol Biol*. **215**:403–410.

Anantharaman V, Koonin EV, Aravind L. 2002. Comparative genomics and evolution of proteins involved in RNA metabolism. *Nucleic Acids Res*. **30**:1427–1464.

Bai R, Wan R, Wang L, Xu K, Zhang Q, Lei J, Shi Y. 2021. Structure of the activated human minor spliceosome. *Science*. **371**:eabg0879.

Baños H, Susko E, Roger AJ, unpublished data, https://www.biorxiv.org/content/10.1101/2022.02.18.481053v1, last accessed December 9, 2022.

Breuer R, Gomes-Filho J-V, Randau L. 2021. Conservation of archaeal C/D box sRNA-guided RNA modifications. *Front Microbiol*. **12**:654029.

Buchfink B, Xie C, Huson DH. 2015. Fast and sensitive protein alignment using DIAMOND. *Nat Methods*. **12**:59–60.

Burge CB, Padgett RA, Sharp PA. 1998. Evolutionary fates and origins of U12-type introns. *Mol Cell*. **2**:773–785.

Califice S, Baurain D, Hanikenne M, Motte P. 2012. A single ancient origin for prototypical serine/arginine-rich splicing factors. *Plant Physiol*. **158**:546–560.

Candales MA, Duong A, Hood KS, Li T, Neufeld RAE, Sun R, McNeil BA, Wu L, Jarding AM, Zimmerly S. 2012. Database for bacterial group II introns. *Nucleic Acids Res*. **40**:D187–D190.

Capella-Gutiérrez S, Silla-Martínez JM, Gabaldón T. 2009. Trimal: a tool for automated alignment trimming in large-scale phylogenetic analyses. *Bioinformatics*. **25**:1972–1973.

Cavalier-Smith T. 1991. Intron phylogeny: a new hypothesis. *Trends Genet*. **7**:145–148.

Charollais J, Dreyfus M, Iost I. 2004. Csda, a cold-shock RNA helicase from *Escherichia coli*, is involved in the biogenesis of 50S ribosomal subunit. *Nucleic Acids Res*. **32**:2751–2759.

Collins L, Penny D. 2005. Complex spliceosomal organization ancestral to extant eukaryotes. *Mol Biol Evol*. **22**:1053–1066.

Csűrös M. 2008. Malin: maximum likelihood analysis of intron evolution in eukaryotes. *Bioinformatics*. **24**:1538–1539.

Csuros M, Rogozin IB, Koonin EV. 2011. A detailed history of intron-rich eukaryotic ancestors inferred from a global survey of 100 complete genomes. *PLoS Comput Biol*. **7**:e1002150.

Delaney KJ, Williams SG, Lawler M, Hall KB. 2014. Climbing the vertebrate branch of U1A/U2B″ protein evolution. *RNA*. **20**:1035–1045.

Derelle R, Philippe H, Colbourne JK. 2020. Broccoli: combining phylogenetic and network analyses for orthology assignment. *Mol Biol Evol*. **37**:3389–3396.

Derelle R, Torruella G, Klimeš V, Brinkmann H, Kim E, Vlček Č, Lang BF, Eliáš M. 2015. Bacterial proteins pinpoint a single eukaryotic root. *Proc Natl Acad Sci U S A*. **112**:E693–E699.

Deutekom ES, Snel B, van Dam TJP. 2021. Benchmarking orthology methods using phylogenetic patterns defined at the base of eukaryotes. *Brief Bioinform*. **22**:bbaa206.

Deutekom ES, Vosseberg J, van Dam TJP, Snel B. 2019. Measuring the impact of gene prediction on gene loss estimates in eukaryotes by quantifying falsely inferred absences. *PLoS Comput Biol*. **15**:e1007301.

de Wolf B, Oghabian A, Akinyi MV, Hanks S, Tromer EC, van Hooff JJE, van Voorthuijsen L, van Rooijen LE, Verbeeren J, Uijttewaal ECH, *et al.* 2021. Chromosomal instability by mutations in the novel minor spliceosome component CENATAC. *EMBO J*. **40**:e106536.

Dlakić M, Mushegian A. 2011. Prp8, the pivotal protein of the spliceosomal catalytic center, evolved from a retroelement-encoded reverse transcriptase. *RNA*. **17**:799–808.

Eddy SR. 2011. Accelerated profile HMM searches. *PLoS Comput Biol*. **7**:e1002195.

Elias M, Brighouse A, Gabernet-Castello C, Field MC, Dacks JB. 2012. Sculpting the endomembrane system in deep time: high resolution phylogenetics of rab GTPases. *J Cell Sci*. **125**:2500–2508.

Farag IF, Zhao R, Biddle JF. 2021. Sifarchaeota," a novel asgard phylum from Costa Rican sediment capable of polysaccharide degradation and anaerobic methylotrophy. *Appl Environ Microbiol*. **87**:e02584–20.

Finn RD, Coggill P, Eberhardt RY, Eddy SR, Mistry J, Mitchell AL, Potter SC, Punta M, Qureshi M, Sangrador-Vegas A, *et al.* 2016. The pfam protein families database: towards a more sustainable future. *Nucleic Acids Res*. **44**:D279–D285.

Finnigan GC, Hanson-Smith V, Stevens TH, Thornton JW. 2012. Evolution of increased complexity in a molecular machine. *Nature*. **481**:360–364.

Guindon S, Dufayard J-F, Lefort V, Anisimova M, Hordijk W, Gascuel O. 2010. New algorithms and methods to estimate Maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. *Syst Biol*. **59**:307–321.

Harris CR, Millman KJ, van der Walt SJ, Gommers R, Virtanen P, Cournapeau D, Wieser E, Taylor J, Berg S, Smith NJ, *et al.* 2020. Array programming with NumPy. *Nature*. **585**:357–362.

Hoang DT, Chernomor O, von Haeseler A, Minh BQ, Vinh LS. 2018. UFBoot2: improving the ultrafast bootstrap approximation. *Mol Biol Evol*. **35**:518–522.

Hogeweg P, Konings DAM. 1985. U1 snRNA: the evolution of its primary and secondary structure. *J Mol Evol*. **21**:323–333.

Huang J-M, Baker BJ, Li J-T, Wang Y. 2019. New microbial lineages capable of carbon fixation and nutrient cycling in deep-sea sediments of the northern South China Sea. *Appl Environ Microbiol*. **85**:e00523–19.

Hudson AJ, McWatters DC, Bowser BA, Moore AN, Larue GE, Roy SW, Russell AG. 2019. Patterns of conservation of spliceosomal intron structures and spliceosome divergence in representatives of the diplomonad and parabasalid lineages. *BMC Evol Biol*. **19**:162.

Huerta-Cepas J, Forslund K, Coelho LP, Szklarczyk D, Jensen LJ, von Mering C, Bork P. 2017. Fast genome-wide functional annotation through orthology assignment by eggNOG-mapper. *Mol Biol Evol*. **34**:2115–2122.

Huerta-Cepas J, Szklarczyk D, Forslund K, Cook H, Heller D, Walter MC, Rattei T, Mende DR, Sunagawa S, Kuhn M, *et al.* 2016. eggNOG 4.5: a hierarchical orthology framework with improved functional annotations for eukaryotic, prokaryotic and viral sequences. *Nucleic Acids Res*. **44**:D286–D293.

Hunter JD. 2007. Matplotlib: a 2D graphics environment. *Comput Sci Eng*. **9**:90–95.

Imachi H, Nobu MK, Nakahara N, Morono Y, Ogawara M, Takaki Y, Takano Y, Uematsu K, Ikuta T, Ito M, *et al.* 2020. Isolation of an

Downloaded from https://academic.oup.com/mbe/article/40/1/msad011/6985000 by Wageningen University and Research – Library user on 27 May 2023

archaeon at the prokaryote–eukaryote interface. *Nature*. **577**: 519–525.
Jain C. 2008. The *E. coli* RhlE RNA helicase regulates the function of related RNA helicases during ribosome assembly. *RNA*. **14**: 381–389.
Jumper J, Evans R, Pritzel A, Green T, Figurnov M, Ronneberger O, Tunyasuvunakool K, Bates R, Žídek A, Potapenko A, *et al.* 2021. Highly accurate protein structure prediction with AlphaFold. *Nature*. **596**:583–589.
Kalyaanamoorthy S, Minh BQ, Wong TKF, von Haeseler A, Jermiin LS. 2017. Modelfinder: fast model selection for accurate phylogenetic estimates. *Nat Methods*. **14**:587–589.
Katoh K, Standley DM. 2013. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol Biol Evol*. **30**:772–780.
Kim D, Lee J, Cho CH, Kim EJ, Bhattacharya D, Yoon HS. 2022. Group II intron and repeat-rich red algal mitochondrial genomes demonstrate the dynamic recent history of autocatalytic RNAs. *BMC Biol*. **20**:2.
Kufel J, Allmang C, Petfalski E, Beggs J, Tollervey D. 2003. Lsm proteins are required for normal processing and stability of ribosomal RNAs*. *J Biol Chem*. **278**:2147–2156.
Lekontseva NV, Stolboushkina EA, Nikulin AD. 2021. Diversity of LSM family proteins: similarities and differences. *Biochem (Mosc)*. **86**:S38–S49.
Liu Y, Makarova KS, Huang W-C, Wolf YI, Nikolskaya AN, Zhang X, Cai M, Zhang C-J, Xu W, Luo Z, *et al.* 2021. Expanded diversity of asgard archaea and their relationships with eukaryotes. *Nature*. **593**:553–557.
Liu Y, Zhou Z, Pan J, Baker BJ, Gu J-D, Li M. 2018. Comparative genomic inference suggests mixotrophic lifestyle for thorarchaeota. *ISME J*. **12**:1021–1031.
Lo Gullo G, De Santis ML, Paiardini A, Rosignoli S, Romagnoli A, La Teana A, Londei P, Benelli D. 2021. The archaeal elongation factor EF-2 induces the release of aIF6 from 50S ribosomal subunit. *Front Microbiol*. **12**:631297.
Makarova KS, Wolf YI, Mekhedov SL, Mirkin BG, Koonin EV. 2005. Ancestral paralogs and pseudoparalogs and their role in the emergence of the eukaryotic cell. *Nucleic Acids Res*. **33**:4626–4638.
Mans B, Anantharaman V, Aravind L, Koonin EV. 2004. Comparative genomics, evolution and origins of the nuclear envelope and nuclear pore Complex. *Cell Cycle*. **3**:1625–1650.
Maris C, Dominguez C, Allain FH-T. 2005. The RNA recognition motif, a plastic RNA-binding platform to regulate post-transcriptional gene expression. *FEBS J*. **272**:2118–2131.
Martin W, Koonin EV. 2006. Introns and the origin of nucleus–cytosol compartmentalization. *Nature*. **440**:41–45.
McKinney W. 2010. Data Structures for Statistical Computing in Python. In Proceedings of the 9th Python in Science Conference (SciPy 2010). Austin (TX). p. 56–61.
Minh BQ, Schmidt HA, Chernomor O, Schrempf D, Woodhams MD, von Haeseler A, Lanfear R. 2020. IQ-TREE 2: new models and efficient methods for phylogenetic inference in the genomic era. *Mol Biol Evol*. **37**:1530–1534.
Miura MC, Nagata S, Tamaki S, Tomita M, Kanai A. 2022. Distinct expansion of group II introns during evolution of prokaryotes and possible factors involved in its regulation. *Front Microbiol*. **13**:849080.
Monzon V, Paysan-Lafosse T, Wood V, Bateman A. 2022. Reciprocal best structure hits: using AlphaFold models to discover distant homologues. *Bioinform Adv*. **2**:vbac072.
Moyer DC, Larue GE, Hershberger CE, Roy SW, Padgett RA. 2020. Comprehensive database and evolutionary dynamics of U12-type introns. *Nucleic Acids Res*. **48**:7066–7078.
Mura C, Randolph PS, Patterson J, Cozen AE. 2013. Archaeal and eukaryotic homologs of hfq. *RNA Biol*. **10**:636–651.
Narrowe AB, Spang A, Stairs CW, Caceres EF, Baker BJ, Miller CS, Ettema TJG. 2018. Complex evolutionary history of translation elongation factor 2 and diphthamide biosynthesis in archaea and parabasalids. *Genome Biol Evol*. **10**:2380–2393.
Nguyen L-T, Schmidt HA, von Haeseler A, Minh BQ. 2015. IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. *Mol Biol Evol*. **32**:268–274.
Nottrott S, Urlaub H, Lührmann R. 2002. Hierarchical, clustered protein interactions with U4/U6 snRNA: a biochemical role for U4/U6 proteins. *EMBO J*. **21**:5527–5538.
Pittis AA, Gabaldón T. 2016. Late acquisition of mitochondria by a host with chimaeric prokaryotic ancestry. *Nature*. **531**:101–104.
Polycarpou-Schwarz M, Gunderson SI, Kandels-Lewis S, Seraphin B, Mattaj IW. 1996. *Drosophila* SNF/D25 combines the functions of the two snRNP proteins U1A and U2B″ that are encoded separately in human, potato, and yeast. *RNA*. **2**:11–23.
Price MN, Dehal PS, Arkin AP. 2010. Fasttree 2—approximately maximum-likelihood trees for large alignments. *PLoS One*. **5**: e9490.
Qu G, Kaushal PS, Wang J, Shigematsu H, Piazza CL, Agrawal RK, Belfort M, Wang H-W. 2016. Structure of a group II intron in complex with its reverse transcriptase. *Nat Struct Mol Biol*. **23**: 549–557.
Russell AG, Charette JM, Spencer DF, Gray MW. 2006. An early evolutionary origin for the minor spliceosome. *Nature*. **443**:863–866.
Sales-Lee J, Perry DS, Bowser BA, Diedrich JK, Rao B, Beusch I, Yates JR, Roy SW, Madhani HD. 2021. Coupling of spliceosome complexity to intron diversity. *Curr Biol*. **31**:4898–4910.
Seemann T. 2014. Prokka: rapid prokaryotic genome annotation. *Bioinformatics*. **30**:2068–2069.
Seitz KW, Dombrowski N, Eme L, Spang A, Lombard J, Sieber JR, Teske AP, Ettema TJG, Baker BJ. 2019. Asgard archaea capable of anaerobic hydrocarbon cycling. *Nat Commun*. **10**:1822.
Shimodaira H. 2002. An approximately unbiased test of phylogenetic tree selection. *Syst Biol*. **51**:492–508.
Sun J, Evans PN, Gagen EJ, Woodcroft BJ, Hedlund BP, Woyke T, Hugenholtz P, Rinke C. 2021. Recoding of stop codons expands the metabolic potential of two novel asgardarchaeota lineages. *ISME Commun*. **1**:30.
The UniProt Consortium. 2019. Uniprot: a worldwide hub of protein knowledge. *Nucleic Acids Res*. **47**:D506–D515.
Toro N, Nisa-Martínez R. 2014. Comprehensive phylogenetic analysis of bacterial reverse transcriptases. *PLoS One*. **9**:e114083.
Tromer E, Bade D, Snel B, Kops GJPL. 2016. Phylogenomics-guided discovery of a novel conserved cassette of short linear motifs in BubR1 essential for the spindle checkpoint. *Open Biol*. **6**:160315.
Tromer EC, van Hooff JJE, Kops GJPL, Snel B. 2019. Mosaic origin of the eukaryotic kinetochore. *Proc Natl Acad Sci U S A*. **116**: 12873–12882.
Tully BJ, Graham ED, Heidelberg JF. 2018. The reconstruction of 2,631 draft metagenome-assembled genomes from the global oceans. *Sci Data*. **5**:170203.
Turunen JJ, Niemelä EH, Verma B, Frilander MJ. 2013. The significant other: splicing by the minor spliceosome. *Wiley Interdiscip Rev RNA*. **4**:61–76.
van Hooff JJE, Tromer E, van Dam TJP, Kops GJPL, Snel B. 2019. Inferring the evolutionary history of your favorite protein: a guide for molecular biologists. *BioEssays*. **41**:1900006.
van Wijk LM, Snel B, unpublished data, https://www.biorxiv.org/content/10.1101/2020.01.27.920793v2, last accessed December 9, 2022.
Veretnik S, Wills C, Youkharibache P, Valas RE, Bourne PE. 2009. Sm/Lsm genes provide a glimpse into the early evolution of the spliceosome. *PLoS Comput Biol*. **5**:e1000315.
von der Dunk SHA, Snel B. 2020. Recurrent sequence evolution after independent gene duplication. *BMC Evol Biol*. **20**:98.
Vosseberg J, Schinkel M, Gremmen S, Snel B. 2022. The spread of the first introns in proto-eukaryotic paralogs. *Commun Biol*. **5**:476.
Vosseberg J, Snel B. 2017. Domestication of self-splicing introns during eukaryogenesis: the rise of the complex spliceosomal machinery. *Biol Direct*. **12**:30.
Vosseberg J, van Hooff JJE, Marcet-Houben M, van Vlimmeren A, van Wijk LM, Gabaldón T, Snel B. 2021. Timing the origin of

Downloaded from https://academic.oup.com/mbe/article/40/1/msad011/6985000 by Wageningen University and Research – Library user on 27 May 2023

eukaryotic cellular complexity with ancient duplications. *Nat Ecol Evol*. **5**:92–100.
Waskom ML. 2021. Seaborn: statistical data visualization. *J Open Source Softw*. **6**:3021.
Watkins NJ, Ségault V, Charpentier B, Nottrott S, Fabrizio P, Bachi A, Wilm M, Rosbash M, Branlant C, Lührmann R. 2000. A common core RNP structure shared between the small nucleoar box C/D RNPs and the spliceosomal U4 snRNP. *Cell*. **103**:457–466.
Weber G, DeKoster GT, Holton N, Hall KB, Wahl MC. 2018. Molecular principles underlying dual RNA specificity in the *Drosophila* SNF protein. *Nat Commun*. **9**:2220.
Wilkinson ME, Charenton C, Nagai K. 2020. RNA splicing by the spliceosome. *Annu Rev Biochem*. **89**:359–388.
Will CL, Schneider C, Hossbach M, Urlaub H, Rauhut R, Elbashir S, Tuschl T, Lührmann R. 2004. The human 18S U11/U12 snRNP contains a set of novel proteins not found in the U2-dependent spliceosome. *RNA*. **10**:929–941.
Williams SG, Hall KB. 2010. Coevolution of *Drosophila snf* protein and its snRNA targets. *Biochemistry*. **49**:4571–4582.
Williams SG, Harms MJ, Hall KB. 2013. Resurrection of an urbilaterian U1A/U2B″/SNF protein. *J Mol Biol*. **425**:3846–3862.
Wong DK, Grisdale CJ, Slat VA, Rader SD, Fast NM. 2022. The evolution of pre-mRNA splicing and its machinery revealed by reduced extremophilic red algae. *J Eukaryot Microbiol*. **n/a**:e12927.
Wu F, Speth DR, Philosof A, Crémière A, Narayanan A, Barco RA, Connon SA, Amend JP, Antoshechkin IA, Orphan VJ. 2022. Unique mobile elements and scalable gene flow at the prokaryote–eukaryote boundary revealed by circularized asgard archaea genomes. *Nat Microbiol*. **7**:200–212.
Zaremba-Niedzwiedzka K, Caceres EF, Saw JH, Bäckström D, Juzokaite L, Vancaester E, Seitz KW, Anantharaman K, Starnawski P, Kjeldsen KU, *et al.* 2017. Asgard archaea illuminate the origin of eukaryotic cellular complexity. *Nature*. **541**:353–358.
Zhao C, Pyle AM. 2016. Crystal structures of a group II intron maturase reveal a missing link in spliceosome evolution. *Nat Struct Mol Biol*. **23**:558–565.
Zhao R, Biddle JF. 2021. Helarchaeota and co-occurring sulfate-reducing bacteria in subseafloor sediments from the Costa Rica margin. *ISME Commun*. **1**:25.
Zimmerly S, Hausner G, Wu X. 2001. Phylogenetic relationships among group II intron ORFs. *Nucleic Acids Res*. **29**:1238–1250.
Zimmerly S, Semper C. 2015. Evolution of group II introns. *Mob DNA*. **6**:7.

Downloaded from https://academic.oup.com/mbe/article/40/1/msad011/6985000 by Wageningen University and Research – Library user on 27 May 2023