The rise of complex life

Tracing the origins of the eukaryotic cell

Julian Vosseberg

Julian Vosseberg
The rise of complex life
Tracing the origins of the eukaryotic cell
PhD thesis, Utrecht University

Cover design: Davi Bos
Printing: Ridderprint, www.ridderprint.nl
ISBN: 978-90-393-7530-3

The rise of complex life
Tracing the origins of the eukaryotic cell

Het verschijnen van complex leven
De oorsprong traceren van de eukaryote cel

(met een samenvatting in het Nederlands)

Proefschrift

ter verkrijging van de graad van doctor aan de Universiteit Utrecht
op gezag van de rector magnificus, prof. dr. H.R.B.M. Kummeling,
ingevolge het besluit van het college voor promoties
in het openbaar te verdedigen op

dinsdag 24 januari 2023
des middags te 4.15 uur

door

Julian Vosseberg

geboren op 9 augustus 1993
te Harderwijk

Promotor:
Prof. dr. B. Snel

Dit proefschrift werd mogelijk gemaakt met financiële steun van de Nederlandse Organisatie voor Wetenschappelijk Onderzoek (NWO) onder projectnummer VICI 016.160.638.

Thesis assessment committee:

Prof. dr. ir. Thijs J. G. Ettema
Wageningen University & Research, the Netherlands

Prof. dr. Naomi M. Fast
University of British Columbia, Canada

Prof. dr. Eugene V. Koonin
National Center for Biotechnology Information, United States of America

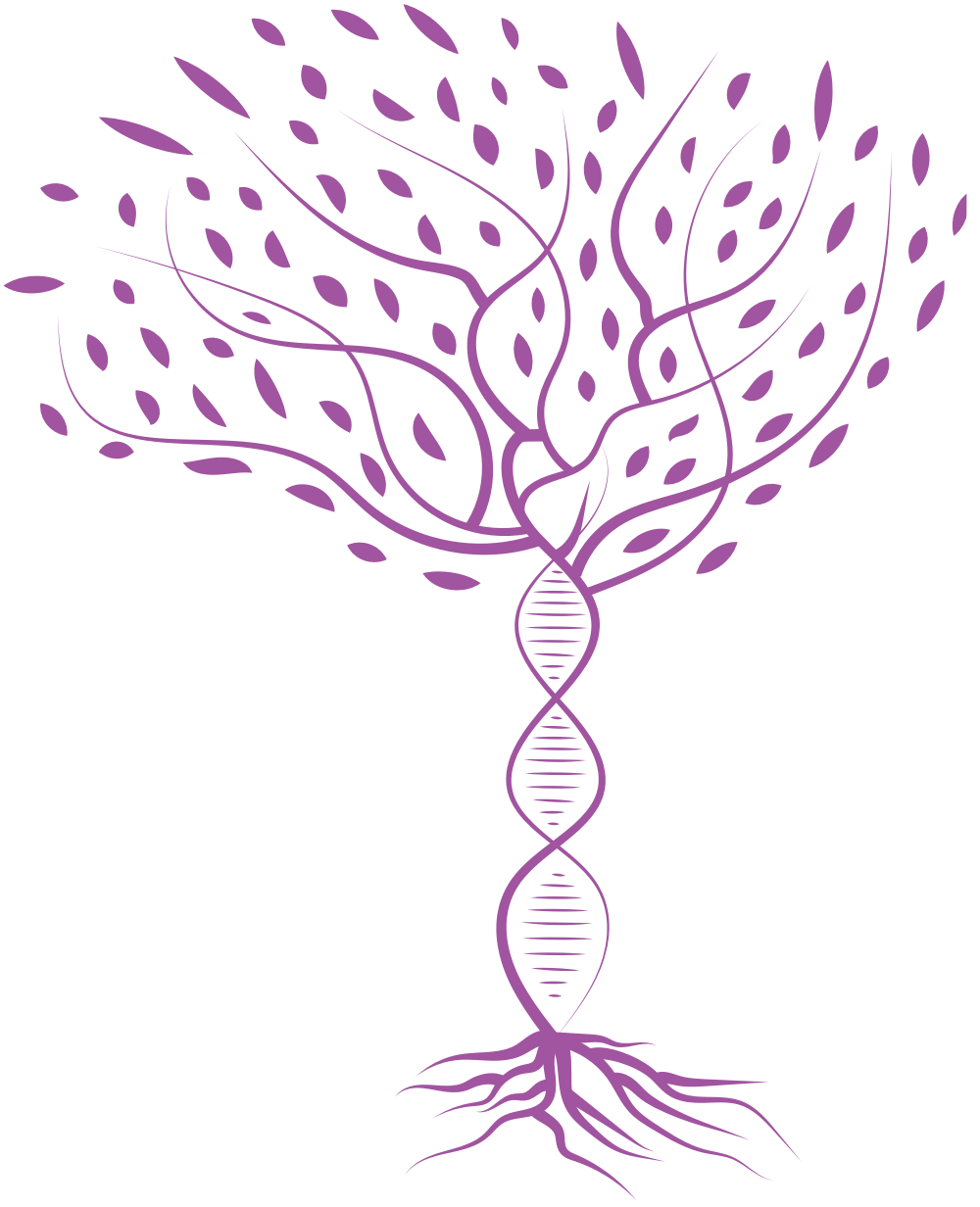
Prof. dr. Geert J. P. L. Kops
Hubrecht Institute, the Netherlands

Prof. dr. Anja Spang
Royal Netherlands Institute for Sea Research, the Netherlands

"There is grandeur in this view of life, with its several powers, having been originally breathed into a few forms or into one; and that, whilst this planet has gone cycling on according to the fixed law of gravity, from so simple a beginning endless forms most beautiful and most wonderful have been, and are being, evolved."

Charles Darwin (1859)

Table of contents



1

General introduction

From the deepest oceanic trench to the top of the highest mountain, life has conquered all corners of the Earth. The variety of organisms present in the numerous ecosystems around the world is simply breath-taking. Our own species, *Homo sapiens*, represents an extremely minor part of the total biodiversity yet has a profound impact on all ecosystems. The ecological role that humans play is worrisome considering the alarming rate at which species go extinct. On the outside, the human species does not seem special. The human power resides in our brains instead. We humans have the creativity, curious mindset and reasoning capacity to study the splendid diversity of life and interpret the commonalities and differences between all these different organisms. In that way biologists can trace the ancestors' tales of extant and extinct organisms. This book is about a specific part of this story of evolution, the origin of the first complex cells, which gave rise to all forms of life that we can see by the naked eye.

Once upon a time...

To appreciate this stunning story, it is necessary to grasp the essence of biological evolution. All forms of cellular life share the same language in which their information is stored. Strands of deoxyribonucleic acid (DNA) contain the instructions that are passed on to the next generation. The DNA alphabet contains just four letters: A, C, G and T. These letters represent the nucleobases adenine, cytosine, guanine and thymine, respectively. Pieces of DNA that encode for a product are called genes. The best-known type of products are proteins, the main workforce of the cell. To make a protein, a transcript of the gene is created in a slightly different kind of nucleic acid, ribonucleic acid (RNA). The messenger RNA (mRNA) that results from transcription is subsequently translated into amino acids by the ribosome. The entire collection of genetic information in a cell is called the genome.

Every time a cell divides, its DNA is copied and transferred to its daughter cells. Nearly all DNA is flawlessly passed on to the next generations, but some random changes occur each time. These changes are called mutations and the resulting genetic diversity is the crucial ingredient evolution is working on. Mutations can be small, for example a particular A becoming a G, or as large as the loss of a substantial part of the genome. Most mutations do not influence the functioning of the organism and can be considered neutral. Mutations that increase the reproduction rate are beneficial to the organism and increase in frequency over the course of generations. Organisms that fit better in the environment they live in (in other words, have adapted) are favoured by natural selection ("survival of the fittest") (Darwin, 1859). Consequently, the adaptive genetic changes become fixed in the population. Fixation of neutral or slightly detrimental mutations can also occur as a result of genetic drift, fluctuations because of chance events.

We usually see evolution as a slow process that takes so many generations that we can rarely see it in action. Although that might be true for species like ours with long generation times, the power of natural selection can be clearly observed for organisms that replicate fast and have large population sizes. For example, the rise of new SARS-CoV-2 variants with one variant outcompeting the other has been tracked throughout the pandemic of the virus causing COVID-19. On larger timescales of millions and billions

of years, life has gone through several transitions. An example of such a transition is the adaptation to living on land in the ancestors of amphibians, reptiles and mammals. John Maynard Smith and Eörs Szathmáry introduced the concept of the major transitions in evolution (Maynard Smith and Szathmáry, 1995). These major transitions relate to a shift in the level of complexity. Examples of major transitions are the origin of life itself and the origin of multicellularity. These are usually easy to appreciate for a lay audience. The major transition that I discuss in this thesis is less visible as it is a transition in the fundaments of cellular biology. The large extent of this transition can be seen in the differences between two types of cells, one is relatively simple, whereas the other is tremendously complex. To introduce this major transition, we will go back in time.

The origin of eukaryotes

Our ancestors' tale started around four billion years ago, with the origin of life (Dacks et al., 2016). We do not know exactly how many times life originated but we do know that only a single cellular lineage survived. All organisms living today are related and can be traced back to the last universal common ancestor. For approximately the first two billion years ago, bacteria and archaea were the only forms of cellular life, at least as far as we know. Archaea look similar to bacteria on first glance, as both are composed of relatively simple cells, but on the molecular level they are fundamentally different. Bacteria and archaea are recognised as separate domains, the highest taxonomic rank. Both groups of micro-organisms are called prokaryotes.

Roughly two billion years ago, according to the current consensus (Betts et al., 2018; Dacks et al., 2016), a new type of cell emerged. Its defining feature is the nucleus, a compartment that contains the genetic material. This type of cell is called eukaryotic; its emergence is called eukaryogenesis. The differences between the eukaryotes and the prokaryotes go much further than the nucleus. Eukaryotic cells are larger and contain multiple compartments besides the nucleus. These compartments (organelles) perform dedicated roles in the eukaryotic cell. The mitochondrion, for example, functions as the powerhouse of the cell and the endoplasmic reticulum functions in protein folding and packing proteins in vesicles for transport to other parts of the cell. The greater cellular complexity of eukaryotes is accompanied with a much larger genome. Whereas a prokaryotic genome encodes on average 3,201 protein-coding genes, the number of genes in an average eukaryotic genome is 17,160 (**Figure 1.1**). Part of these additional genes enable the compartments in eukaryotic cells.

Evolutionary relationships between organisms are represented as a tree, the "tree of life". The position of eukaryotes in the tree of life remained largely elusive until a decade ago. Two main hypotheses had been postulated. In the first, eukaryotes were the sister group of the Archaea (the three-domain hypothesis); in the second, the eukaryotes originated from within the Archaea (the two-domain hypothesis). Resolving such ancient relationships is notoriously difficult due to the limited phylogenetic signal in sequences to resolve deep splits in the tree of life. This makes these analyses prone to systematic biases that favour incorrect topologies (Williams et al., 2020). A well-known artefact is the grouping together of divergent taxa on long branches (long-branch attraction).

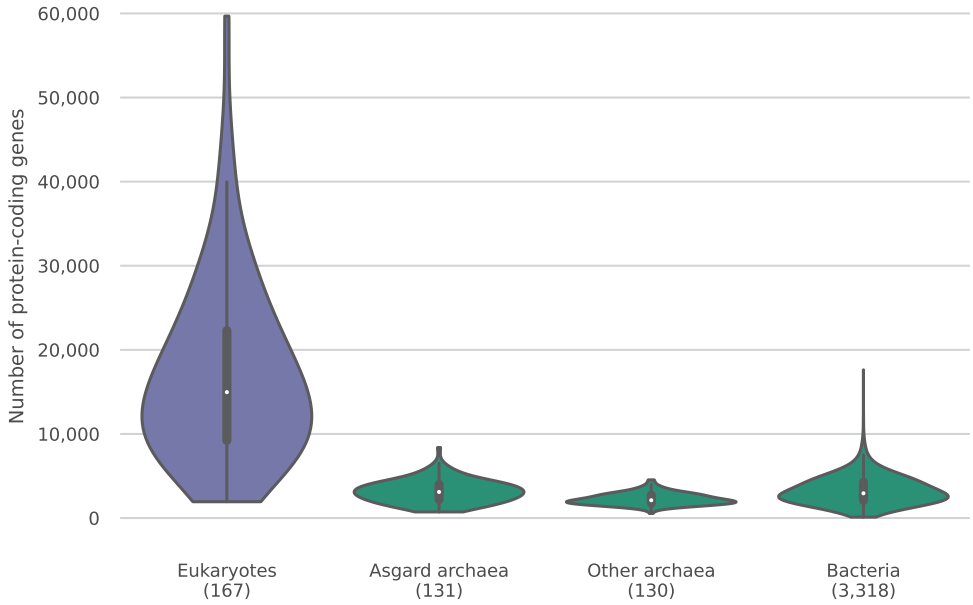


Figure 1.1 | Difference in the number of genes between eukaryotes and prokaryotes. The distributions of the number of protein-coding genes are shown as a violin plot with box plots inside. The widths of the violins are the same for all groups. Virtually all Asgard archaeal genomes are metagenome-assembled genomes and are hence not fully complete and might experience strain heterogeneity or contamination. The eukaryotic genomes are from our in-house dataset (Deutekom et al., 2021), the Asgard archaea are from the expanded set described in chapter 5 and the other prokaryotes are from eggNOG 4.5 (Huerta-Cepas et al., 2016a).

Initial trees were in favour of the three-domain hypothesis (Woese, 1987; Woese et al., 1990). However, with the use of more sophisticated models of evolution and the increased sampling of prokaryotic diversity, the support for the two-domains tree of life increased (Guy and Ettema, 2011; Williams et al., 2012). The discovery of a new group of archaea that both confidently grouped together with eukaryotes and encoded multiple homologs of genes that had previously been considered eukaryote-specific, made the case for the eukaryotes-within-archaea hypothesis compelling (Spang et al., 2015; Zaremba-Niedzwiedzka et al., 2017). The first described members of this group were found close to a hydrothermal vent named Loki's castle and were hence named Lokiarchaeota (Spang et al., 2015). Other Norse deities were used to name related archaeal phyla, giving the name Asgard archaea to the group of archaea that is closely related to eukaryotes (Zaremba-Niedzwiedzka et al., 2017).

The exact phylogenetic position of eukaryotes within the Asgard archaea is not completely resolved, although there is some support for a specific clade, Heimdallarchaeota LC3, which has recently been renamed to Hodarchaeota (Liu et al., 2021), being the sister group of eukaryotes (Narrowe et al., 2018; Williams et al., 2020; Zaremba-Niedzwiedz-

ka et al., 2017). Until recently, only partial genomes of Asgard archaea were available, based on metagenome assemblies. Using additional long-read sequencing data, closed *Heimdallarchaeum* and *Odinarchaeum* genomes have been obtained (Tamarit et al., 2022; Wu et al., 2022) and the first successful cultivation of an Asgard archaeon, more specifically a lokiarchaeote, resulted in a complete *Prometheoarchaeum* genome (Imachi et al., 2020). This cultivation effort also provided the first microscopy pictures of these archaea (Imachi et al., 2020). Further visualisations (Avcı et al., 2022), biochemical studies (Akıl and Robinson, 2018; Akıl et al., 2020, 2022; Hatano et al., 2022; Neveu et al., 2020; Survery et al., 2021) and bioinformatical analyses (James et al., 2017; Klinger et al., 2016; Liu et al., 2021) have sharpened the views on the cell biology of the common ancestor of Asgard archaea and eukaryotes. The presence of several building blocks of complex eukaryotic cells and their inferred functions in Asgard archaea suggest that some sort of endomembrane system and dynamic cytoskeleton was already present in the archaeal ancestor of eukaryotes.

Eukaryotes are not just highly derived archaea. In the transition from a *bona fide* archaeon to a eukaryote, a bacterium was taken up that evolved into the mitochondria mentioned above. The key role of endosymbiosis (i.e., a cell living within another cell) in the origin of eukaryotes was proposed in a seminal paper by Lynn Margulis (Sagan, 1967). Multiple lines of evidence have been compiled for the bacterial origin of mitochondria since (Roger et al., 2017), such as the presence of a separate bacteria-like genome in mitochondria and separate transcription and translation machineries. The specific group of bacteria that are most closely related to mitochondria was identified earlier than the host-related archaeal clade (Yang et al., 1985). Recent work has shown that the mitochondria likely form a sister group to all alphaproteobacteria (Martijn et al., 2018; Muñoz-Gómez et al., 2022).

Mitochondrial endosymbiosis was a key event in eukaryogenesis. Different hypotheses have been proposed about the nature of the initial relationship between the host and endosymbiont. Most infer an exchange of metabolites between them, such as the hydrogen (Martin and Müller, 1998; Sousa et al., 2016), syntrophy (López-García and Moreira, 2020; Moreira and López-García, 1998) and the reverse flow hypothesis (Spang et al., 2019). In those cases, the symbiosis was mutualistic from the start. An example of non-mutualistic scenario is the farming scenario, in which the protomitochondrion was captured as prey by a phagocytosing host and was not completely digested but farmed instead (Zachar et al., 2018). In this scenario it is only the host that benefits. The protomitochondrion has also been viewed as an energy parasite by some (Wang and Wu, 2014), a scenario in which only the endosymbiont benefits.

Regardless of the initial relationship between the host and the endosymbiont, the endosymbiont was fully integrated into the host cell and lost many genes (Poole and Gribaldo, 2014; Roger et al., 2017). The archaeal type of cells had evolved into what we would recognise as eukaryotic cells. The subsequent eukaryotic radiation resulted in the diverse eukaryotic groups that are present today.

Eukaryotic evolution

Plants, animals and fungi are the well-known examples of eukaryotic life. However, it should be emphasised that these groups are merely twigs of the eukaryotic tree of life. Most eukaryotic diversity can be found among eukaryotes that are neither of these three groups and that are collectively referred to as protists. These protists are mostly unicellular and include ciliates, slime moulds and all kinds of algae. Eukaryotes play important and diverse ecological roles on all trophic levels, as partners in symbioses and in different habitats (Burki et al., 2021; Jamy et al., 2022). The morphological diversity of unicellular eukaryotes alone is enormous with differences in cell size, shape and organisation; their feeding behaviours, life cycles and motility consequently differ greatly (Adl et al., 2019).

In initial inferences of the tree of life based on molecular data simple eukaryotes with parasitic lifestyles such as microsporidia were close to the root (Woese, 1987). This branching pattern was consistent with a gradual increase in complexity during eukaryotic evolution. Later trees revealed that these parasitic lineages do not represent a primitive state but that they represent a reduced state from a more complex ancestor, in agreement with the emerging recognition of features of these simpler organisms being derived instead of primitive (Embley and Hirt, 1998). The microsporidia, for example, are now recognised as fungal relatives (Galindo et al., 2021). The relationship between eukaryotes supports a scenario with rare bursts of increased complexity, followed by a gradual decrease in complexity throughout eukaryotic evolution. This pattern of a rapid increase in complexity followed by a longer period of streamlining corresponds with a biphasic model of the evolution of complexity (Cuypers and Hogeweg, 2012; Wolf and Koonin, 2013).

The eukaryotic tree has not fully been resolved yet and the exact phylogenetic position of the root is uncertain (Al Jewari and Baldauf, 2022; Burki et al., 2020). The rapidly increasing amount of sequencing data from genomically uncharacterised taxa illuminates new parts of the eukaryotic tree. Recent studies have uncovered novel deep branches (Brown et al., 2018; Lax et al., 2018) and the accepted position of these and other deep branches has changed following those studies or remains under debate (Adl et al., 2019; Burki et al., 2020). The current consensus (Adl et al., 2019) about groupings of eukaryotes is shown in **Figure 1.2**. Notwithstanding the uncertainties about the position of the root, its impact on the inferred nature of the last eukaryotic common ancestor (LECA) is subtle. The complex cellular characteristics of present-day eukaryotes throughout the tree of life provide convincing support for a complex LECA (Koumandou et al., 2013). The complex LECA illustrates the fundamental gap between prokaryotes and eukaryotes that was bridged during eukaryogenesis (Dacks et al., 2016).

Biology of the eukaryotic cell

A list of characteristics of a eukaryotic cell compared with a prokaryotic cell can easily fill multiple pages. For clarity, I will limit the list to the main ones. Note that these are not present in all present-day eukaryotes but they were very likely present in LECA and hence originated during eukaryogenesis (Koumandou et al., 2013). Moreover, some prokaryotes have evolved eukaryote-like features in parallel (see chapter 6). Eukaryotic cells have an elaborate endomembrane system with a nucleus, endoplasmic reticulum and Golgi

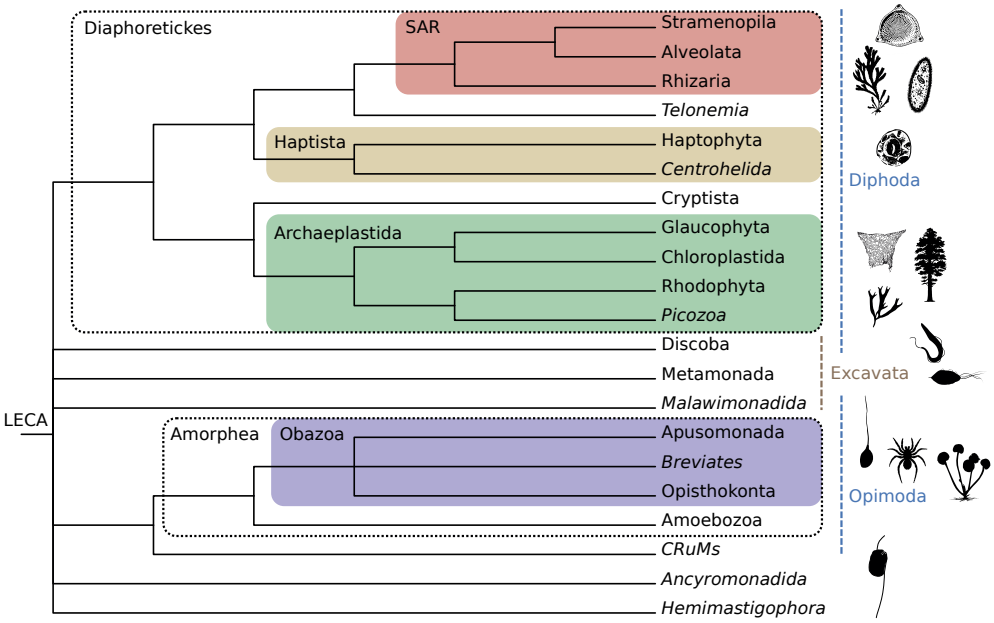


Figure 1.2 | The eukaryotic tree of life. The groups are based on a recent review (Burki et al., 2020); the branching order within Diaphoretickes is based on recent studies (Schön et al., 2021; Strassert et al., 2021). Taxa that are not represented in our in-house dataset are italicised. One of the proposed positions of the root is between Opimoda and Diphoda (Derelle et al., 2015). The monophyly of Excavata is debated (Burki et al., 2020). Representatives of Stramenopila, Alveolata, Haptophyta, Chloroplastida, Rhodophyta, Discoba, Metamonada, Breviates, Opisthokonta and Ancyromonadida are depicted as silhouette images, obtained from PhyloPic (https://beta.phylopic.org).

apparatus. A dynamic cytoskeleton with actin and tubulin filaments and motor proteins in the cytosol and cilia enables cell shape flexibility, motility and intracellular trafficking of vesicles. An intricate signalling system of protein kinases, small GTPases and cyclins relay information through the large cell. Two different types of cell divisions can occur: mitosis and meiosis. The latter is a reductional division and results in haploid gametes. The fusion of two gametes is the base of sexual reproduction. The linear chromosomes in eukaryotes require a protein complex, the kinetochore, for the separation of the doubled genetic material during cell division (van Hooff et al., 2017). The kinetochore is an example of a new protein complex that originated during eukaryogenesis.

Eukaryotic gene regulation occurs on multiple levels. A wide repertoire of transcription factors and histone modifications regulate transcription. These histone modifications are scarce in archaea (Grau-Bové et al., 2022). The eukaryote-specific mRNA processing steps capping, splicing and polyadenylation take place in the nucleus. Incorrectly spliced mRNA molecules are degraded by non-sense mediated RNA decay. Small RNA molecules can inhibit gene expression by means of RNA interference. Regulation after translation is facilitated by the stupendous variety of post-translational modifications.

Eukaryotic genes are characterised by their exon-intron structure. These introns are recognised in the pre-mRNA by the spliceosome. During splicing the introns are excised and the exons are ligated to form the mature mRNA. The spliceosome is a large complex, made up of small nuclear RNAs (snRNAs) and associated proteins. The snRNAs recognise the exon-intron boundaries and catalyse the splicing reaction (Wilkinson et al., 2020). The spliceosome is a dynamic complex and its composition changes during the splicing cycle (Wilkinson et al., 2020).

The highly orchestrated splicing cycle involves the modular recruitment and release of different subcomplexes. The first steps in the assembly of the spliceosome include the recognition of the 5′ splice site by U1 snRNA, together with its associated proteins called the U1 small nuclear ribonucleoprotein (snRNP), and the binding of U2AF and SF1 to the sequence around the branch point nucleotide (**Figure 1.3**). The latter two subunits are subsequently replaced by the U2 snRNP and the U4/U6.U5 tri-snRNP is recruited to the intron. The catalytic centre is formed by rearrangements that include the transfer of the 5′ splice site from U1 snRNA to U6 snRNA, the release of U1 snRNP and U4 snRNA together with the U4/U6 di-snRNP proteins and the binding of the NTC-NTR and RES complexes. The splicing reaction consists of two steps. In the first reaction a covalent bond between the 5′ splice site and the branch point nucleotide is created; in the second reaction, the two exon ends are ligated. The exon-junction complex (EJC) is deposited on the connected exons and the intron is released. The spliceosome disassembles and the subunits are recycled for another splicing cycle.

The spliceosome is one of the most complex molecular machineries that originated during eukaryogenesis (see chapters 4 and 5). Prokaryotes do not have introns that interrupt protein-coding genes and therefore do not require such a machinery. The emergence

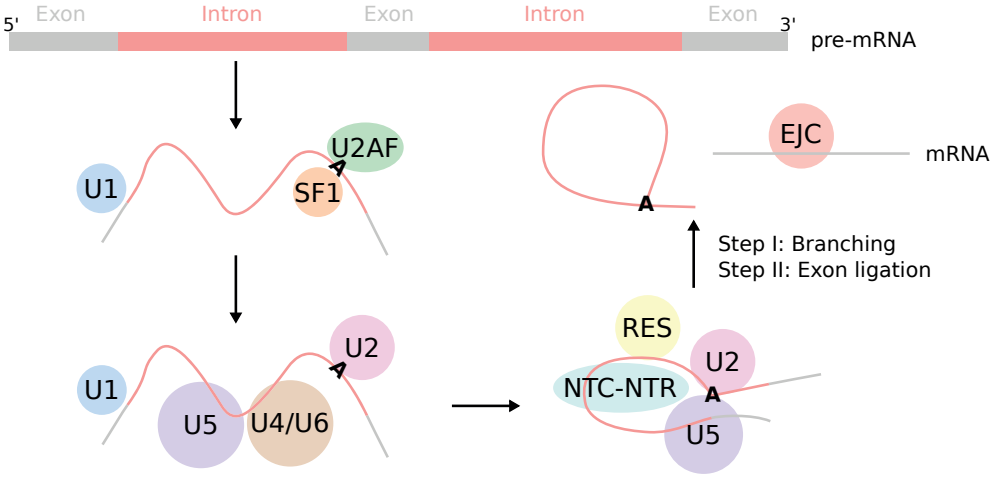


Figure 1.3 | Steps of the splicing cycle to remove the introns from the pre-mRNA. The branch point nucleotide, an adenosine, is indicated in bold. The small nuclear ribonucleoproteins (U1-U6) and several additional complexes are shown. U2AF: U2 auxiliary factor; NTC: Prp19-associated complex; NTR: Prp19-related complex; RES: retention and splicing complex; EJC: exon-junction complex.

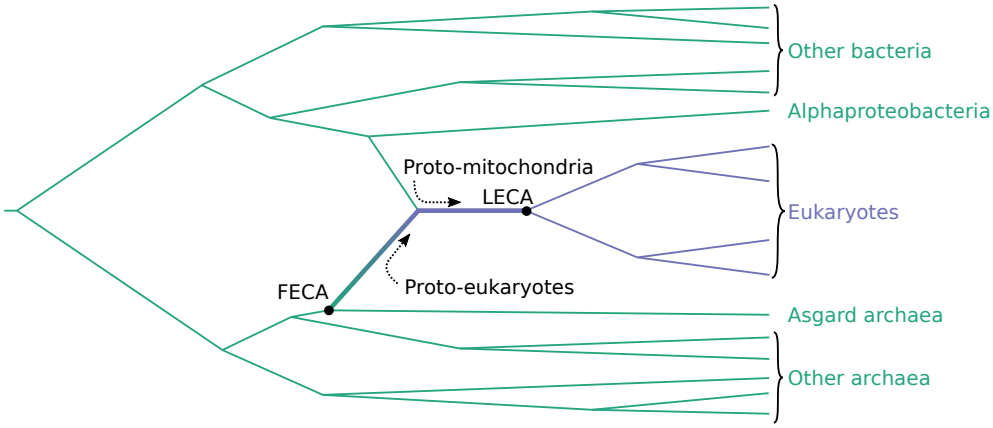


Figure 1.4 | Eukaryogenesis. Because it is not clear when the first *bona fide* eukaryote originated exactly, we define the entire period between the first (FECA) and last eukaryotic common ancestor (LECA) as eukaryogenesis (thick branch). The organisms comprising this stem lineage are called proto-eukaryotes. The alphaproteobacteria-related endosymbionts in proto-eukaryotes are referred to as proto-mitochondria. It should be noted that FECA and LECA are ancestral reconstructions that do not reflect a particular organism. Lineages that have gone extinct are branching off from all branches (not shown).

of intragenic introns and the large spliceosomal complex starkly illustrate the increase in complexity during eukaryogenesis.

Order of events during eukaryogenesis

As described above, many cellular changes occurred during the transition from relatively simple prokaryotes to complex eukaryotes. As far as we currently know, all intermediate lineages have gone extinct. The fossil record of proto-eukaryotes, lineages between the divergence from the closest archaeal group and LECA (**Figure 1.4**), is limited and its interpretation is debated (Porter, 2020). This, combined with the observation that it happened only once, makes eukaryogenesis an evolutionary conundrum. Questions about the specific environment and selective pressures that played a role are of utmost interest to make informed estimates about the probability that such a huge leap in cellular complexity could happen again. On the cellular level, the enormous list of features and the lack of intermediate stages makes it difficult to disentangle the emergence of the different parts of the eukaryotic cell and to reconstruct an order of events.

Numerous scenarios about the order of events during eukaryogenesis have been proposed, based on different lines of reasoning. In particular, the timing of endosymbiosis relative to other events is heavily debated (Poole and Gribaldo, 2014). In mitochondria-late scenarios the host was a fully complex eukaryotic cell that merely missed a mitochondrion. The uptake of the protomitochondrion could have been "a crucial late step in eukaryogenesis, which brought about the definitive selective advantage that facilitated the dominance and radiation of the eukaryotic groups that have survived to the present day" (Pittis and Gabaldón, 2016). In these scenarios it is usually proposed that the cellular complexity, especially phagocytosis, was needed to ingest the protomitochon-

drion. It should be noted that a very late endosymbiosis is not plausible because the protomitochondria were fully integrated into the host cell as organelles before LECA (Poole and Gribaldo, 2014; Roger et al., 2017). Opposing mitochondria-early scenarios typically postulate that the endosymbiosis was the key event that started eukaryogenesis. The host was a simple archaeon and the mitochondrion was absolutely necessary for the rise in complexity (Lane and Martin, 2010). Besides these more extreme cases, the host had already evolved some but not all complex features before endosymbiosis, such as an endomembrane system or dynamic cytoskeleton, in mitochondria-intermediate scenarios.

Studying molecular evolution

Underlying these changes in cell biology are changes in the genome. Mutations can result in genetic innovations and the gain of a novel function. Traces of ancient genetic innovations during eukaryogenesis can be detected in the genomes of present-day organisms. In this thesis we search for these traces to illuminate the evolution of early eukaryotes.

Different types of genetic innovations can be distinguished. The first type of genetic innovation is the birth of a completely novel gene from a non-coding part of the genome, a gene invention. Small mutations in an already existing gene can also result in new functions. The complete fusion of two genes or only the addition of a domain from another gene is the third example of genetic innovations. An organism can also acquire a gene from another organism than its parent via horizontal gene transfer. A powerful source of genetic innovations is gene duplications (Ohno, 1970). The duplication of a gene results in two identical genes. Because keeping both is rarely beneficial (except when increased dosage is adaptive), there are three main outcomes: one of them is lost, one of them gains a new function through additional mutations (neofunctionalization) or both acquire additional mutations that result in a division of tasks between the duplicates (subfunctionalization) (Force et al., 1999). In the latter case, the complementary loss of functions that the ancestral gene performed means that neither of the duplicates on their own represent the ancestral preduplication state. Finally, although usually ignored in this context, the loss of a gene can also be seen as an innovation.

Genes that share a common ancestor are called homologs (Koonin, 2005). If the separation after their last common ancestor was due to a speciation event, the homologs are called orthologs (Fitch, 1970). If the separation was due to a gene duplication, the homologs are called paralogs (Fitch, 1970). When studying eukaryogenesis it is useful to group orthologs together that can be traced back to a single gene (an ortholog) in LECA. All the descendants of this gene comprise an orthogroup (OG). Note that duplications in a eukaryotic lineage after LECA do not break up this group (**Figure 1.5**). These paralogs inside an OG are called inparalogs (Sonnhammer and Koonin, 2002). Orthologs are expected to perform very similar functions (Gabaldón and Koonin, 2013). For example, the orthologous actin proteins form filaments in plant, yeast and animal cells. It is very likely that the orthologous actin protein in a newly discovered eukaryote performs the same function.

The relationship between homologous genes is visualised using phylogenetic trees. A typical phylogenetics workflow consists of (1) collecting homologous sequences, (2) in-

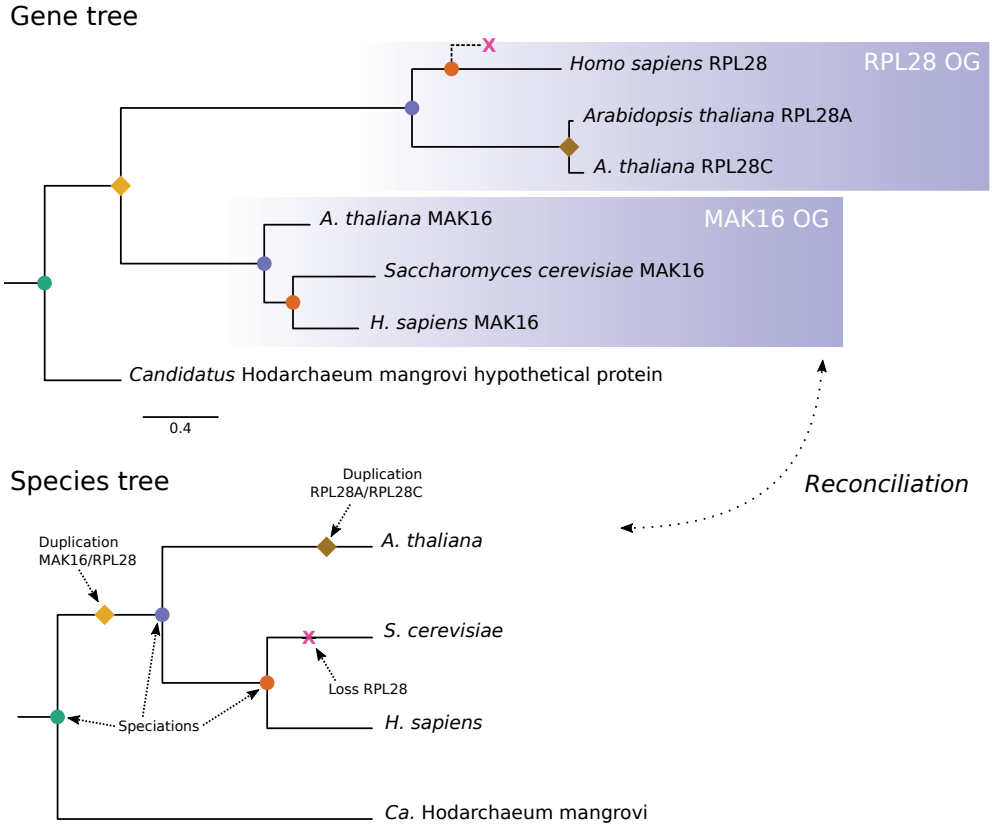
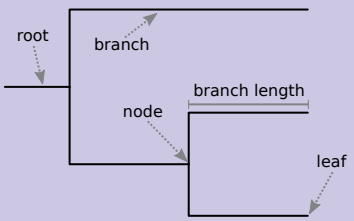


Figure 1.5 | Reconciled gene tree showing the relationships between proteins of the RPL28/MAK16 family. The species tree is used to annotate nodes in the gene tree as speciation (circles) or duplication events (diamonds). A gene loss event is also shown (X). The events are also indicated on the species tree. The gene tree was rooted on the archaeal lineage. Rooting is also possible on the MAK16 or RPL28 branch. Without considering horizontal gene transfers, this would imply that the MAK16/RPL28 duplication occurred before the speciation event that separated the Hodarchaea and proto-eukaryotes. One of the two copies would have been lost in the hodarchaeal lineage, which makes this scenario less parsimonious as it involves an additional event.

ferring a multiple sequence alignment and (3) inferring a phylogenetic tree. One can use nucleotide sequences or amino acid sequences for phylogenetic inference. When studying deep evolutionary events such as eukaryogenesis, amino acid sequences contain more phylogenetic information as they change more slowly than the encoding genetic sequences. Performing phylogenetic analyses on a genome-wide scale is called phylogenomics.

First, homologs need to be detected. Sequences are considered homologous when they are similar to a degree that is no longer expected from unrelated sequences, hence indicating shared ancestry. An often-employed approach is to search with a single sequence (the query) in a sequence database. The basic local alignment score tool (BLAST) (Altschul et al., 1990) is the standard tool used for this. If you already have a set of ho-

Box 1.1



A phylogenetic tree consists of nodes connected by branches. Trees are typically bifurcating, which means that two daughter branches originate from a node. In most cases, the output of phylogenetic programs is an unrooted tree. One of the branches has to be selected to root the tree. The newly created root node provides a direction of time. The terminal nodes are called leaves. A node and all its descendants form a monophyletic group, which is called a clade. Branch support values, usually obtained from bootstrap replicates, reflect the support for the descendants forming a clade.

The lengths of the branches in a tree represent the estimated number of substitutions, normalised for the alignment length. Scale bars indicate this estimated number of substitutions per site. It is the product of the time that has passed between the nodes and the rate of evolution. In **Figure 1.5**, for example, 0.35 substitutions per site likely occurred between the human-yeast MAK16 ancestor and human MAK16. For the used alignment, that would be 62 substitutions in total. Leaves in this thesis correspond with the same time (present) and the same speciation events, such as LECA, can be approximated as the same time, depending on the population structure and diversity of the ancestral population (O'Malley et al., 2019). The shorter branch for *A. thaliana* MAK16 compared with the human and yeast ones cannot reflect differences in time because the time that has passed is the same. Therefore, it is caused by a relatively slower evolutionary rate for the *A. thaliana* MAK16 branch. With the chosen root, substantially more changes occurred in the RPL28 branch between the yellow duplication node and the purple speciation node compared with the corresponding MAK16 branch.

mologs, you can make a profile of these sequences to perform more sensitive homology searches. The best-known kind of profile is the hidden Markov model (HMM). The higher sensitivity is useful in case of high rates of sequence divergence, which may complicate the recognition of homologs. Comparisons between profiles are even more sensitive (Söding, 2005). When sequence similarity has been reduced to random levels, homology could still be inferred based on structural similarity. Large-scale protein structure predictions that have been made possible with the recently developed AlphaFold tool (Jumper et al., 2021) are highly promising to uncover deep homologies.

The next step after you have collected homologous sequences is aligning these sequences. This step detects nucleotides or amino acids that are probably homologous between the different sequences. The resulting multiple sequence alignment is a hypothesis on the homology of characters. Gaps that are introduced in the alignment reflect insertions and deletions. In most cases the alignment is trimmed by removing poorly aligned regions before a tree is inferred. Also positions that are evolving fast or are compositionally biased can be removed because these are prone to phylogenetic artefacts. Removing positions, however, also reduces the phylogenetic signal and should be applied with caution.

Modern phylogenetic tools use maximum likelihood or Bayesian inference because of their well-defined statistical framework. For these methods, a model of sequence evolution that describes the substitution rates between different amino acids and the equilibrium frequencies of the amino acids has to be specified. In the LG model, for example, the equilibrium frequency of serine is 6.1% and its substitution rate with threonine is 6.47, whereas it is 0.064 with isoleucine (Le and Gascuel, 2008). Different rates across sites are often implemented using a gamma distribution. Even more sophisticated models with for example mixture models with different equilibrium frequencies across sites can also be chosen (Le et al., 2008). Tools exist to detect the best model for your multiple sequence alignment (Kalyaanamoorthy et al., 2017).

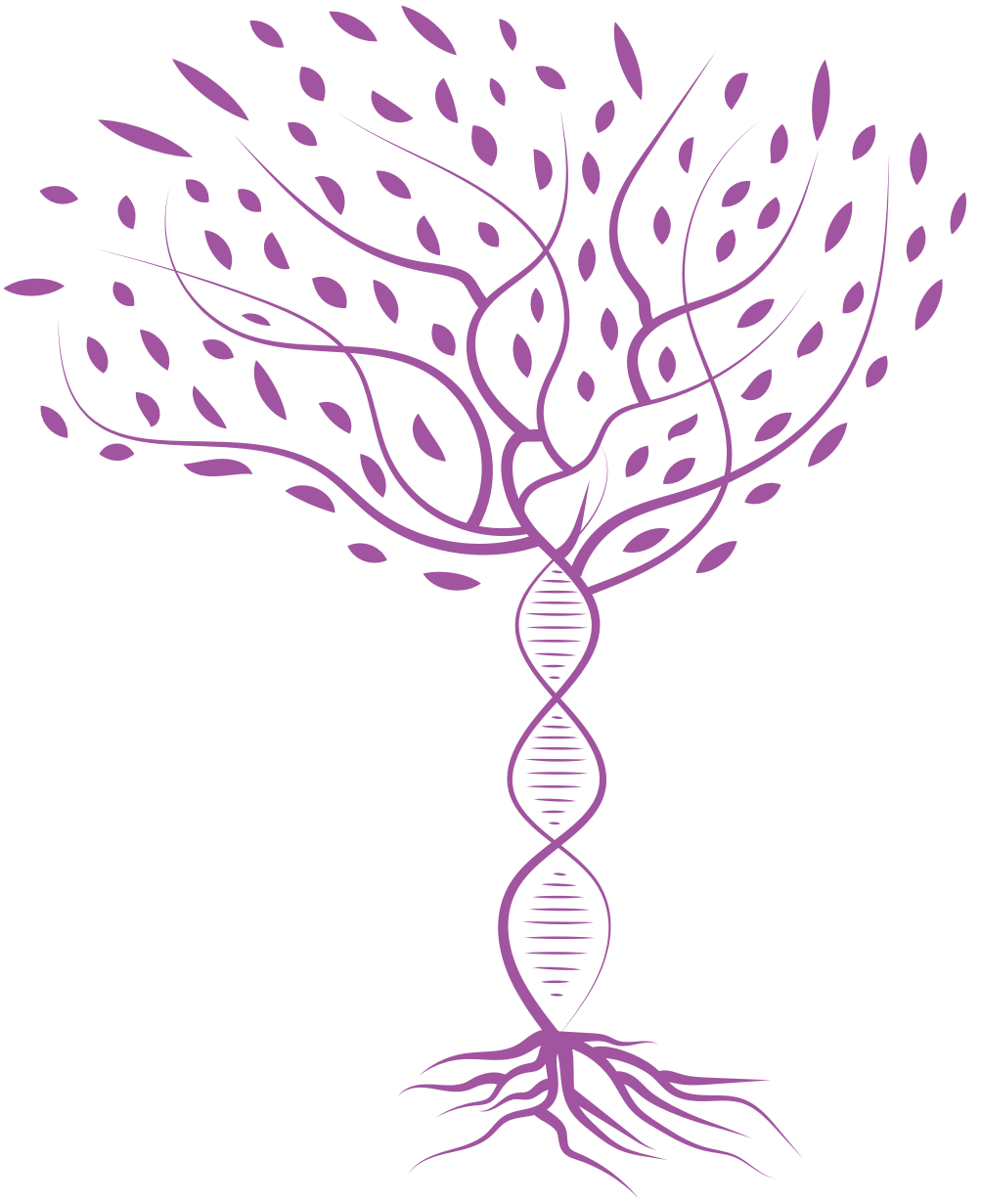
Phylogenetic programs test different tree topologies and optimise model parameter values, including branch lengths (**Box 1.1**). Maximum-likelihood-based programs return the tree and model parameters with the highest likelihood. The likelihood represents the probability that the given topology, branch lengths and model of evolution generated the alignment. In other words, from all hypotheses considered, the maximum-likelihood tree matches the recorded evolutionary outcome best. By comparing the gene tree with the known species tree, gene duplication, loss and transfer events can be inferred (**Figure 1.5**). The inference of these events from a gene tree is called tree reconciliation.

The large amount of sequence data that is currently available provides more evolutionary signal to reveal the relationships between sequences. However, phylogenetic inference on large datasets is computationally demanding because of the high number of possible tree topologies and parameters that need to be estimated. Selecting a subset of sequences, while keeping the diversity of lineages high, improves the computational feasibility. The resolution of the resulting trees can even be improved by limiting the analysis to slowly evolving sequences for deep phylogenetic inference (Elias et al., 2012; van Wijk and Snel, 2020).

This thesis

The long list of features that have arisen during eukaryogenesis raises the question how the evolutionary transition from archaeon to eukaryote took place. Many scenarios have been proposed on the order of events but empirical data supporting a specific scenario are scarce. The research described in this thesis contributes to these data to unravel the rise of complex life. I have reconstructed proto-eukaryotic genetic innovations to illuminate the order of events that resulted in the first complex eukaryotic cells. To perform these reconstructions, I used both large-scale phylogenomic and small-scale phylogenetic methods on a diverse set of eukaryotes and prokaryotes. In chapter 2, I focus on the gene duplications that occurred during the transition. We reconstructed and characterised these proto-eukaryotic duplications. By analysing the branch lengths in phylogenetic trees, we obtained relative time estimates for the duplication events, mitochondrial endosymbiosis and horizontal gene transfers from other prokaryotes. The focus of chapter 3 is on another proto-eukaryotic genetic innovation: the emergence of intragenic introns. The duplication data from chapter 2 is combined with the inferred presence of shared intron positions between paralogs to trace the spread of introns through the

proto-eukaryotic genome, in relation to the duplication of genes. Introns are removed by the spliceosome, one of the most complex molecular machines that originated during eukaryogenesis. Chapter 4 is a literature review on the origin of the spliceosome during eukaryogenesis and the subsequent evolution of this machinery after LECA. The origins of the spliceosomal proteins in LECA is the topic of chapters 5. We specifically related the origin of the complex spliceosome to the spread of the introns it functions on. The evolutionary histories that we inferred using phylogenetic analyses and intron analyses provide insight into the emergence of complex machines during eukaryogenesis. In the final chapter I discuss the implications of our findings on eukaryogenesis and the origin of eukaryotic complexity.



2

Timing the origin of eukaryotic cellular complexity with ancient duplications

Julian Vosseberg*, Jolien J. E. van Hooff*, Marina Marcet-Houben, Anne van Vlimmeren, Leny M. van Wijk, Toni Gabaldón, Berend Snel

*These authors contributed equally to this work

Nature Ecology and Evolution, 2021

Abstract

Eukaryogenesis is one of the most enigmatic evolutionary transitions, during which simple prokaryotic cells gave rise to complex eukaryotic cells. While evolutionary intermediates are lacking, gene duplications provide information on the order of events by which eukaryotes originated. Here we use a phylogenomics approach to reconstruct successive steps during eukaryogenesis. We found that gene duplications roughly doubled the proto-eukaryotic gene repertoire, with families inherited from the Asgard archaea-related host being duplicated most. By relatively timing events using phylogenetic distances, we inferred that duplications in cytoskeletal and membrane-trafficking families were among the earliest events, whereas most other families expanded predominantly after mitochondrial endosymbiosis. Altogether, we infer that the host that engulfed the proto-mitochondrion had some eukaryote-like complexity, which drastically increased upon mitochondrial acquisition. This scenario bridges the signs of complexity observed in Asgard archaeal genomes to the proposed role of mitochondria in triggering eukaryogenesis.

Introduction

Compared with prokaryotes, eukaryotic cells are tremendously complex. Eukaryotic cells are larger, contain more genetic material, have multiple membrane-bound compartments and operate a dynamic cytoskeleton. Although certain prokaryotes have some eukaryote-like complexity, such as a large size, internal membranes and even phagocytosis-like cell engulfment (Dacks et al., 2016; Shiratori et al., 2019), a fundamental gap remains. The last eukaryotic common ancestor (LECA) already had the intracellular organisation and gene repertoire characteristic of present-day eukaryotes (Koumandou et al., 2013), making the transition from prokaryotes to eukaryotes (eukaryogenesis) one of the main unresolved puzzles in evolutionary biology (Dacks et al., 2016; Szathmáry, 2015).

Most eukaryogenesis scenarios posit that a host, related to the recently discovered Asgard archaea (Spang et al., 2015; Zaremba-Niedzwiedzka et al., 2017), took up an Alphaproteobacteria-related endosymbiont (Martijn et al., 2018; Roger et al., 2017) that gave rise to the mitochondrion. However, the timing and impact of this endosymbiosis event in the evolution of eukaryotic complexity are hotly debated and at the heart of different scenarios on eukaryogenesis (Poole and Gribaldo, 2014).

Besides the acquisition of genes via the endosymbiont, the proto-eukaryotic genome expanded through gene inventions, duplications and horizontal gene transfers during eukaryogenesis (Makarova et al., 2005; Pittis and Gabaldón, 2016a). Previous work has suggested that gene duplications nearly doubled the ancestral proto-eukaryotic genome (Makarova et al., 2005). Gene families such as small GTPases, kinesins and vesicle coat proteins greatly expanded, which enabled proto-eukaryotes to employ an elaborate intracellular signalling network, a vesicular trafficking system and a dynamic cytoskeleton (Dacks and Field, 2018; Elias et al., 2012; Jékely, 2003; Wickstead et al., 2010).

Uncovering the order in which these and other eukaryotic features emerged is complicated due to the absence of intermediate life forms. However, duplications occurred during the transition and are likely to yield valuable insights into the intermediate steps

of eukaryogenesis. In this study we attempt to reconstruct the successive stages of eukaryogenesis by systematically analysing large sets of phylogenetic trees inferred from prokaryotic and eukaryotic sequences. We determined the scale of gene inventions and duplications during eukaryogenesis and how different functions and phylogenetic origins had contributed to these eukaryotic innovations. Furthermore, we timed the prokaryotic donations and duplications relative to each other using information from phylogenetic branch lengths.

Results

Unprecedented resolution of duplications during eukaryogenesis

To obtain a comprehensive picture of duplications during eukaryogenesis we made use of the Pfam database (Finn et al., 2016) (see Methods). We took a phylogenomics approach inspired by the ScrollSaw method (Elias et al., 2012), which limits phylogenetic analyses to slowly evolving sequences and collapses duplications after LECA, thereby increasing the resolution of deep tree nodes. We constructed phylogenetic trees and detected 10,233 nodes in these trees that represent a single Pfam domain in LECA (LECA families; **Figure 2.1a**). These 10,233 LECA families do not include genes having only small Pfam domains, which we excluded for computational reasons, or genes without any domains. Therefore, we used a linear regression analysis to obtain an estimated LECA genome containing 12,753 genes (95% prediction interval: 7,447–21,840; **Extended Data Figure 2.1**).

Comparing the number of inferred LECA families to extant eukaryotes showed that the genome size of LECA reflected that of a typical present-day eukaryote (**Figure 2.1a**), which is in line with the inferred complexity of LECA, but in contrast with lower estimates obtained previously (Fritz-Laylin et al., 2010; Makarova et al., 2005). We used the split between Opimoda and Diphoda as root position of the eukaryotic tree of life (Derelle et al., 2015). As the exact position of the eukaryotic root is under debate (Burki et al., 2020), we tested alternative root positions and obtained very similar numbers of LECA families, except for the root positions at the base of and within the Excavates (15 – 46% fewer families compared with an Opimoda-Diphoda root; **Extended Data Figure 2.2a**). In case of a true excavate root, this could reflect fewer genes in LECA. However, given the sampling imbalance between both sides of an excavate root and the reduced nature of sampled excavate genomes, we consider a gene-rich LECA and subsequent gene losses a more likely scenario.

The multiplication factor (the number of LECA families divided by the number of acquired and invented genes or domains) was 1.8, approximating the near doubling reported before (Makarova et al., 2005). The observed doubling was validated in an additional dataset (**Supplementary Table 2.1**), despite a recent study (Tria et al., 2019) that has inferred very few duplications during eukaryogenesis (see Supplementary Information). Although on average genes duplicated once, the distribution of duplications is heavily skewed with many acquisitions from prokaryotes or eukaryotic inventions not having undergone any duplication (**Figure 2.1b**). The enormous expansion of the proto-eukaryotic genome was dominated by massive duplications in a small set of families (**Supplementary Table 2.2**).

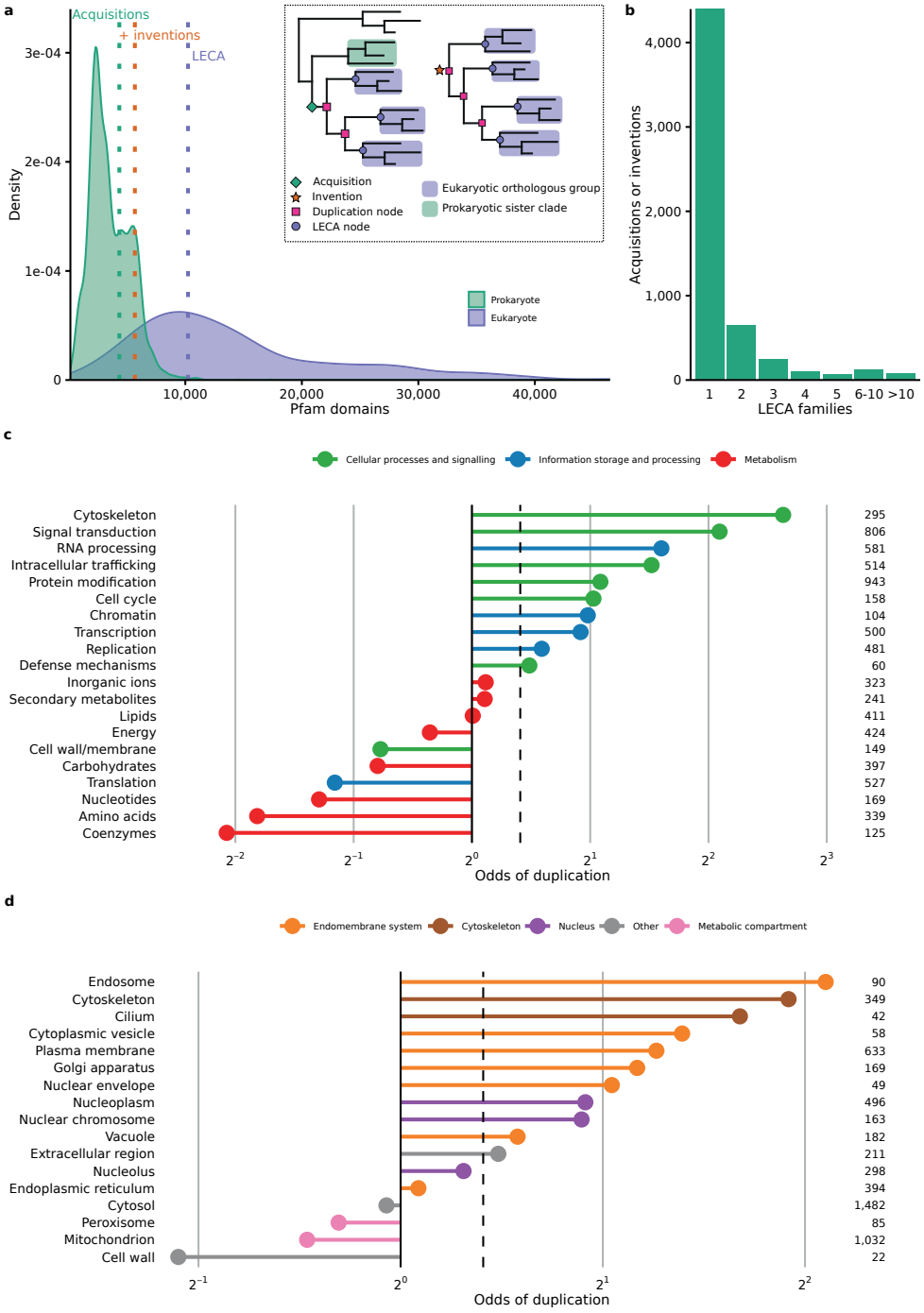

a
Acquisitions
+ inventions
LECA
Density
Pfam domains
Acquisition
Invention
Duplication node
LECA node
Eukaryotic orthologous group
Prokaryotic sister clade
Prokaryote
Eukaryote
b
Acquisitions or inventions
LECA families
c
Cellular processes and signalling
Information storage and processing
Metabolism
Cytoskeleton
Signal transduction
RNA processing
Intracellular trafficking
Protein modification
Cell cycle
Chromatin
Transcription
Replication
Defense mechanisms
Inorganic ions
Secondary metabolites
Lipids
Energy
Cell wall/membrane
Carbohydrates
Translation
Nucleotides
Amino acids
Coenzymes
Odds of duplication
d
Endomembrane system
Cytoskeleton
Nucleus
Other
Metabolic compartment
Endosome
Cytoskeleton
Cilium
Cytoplasmic vesicle
Plasma membrane
Golgi apparatus
Nuclear envelope
Nucleoplasm
Nuclear chromosome
Vacuole
Extracellular region
Nucleolus
Endoplasmic reticulum
Cytosol
Peroxisome
Mitochondrion
Cell wall
Odds of duplication

Figure 2.1 | Characterisation of duplications during eukaryogenesis. a, Density plot showing the distribution of the number of Pfam domains in present-day prokaryotes (green) and eukaryotes (purple) compared with the acquisition, acquisition plus invention and LECA estimates (dashed lines) obtained from phylogenetic trees (see inset). **b**, Number of acquisitions or inventions that gave rise to a particular number of LECA families, demonstrating the skewedness of duplications across protein families. **c**, Odds of duplication for LECA families according to KOG functional categories. Eighty-one percent of pairwise comparisons were significantly different (**Supplementary Figure 2.1**). The poorly characterised categories and functions of very few families (cell motility, extracellular structures and nuclear structure) are not depicted. **d**, Odds of duplication for LECA families according to cellular localisation. Fifty-four percent of pairwise comparisons were significantly different (**Supplementary Figure 2.2**). **c-d**, Numbers on the right side indicate the number of LECA families and dashed lines indicate the odds of duplication of all LECA families in total.

Duplicated and non-duplicated LECA families differed considerably in their functions and cellular localisations. Metabolic LECA families rarely had a duplication history, whereas LECA families involved in information storage and processing and in cellular processes and signalling were more likely to descend from a duplication ($\chi^2 = 572$, d.f. = 2 and $P = 7.7 \times 10^{-125}$; **Figure 2.1c** and **Supplementary Figure 2.1**). Notable exceptions to this pattern were families involved in cell wall or membrane biogenesis and translation, which were rarely duplicated. The observed differences in functions were reflected by differences between cellular localisations, with proteins in the endomembrane system and cytoskeleton mostly resulting from a duplication ($\chi^2 = 262$, d.f. = 4 and $P = 1.6 \times 10^{-55}$; **Figure 2.1d** and **Supplementary Figure 2.2**). Like duplications, inventions primarily occurred to families involved in informational and cellular processes ($\chi^2 = 226$, d.f. = 2 and $P = 8.8 \times 10^{-50}$ (function); $\chi^2 = 186$, d.f. = 4 and $P = 4.9 \times 10^{-39}$ (localisation); **Extended Data Figure 2.3** and **Supplementary Figures 2.3-2.6**). For complex eukaryotes to emerge, most innovations occurred in nuclear processes, the endomembrane system, intracellular transport and signal transduction, especially due to gene duplications.

Relatively large contribution of the host to duplicated LECA families

For the Pfams that were donated to the eukaryotic stem lineage we identified the prokaryotic sister group, which represents the best candidate for the Pfam's phylogenetic origin (**Extended Data Figure 2.4a**). Most acquisitions had a bacterial sister group (77%), of which only a small proportion was alphaproteobacterial (7% of all acquisitions), in agreement with previous analyses (Esser et al., 2004; Pisani et al., 2007; Pittis and Gabaldón, 2016a). The acquisitions from archaea (16%) predominantly had an Asgard archaeal sister (7% of all acquisitions). Moreover, the most common Asgard archaeal sister group was solely composed of Heimdallarchaeota (**Extended Data Figure 2.4b**); especially Heimdallarchaeote LC3 was frequently the sister group. This is in line with previous analyses providing support for either all Heimdallarchaeota or specifically LC3 being the currently known archaeal lineage most closely related to eukaryotes (Narrowe et al., 2018; Williams et al., 2020). The species in alphaproteobacterial sister groups, on the other hand, came from different orders (**Extended Data Figure 2.4c**), consistent with the recently proposed deep phylogenetic position of mitochondria (Martijn et al., 2018). The remaining acquisitions (7%) had an unclear prokaryotic ancestry (see Supplementary Discussion).

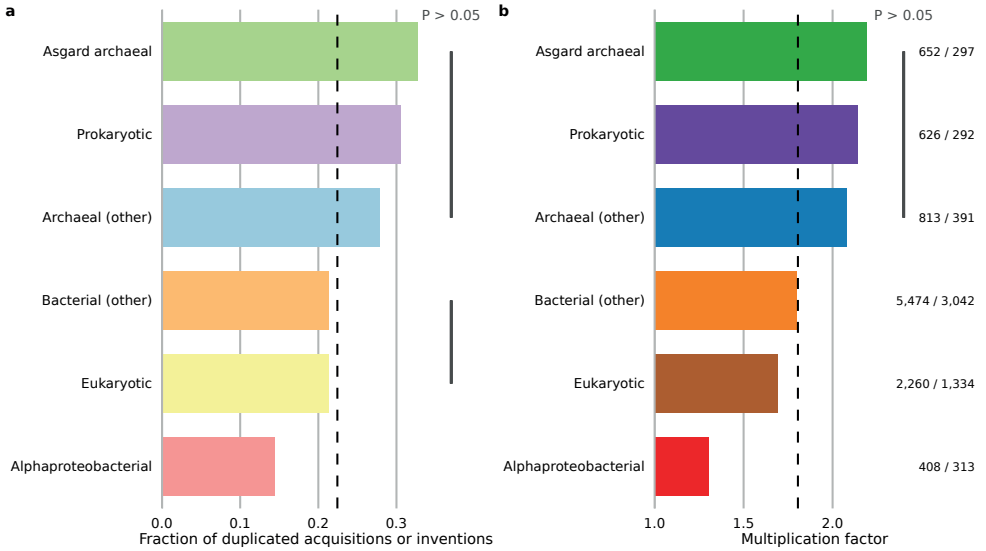


Figure 2.2 | Contribution of different phylogenetic origins to duplications during eukaryogenesis. **a**, Duplication tendency as fraction of clades having undergone at least one duplication. **b**, Multiplication factors, defined as the number of LECA families divided by the number of acquisitions or inventions. These numbers are shown beside the corresponding bar. **a**, **b**, Dashed lines indicate the duplication tendency and multiplication factor for all acquisitions and LECA families. The four (**a**) and three (**b**) pairwise comparisons that did not give a significant *P* value (χ^2 contingency table test) are shown by the grey lines. Prokaryotic: unclear prokaryotic ancestry (could not be assigned to a domain or lower taxonomic level).

Families with different sister clades varied substantially in the number of gene duplications they experienced during eukaryogenesis ($\chi^2 = 50$, d.f. = 5 and $P = 1.2 \times 10^{-9}$ (duplication tendency); $\chi^2 = 190$, d.f. = 5 and $P = 4.3 \times 10^{-39}$ (LECA families from duplication); **Figure 2.2**). The multiplication factor of 2.2 for families likely inherited from the Asgard archaea-related host was strikingly high compared with the invented families and families acquired from bacteria (between 1.3 and 1.8). Duplications related to the ubiquitin system and trafficking machinery especially contributed to the relatively large number of host-related paralogues (**Supplementary Table 2.2**). In contrast, there was a clear deficit of duplications in families with an alphaproteobacterial sister group (multiplication factor of 1.3). Hence, the endosymbiont marginally contributed to the near doubling of the genetic material via duplications during eukaryogenesis, whereas the host's relative contribution was largest.

Using branch lengths to time acquisitions and duplications

The remarkable differences in duplication dynamics between families with different affiliations could tentatively stem from differences in timing of these acquisitions and subsequent duplications. For example, the low number of alphaproteobacterial-associated duplications could be the result of a late mitochondrial acquisition. To research this, branch lengths in phylogenetic trees can be used. They serve as a good proxy for relative time and

have previously been used to time the acquisition of genes from the different prokaryotic donors (Pittis and Gabaldón, 2016a). Shorter branch lengths, corrected for differences in evolutionary rates across families, reflect more recent acquisitions. Duplications were not included in the previous timing analysis, but they can be timed in a similar way using the length of the branch connecting the duplication and LECA nodes (**Figure 2.3**). Although the measure has been criticised for its assumption that evolutionary rates pre- and post-LECA are correlated (Lane, 2017; Martin et al., 2017), it has yielded correct timings for specific post-LECA events (Pittis and Gabaldón, 2016a, 2016b). The observed trends can either be created by a common rate change in proteins of the same phylogenetic origin or can be due to different time points of acquisitions. Previous studies (Pittis and Gabaldón, 2016a, 2016b) have shown that the latter explanation is most plausible.

Although the inclusion of duplications in branch length analyses provides potentially valuable information, duplications could have affected the branch lengths by causing a shift in evolutionary rate. The stem lengths of acquisitions that happened simultaneously should approximate the same value, enabling us to assess the effect of duplication on branch lengths. We observed slight but notable increases in stem lengths for duplicated families from alphaproteobacterial origin (**Extended Data Figure 2.5a**) and for more recent duplications in vertebrates (**Extended Data Figure 2.5f**), but not for duplicated families from Asgard archaeal origin (**Extended Data Figure 2.5b**). It is therefore possible that in some families an accelerated rate could result in a slightly too early inferred duplication event according to our branch length analysis. We further checked whether there was a rate change after duplication in different functional groups of proteins and looked for an effect of homomer to heteromer transitions but we could not detect a clear pattern of rate shifts for different groups of proteins (**Extended Data Figure 2.5c-e**). We validated the use of duplication lengths by examining phylogenetic trees containing more recent duplications in the primate lineage, for which we have multiple intermediate speciation events. The distributions of duplication lengths followed the speciation events (**Extended Data Figure 2.5g**), demonstrating the validity of using duplication lengths to obtain an order of events. We also observed a small effect of function but the effect of time was much larger (**Extended Data Figure 2.5h**). Although duplications themselves and function can have an influence, time is the predominant factor explaining the differences in branch lengths. Thus, analysing branch lengths, including in duplicated families, is a valid and effective approach to infer an order of events.

Branch lengths point to a mitochondria-intermediate scenario

For the timing of acquisitions we obtained similar results to previous work (Pittis and Gabaldón, 2016a), with archaeal stems being longer than bacterial stems ($P = 4.5 \times 10^{-98}$, two-sided Mann-Whitney *U* test; **Figure 2.3**). Among the archaeal stem lengths the Asgard archaeal stems were shortest, as were the alphaproteobacterial stems among the bacterial stems, although for the former the difference failed to reach statistical significance ($P = 0.88$ and $P = 4.0 \times 10^{-4}$, respectively). This pattern is independent of the normalisation by post-LECA branches, the presence of duplications and functional divergence between the acquisition and LECA (**Extended Data Figure 2.6**). **Figure 2.3** shows that

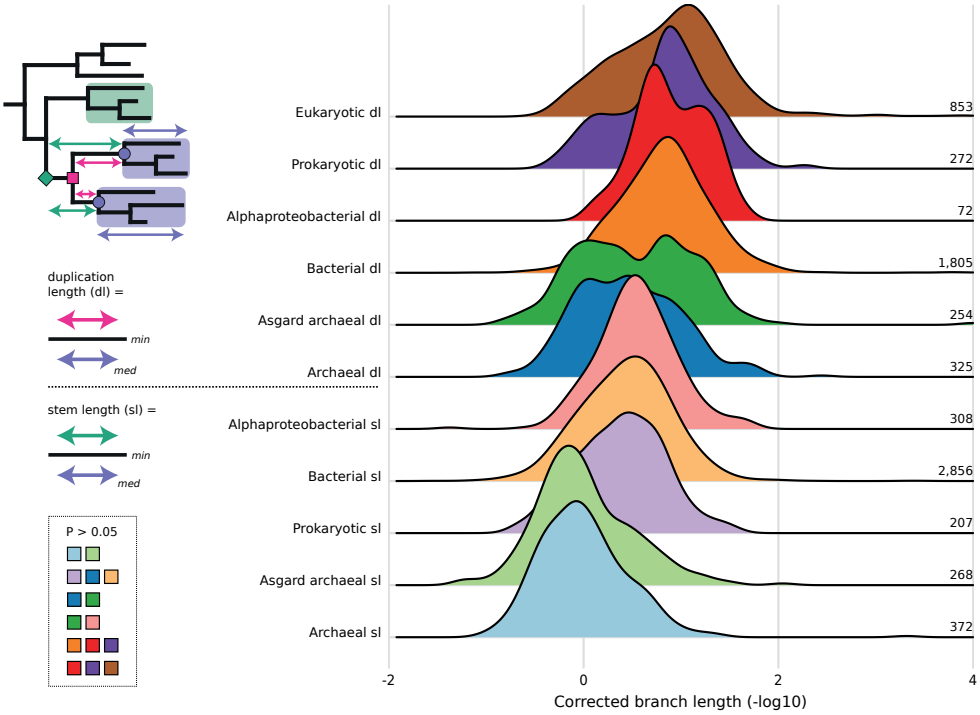


Figure 2.3 | Timing of acquisitions and duplications from different phylogenetic origins during eukaryogenesis. Ridgeline plot showing the distribution of corrected stem or duplication lengths, depicted as the additive inverse of the log-transformed values. Consequently, longer branches have a smaller value and vice versa. For clarity, a peak of near-zero branch lengths is not shown (see **Extended Data Figure 2.6**). Numbers on the right side of the plot indicate the number of acquisitions or duplications for which the branch lengths were included. Groups of stem and duplication lengths are ordered based on the median value. The tree illustrates how the stem and duplication lengths were calculated; the symbols and colour schemes are identical to **Figure 2.1a**. The phylogenetic distances between the acquisition or duplication and LECA were normalised by dividing them by the median branch length between LECA and the eukaryotic terminal nodes. In case of duplications the shortest of the possible normalised paths was used. Pairwise comparisons that did not give a significant *P* value (Mann-Whitney *U* test) are shown (bottom-left inset).

there is a wide distribution of host-related duplication lengths, with a substantial number of duplication lengths both longer and shorter than (alphaproteo)bacterial stem lengths. Bacteria-affiliated, endosymbiont-related and invented families showed the shortest duplication lengths. These duplication lengths were not affected by the position of the eukaryotic root (**Extended Data Figure 2.2b**). The differences in branch lengths indicate that an increase in genomic complexity via duplications probably had already occurred before the mitochondrial acquisition.

To shed light on the evolution of cellular complexity we categorised the duplications according to their functional annotations and cellular localisations. A marked distinction in duplication lengths between different functions can be observed, with duplications

in metabolic functions corresponding to shorter branches ($P = 8.0 \times 10^{-5}$, Kruskal-Wallis test; **Figure 2.4a** and **Supplementary Figure 2.7**). Moreover, a substantial number of duplication lengths in information storage and processes and in cellular processes and signalling functions were longer than the alphaproteobacterial stem length and duplications related to energy production, which mainly involve the mitochondria. These long duplication lengths include multiple duplications assigned to the cytoskeleton and intracellular trafficking. Duplications in signal transduction and transcription families mainly had shorter branch lengths, indicating that these regulatory functions evolved and diversified relatively late. With respect to cellular localisation, nucleolar and cytoskeletal duplication lengths were longest. Most duplications related to the endomembrane system had duplication lengths similar to those of mitochondrial duplications (**Figure 2.4b** and **Supplementary Figure 2.8**). These findings indicate that the increase in cellular complexity before the mitochondrial acquisition mainly comprised the evolution of cytoskeletal, intracellular trafficking and nucleolar components.

Discussion

This large-scale analysis of duplications during eukaryogenesis provides compelling evidence for a mitochondria-intermediate eukaryogenesis scenario. The results suggest that the Asgard archaea-related host already had some eukaryote-like cellular complexity, such as a dynamic cytoskeleton and membrane trafficking. Upon mitochondrial acquisition there was an even further increase in complexity with the establishment of a complex signalling and transcription regulation network and by shaping the endomembrane system. These post-endosymbiosis innovations could have been facilitated by the excess of energy allegedly provided by the mitochondrion (Lane, 2014; Lane and Martin, 2010).

A relatively complex host is in line with the presence of homologues of eukaryotic cytoskeletal and membrane trafficking genes in Asgard archaeal genomes (Klinger et al., 2016; Spang et al., 2015; Zaremba-Niedzwiedzka et al., 2017). Moreover, some of them, including ESCRT-III homologues, small GTPases and (loki)actins, have duplicated in these archaea as well, either before eukaryogenesis or more recently (Klinger et al., 2016; Spang et al., 2015; Zaremba-Niedzwiedzka et al., 2017). This indicates that there has already been a tendency for at least the cytoskeleton and membrane remodelling to become more complex in Asgard archaeal lineages. A dynamic cytoskeleton and trafficking system, perhaps enabling primitive phagocytosis (Martijn and Ettema, 2013), might have been essential for the host to take up the bacterial symbiont. Molecular and cell biology research on these archaea, from which the first results have recently become public (Akıl and Robinson, 2018; Imachi et al., 2020), is very likely to yield more insight into the nature of the host lineage. In addition to a reconstruction of the host, further exploration of the numerous acquisitions, inventions and duplications during eukaryogenesis is key to fully unravelling the origin of eukaryotes.

Methods

In this study we inferred and analysed two different sets of phylogenetic trees. The first set (Pfam–ScrollSaw) was used for the main analysis, whereas the second set (eukaryotic

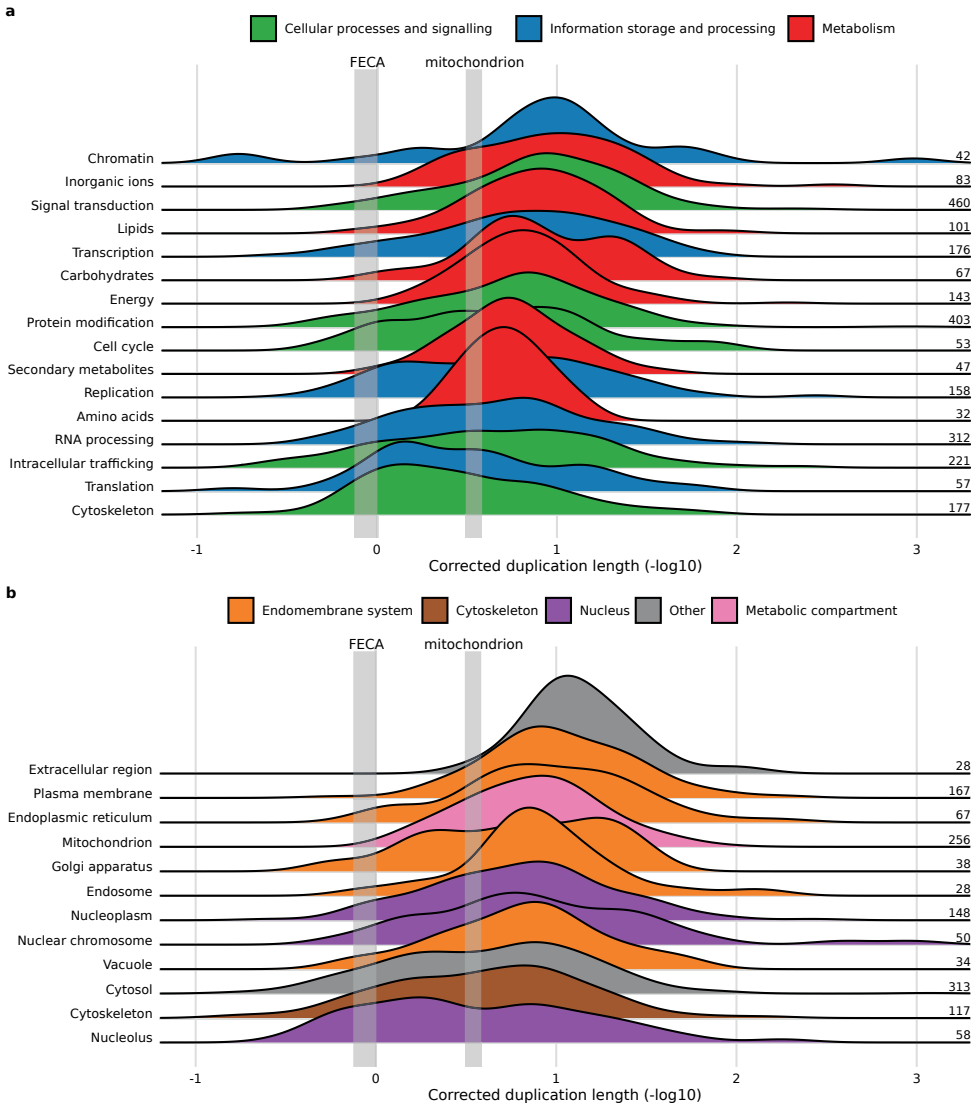


Figure 2.4 | Timing of duplications during eukaryogenesis according to function and localisation. **a**, **b**, Ridgeline plots showing the distribution of duplication lengths for different functional categories (**a**) and cellular localisations (**b**). Numbers on the right side of the plots indicate the number of duplications for which the duplication lengths were included. To enable a comparison with the timing of acquisitions, the binomial-based 95% confidence interval of the median of the Asgard archaeal (first eukaryotic common ancestor (FECA)) and alphaproteobacterial stem lengths (mitochondrion) are depicted in grey, indicating the divergence of eukaryotes from their Asgard archaea-related and Alphaproteobacteria-related ancestors, respectively. Groups are ordered based on the median value. For significant differences between groups, see **Supplementary Figures 2.7 and 2.8**.

orthologous groups (KOGs) mapped to homologous prokaryotic clusters of orthologous groups (COGs); KOG-to-COG clusters) was used to verify our method to infer duplications during eukaryogenesis. We also used a third, already existing set of gene trees (human phylome) to validate the use of branch lengths in case of duplications. Below we describe how we created and analysed the main set of phylogenetic trees. The second and third sets of gene trees are described in the Supplementary Methods.

Data

We used 209 eukaryotic (predicted) proteomes from an in-house dataset that has been used and described in previous work (Deutekom et al., 2019). Prokaryotic proteomes (3,457 in total) were extracted from eggNOG 4.5 (Huerta-Cepas et al., 2016a). The prokaryotic dataset was supplemented with nine predicted proteomes from the recently described Asgard superphylum (Zaremba-Niedzwiedzka et al., 2017).

Pfam assignment

We used hmmsearch (HMMER v3.1b2 (Eddy, 2011)) with the Pfam 31.0 profile hidden Markov models (HMMs) (Finn et al., 2016) and the corresponding gathering thresholds to assess to which Pfam what part of each prokaryotic and eukaryotic sequence should be assigned. We opted for Pfam profile HMMs to collect homologous sequences because of their sensitivity to detect homology. The domains that were hit were extracted from the sequences based on the envelope coordinates. If a sequence had hits to multiple Pfams and these hits were overlapping for at least 15 amino acids only the best hit was used. If the same Pfam had multiple hits in the sequence due to an insertion relative to the model the different hits were artificially merged. Since the latter is more prone to errors for short models and since short sequences contain less phylogenetic signal, profile HMMs shorter than fifty amino acids were not considered for further analysis.

Reduction of sequences

For each Pfam, the number of prokaryotic sequences was reduced with kClust v1.0 (Hauser et al., 2013) using a clustering threshold of 2.93, which corresponds to a sequence identity of 60%. We chose this threshold because we expect it to retain sufficient prokaryotic diversity while removing sequences from related species to keep the analysis computationally feasible. However, because of horizontal gene transfer (HGT), it will also remove sequences from more distantly related species in some cases.

The number of eukaryotic sequences was reduced with a novel method (van Wijk and Snel, 2020) based on the ScrollSaw approach (Elias et al., 2012). The idea behind ScrollSaw is that instead of selecting a species subset *a priori*, the slowest evolving sequences are selected. In that way the resolution of deep nodes in trees from expanded families is drastically improved. Although in the original paper (Elias et al., 2012) the distances between sequences were calculated with a maximum likelihood method, we used the bit score in the Basic Local Alignment Search Tool (BLAST) (Altschul et al., 1990) as a proxy to obtain genetic distances. For each Pfam an all species versus all species BLAST was performed. Because we were only interested in the best hit the max_target_seqs option

was set to 1. Although this option has raised some attention recently (González-Pech et al., 2019; Shah et al., 2019), we only used it as a proxy for evolutionary distance and our analysis would not be seriously impacted by this option given the overall small sizes of our databases. Subsequently, bidirectional best hits (BBHs) between sequences from different eukaryotic groups were identified. Eukaryotic species can be grouped into different 'supergroups', whose names and definitions have changed following new findings (Adl et al., 2019; Burki et al., 2020). The species in our dataset are from the following six groups: Archaeplastida + Cryptista, SAR + Haptista, Discoba, Metamonada, Obazoa and Amoebozoa. For our main analysis we used BBHs between sequences from two groups, because that provided the best resolution (van Wijk and Snel, 2020). Although the exact position of the root of the eukaryotic tree of life is uncertain (Burki et al., 2020), a likely position is between Opimoda (Obazoa and Amoebozoa in our set) and Diphoda (other supergroups) (Derelle et al., 2015). Therefore, BBHs between Opimoda and Diphoda sequences were identified and the corresponding sequences were used for phylogenetic analysis.

To assess the impact of a different position of the eukaryotic root, we also identified BBHs between five groups, merging Metamonada and Discoba into Excavata, and four groups, in which Archaeplastida + Cryptista and SAR + Haptista were combined as Diaphoretickes and Obazoa and Amoebozoa were together as Amorphea (see 'Effect of the position of the eukaryotic root').

Phylogenetic analysis

Multiple sequence alignments were made with MAFFT v7.310 (Katoh and Standley, 2013) (auto option) and trimmed with trimAl v1.4.rev15 (Capella-Gutiérrez et al., 2009) (gap threshold 10%). Phylogenetic trees were inferred with IQ-TREE v1.6.4 (Nguyen et al., 2015) (LG4X model (Le et al., 2012), 1,000 ultrafast bootstraps (Minh et al., 2013)). If the consensus tree had a higher likelihood than the best tree from the search, the former was used for further analysis. Because inferring trees for PF00005 (ABC transporter), PF00072 (response regulator receiver domain), PF00528 (binding-protein-dependent transport system inner membrane component), PF02518 (histidine kinase-, DNA gyrase B-, and HSP90-like ATPase) and PF07690 (major facilitator superfamily) in this way was too computationally demanding, we used FastTree v2.1.10 (Price et al., 2010) with the LG model to construct trees for these Pfams. These Pfams were not considered for branch length analysis.

Tree analyses

Removal of interspersing prokaryotes

Trees were analysed with an in-house ETE3 (Huerta-Cepas et al., 2016b) script. We examined whether the tree contained prokaryotic sequences that probably reflect recent HGT events and that might interfere with our analysis. Prokaryotic sequences from a single genus that were in between eukaryotic sequences were pruned from the tree. If there was only one prokaryotic sequence in the tree it was kept only if it was an Asgard archaeal sequence, because it has been reported that sometimes a single sequenced Asgard archaeon contains a homologue to sequences otherwise only present in eukaryotes (Zarem-

ba-Niedzwiedzka et al., 2017). This was the case for 16 trees containing LECA families (see below), including RPL28/MAK16, Sec23/24, UFM1 and the C-terminal domain of tubulins, for which the Asgard archaeal origin has been shown before. Because another prokaryotic outgroup to root these trees was lacking, they were not used to calculate stem lengths (see 'Branch length analysis').

Annotation of eukaryotic nodes

For each eukaryotic clade the nodes were annotated as duplications before LECA, LECA nodes, post-LECA nodes or unclassified. Only clades that contained at least one LECA node were of interest. The node combining the eukaryotic clade with the rest of the tree (if present) was annotated as acquisition node.

For the annotation of nodes in trees the information from the eukaryotic sequences that were not in the BBHs was included, since the number of eukaryotic sequences in the trees had been reduced. To correctly assign in-paralogues we additionally performed an own species versus own species BLAST for each Pfam (max_target_seqs = 2). The sequences that were not in a tree were mapped onto their best hits in the tree according to the BLAST score.

To infer reliable duplication nodes in the tree, duplication consistency scores were calculated for all internal nodes starting from the root of a eukaryotic clade. This score is the overlap of species at both sides of a node divided by the total number of species at both sides, taking both sequences in the tree and assigned sequences (as described above) into account. If the duplication consistency score was at least 0.2 and both daughter nodes fulfilled the LECA criteria, this node was annotated as a duplication node. The first LECA criterion was that a node had to have both Opimoda and Diphoda tree sequences in the clade. Secondly, to take care of post-LECA HGT events among eukaryotes and of tree uncertainties, the mean presence of a potential LECA family in the five different supergroups (Obazoa, Amoebozoa, SAR + Haptista, Archaeplastida + Cryptista and Excavata) had to be at least 15%. If a node did not fulfil the LECA criteria it was annotated as a post-LECA node.

The above-mentioned thresholds were chosen based on manual inspection of a selection of trees. Using different thresholds for duplication consistency (0, 0.1, 0.2 and 0.3) and LECA coverage scores (0, 5, 10, 15, 20 and 25%) had a gradual impact on the absolute numbers and quality measures, such as the fraction of well-supported LECA and duplication nodes (**Supplementary Table 2.3**). This underlines that the reported results were not contingent on the specific set of thresholds chosen and that for most nodes the duplication consistency and LECA coverage were high.

After this first annotation round all LECA nodes in the trees were re-evaluated. If there were duplication nodes in both daughters, the node connecting these duplications had to be a duplication node as well even though its duplication consistency score was below the threshold. This was only the case for two nodes in total. If there were duplication nodes in only one daughter lineage, the LECA node was annotated as unclassified. It could reflect a duplication event or a tree artefact due to rogue taxa. If there were no duplication nodes in either daughter lineage, all LECA nodes in the daughter lineages of this LECA node were reannotated as post-LECA nodes.

Rooting eukaryote-only trees

For trees with only eukaryotic sequences and trees for which all prokaryotic sequences had been removed, inferring the root poses a challenge. For these trees duplication and LECA nodes were called in unrooted mode. The distances between the LECA nodes were calculated and the tree was rooted in the middle of the LECA nodes that were furthest apart, resulting in an additional duplication node at this root. If there were no duplications found in this way because there were less than two duplications in the tree, rooting was tried on each internal node. The node that fulfilled the duplication criteria and that maximised the species overlap was chosen. If none fulfilled the criteria, it was checked if the entire tree fulfilled the LECA criteria. For Pfams for which we could not infer a tree because there were only two or three sequences selected, we also checked if this Pfam itself fulfilled the LECA criteria. These Pfams correspond to eukaryote-specific families that did not duplicate.

Sister group identification

For each eukaryotic clade in trees also containing prokaryotic sequences the sister group was identified in unrooted mode. By doing so, the eukaryotic clade initially had two candidate sister groups. Eukaryotic sequences in a sister group, if present, were ignored, as they could reflect HGT events, contaminations, tree artefacts or true additional acquisitions. To infer the actual sister group it was first checked if one of the two candidate sister groups was more likely by checking if one of them consisted only of Asgard archaea, TACK archaea, Asgard plus TACK archaea, alphaproteobacteria, beta/gammaproteobacteria, or alpha/beta/gammaproteobacteria. If so, that clade was chosen as the actual sister group. If both sister groups had the same identity or if both groups had another identity than the ones described above, the tree was rooted on the farthest leaf from the eukaryotic clade. In many cases the last common ancestor of the taxa in the sister group was Bacteria, Archaea or cellular organisms according to the National Center for Biotechnology Information (NCBI) taxonomy. Such wide taxonomic assignments probably reflect extensive HGT among distantly related prokaryotes. In these cases, it was checked whether one of the previously mentioned groups or otherwise a particular phylum or proteobacterial class comprised a majority of the prokaryotic taxa to get a more precise sister group classification.

We observed that in a substantial number of cases there was another eukaryotic clade with LECA nodes in the sister group of a eukaryotic clade. These cases could reflect a duplication and subsequent loss in prokaryotes but probably reflect tree artefacts. Therefore, these clades were ignored for the branch length analysis. Acquisitions that were nested, that is, they shared the same prokaryotic sister group because one acquisition had in its sister clade only one prokaryotic clade and one or multiple other acquisitions, were merged for further analysis.

Branch length analysis

Multiple branch lengths were calculated in clades containing LECA nodes. For the stem length (sl) the distance to the acquisition node (the node uniting the eukaryotic clade and its prokaryotic sister) was calculated for each LECA node. This distance was divided

by the median of the distances from the LECA node to the eukaryotic leaves (eukaryotic branch lengths (ebl)) to correct for rate differences between orthologous groups as done in previous work (Pittis and Gabaldón, 2016a). In case of multiple possible paths due to duplications, the minimum of these distances was used as the sl, since it was closest to sl values from zero-duplication clades. To calculate the duplication length (dl) a similar approach was followed, using the duplication node instead of the acquisition node.

To investigate the impact of rates after duplication in both paralogue lineages within a family, we also calculated for all duplication nodes in Asgard archaea-derived families the minimal sl going through these duplications (**Extended Data Figure 2.5c**, **d**). In this way, we obtained an sl value for each duplication, in addition to the aforementioned sl value for each acquisition. These values were also divided into duplications in families that had undergone a transition from homomers to heteromers (proteasome, Snf7, TRAPP, Vps36 and OST3/OST6) and families that had not (**Extended Data Figure 2.5e**).

Combining eukaryote-only Pfam families with prokaryotic donations in their clan

The classification of protein families into Pfams is not based on taxonomic levels. A Pfam present only in eukaryotes can therefore be the result of a duplication event instead of a *bona fide* invention. To distinguish these possible scenarios we used the Pfam clans, in which related Pfam families are combined. If there were only eukaryote-only Pfams in a clan based on our analysis, these Pfams were merged into one invention event. If there was only one Pfam with an acquisition from prokaryotes and for this Pfam there was only one acquisition, the eukaryote-only Pfams were combined with this acquisition. If there were multiple acquisitions in a clan, a profile-profile search with HH-suite3 v3.0.3 (Steinegger et al., 2019) was performed to assign eukaryote-only Pfams to an acquisition. Per acquisition in a clan an alignment was made from the tree sequences in the corresponding eukaryotic clade with MAFFT L-INS-i v7.310 (Katoh and Standley, 2013). Profile HMMs were made of these alignments (hhmake –M 50) and they were combined in a database (ffindex_build). The eukaryote-only Pfam HHMs were searched against the acquisition HHM database per clan with hhsearch. Each Pfam was assigned to the acquisition that had the best score.

Functional annotation

Functional annotation of sequences was performed using emapper-1.0.3 (Huerta-Cepas et al., 2017) based on eggNOG orthology data (Huerta-Cepas et al., 2016a). Sequence searches were performed using DIAMOND v0.8.22.84 (Buchfink et al., 2015).

The most common KOG functional category among the tree sequences of a LECA node was chosen as the function of the LECA node. If there was not one most common function, the node was annotated as S (function unknown). For the functional annotation of duplication nodes a Dollo parsimony approach was used. For this we checked whether there was one single annotation shared between LECA nodes at both sides, ignoring unknown functions. If this was not the case but the parent duplication node (if present) had a function, this function was also used for the focal duplication node. The

functional annotation of the prokaryotic sister group was performed the same way as for a LECA node. In the figures the names of most categories were shortened for increased readability: translation (translation, ribosomal structure and biogenesis), RNA processing (RNA processing and modification), replication (replication, recombination and repair), chromatin (chromatin structure and dynamics), cell cycle (cell cycle control, cell division, chromosome partitioning), signal transduction (signal transduction mechanisms), cell wall/membrane (cell wall/membrane/envelope biogenesis), intracellular trafficking (intracellular trafficking, secretion, and vesicular transport), protein modification (posttranslational modification, protein turnover, chaperones), energy (energy production and conversion), carbohydrates (carbohydrate transport and metabolism), amino acids (amino acid transport and metabolism), nucleotides (nucleotide transport and metabolism), coenzymes (coenzyme transport and metabolism), lipids (lipid transport and metabolism), inorganic ions (inorganic ion transport and metabolism) and secondary metabolites (secondary metabolites biosynthesis, transport and catabolism).

The same approach was used to assign cellular components to LECA and duplication nodes, using a custom set of gene ontology (GO) terms: extracellular region (GO:0005576), cell wall (GO:0005618), cytosol (GO:0005829), cytoskeleton (GO:0005856), mitochondrion (GO:0005739), cilium (GO:0005929), plasma membrane (GO:0005886), endosome (GO:0005768), vacuole (GO:0005773), peroxisome (GO:0005777), cytoplasmic vesicle (GO:0031410), Golgi apparatus (GO:0005794), endoplasmic reticulum (GO:0005783), nuclear envelope (GO:0005635), nucleoplasm (GO:0005654), nuclear chromosome (GO:0000228) and nucleolus (GO:0005730).

Predicting the number of genes in LECA

We used a linear regression model to predict the number of genes in LECA based on the inferred number of Pfam domains in LECA. For this we used the number of sufficiently long Pfam domains (see 'Pfam assignment' above) and the number of protein-coding genes in the eukaryotes in our dataset. The assumptions of a normal distribution of gene values and equal variance at each Pfam domain value were reasonably met after log transformation. Based on the relationship between the number of Pfam domains and genes in present-day eukaryotes, the number of protein-coding genes in LECA was estimated.

Effect of the position of the eukaryotic root

The eukaryotic phylogeny and the position of its root are incorporated in our analysis at two points: in the ScrollSaw step during the identification of BBHs between eukaryotic taxa and in the LECA criteria in the tree analyses. For computational reasons we limited the analysis of the impact of the eukaryotic phylogeny on our results to the Pfams that were only present in eukaryotes. In addition to the Opimoda-Diphoda BBHs, we selected the sequences from BBHs between either five or four supergroups, as described above, and inferred phylogenetic trees. The three different sets of trees were analysed using all seven root possibilities, given the monophyly of Amorphea, Diaphoretickes, Discoba and Metamonada. To fulfil the LECA criteria a node had to contain tree sequences from both sides of the root and the mean presence of a potential LECA family in the four different

groups had to be at least 15%.

Statistical analysis

Overrepresentations of functions and localisations in duplications, inventions and innovations, and overrepresentations of sister groups in duplications and duplication tendencies were tested by comparing odds ratios with Fisher's exact tests (only pairwise comparisons of functions for inventions and localisations for innovations due to small sample sizes) or χ^2 contingency table tests (all other comparisons). Differences in branch lengths were assessed with a Kruskal-Wallis test, followed by Mann-Whitney *U* tests upon a significant outcome of the Kruskal-Wallis test. Only one Kruskal-Wallis test did not give a significant result (**Extended Data Figure 2.2b**). Differences between two groups were assessed with Mann-Whitney *U* tests. All performed tests were two-sided. In all cases of multiple comparisons, the *P* values were adjusted to control for the false discovery rate.

The ridgeline plots were drawn with the ggridges v0.5.1 R package (https://github.com/wilkelab/ggridges).

Data availability

Fasta files, phylogenetic trees and their annotations are available in figshare with the identifier https://doi.org/10.6084/m9.figshare.10069985 (Vosseberg et al., 2020).

Code availability

The code used to annotate the phylogenetic trees can be accessed in Github (https://github.com/JulianVosseberg/feca2leca).

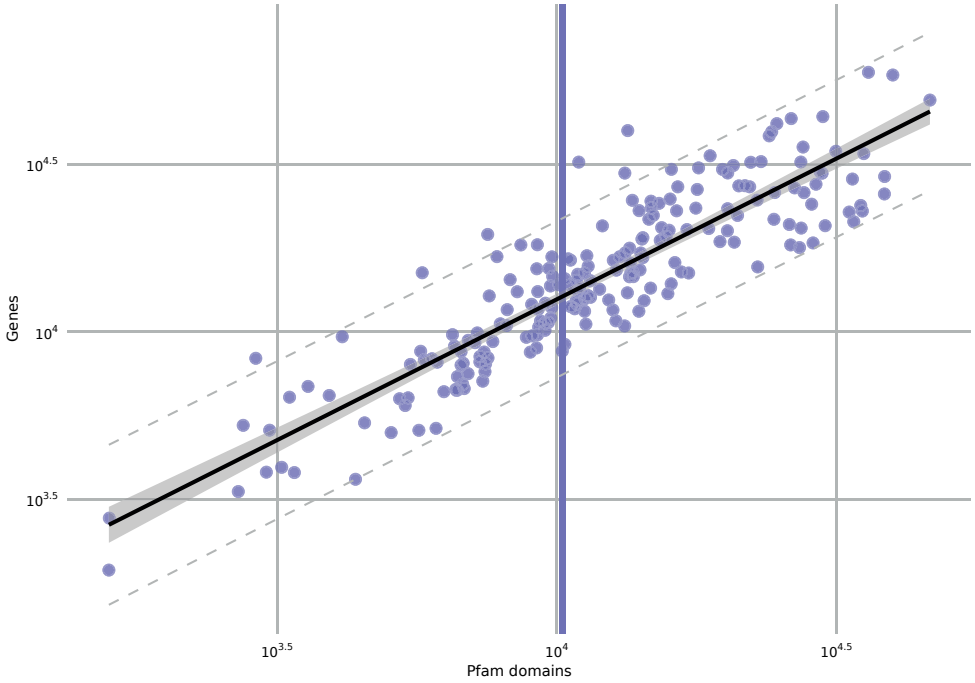
Acknowledgements

We thank K. S. Marakova and E. V. Koonin for sharing their KOG-to-COG protein clusters with us. We are grateful to T. J. P. van Dam, E. S. Deutekom and G. J. P. L. Kops for useful advice and discussions. This work is part of the research programme VICI with project number 016.160.638, which is (partly) financed by the Netherlands Organisation for Scientific Research (NWO). T.G. acknowledges support from the Spanish Ministry of Science and Innovation for grant PGC2018-099921-B-I00 and from the European Union's Horizon 2020 research and innovation programme under grant agreement ERC-2016-724173.

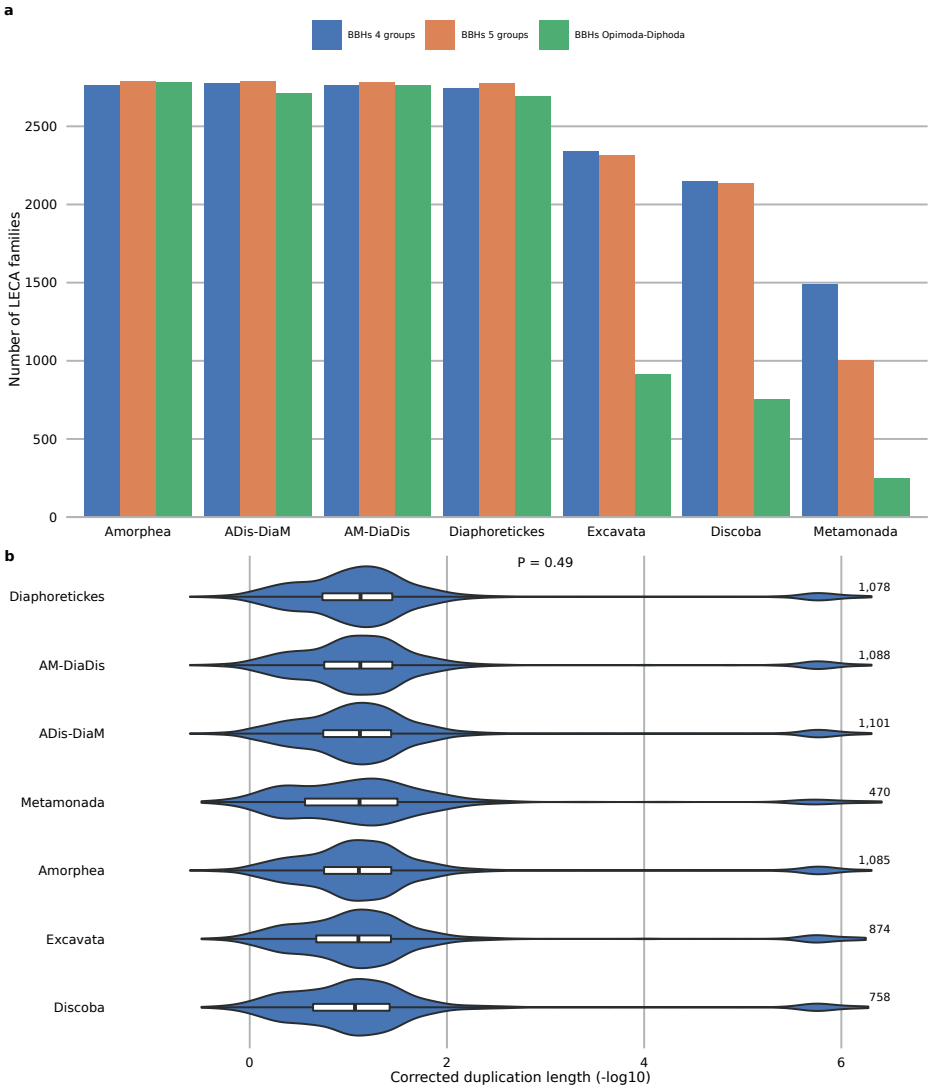
Author contributions

J.J.E.v.H., T.G. and B.S. conceived the study. J.V. and J.J.E.v.H. performed the research. J.V., J.J.E.v.H., T.G. and B.S. analysed and interpreted the results. M.M.H. performed the analysis on the human phylome. M.M.H. and A.v.V. aided in the development of the tree analysis pipeline. L.M.v.W. implemented the ScrollSaw-based method. J.V., J.J.E.v.H. and B.S. wrote the manuscript, which was edited and approved by all authors.

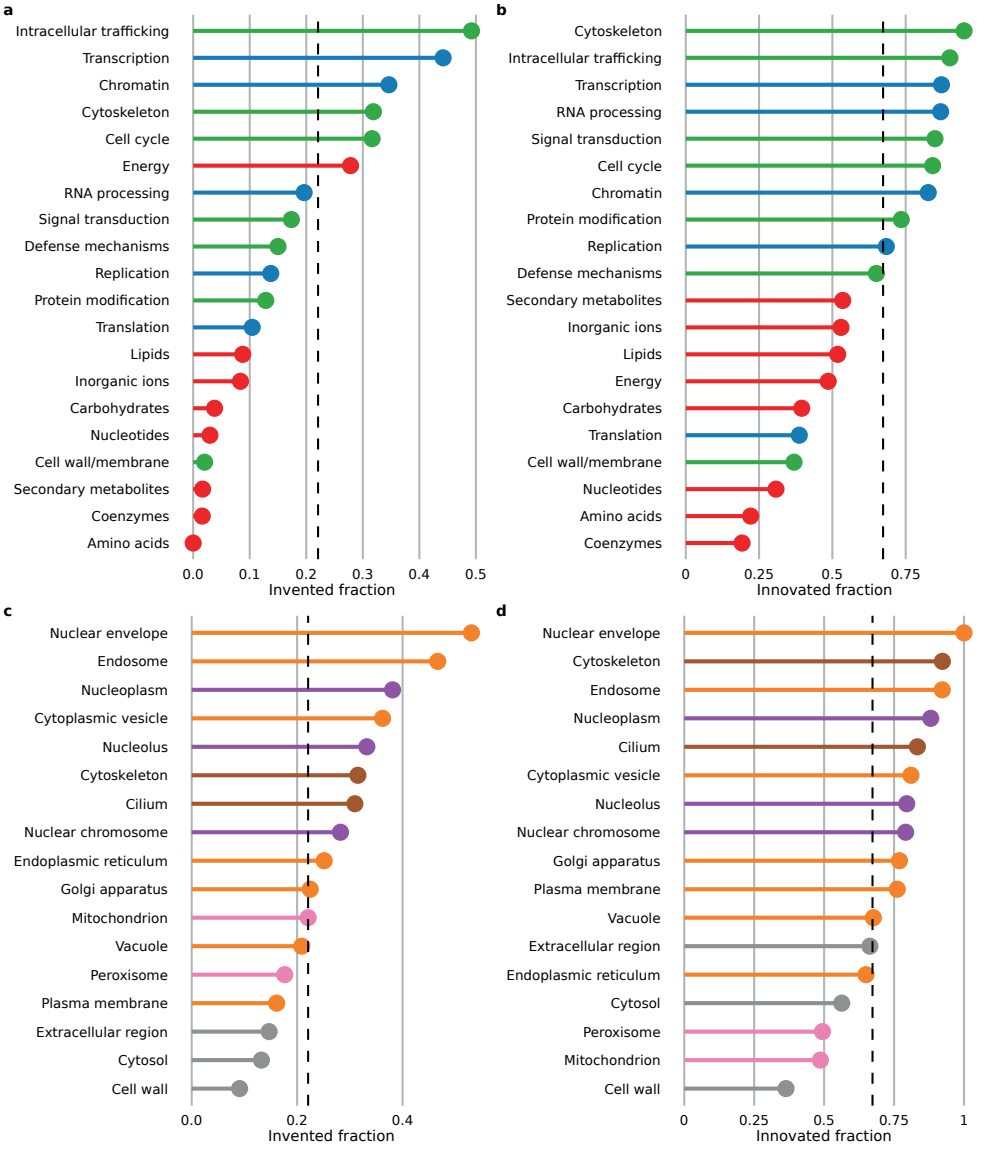
Extended Data



Extended Data Figure 2.1 | Estimating the number of LECA genes from the number of Pfam domains with linear regression. Scatter plot showing the number of Pfam domains and protein-coding genes in present-day eukaryotes, with each dot representing one genome. The regression line (black) and its 95% confidence (filled grey) and prediction intervals (dashed grey) are depicted. The vertical line corresponds to the obtained number of LECA Pfam domains.

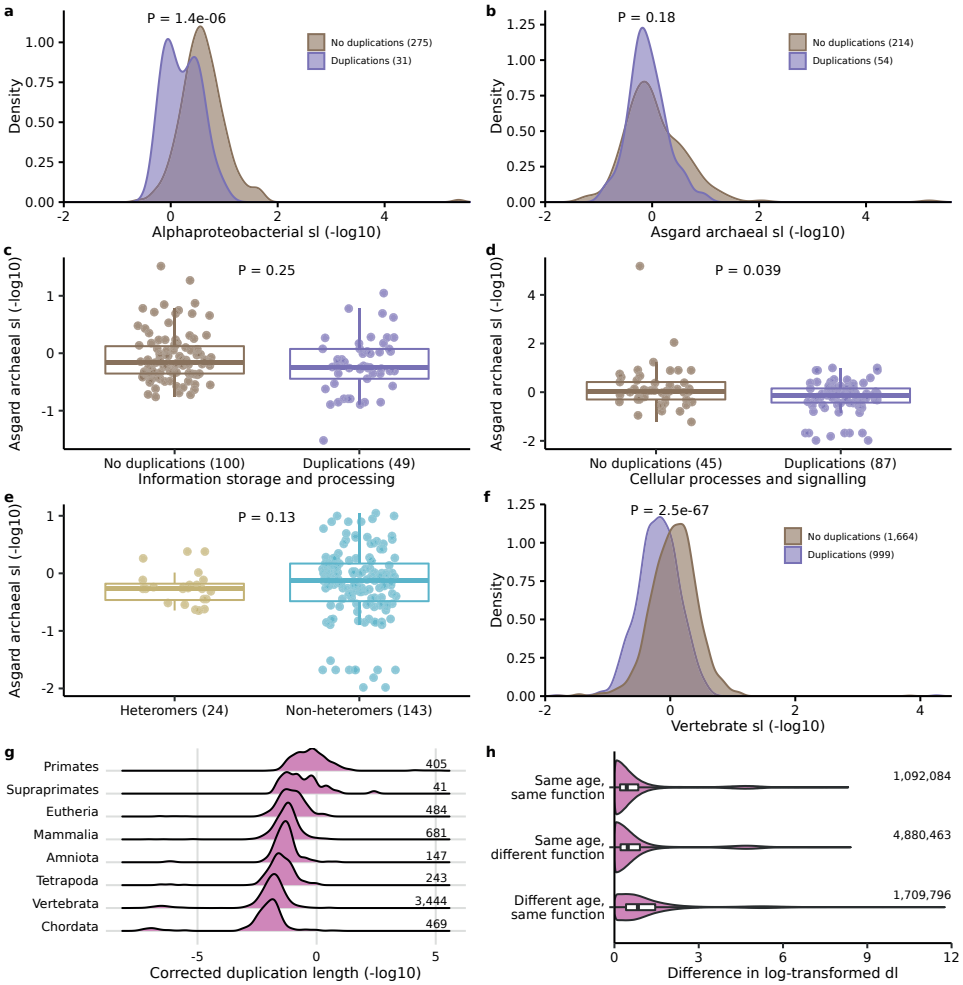


Extended Data Figure 2.2 | Effect of a different phylogenetic position of the eukaryotic root. a, Number of inferred LECA families considering different root positions. These numbers are based on phylogenetic trees from Pfams that are only present in eukaryotes. Besides the Opimoda and Diphoda groups, two other group definitions were used to identify bidirectional best hits (BBHs) and select sequences for tree inference. Names of root positions indicate either the lineage at one side of the root or the position of the split (ADis-DiaM: Amorphea+Discoba – Diaphoretickes+Metamonada; AM-DiaDis: Amorphea+Metamonada – Diaphoretickes+Discoba). Excavate sequences, especially from Metamonada species, are rarely involved in BBHs, unless specifically searched for (Excavata in BBHs 5 groups; Discoba and Metamonada in BBHs 4 groups). **b**, Distribution of duplication lengths obtained using different root positions for eukaryote-only trees based on the four group BBHs. The difference between distributions is not statistically significant according to the Kruskal-Wallis test.

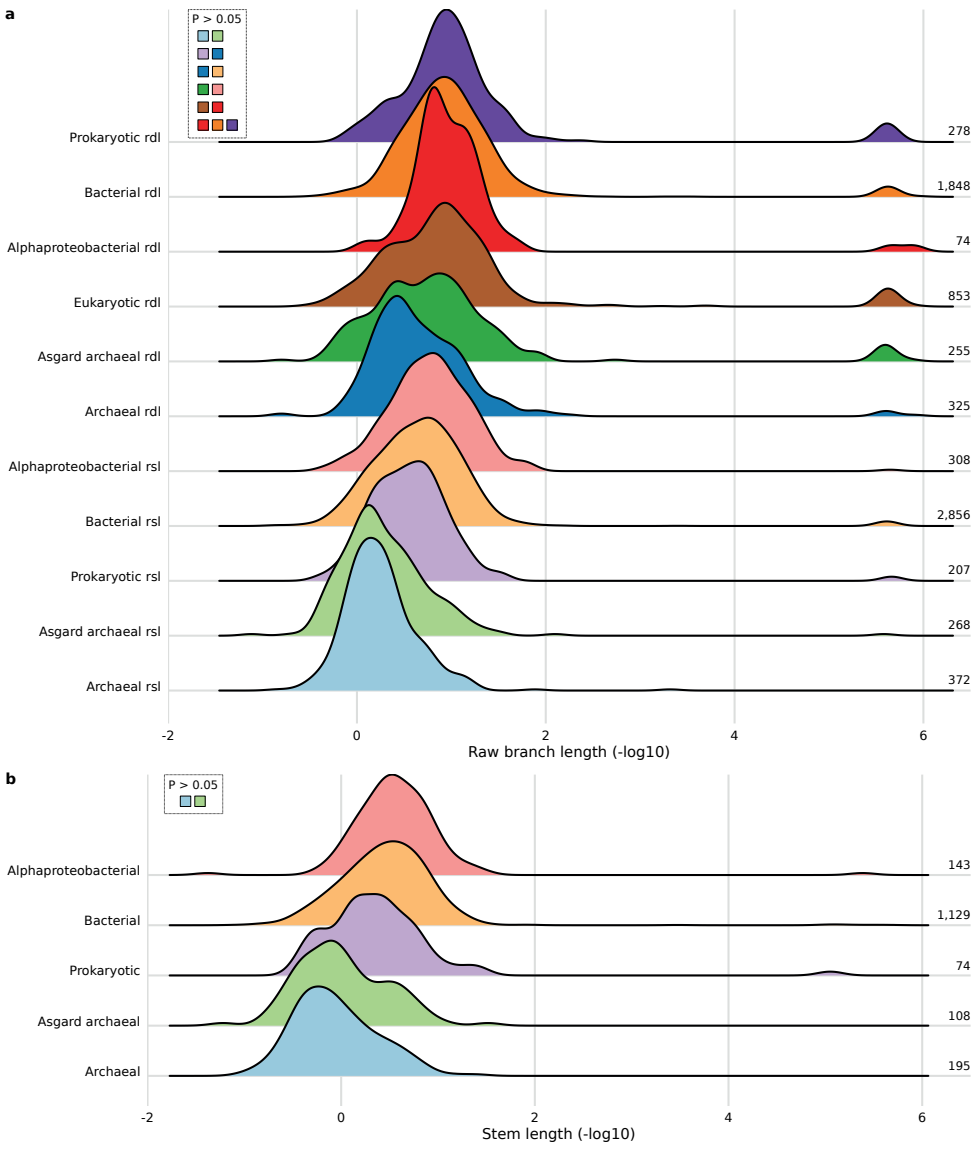


Extended Data Figure 2.3 | Fraction of LECA families resulting from inventions. a, Contribution of inventions to LECA families performing different functions. 82% of pairwise comparisons were significantly different (**Supplementary Figure 2.3**). **b**, Fraction of LECA families resulting from either an invention or duplication – a eukaryotic innovation – according to functional category. 84% of pairwise comparisons were significantly different (**Supplementary Figure 2.5**). **c**, Contribution of inventions to LECA families performing their function in different cellular components. 51% of pairwise comparisons were significantly different (**Supplementary Figure 2.4**). **d**, Fraction of LECA families resulting from an innovation according to cellular localisation. 74% of pairwise comparisons were significantly different (**Supplementary Figure 2.6**). **a-d**, Dashed lines indicate the overall invented or innovated fraction.

Extended Data Figure 2.4 | Phylogenetic origin of acquired Pfams. a, **b**, Phylogeny of the prokaryotes (**a**) and Asgard archaea (**b**) present in our dataset based on the NCBI taxonomy. The branch widths and numbers indicate the number of acquisitions from a group. **c**, Number of acquisitions from different alphaproteobacterial orders or a combination of multiple orders ('Alphaproteobacteria').



Extended Data Figure 2.5 | Effect of duplications on branch lengths. a, **b**, Distribution of alphaproteobacterial (**a**) and Asgard archaeal (**b**) stem lengths (sl's) for acquisitions without and with duplications. Two alphaproteobacterial sl's from acquisitions with Magnetococcales as sister group were removed based on the previously inferred phylogenetic position of mitochondria (Martijn et al., 2018). **c**, **d**, Distribution of Asgard archaeal sl's for information storage and processing (**c**) and cellular processes and signalling families (**d**), comparing those without and with duplications. Upon removal of the outliers, the difference in cellular processes and signalling families no longer reached statistical significance. **e**, Distribution of Asgard archaeal sl's for duplicated acquisitions, in which homomer-to-heteromer transitions had occurred compared to the other duplicated acquisitions. **f**, Distribution of vertebrate sl's for families without and with duplications. **g**, Distribution of duplication lengths (dl's) grouped according to the lineage in which the duplication occurred. All pairwise comparisons were significantly different (Mann-Whitney *U* tests). **h**, Distribution of differences in log-transformed dl values for all pairwise comparisons between chordate duplications according to age and functional annotation. All groups are significantly different (Mann-Whitney *U* tests). **a-f**, *P* values of Mann-Whitney *U* tests are shown. **c-e**, The minimal sl via each duplication node is plotted.



Extended Data Figure 2.6 | Effect of branch length normalisation and functional divergence. a, Ridgeline plot showing the distribution of uncorrected stem (rsl) or duplication lengths (rdl). Numbers indicate the number of acquisitions or duplications for which the branch lengths were included. The low peaks at very short branch lengths are an artefact from near-zero branch lengths. Groups are ordered based on the median value of rsl's and rdl's. **b**, Ridgeline plot showing the distribution of sl's for non-duplicated acquisitions that share the same functional annotation of the prokaryotic sister group and are therefore expected to have undergone little functional divergence during eukaryogenesis. **a**, **b**, Branch lengths are depicted as the additive inverse of the log-transformed values. Pairwise comparisons that did not give a significant *P* value (Mann-Whitney *U* tests) are shown.

Supplementary Methods

KOG-to-COG clusters analysis

Selecting sequences and generating clusters

In order to compare our phylogenomics approach to previously reported accounts of duplications during eukaryogenesis, we applied it to the clusters of homologous sequences established by Makarova *et al.* (Makarova et al., 2005). Briefly, they mapped eukaryotic orthologous groups (KOGs) to homologous prokaryotic orthologous groups (COGs). In many cases, multiple KOGs mapped to a single COG, which often reflects a duplication during eukaryogenesis. Furthermore, KOGs had been clustered together if they are homologous to each other but lack a homologous COG. We used these KOG-to-COG clusters to assess if we, using a phylogenomics approach, were able to recapture the prevalence of gene duplications during eukaryogenesis that Makarova *et al.* observed by calculating ratios of KOGs to their affiliated COGs. Moreover, we took advantage of the current wealth of sequenced biodiversity by using an alternative, more representative species and sequence dataset compared to the original study. The results of this KOG-to-COG analysis can be found in **Supplementary Table 2.1**.

To recreate the KOG-to-COG clusters we used the COG assignment of the non-Asgard archaeal prokaryotic sequences provided by eggNOG and performed sequence profile searches with the Asgard archaeal and eukaryotic sequences. For the Asgard archaea, we downloaded profile HMMs of all COGs from eggNOG 4.5 (Huerta-Cepas et al., 2016a) and assigned the Asgard protein sequences to COGs using hmmscan (HMMER v3.1b1 (Eddy, 2011)). For eukaryotes, we selected ten species to obtain a good representation of eukaryotic diversity: *Naegleria gruberi* and *Euglena gracilis* (Excavata), *Cladospihon okarmurans* and *Bigelowiella natans* (SAR+Haptista), *Guillardia theta* and *Klebsormidium flaccidum* (Archaeplastida+Cryptista), *Acanthamoeboa castellanii* and *Acytostelium subglobosum* (Amoebozoa), and *Capsaspora owczarzaki* and *Nuclearia* sp. (Obazoa). We specifically opted for these species, because they were often involved in BBHs in the Pfam sequence selection (see Methods, 'Reduction of sequences'). Subsequently, we downloaded profile HMMs for orthologue clusters at the level of eukaryotes from eggNOG 4.5 (Huerta-Cepas et al., 2016a). These contained both the supervised KOGs and non-supervised orthologous groups (ENOGs). The original KOG-to-COG clusters from Makarova *et al.* (Makarova et al., 2005) did not include these ENOGs, but instead included candidate orthologous groups (TWOGs). Because these TWOGs are now obsolete, we sought to find the best matching ENOG based on the original sequence members of each TWOG. We combined the profile HMMs of these ENOGs with those of the KOGs and created a profile database. We performed hmmscan to assign protein sequences from the eukaryotic species to these KOGs/ENOGs.

Subsequently, for all KOGs/ENOGs and COGs, we reduced the number of sequences with kClust v1.0 (Hauser et al., 2013), using a score per column of 3.53 (approximately 70% sequence identity). We subsequently merged homologous sequences from eukaryotes, prokaryotes and Asgard archaea according to the KOG-to-COG mapping, resulting in updated KOG-to-COG clusters comprising sequences from diverse and informative

eukaryotic and prokaryotic clades.

Phylogenetic analyses
For each KOG-to-COG cluster, we generated phylogenetic trees using an in-house pipeline also used previously (Pittis and Gabaldón, 2016a). The sequences were aligned using MAFFT v6.861b (Katoh and Toh, 2008), option –auto, and subsequently trimmed using trimAl v1.4 (Capella-Gutiérrez et al., 2009) with a gap threshold of 0.1. From these alignments, we constructed phylogenetic trees using FastTree v2.1.8 (Price et al., 2010) with WAG as evolutionary model.

Tree analyses
For the annotation of nodes in KOG-to-COG trees a similar approach as for the Pfam-ScrollSaw trees was followed. Only the criteria for LECA and duplication nodes were slightly different. Because of the lower number of eukaryotic species we here simply annotated a node as a LECA node if it contained both Opimoda and Diphoda sequences, and instead of a consistency score, we used a species overlap criterion of two to annotate duplication nodes: if the daughters both fulfilled the LECA criterion and shared at least two out of the in total ten eukaryotic species, it was annotated as a duplication node.

Human phylome analysis
To validate the use of branch lengths to time gene duplications, we also applied this approach to the numerous duplications in chordates. We inferred these from the human phylome, which we downloaded from PhylomeDB (Huerta-Cepas et al., 2014) (Phylome ID 76: http://phylomedb.org/phylome_76). The results of this validation can be found in **Extended Data Figure 2.5f-h**.

In this collection of phylogenetic trees, we calculated the normalised vertebrate stem lengths by dividing the branch length between the common ancestors of chordates and vertebrates by the median branch length between the latter and present-day vertebrates. In case of duplications the stem length was included if the human seed protein was in the shortest possible stem length.

To obtain duplication lengths for duplications that occurred at different phylogenetic time points, we scanned in each tree the lineage of the human seed protein between the common ancestors of bilaterians and primates for the presence of duplications. Nodes connecting the seed with a human paralogue were annotated as duplication nodes. The phylogenetic time point ('age') of the duplication was obtained using the common ancestor of all species involved in the duplication event. Duplication lengths were calculated by dividing the branch length between the duplication node and the common ancestor of primates by the median branch length between the latter and present-day primates.

KOG functional categories were assigned to each protein in the phylome using emapper-2.0.1 (Huerta-Cepas et al., 2017) based on eggNOG orthology data (Huerta-Cepas et al., 2019). Functional annotation of the nodes in the trees were performed as described for duplication nodes before (see 'Functional annotation'). For each pair of duplications it was checked if they performed the same function and had the same age, performed the

same function but had a different age or performed a different function but had the same age. For these pairs the difference in log-transformed duplication lengths was calculated.

Supplementary Discussion

Data sets used

We tested two different data sets. The KOG-to-COG gene family clusters (Makarova et al., 2005) are a set specifically constructed to study duplications during eukaryogenesis and were therefore an ideal starting point. To get an even more complete picture of duplications we decided to use the Pfam database. By using this database, we circumvented the need to use orthologous groups or infer homology. For certain families the Pfam domains correspond to full-length genes, whereas for others it is only a domain or even a motif. Although certain domain duplications are not fully independent of each other due to their presence in a single gene upon duplication, it is not unlikely that truly separated genes co-duplicated as well. Ideally, one would want to define the unit, either a domain or full-length gene, that evolved as an individual entity during eukaryogenesis. However, for various domains/genes it would be simply impossible to identify such a single entity, for example for domains that were independent upon acquisition or invention but fused during eukaryogenesis and were therefore interdependent in LECA. This is especially probable given the abundance of gene fusion events during eukaryogenesis (Méheust et al., 2018).

Sister group identity

7% of the acquisitions had an unclear prokaryotic ancestry. Both bacteria and archaea were present in the sister group with no phylum comprising a majority. A tentative explanation is that the identity of the donor is obscured due to post-acquisition HGT among distantly related prokaryotes. The tendency of these acquisitions to duplicate was similar to the Pfams with an archaeal ancestry (**Figure 2.2**). This suggests that a large fraction of this group reflect genes present in the host lineage. Furthermore, a relatively large fraction of these acquisitions had another eukaryotic clade with LECA families in their sister group (34%, between 3 and 10% for the other groups), indicating that some of these acquisitions are placed in an incorrect, deep phylogenetic position. The stem and duplications lengths of these families with an unclear prokaryotic ancestry, however, were similar to those from families acquired from bacteria. Further research into these families is needed to elucidate their phylogenetic origin.

Branch lengths analysis

The stem lengths of acquisitions that happened simultaneously should approximate the same value, enabling us to assess the effect of duplications on branch lengths. Assuming the deep mitochondrial origin outside the alphaproteobacteria (Martijn et al., 2018), all acquisitions with alphaproteobacteria as sister group should correspond to the same event, namely the divergence of the pre-mitochondrial and alphaproteobacterial lineages. We observed a difference in stem lengths between duplicated and non-duplicated fam-

ilies from alphaproteobacterial origin, with duplicated families corresponding to longer stems (**Extended Data Figure 2.5a**). Even using the shortest branch as stem, which we chose in case of duplications, could not fully account for the difference in stem lengths in these few duplicated families. In contrast, no difference in stem lengths with duplications was seen for acquisitions with an Asgard archaeal sister group (**Extended Data Figure 2.5b**). We also looked at the effect of duplications on the stem lengths for the numerous duplications that occurred in the vertebrate stem. For these more recent duplications we observed a longer vertebrate stem in case of duplications (**Extended Data Figure 2.5f**), in line with the alphaproteobacterial-related duplications. The presence of duplications can result in a subtle yet significant accelerated evolutionary rate in both daughter lineages.

Because we had detected more duplicated families with an Asgard archaeal sister group than an alphaproteobacterial one, we looked more in depth into the first. We could not detect a clear pattern of acceleration after duplications in both daughter lineages for different functional groups (**Extended Data Figure 2.5c-d**). The barely significant difference for duplications related to cellular processes and signalling was dependent on the presence of outliers. Duplications that resulted in the transition from a homomer to a heteromer could have had a different effect on evolutionary rate as the selection pressures on the protein interface has changed. We did not observe a difference between duplications in families that underwent such a transition and other families (**Extended Data Figure 2.5e**). However, the number of the first group was low and involved all duplications in these families, not only those resulting in the homomer-heteromer transition. Further research into these different effects of duplications is warranted. In conclusion, we could not confidently distinguish differences in rates for different groups of proteins upon duplication that could bias our results.

The inferred timing of acquisitions represents the *earliest* possibility of the actual acquisition because they are the result of taxon sampling (i.e., which of the present-day organisms have been discovered, sequenced and/or included in the analysis) and historical contingency (i.e., which lineages have not gone extinct). Duplication nodes, on the other hand, represent the latest possibility of the actual acquisition, and therefore they could be used to attenuate the inferred acquisition time point.

Comparison with Tria et al.

Our conclusions are in stark contrast with a recent preprint (Tria et al., 2019), which reported remarkably fewer gene duplications and relatively many duplications in bacterial-related genes (compared to archaeal-related genes), which they interpret as being derived from the proto-mitochondrion. Based on their findings, the authors concluded that gene duplications support a eukaryogenesis model in which mitochondria entered early in eukaryogenesis, into a relatively simple, prokaryote-like host. We think this conclusion is insufficiently supported by their approach and resulting observations, because these have some clear deficits.

First and foremost, they infer very few eukaryogenesis duplications: 713 compared to 4,564 in our main dataset (see **Supplementary Table 2.1**). As an illustration: they did not recover well-documented greatly expanded protein families such as protein kinases and

small GTPases (Elias et al., 2012; Jékely, 2003), which we were able to recover (see **Supplementary Table 2.2**). The family that according to this preprint was most duplicated during eukaryogenesis was the dynein light chain family with 12 duplications.

Second, because they only inferred gene trees for eukaryotic sequences, they could not distinguish between duplications that happened during eukaryogenesis, those that happened before and pseudoparalogues (e.g., cytosolic and mitochondrial ribosomal proteins). Moreover, their limited usage of gene phylogenies also prohibits them from specifying the potential identity of the prokaryotic donor lineage.

Third, they do not discriminate between genes with alphaproteobacterial and another bacterial origin, but instead label all eukaryotic genes with bacterial affiliations as coming from the mitochondrial endosymbiont. Some, if not most, of these genes might in fact have been acquired through HGT from other bacterial lineages. Potentially, mixing these contributes to the relatively high number of gene duplications that count for endosymbiont-derived genes.

Fourth, they did not include the Asgard archaea in their analysis, which are crucial for any inference about eukaryogenesis. This might explain why the duplications in the cytoskeletal and ubiquitin systems were not correctly identified as duplications associated to archaeal acquisitions (Spang et al., 2015; Zaremba-Niedzwiedzka et al., 2017) in their analysis. This may have led to an underestimation of the duplications in host-related genes.

Supplementary Tables

Supplementary Table 2.1 | Comparison of different datasets.

	Pfam-ScrollSaw trees	Trees from recreated KOG-to-COG clusters	Original KOG-to-COG clusters (no trees) (Makarova et al., 2005)
Acquisitions	4,335	3,460	1,092
Inventions	1,334	883	1,058
Duplications	4,564	4,888	1,987
LECA families	10,233	9,231	4,137
Multiplication factor	1.81	2.12	1.92

Supplementary Table 2.2 | Most expanded acquisitions or inventions during eukaryogenesis.

Pfam	Ancestry	Number of LECA families
Total		
Mitochondrial carrier*	Invention	123
Protein kinase	Planctomycetes	106
RING-finger/U-box	Actinobacteria	92
PH domain	Haloplasma	82
Ubiquitin	Asgard archaea	76
C2 domain	Prokaryotes	72
RNA recognition motif	Aβγ-proteobacteria	71
Tetratricopeptide repeat	Firmicutes	66
POZ domain	Chlamydiae	50
FYVE/PHD zinc finger	Invention	46
Asgard archaea		
Ubiquitin	Asgard archaea	76
Vps51 domain superfamily	Asgard archaea	19
Cyclin	Asgard archaea	19
Helix-turn-helix	Asgard archaea	16
Thioredoxin	Asgard archaea	15
Helix-turn-helix	Asgard archaea	11
Golgi-transport	Asgard archaea	10
Helix-turn-helix	Asgard archaea	10
Gelsolin repeat	Asgard archaea	10
Gelsolin repeat	Asgard archaea	10
Alphaproteobacteria		
Sterile alpha motif	Alphaproteobacteria	10
Galactosyltransferase	Alphaproteobacteria	9
EF-hand 8	Alphaproteobacteria	8
Iron/zinc purple acid phosphatase-like protein C	Alphaproteobacteria	5
DDE superfamily endonuclease	Alphaproteobacteria	5
ABC transporter	Alphaproteobacteria	5
Alpha/beta hydrolase fold	Alphaproteobacteria	5
Ferric reductase	Alphaproteobacteria	4

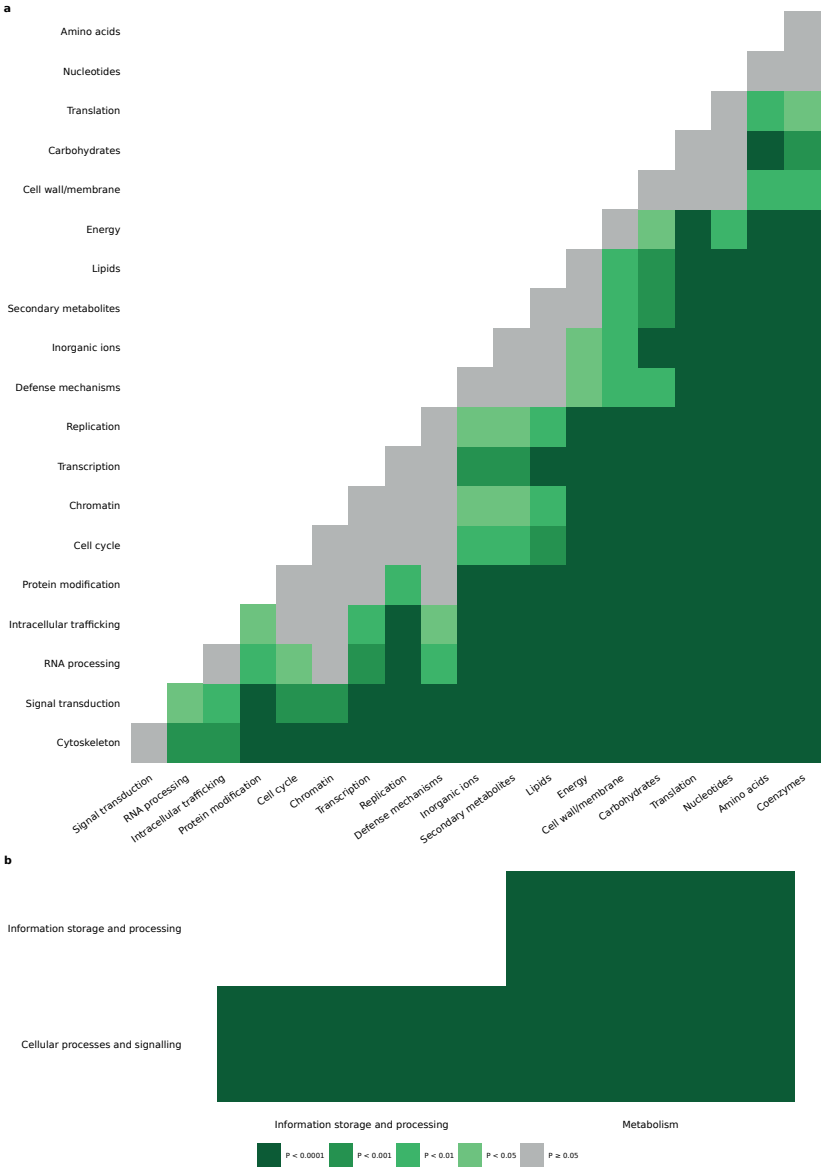
*A mitochondrial carrier protein typically contains three of these domains.

Supplementary Table 2.3 | Effect of different duplication consistency and LECA coverage thresholds.

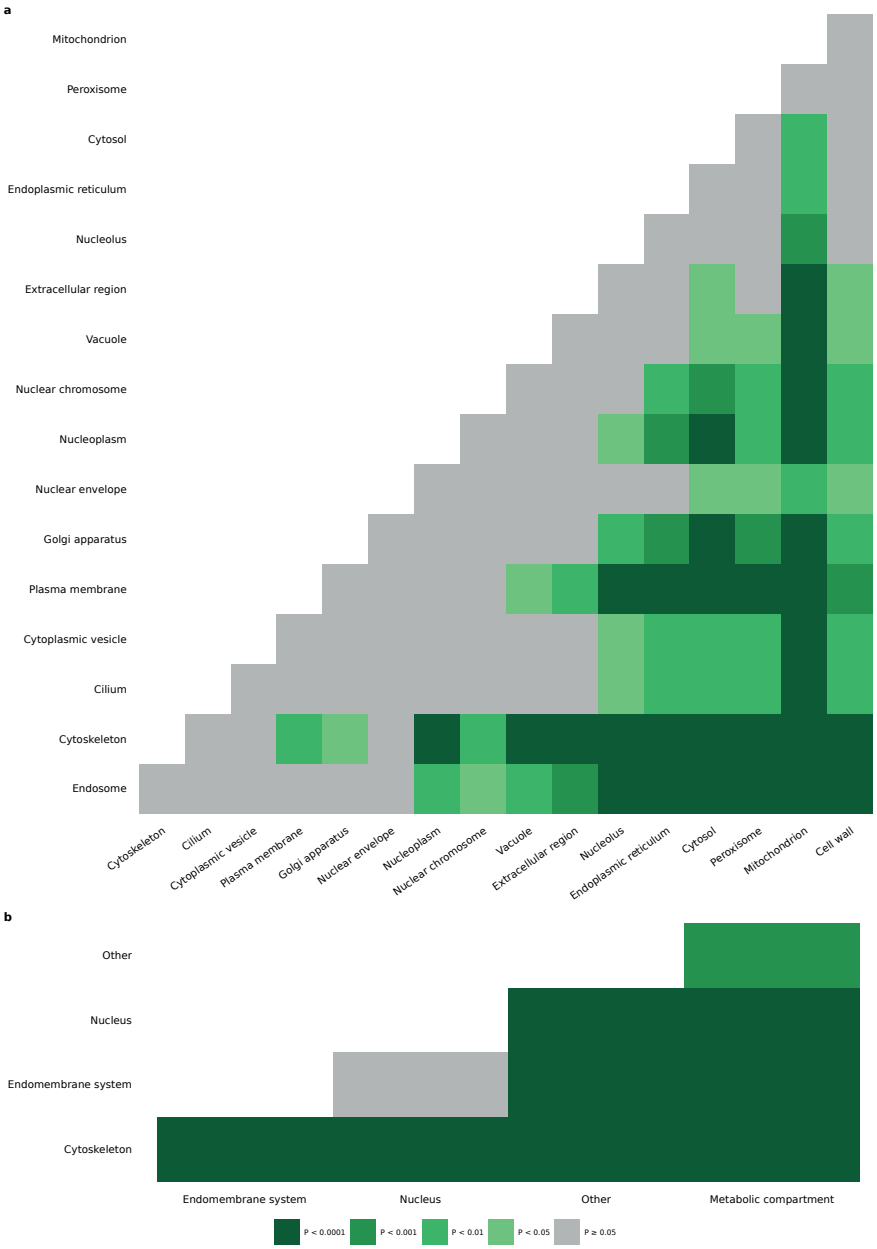
Duplication consistency score	LECA coverage score	Number of LECA families	Number of unclassified nodes	Number of eukaryotic clades without LECA families	Fraction well-supported* LECA nodes	Fraction well-supported* duplication nodes
0	0	23,567	5,304	19,661	0.47	0.26
	5	19,724	4,801	21,556	0.43	0.26
	10	15,671	4,013	23,095	0.41	0.27
	15	12,531	3,205	24,314	0.42	0.28
	20	10,248	2,591	25,145	0.43	0.29
	25	8,648	2,000	25,731	0.45	0.30
10	0	18,588	2,928	19,661	0.53	0.24
	5	16,028	3,221	21,556	0.51	0.24
	10	13,317	2,522	23,095	0.49	0.26
	15	11,048	2,137	24,314	0.50	0.26
	20	9,339	1,916	25,145	0.51	0.28
	25	8,083	1,651	25,731	0.52	0.28
20	0	16,547	2,354	19,661	0.55	0.24
	5	14,335	2,514	21,556	0.53	0.24
	10	12,092	2,029	23,095	0.52	0.25
	15	10,233	1,772	24,314	0.52	0.26
	20	8,821	1,586	25,145	0.53	0.27
	25	7,764	1,397	25,731	0.54	0.28
30	0	15,241	1,976	19,661	0.56	0.25
	5	13,161	1,924	21,556	0.54	0.25
	10	11,147	1,673	23,095	0.54	0.26
	15	9,523	1,490	24,314	0.54	0.27
	20	8,306	1,360	25,145	0.55	0.28
	25	7,420	1,235	25,731	0.55	0.29

*Ultrafast bootstrap support value 95 or higher.

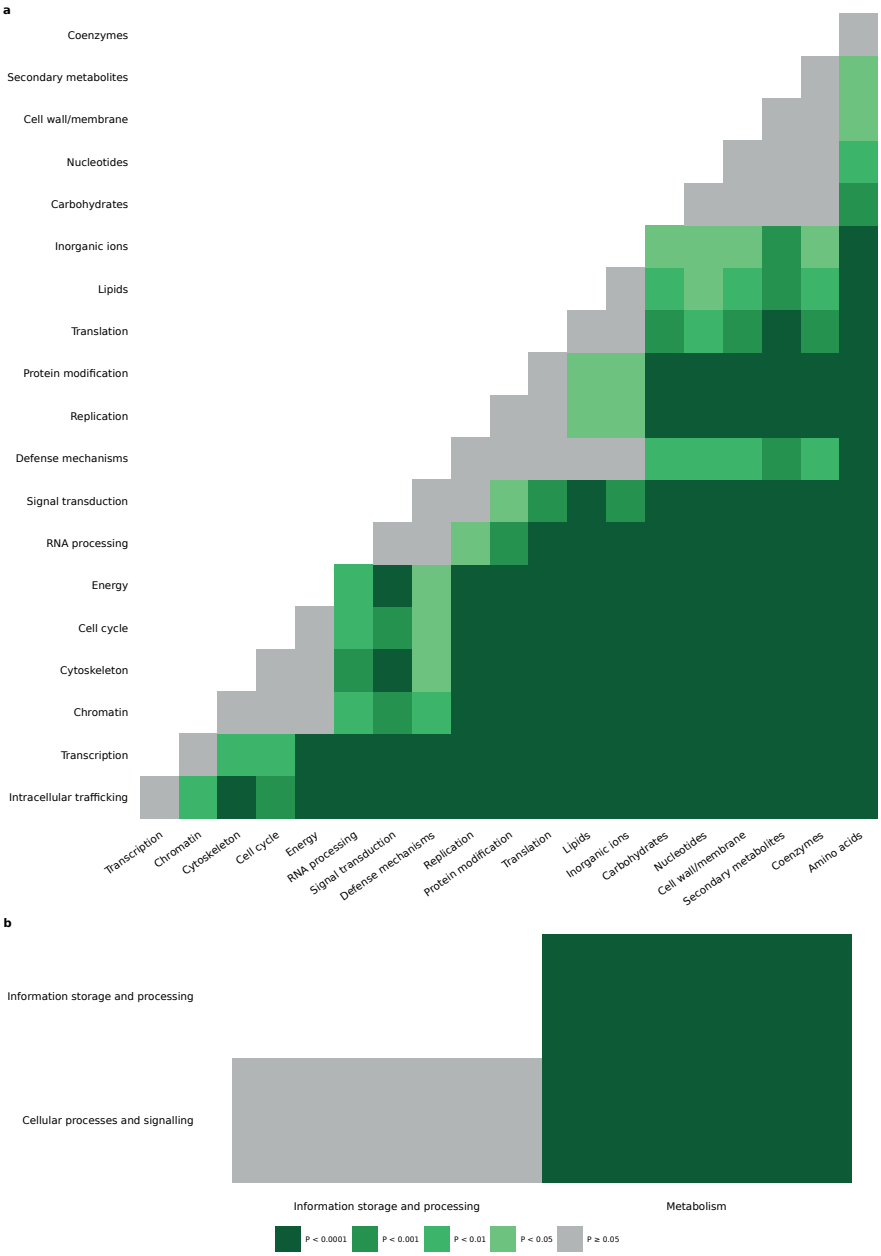
Supplementary Figures



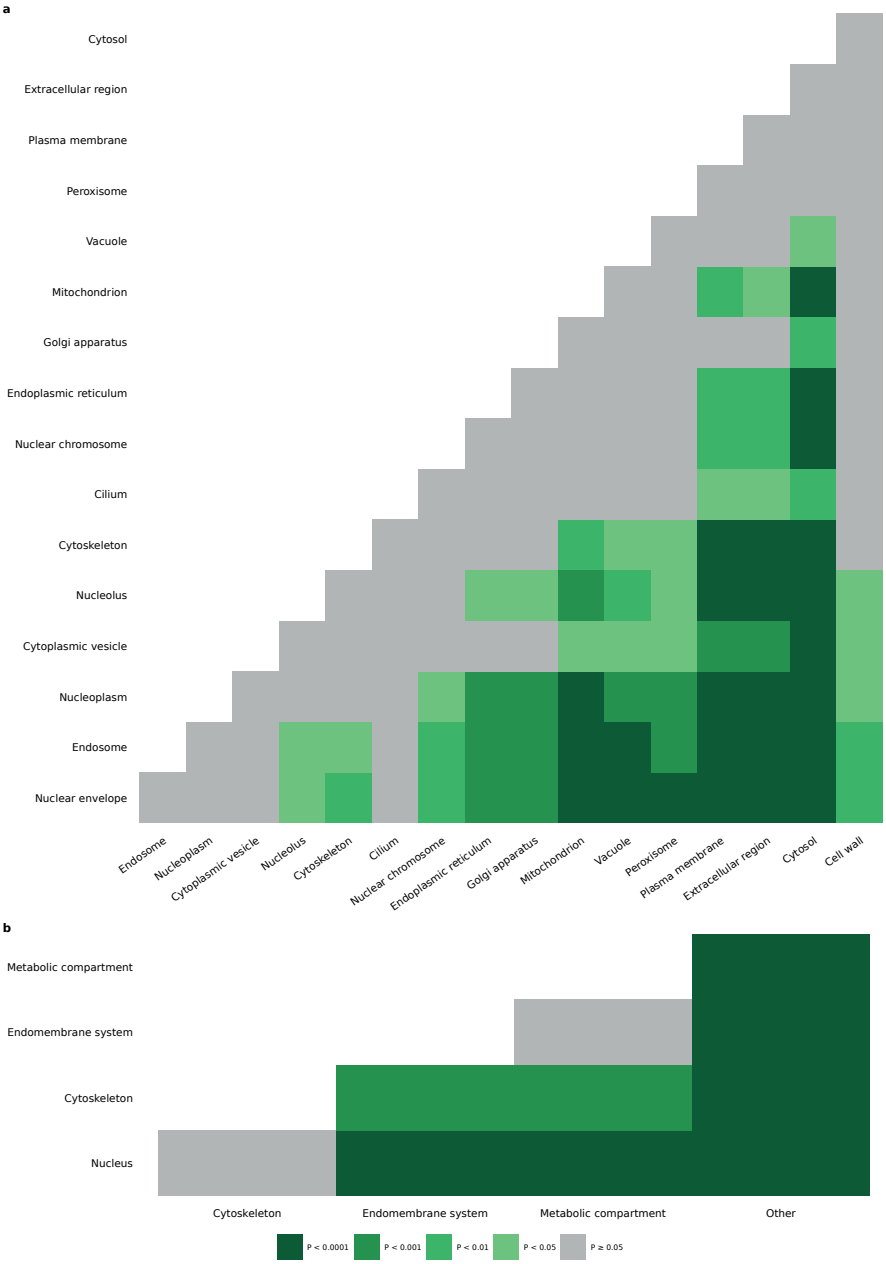
Supplementary Figure 2.1 | Contribution of duplications to families with a particular function. Statistical significance of pairwise comparisons (χ^2 contingency table tests) between the proportions of LECA families being derived from duplications for different functional categories (**a**) and the corresponding broad categories (**b**). The values for each functional category are shown in **Figure 2.1c**. The axis labels are ordered based on the odds of duplication.



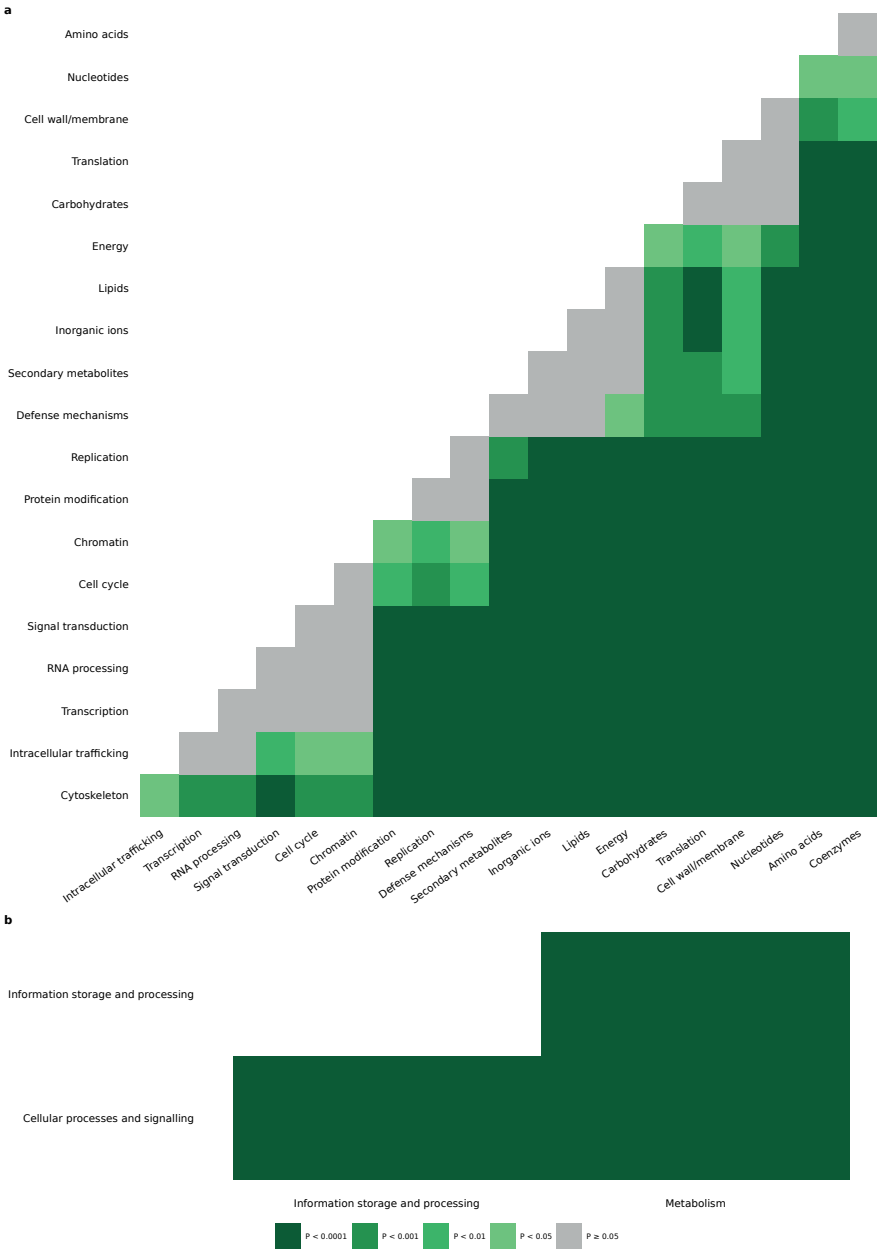
Supplementary Figure 2.2 | Contribution of duplications to families with a particular cellular localisation. Statistical significance of pairwise comparisons (χ^2 contingency table tests) between the proportions of LECA families being derived from duplications for different localisations (**a**) and the corresponding broad categories (**b**). The values for each localisation are shown in **Figure 2.1d**. The axis labels are ordered based on the odds of duplication.



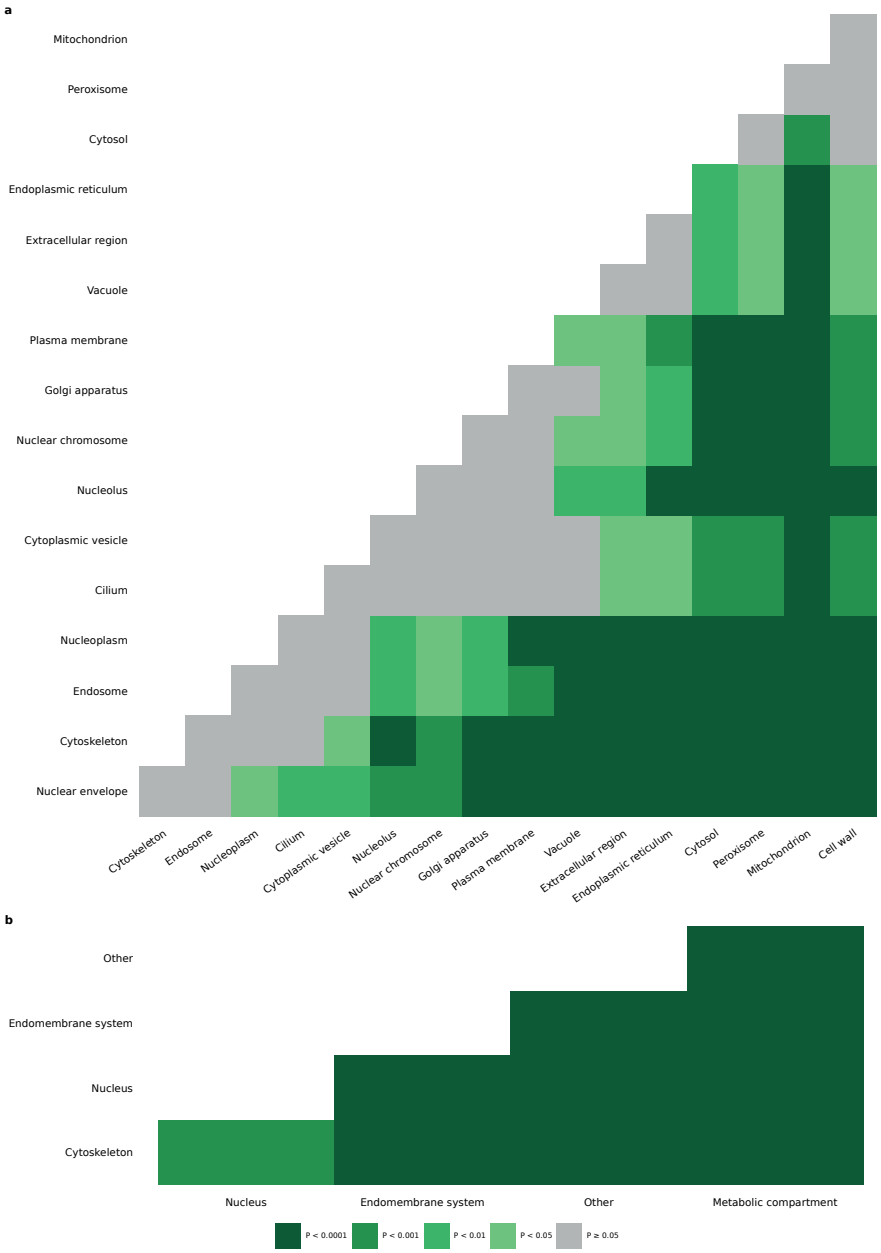
Supplementary Figure 2.3 | Contribution of inventions to families with a particular function. Statistical significance of pairwise comparisons (Fisher's exact tests) between the proportions of LECA families being derived from inventions for different functional categories (**a**) and the corresponding broad categories (**b**). The values for each functional category are shown in **Extended Data Figure 2.3a**. The axis labels are ordered based on the invented fraction.



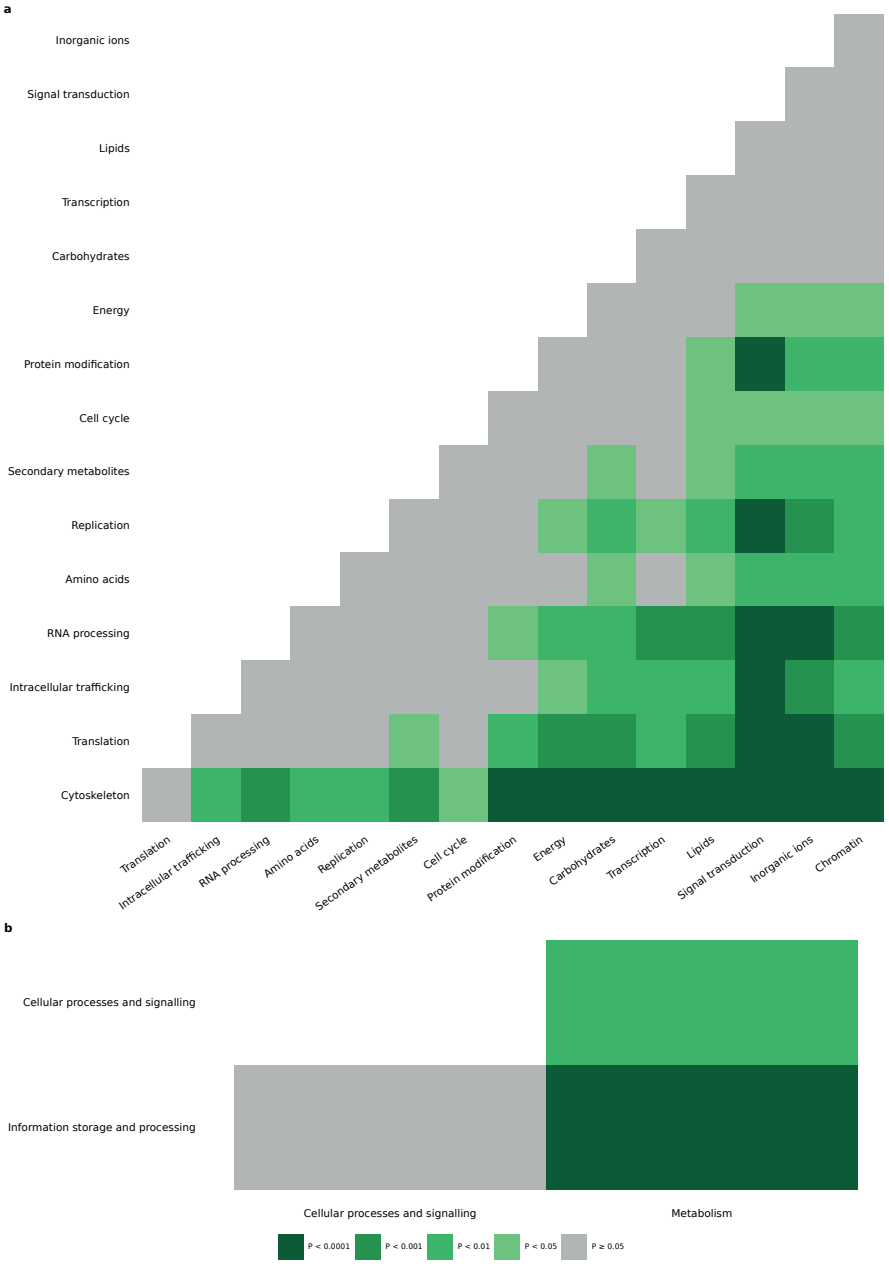
Supplementary Figure 2.4 | Contribution of inventions to families with a particular cellular localisation. Statistical significance of pairwise comparisons (χ^2 contingency table tests) between the proportions of LECA families being derived from inventions for different localisations (**a**) and the corresponding broad categories (**b**). The values for each localisation are shown in **Extended Data Figure 2.3c**. The axis labels are ordered based on the invented fraction.



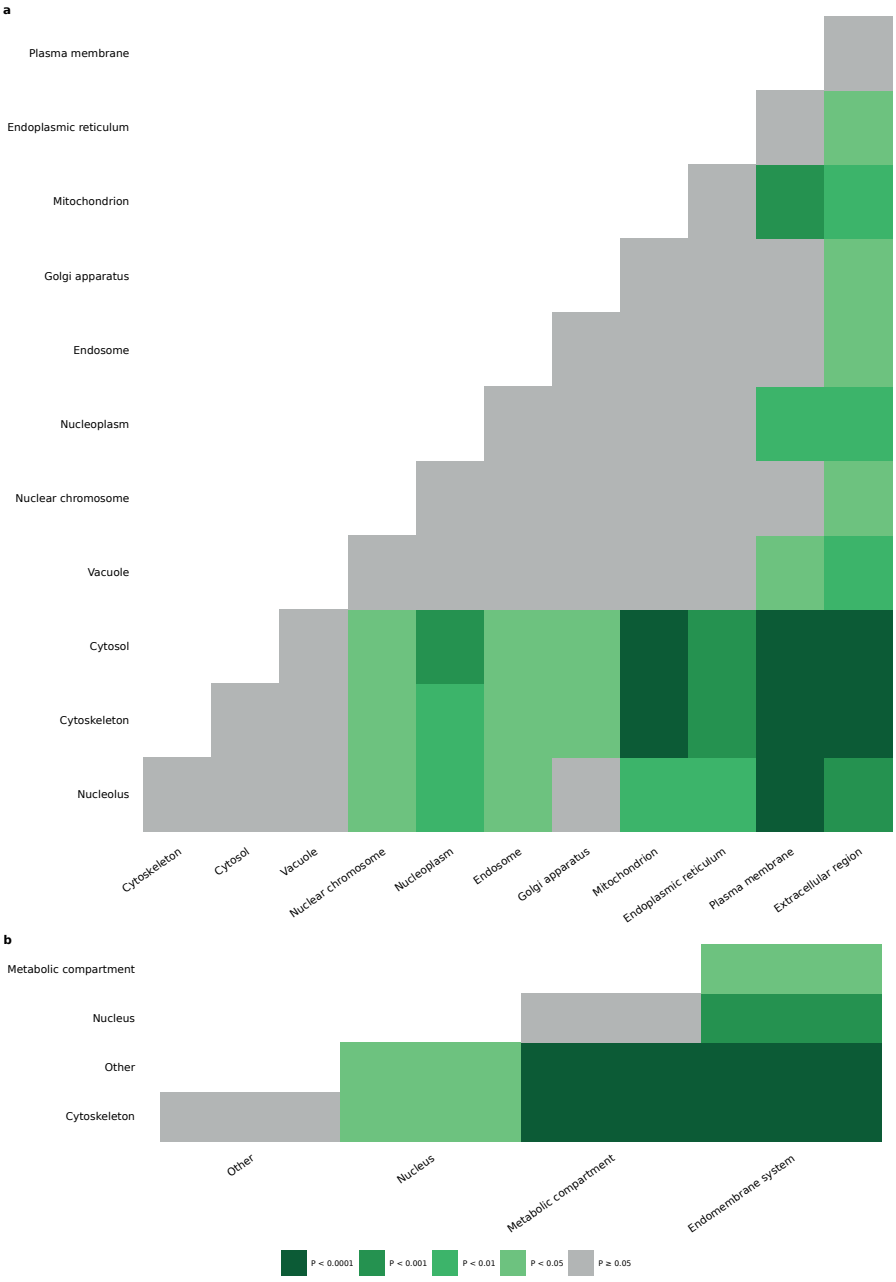
Supplementary Figure 2.5 | Contribution of innovations to families with a particular function. Statistical significance of pairwise comparisons (χ^2 contingency table tests) between the proportions of LECA families being derived from a eukaryotic innovation (invention or duplication) for different functions (**a**) and the corresponding broad categories (**b**). The values for each functional category are shown in **Extended Data Figure 2.3b**. The axis labels are ordered based on the innovated fraction.



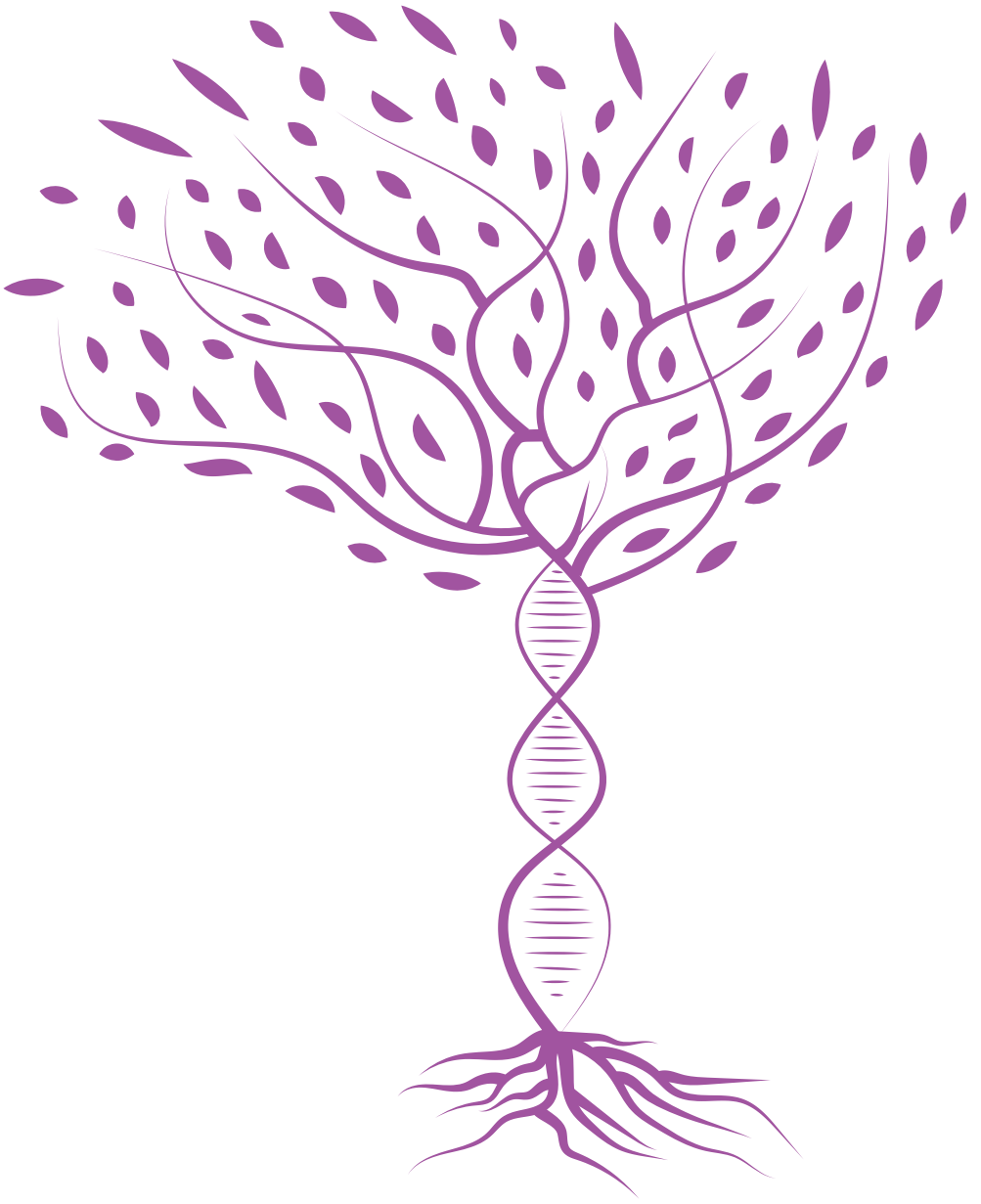
Supplementary Figure 2.6 | Contribution of innovations to families with a particular cellular localisation. Statistical significance of pairwise comparisons (Fisher's exact tests) between the proportions of LECA families being derived from a eukaryotic innovation (invention or duplication) for different localisations (**a**) and the corresponding broad categories (**b**). The values for each localisation are shown in **Extended Data Figure 2.3d**. The axis labels are ordered based on the innovated fraction.



Supplementary Figure 2.7 | Comparison of duplication lengths between different functions. Statistical significance of pairwise comparisons (Mann-Whitney *U* tests) between the duplication lengths for different functions (see **Figure 2.4a**) (**a**) and the corresponding broad categories (**b**). The axis labels are ordered based on the median of duplication lengths.



Supplementary Figure 2.8 | Comparison of duplication lengths between different cellular localisations. Statistical significance of pairwise comparisons (Mann-Whitney *U* tests) between duplication lengths for different localisations (see **Figure 2.4b**) (**a**) and the corresponding broad categories (**b**). The axis labels are ordered based on the median of duplication lengths.



3

The spread of the first introns in proto-eukarotic paralogs

Julian Vosseberg, Michelle Schinkel, Sjoerd Gremmen, Berend Snel

Communications Biology, 2022

Abstract

Spliceosomal introns are a unique feature of eukaryotic genes. Previous studies have established that many introns were present in the protein-coding genes of the last eukaryotic common ancestor (LECA). Intron positions shared between genes that duplicated before LECA could in principle provide insight into the emergence of the first introns. In this study we use ancestral intron position reconstructions in two large sets of duplicated families to systematically identify these ancient paralogous intron positions. We found that 20-35% of introns inferred to have been present in LECA were shared between paralogs. These shared introns, which likely preceded ancient duplications, were widespread across different functions, with the notable exception of nuclear transport. Since we observed a clear signal of pervasive intron loss prior to LECA, it is likely that substantially more introns were shared at the time of duplication than we can detect in LECA. The large extent of shared introns indicates an early origin of introns during eukaryogenesis and suggests an early origin of a nuclear structure, before most of the other complex eukaryotic features were established.

Introduction

Protein-coding genes in eukaryotic genomes are characterised by the presence of introns (Gilbert, 1978). Upon transcription, the introns are removed from the pre-mRNA by the spliceosome and the exons are spliced together to form mature mRNA, which is subsequently exported from the nucleus and translated into a functional protein. There are two types of introns; the vast majority of introns is of U2-type (Moyer et al., 2020), which are recognised and spliced out by the major spliceosome. U12-type introns are removed by the minor spliceosome and comprise less than a percent of introns in eukaryotic genomes (Moyer et al., 2020), with a recently discovered exception of 12% in *Physarium polycephalum* (Larue et al., 2021); in many species U12-type introns are completely absent (Bartschat and Samuelsson, 2010).

Ancestral reconstructions have revealed that the last eukaryotic common ancestor (LECA) had a genome with a relatively high intron density compared with present-day eukaryotes (Carmel et al., 2007; Csuros et al., 2011) and a complex major spliceosome with approximately 80 proteins (Collins and Penny, 2005). LECA also had U12-type introns and a minor spliceosome (Russell et al., 2006). Eukaryotic evolution after LECA predominantly involved the loss of introns, while only certain lineages including plants and animals experienced net intron gain (Csuros et al., 2011).

It has been established that spliceosomal introns originated from prokaryotic self-splicing group II introns during the prokaryote-to-eukaryote transition (Lambowitz and Belfort, 2015). These self-splicing introns can proliferate in the host genome but are rarely present within genes in prokaryotes. The most widely assumed scenario is that the self-splicing introns were introduced in the host genome from the protomitochondrion (Cavalier-Smith, 1991; Martin and Koonin, 2006) but we previously called other sources possible as well (Vosseberg and Snel, 2017). The emergence of intragenic introns underlined the importance of a nucleus – the defining feature of eukaryotes – to separate transcription and translation for splicing to take place completely prior to protein synthesis

(López-García and Moreira, 2006; Martin and Koonin, 2006). Furthermore, the origin of nonsense-mediated decay and the elaboration of ubiquitin signalling are proposed to be defence mechanisms against aberrant transcripts and proteins caused by the spread of introns (Koonin, 2006).

Eukaryotes are considered more complex than prokaryotes: cells are much larger and contain multiple membrane-bound compartments. Underlying the increase in cellular complexity during the transition to eukaryotes (eukaryogenesis) was an increase in the number of genes caused by gene transfers and gene duplications (Makarova et al., 2005; Tria et al., 2021; Vosseberg et al., 2021a). Mainly genes involved in establishing and regulating a complex cell and relatively few metabolic genes duplicated during eukaryogenesis (Vosseberg et al., 2021a).

Both the numerous gene duplications and the spread of introns through the genome occurred during eukaryogenesis and their interaction could inform the reconstruction of intermediate stages of this still largely unresolved transition. The relation between proto-eukaryotic gene duplications and introns can be researched by identifying positions of introns that are shared between ancient paralogs. An analysis performed on six eukaryotic genomes almost fifteen years ago identified very few shared intron positions that could represent intron insertions predating gene duplication events (Sverdlov et al., 2007). However, a study investigating the evolutionary history of a specific protein family, the spliceosomal Lsm and Sm proteins, found introns shared between multiple pre-LECA paralogs (Veretnik et al., 2009). This implies that introns had spread through the genome before the duplications resulting in these paralogs took place. It also suggests that more of these shared intron positions could be detected in other duplicated families.

In this study we utilise the greatly expanded set of eukaryotic genomes currently available to reassess the relation between the emergence of introns and gene duplications during eukaryogenesis. We detected many more shared intron positions than previous estimates. Our findings have implications for the dynamics of intron evolution and the timeline of events during eukaryogenesis.

Results

Intron-rich LECA and predominantly loss after

To investigate the interaction between the spread of introns and gene duplications during eukaryogenesis we used sets of proto-eukaryotic duplications inferred by two independent approaches: the Pfam domain trees from a recent study (Vosseberg et al., 2021a) (**Figure 3.1a**) and the clusters of eukaryotic orthologous groups (KOGs) that have been used before (Sverdlov et al., 2007) (**Figure 3.1b**). Intron positions were mapped onto protein alignments and ancestral intron reconstructions were performed using maximum likelihood for each KOG and Pfam domain orthogroup (OG). These reconstructions showed intron-rich ancestors of the eukaryotic supergroups (**Supplementary Figure 3.1**). We estimated an intron density in LECA of 10.8 introns per KOG and 1.9 introns per Pfam OG. Similar intron densities of LECA were obtained when using a tree topology with an unresolved root instead of a root between Opimoda and Diphoda (Derelle et al., 2015)

(9.9 and 1.7, respectively). Intron loss occurred frequently throughout eukaryotic evolution, with some lineages losing all introns in our set of genes. Intron gain only had a substantial contribution at certain branches, especially dinoflagellates, which has been described before (Roy et al., 2020). These findings for the dynamics of introns from LECA to present-day eukaryotes are fully consistent with a previous study that also reconstructed intron-rich ancestors (Csuros et al., 2011) and with the complexity of the spliceosome in LECA and subsequent simplification in most eukaryotic lineages, as inferred before (Hudson et al., 2015).

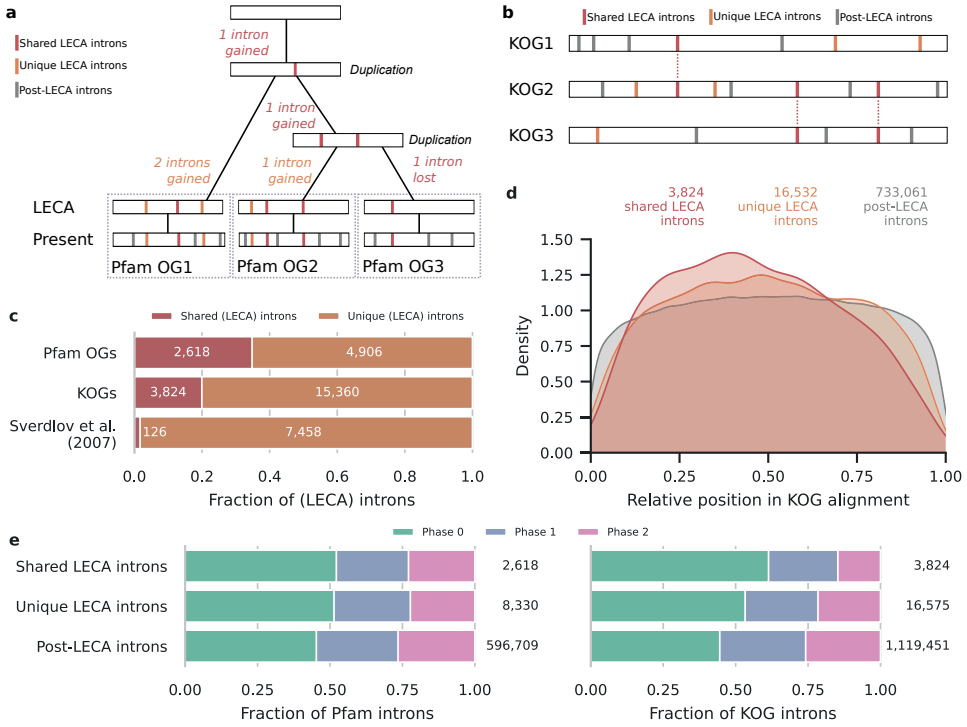


Figure 3.1 | Characteristics of unique and shared LECA introns. a, The reconstruction of introns in LECA (Pfam orthogroups (OGs)), distinguishing unique LECA introns and shared LECA introns that likely originated before a duplication. Pfam OG1-3 represent three paralogous Pfam OGs that resulted from two duplications during eukaryogenesis, as indicated. **b**, The comparisons of intron positions in KOGs within a cluster, identifying post-LECA introns, unique LECA introns and LECA introns shared between KOGs. KOG1-3 represent three paralogous KOGs in a cluster that resulted from two gene duplications during eukaryogenesis. **c**, Fraction of shared LECA introns in the two datasets used in this study, in comparison with the fraction of shared introns as calculated in Sverdlov *et al.* (Sverdlov et al., 2007). **d**, Density plot showing the relative positions of introns in the alignment of a KOG. The three distributions are all significantly distinct from one another according to Kolmogorov-Smirnov tests. **e**, Intron phase distributions in Pfam OGs and KOGs. All pairwise comparisons were significant, except shared LECA versus unique LECA Pfam introns (**Supplementary Tables 3.1 and 3.2**). Numbers in **c-e** indicate the number of introns considered.

Many intron positions in LECA shared between proto-eukaryotic paralogs

The relatively high number of introns that could be traced back to LECA underlined the potential to find LECA introns that are in the same position in OGs that stem from a proto-eukaryotic duplication, which we refer to as proto-eukaryotic paralogs. For the KOGs, 19.9% of the 19,184 LECA introns considered had an equivalent LECA intron in at least one paralog (**Figure 3.1c**). This is in sharp contrast to the 1.7% of shared introns found in Sverdlov *et al.* (Sverdlov et al., 2007), which is probably due to a combination of the low number of six available genomes that were used and the frequent loss of introns. The percentage of shared introns was even higher for the Pfams, with 34.8% of the 7,524 LECA introns in paralogous OGs being shared.

Intron positions shared between proto-eukaryotic paralogs could result from the intron being present prior to duplication and subsequently being passed on to both paralogs. It could also result from two parallel intron gains in the same position (**Supplementary Figure 3.2a**). A shared intron position between two homologous genes that were acquired as two separate genes during eukaryogenesis (e.g., cytoplasmic and mitochondrial ribosomal proteins (Yoshihama et al., 2006)) must have been the result of parallel intron gains. We compiled a set of separately acquired genes for both datasets and obtained percentages of 5.0% and 5.4% shared LECA introns between separate acquisitions for the KOGs and Pfams, respectively. Notwithstanding the influence of incorrect OG assignment inflating the estimated number of LECA introns shared between separate acquisitions (for example, nearly all sequences with introns shared between KOG0806 and KOG0807 correspond to one Pfam OG (NIT2)), this shows that parallel intron gain is a real phenomenon (**Supplementary Figure 3.2b**). However, the introns shared between proto-eukaryotic paralogs were very likely not only the result of parallel gains (Fisher's exact tests, $P = 5.8 \times 10^{-77}$ (KOGs), $P = 3.1 \times 10^{-220}$ (Pfams)). Another potential explanation for shared introns is transfer of introns between paralogs due to gene conversion (Yenerall and Zhou, 2012). However, this scenario cannot account for the frequent presence of multiple LECA introns in the same position between multiple paralogs (Supplementary Note 3.1). Instead, most of the shared introns probably represent paralogous introns, which hint at a strong association between duplications and intron spread and could elucidate the early spread of introns.

Intron loss was likely also pervasive before LECA

To characterise the detected LECA introns shared between proto-eukaryotic paralogs, we compared them with non-shared (which we refer to as unique) LECA introns and post-LECA introns with respect to the relative position of the introns in the gene. Whereas the relative positions of post-LECA introns showed a fairly uniform distribution, unique LECA introns were more at the 5′ end of the gene (compared with post-LECA introns, Kolmogorov-Smirnov (KS) statistic = 0.034, $P_{adj} = 4.1 \times 10^{-16}$) and shared LECA introns were even more biased towards the start of the gene (compared with unique LECA introns, KS statistic = 0.062, $P_{adj} = 5.6 \times 10^{-11}$; **Figure 3.1d**). This bias could reflect preferential intron insertion at the 5′ end specifically during eukaryogenesis or predominant intron loss at the 3′ end. A well-described mechanism of intron loss is by reverse

transcription of the intronless mRNA followed by homologous recombination (Roy and Gilbert, 2005). This mainly affects the 3′ end of the gene, resulting in intron losses from the 3′ to 5′ end. Intron losses before LECA is therefore likely to explain the 5′ bias of LECA introns.

We also compared the phases of the introns, which refers to the three possible positions of an intron in a codon. The phase distribution of the three different categories of introns was also different (**Figure 3.1e**; $\chi^2 = 966$, d.f. = 4, $P = 8.1 \times 10^{-208}$ (KOGs); $\chi^2 = 182$, d.f. = 4, $P = 2.4 \times 10^{-38}$ (Pfams); **Supplementary Tables 3.1 and 3.2**). LECA introns were more often in phase 0 and less in phase 2 than post-LECA introns. For the shared LECA introns in KOGs this bias was even stronger but in Pfams there was no significant difference between unique and shared LECA introns. The phase distribution differences point to different intron gain or loss dynamics also with respect to phase before and after LECA.

Published phylogenetic trees that were created for the Pfams dataset could in principle help to evaluate the prevalence of intron loss, which in turn might explain the phase and positional biases. We used the topology information in the trees to reconstruct for each duplication node the introns that were likely gained or lost before the duplication. In total, we inferred 999 intron gains and 986 losses before duplications and a further 4,906 gains and 1,214 losses on the branches that resulted in the LECA families. The phases of gained and lost introns differed slightly ($\chi^2 = 20.0$, d.f. = 2, $P = 4.6 \times 10^{-5}$; **Supplementary Figure 3.3**) but the typical phase bias was inferred for both. This strongly suggests that the phase bias originated from the preferential insertion or fixation of especially introns between codons (i.e., phase 0). For 38% of the duplications, we reconstructed introns being present prior to duplication. For an additional 5% we did not infer the presence of introns in those duplications but we had traced introns in more ancestral duplications. These reconstructions strengthen the inference of the dynamic nature of early intron evolution, including the pervasiveness of intron loss already before LECA.

The indications of considerable intron loss prior to LECA suggest that there were initially more pre-duplication introns that were lost and that can no longer be detected. This would mean that the numbers of introns stemming from duplications are underestimates. Moreover, gene families could have been experienced different intron gain and loss dynamics, which means that the absence of detected shared introns should not be seen as evidence that no introns were present prior to duplication.

Shared introns widespread across different functions

Eukaryogenesis was characterised by the complexification of multiple cellular processes and the presence of shared introns between paralogs of a certain function could illuminate how the duplications in that process relate to the spread of introns. Because information on the phylogenetic relationship between the KOGs was not available, we compared the LECA intron positions between KOGs of the same function in a cluster. Intron positions were shared between paralogs of most functions (**Figure 3.2**), mirroring the strong association of duplications and introns. However, appreciable differences between functions could be seen. A relatively large fraction of introns was shared between paralogs in cellular processes and signalling functions, compared with metabolic and informational

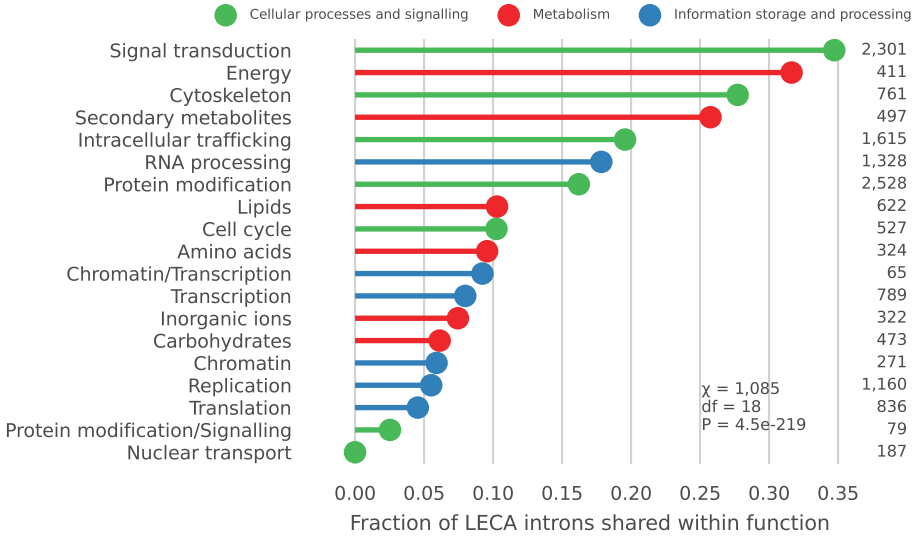


Figure 3.2 | Fraction of shared LECA introns between pairs of KOGs in a cluster with the same function. Sixty-nine percent of pairwise comparisons were significant, including all but one with nuclear transport (**Supplementary Data 3.3**). Numbers indicate the number of LECA introns. Only functions with at least ten LECA introns and ten pairs are shown.

functions ($\chi^2 = 340$, d.f. = 2, $P = 1.9 \times 10^{-74}$; **Supplementary Table 3.3**). The lack of any shared introns out of the 187 LECA introns between the eleven nuclear transport paralogs in the dataset is the most remarkable. The absence of shared nuclear transport introns seems to suggest that a large fraction of these duplications occurred prior to the spread of introns.

For Pfams, the aforementioned reconstruction of intron presence before duplications was used to compare duplications related to different functions. A similar pattern as for the KOGs was observed. Fewer introns preceding a Pfam duplication were inferred for informational and metabolic paralogs than paralogs in cellular processes and signalling functions ($\chi^2 = 46.7$, d.f. = 2, $P = 7.1 \times 10^{-11}$; **Figure 3.3a**; **Supplementary Table 3.4**). The large fraction of duplications in energy metabolism with shared introns is almost exclusively due to mitochondrial carrier proteins. Differences in cellular localisation were more subtle ($\chi^2 = 11.0$, d.f. = 4, $P = 0.026$; **Figure 3.3b**; **Supplementary Table 3.5**), except for the high fraction of endosome duplications and the absence of shared introns in cilium duplications. Because nuclear transport is a combination of two functions (nuclear structure and intracellular trafficking), this category is absent in the Pfams set. The ancestral intron reconstructions in the adaptin (**Figure 3.3c**) and SNF2 families (**Figure 3.3d**) illustrate the extent of shared introns, the presence of introns prior to the most ancestral duplication and frequent intron losses in duplicated families. The large extent of shared introns across different functions and localisations implies that introns were present in the genome before much of the complexification of the signalling system, cytoskeleton

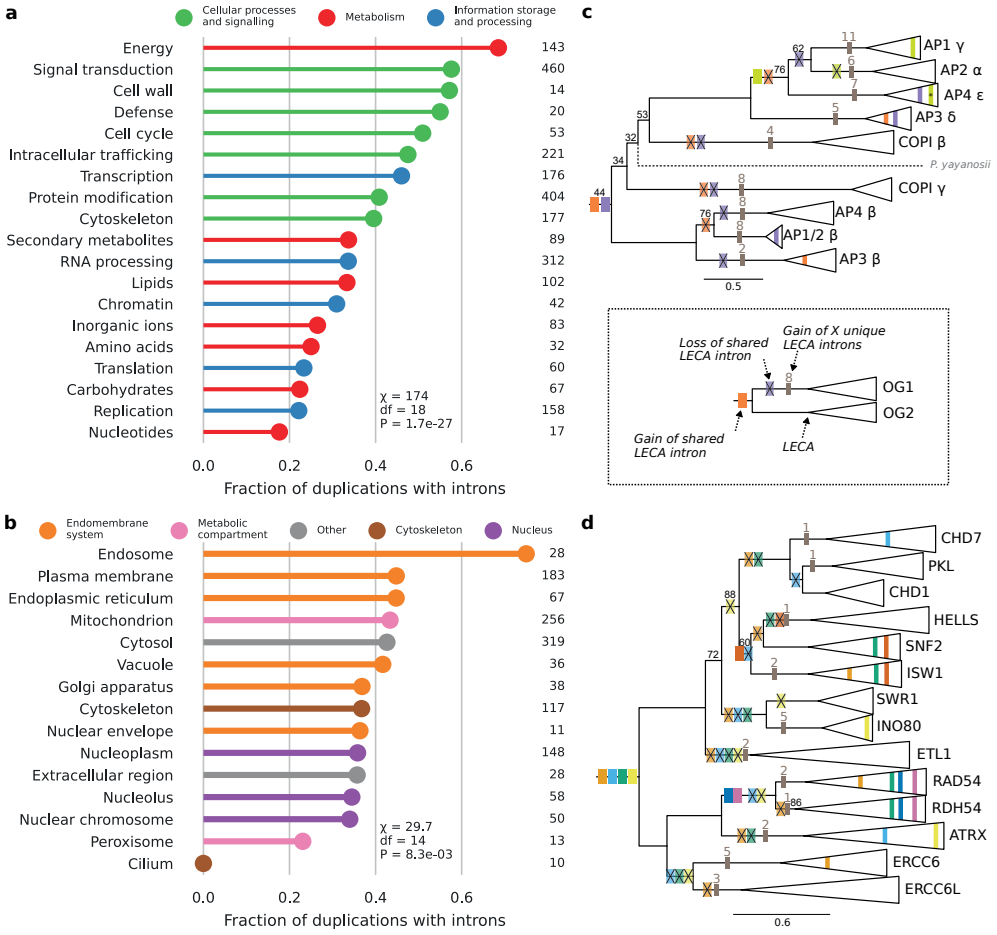


Figure 3.3 | Reconstruction of pre-duplication introns in Pfam duplications of different functions and cellular localisations. a, **b**, Fraction of duplications with introns traced to their pre-duplication state according to functional category (**a**) and cellular localisation (**b**). Thirty percent of pairwise comparisons of functions were significant (**Supplementary Data 3.4**). Fourteen percent of pairwise comparisons of localisations were significant, which were only comparisons including the endosome and cilium (**Supplementary Data 3.5**). Numbers indicate the number of duplications. Only functions and localisations with at least ten duplications are shown. **c**, **d**, Excerpts from the gene trees of the adaptin (PF01602) (**c**) and SNF2 family (PF00176) (**d**) with the reconstructed presence of introns depicted. The triangles and names correspond to the Pfam OGs. The shared LECA introns in a Pfam OG are coloured and the gain and loss of these introns is mapped onto the phylogeny. The number of unique LECA introns is indicated in grey. Ultrafast bootstrap support values lower than 100 are shown. The branch with a prokaryotic sequence that fell between the Pfam OGs in (**c**) is shown as a dotted line. The shared intron in AP4 ε that is marked with an asterisk was classified as a U12-type intron. Although the phylogenetic position of the two COPI subunits is probably incorrect, the inferred intron gains and losses in these trees are largely unaffected by topology changes.

and endomembrane system and before the full integration of the protomitochondrion into the host.

Few shared introns between mitochondria-derived paralogs

During eukaryogenesis, genes were acquired from different prokaryotic sources or arose *de novo* (i.e., a gene invention). Pfam clades inferred to have been a proto-eukaryotic invention or with an Asgard archaeal or diverse prokaryotic sister group in the phylogenetic tree had the highest fraction of duplications with introns (χ^2 = 53.6, d.f. = 5, P = 2.5 × 10^{-10}; **Figure 3.4a**; **Supplementary Table 3.6**). Duplications in Pfam domains that had an alphaproteobacterial sister group were noticeably less likely to have reconstructed introns. When looking only at the first duplication in an acquisition or invention (i.e., the most ancestral one) a similar though not significant pattern was observed (χ^2 = 9.2, d.f. = 5, P = 0.10; **Figure 3.4b**). A substantial fraction of the Pfams that were very likely inherited from the Asgard archaea-related host (0.21, [95% CI using the Wilson score interval: 0.13-0.32]) had introns traced back prior to the most ancestral duplication.

The lack of shared introns in certain gene families could be because these duplications had occurred at an early stage during eukaryogenesis when there were no or few introns or could be due to factors such as extensive domain accretion and loss, a low number of LECA introns and intron loss (**Supplementary Figure 3.4**). Due to the inferred pervasive intron loss, it is also more likely to detect shared introns in case of multiple paralogs. By comparing differences between functions and phylogenetic origin in the fraction of duplications with shared introns, the fraction of introns that are shared and the number of introns per Pfam OG, the contribution of these factors can be elucidated (**Supplementary Figure 3.5**, **Supplementary Tables 3.7-3.9**, **Supplementary Data 3.6 and 3.7**). For exam-

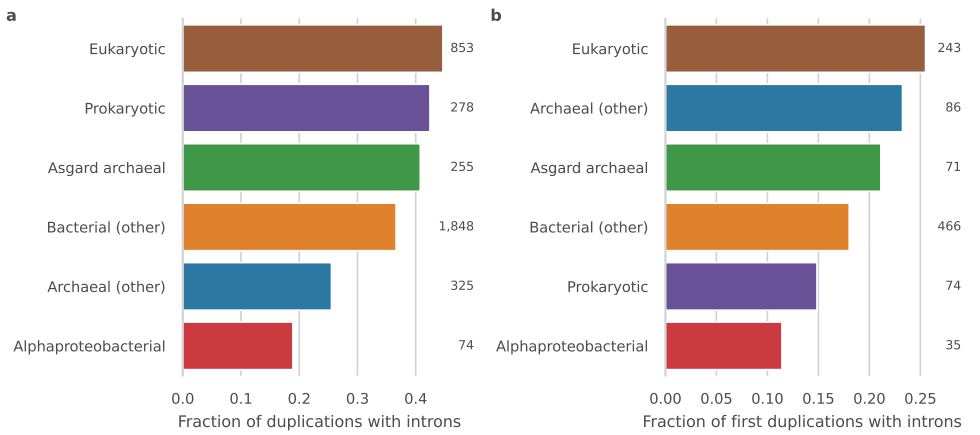


Figure 3.4 | Reconstruction of pre-duplication introns in Pfam duplications of different phylogenetic origins. a, Fraction of duplications with introns traced to their pre-duplication state for different phylogenetic origins. Sixty percent of pairwise comparisons were significant, including all but one comparison with alphaproteobacterial duplications (**Supplementary Table 3.6**). **b**, Fraction of the most ancestral duplication in an acquisition or invention with introns traced to their pre-duplication state. Differences between groups were not significant. Numbers indicate the number of duplications.

ple, fewer LECA introns were present in alphaproteobacteria-related OGs, which would make the detection of shared introns less likely and could explain the low fraction of alphaproteobacteria-related duplications with introns. Despite the relatively large influence of a few clades on differences between groups (Supplementary Note 3.2), an appreciable number of shared introns was detected for most functions, localisations and phylogenetic origins.

In a previous study, we estimated the timing of duplication events with branch lengths (Vosseberg et al., 2021a). Based on this branch length analysis, duplications without shared introns were in general older than those with shared introns (KS statistic = 0.064, $P = 0.0016$; **Supplementary Figure 3.6a**). However, the distributions overlap to a very large extent and a considerable number of duplications with introns were relatively old. Almost one-fourth of duplications with introns were estimated to be older than the mitochondrial acquisition. Notwithstanding the uncertainties and limitations of these branch lengths analyses (Susko et al., 2021), introns seemed to have been present in the proto-eukaryotic genome from an early stage in eukaryogenesis.

Discussion

In this study, we investigated the intersection of the emergence of introns and gene duplications during eukaryogenesis. We detected a 12-fold higher fraction of shared intron positions between proto-eukaryotic paralogs in the KOGs dataset than Sverdlov *et al.* (Sverdlov et al., 2007) and an even higher fraction in a second independent dataset. The numbers of shared introns were no longer in the range of what is expected from parallel intron insertions, which means that that the vast majority of these shared introns were very likely obtained before the duplication. Because our observations hint at a pattern of pervasive intron loss during eukaryogenesis, the number of introns stemming from duplications that we present are still underestimates.

The higher fraction of shared introns in the Pfams dataset could result from KOGs being undersplit compared with Pfam OGs, which means that multiple *bona fide* OGs are combined into one OG. This is illustrated by the SNF2 and adaptin examples, for which several Pfam OGs correspond to a single KOG (e.g., AP1G and AP4E in KOG1062, and ERCC6 and ERCC6L in KOG0387). An additional explanation could be that Pfam domains are more conserved. Consequently, it would be more likely for an intron to be paired with an intron in a paralog. Notwithstanding the subtle differences between the two datasets, both revealed consistent findings.

Introns in nearly all species have a phase bias, with most introns in phase 0 and fewest in phase 2. This bias was also present among LECA introns and the bias tended to be even stronger in shared LECA introns. The bias could be due to the preferential insertion or fixation of introns in a certain phase resulting from the overrepresentation of proto-splice sites (Dibb and Newman, 1989) in a certain phase. A biased loss could also explain the phase bias (Long and Deutsch, 1999). Another plausible explanation that has been put forward is that the initial distribution was uniform and that the eventual phase bias was due to a combination of massive U12-type intron loss and the directed conversion of phase 0 U12-type to U2-type introns (Moyer et al., 2020). The phase bias of recent U2-type intron gains in the dinoflagellate lineage and recent U12-type intron gains in

Physarium lends support to the protosplice site model in at least these eukaryotic lineages (Larue et al., 2021; Roy et al., 2020). The similar phase bias of pre-LECA gains and losses (**Supplementary Figure 3.3**) points to differences already during intron gain, which are then reflected in phase differences during intron loss. Furthermore, the relatively low number of inferred U12-type intron losses and conversions (Supplementary Note 3.3; **Supplementary Figures 3.7 and 3.8**, **Supplementary Tables 3.10-3.12**, **Supplementary Data 3.8 and 3.9**) tends to refute a major role of U12-type introns in the phase distribution of all introns. Both observations provide support for the protosplice site model during eukaryogenesis as well.

Our data on the strong association between proto-eukaryotic duplications and introns as compiled here have several implications for the order of events during eukaryogenesis. The large extent of shared introns between ancient paralogs across different functions, subcellular localisations and phylogenetic origins as well as the branch lengths provide consistent evidence for an early origin of intragenic introns during eukaryogenesis, before most of the complex eukaryotic features emerged. In fact, it seems unambiguous that numerous gene families expanded primarily after the spread of introns through the proto-eukaryotic genome (e.g., SNF2 and adaptin). Other families seem conspicuously devoid of shared introns (e.g., those involved in nuclear transport and the cilium). The conservation of introns upon duplication challenges the main role of retrotransposition in creating proto-eukaryotic paralogs, which results in intronless paralogs and was proposed based on the initial lack of detected shared introns (Sverdlov et al., 2007). An early origin of introns should have entailed an early origin of a structure to separate transcription and translation, preventing the erroneous translation of introns into protein. The recent observation of spatial separation between DNA and ribosomes in Asgard archaeal cells tentatively suggests that a separating mechanism may have already been present before eukaryogenesis (Avcı et al., 2022). The lack of introns shared between nuclear transport paralogs seems to indicate that the emergence of a nucleus with an elaborate nuclear transport system occurred before the wide spread of introns.

A notable exception to the described pattern of shared introns between most categories is the low number of duplications with shared introns in alphaproteobacterial acquisitions, which were very likely present in the protomitochondrion. It is tempting to speculate that these duplications were due to another mechanism; for example, they may have been the result of serial endosymbiotic gene transfers (Tria et al., 2021). The protomitochondrion has been widely considered to be the source of introns, even though direct phylogenetic evidence is lacking (Vosseberg and Snel, 2017). Based on the analysis of shared introns, the close integration of the endosymbiont within the host by means of mitochondrial transport seemed to have occurred after substantial spread of introns. Although the symbiosis must have started before the close integration, this observation combined with the inferred timing from our branch lengths analysis is not easy to reconcile with the hypothesis that spliceosomal introns originated from mitochondrial self-splicing group II introns. The self-splicing introns could have come from another lineage instead.

Our analysis was to the best of our knowledge the second large-scale investigation on the association between introns and proto-eukaryotic duplications, yet it was the first to

encounter a large-scale occurrence of introns shared between proto-eukaryotic paralogs. Besides the potential implications on the order and causality of events during eukaryogenesis, the strong association between proto-eukaryotic duplications and introns also sheds unique light on the origin and evolution of intron phases and positional biases as well as the discussion on the emergence of U2- and U12-type introns (Supplementary Note 3.3). Thus, going forward we expect that further utilisation and understanding of these intertwined processes could be of great help to understand the evolutionary history of individual gene families as well as eukaryogenesis.

Methods

Data

To reconstruct ancestral intron positions we used a diverse set of 167 eukaryotic (predicted) proteomes, as compiled for a previous study (Deutekom et al., 2021). In that study, these proteins had been assigned to the different eukaryotic eggNOG families (euNOGs) (Huerta-Cepas et al., 2016a) using hidden Markov model profile searches (Deutekom et al., 2021). Sverdlov *et al.* (Sverdlov et al., 2007) used the homologous clusters of eukaryotic orthologous groups (KOGs) and candidate orthologous groups (TWOGs) from Makarova *et al.* (Makarova et al., 2005). KOGs are included in the euNOGs and we used the euNOG corresponding to a TWOG, if present, as determined in Vosseberg *et al.* (Vosseberg et al., 2021a). Both types of euNOGs are referred to as "KOG" in the main text. We detected a few differences between the Makarova *et al.* and Sverdlov *et al.* clusters and chose one clustering over the other on a case-by-case basis after manual inspection (**Supplementary Data 3.1**). The sequences corresponding to these clusters of KOGs were selected and combined per KOG.

We also used the Pfam LECA families and duplications that we published recently (Vosseberg et al., 2020). In short, we selected eukaryotic sequences based on best bidirectional hits between Opimoda and Diphoda for tree inference and supplemented these with prokaryotic sequences. In the resulting phylogenetic trees, acquisition, duplication and LECA nodes were inferred. The tree sequences belonging to a LECA node were complemented with the eukaryotic sequences that had one of these tree sequences as their best BLAST (Altschul et al., 1990) hit, resulting in an OG. Sequences from species that are not in the set of 167 species and human sequences that are not in the primary assembly were removed from the OG. If there was only one OG for a Pfam, it was not included.

For predicting the type of introns, genome sequence files of the species in our set were obtained using the links in Supplementary Table 1 of Deutekom *et al.* (Deutekom et al., 2019), with the exception of *Homo sapiens*, whose genome was replaced with the corresponding primary assembly (ftp://ftp.ensembl.org/pub/release-87/fasta/homo_sapiens/dna/Homo_sapiens.GRCh38.dna.primary_assembly.fa.gz), and *Stentor coereleus*, for which we used the file from NCBI (ftp://ftp.ncbi.nlm.nih.gov/genomes/all/GCA/001/970/955/GCA_001970955.1_ASM197095v1/GCA_001970955.1_ASM197095v1_genomic.fna.gz) to be able to match the sequence file with the genome features file.

Multiple sequence alignments

All multiple sequence alignments were performed with MAFFT v7.310 (Katoh and Standley, 2013). Each KOG was aligned separately using the E-INS-i algorithm and the resulting KOG alignments for a cluster of KOGs were merged into a single alignment (merge option with E-INS-i). If this was not feasible due to memory issues, the alignment was made with the FFT-NS-i option, or FFT-NS-2 if that was also not feasible. Each Pfam OG was aligned separately, followed by a merged alignment of all OGs per Pfam, both with the L-INS-i algorithm. For PF00001 and PF00069, alignments had to be performed with the FFT-NS-i option.

Mapping intron positions onto the alignments

We downloaded the genome annotation files from 156 species from our set that we could extract intron information from (**Supplementary Data 3.2**). The location of introns was mapped onto the protein alignments using a custom Python script. For each intron position detected in the alignment of an OG, taking into account the three different possible phases, it was determined if at least one sequence of each species had an intron at that position. An intron table was created with per species a string of intron presences ("1") and absences ("0") and a mapping to the position in the alignment of an OG. If an ortholog was missing or intron mapping was not successful, question marks were inserted.

To calculate the relative position of an intron in a gene, sites with 90% or more gaps in the alignment of a KOG were masked. These gap scores were calculated with trimAl v1.4.rev15 (Capella-Gutiérrez et al., 2009).

Intron gain and loss rates across the eukaryotic tree of life

For each branch in the species phylogeny, maximum-likelihood estimates of intron gain and loss rates were obtained using Malin (Csűrös, 2008) with default settings. The used species phylogeny can be found in Supplementary Figure 1 of Deutekom *et al.* (Deutekom et al., 2021). Because the position of the eukaryotic root remains under debate (Burki et al., 2020), we also used a tree with an unresolved root between Diaphoretickes, Amorphea, Metamonada and Discoba.

Ancestral intron reconstructions

The number of introns per ancestral node including missing sites were estimated in Malin and per intron position the probability of the intron being present at a node and gained or lost on the branch leading to a node was inferred. The distribution of posterior intron presence probabilities at the LECA node showed a clear divide between most introns with a very low and a small fraction with a very high LECA probability (**Supplementary Figure 3.1c**, **d**). For choosing an appropriate threshold to consider an intron a LECA intron, we tried to minimise the effect of misalignment of residues and incorrect OG assignment on the one hand and to be not too strict on the other hand. This was because an intron with a lower LECA probability that is shared with a paralog, makes it more likely that the intron was in fact present in LECA. Therefore, intron positions with a probability of at least 0.5 were considered LECA introns.

The different KOGs in the KOG clusters do not all represent gene duplications during eukaryogenesis; some were acquired as separate genes (Makarova et al., 2005). This could be due to separate acquisition events (pseudoparalogs) or the acquisition of already duplicated genes. Shared intron positions in these had to be the result of parallel intron insertions. To identify these, we used the phylogenetic trees of these clusters that we inferred before (Vosseberg et al., 2020). If the sequences corresponding to different KOGs were in separate acquisitions and none of the acquisitions in the tree had another acquisition in the inferred sister group, the intron positions in these separately acquired KOGs were compared and shared introns were identified. Introns that that were only shared between separate acquisitions were not included in the shared introns analysis. All introns of KOGs that were acquired separately from all other KOGs in the cluster were not used for calculating the fraction of shared introns.

For the Pfams, separate acquisitions were identified based on phylogenetic trees as well using the same approach. For each duplication node in the trees, the intron positions that were present before the duplication event were inferred from the LECA introns using a Dollo parsimony approach. The inferred sister groups of acquisitions and the functional annotation and duplication length information were extracted from Vosseberg *et al.* (Vosseberg et al., 2020).

U12-type intron predictions

Spliceosomal snRNA genes were searched for in the genomes using Infernal v1.1.2 (Nawrocki and Eddy, 2013) (command used: cmscan --nohmmonly --rfam --cut_ga) with the spliceosomal snRNA Rfam 14.2 (Kalvari et al., 2021) covariance models RF00003, RF00004, RF00007, RF00015, RF00020, RF00026, RF00488, RF00548, RF00618, RF00619, RF02491, RF02492, RF02493 and RF02494. Introns from species for which none of the snRNA genes specific for the minor spliceosome (U11, U12, U4atac or U6atac) were detected in the genome were annotated as U2. Intron types from the remaining 70 species were predicted with intronIC v1.0.11 + 2.gf7ac7be (Moyer et al., 2020), using all isoforms if needed. Intron positions that were predicted as U12 in at least three species were annotated as U12-type introns.

Statistics and reproducibility

Associations between two categorical variables were tested with χ^2 contingency table tests or Fisher's exact tests (in case of 2×2 tables). When testing the overrepresentation of functional categories, KOGs with multiple categories spanning the three main groups (information storage and processing, cellular processes and signalling, metabolism) were excluded for comparisons between these groups. The numbers of unique LECA introns and shared LECA introns or duplications with and without introns traced back to the pre-duplication state were compared between different functions, cellular localisations and phylogenetic origins. Differences in relative position and branch lengths were assessed with Kolmogorov-Smirnov tests. All performed tests were two-sided and *P* values from multiple comparisons were adjusted for the false discovery rate. Statistical analyses were performed in Python using NumPy v1.21.1 (Harris et al., 2020), pandas v1.3.1

(McKinney, 2010), SciPy v1.7.0 (Virtanen et al., 2020) and statmodels v0.11.2 (Seabold and Perktold, 2010). Figures were created with Matplotlib v3.4.2 (Hunter, 2007), seaborn v0.11.1 (Waskom, 2021), ETE v3.1.1 (Huerta-Cepas et al., 2016b) and Jalview v2.11.1.4 (Waterhouse et al., 2009).

Data availability

The data underlying this article are available in figshare, at https://doi.org/10.6084/m9.figshare.16601744 (Vosseberg et al., 2021b). The accession information for the public datasets used in this study is presented in **Supplementary Data 3.2**. The source data behind the graphs in the paper are provided as **Supplementary Data 3.10**.

Code availability

The code used to map the intron positions onto the alignments and create the intron tables is available on Github (https://github.com/JulianVosseberg/imapper) and figshare, at https://doi.org/10.6084/m9.figshare.19411820.v1 (Vosseberg et al., 2022a).

Acknowledgements

We thank Eva Deutekom for providing the eggNOG annotations and the members of the Theoretical Biology & Bioinformatics group for useful discussions. This work is part of the research programme 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 M.S. performed the research. J.V. and B.S. analysed and interpreted the results. J.V., S.G. and M.S. developed the intron mapping pipeline. J.V. wrote the manuscript, which was edited and approved by all authors.

Supplementary Notes

Supplementary Note 3.1: Potential intron transfer between paralogs

A third explanation for the presence of a shared intron between paralogs – next to vertical inheritance from a pre-duplication intron and parallel insertion in the same position – is the transfer of an intron from the intron-containing paralog to the other intron-lacking paralog. This transfer is proposed to occur via homologous recombination, resulting in ectopic gene conversion (Hankeln et al., 1997; Yenerall and Zhou, 2012). Three likely cases have been described in literature: a transfer of an intron between three globin paralogs in the insect *Chironomus* (Hankeln et al., 1997), between two ABC transporter paralogs in the ascomycete *Aspergillus* (Zhang et al., 2010) and between two metalloprotease paralogs in the ascomycete *Mycosphaerella* (Torriani et al., 2011). Although these examples demonstrate that intron transfer between paralogs is a plausible mechanism, its relative contribution to shared introns between paralogs remains elusive.

To assess the possible impact of intron transfers between proto-eukaryotic paralogs, we looked more closely at the KOG clusters for which only one LECA intron position was shared between only two paralogs. The rationale for these criteria was that we think it is unlikely for multiple introns to be transferred or for an intron to be transferred between more than two paralogs. 148 LECA introns fulfilled these criteria for the KOGs, corresponding to 3.9% of the total number of shared LECA introns. For the Pfam OGs, 294 LECA introns were intron transfer candidates (11% of the total number of shared LECA introns). Given that intron transfers between paralogs and parallel insertions could have accounted for only a small number of shared LECA introns, we infer that most shared introns likely represent paralogous introns.

Supplementary Note 3.2: Differences in the detection of shared introns between different groups

For the Pfam OGs we analysed both the fraction of duplications with shared LECA introns and the fraction of LECA introns shared with paralogs. Differences can arise due to the phylogenetic relationships between OGs, which is ignored in calculating the fraction of LECA introns, and the large impact of a few clades on both numbers. For example, two intron-rich OGs with many shared introns can dominate the fraction of shared LECA introns of a category, while only reflecting a single duplication. Multiple paralogs that all share introns (i.e., a large fraction of duplications with pre-duplication introns) yet also have many OG-specific introns results in a low fraction of shared introns. When comparing both approaches, most differences are quite subtle.

For function (**Figure 3.2a** and **Supplementary Figure 3.5a**) two categories were notably different. Energy metabolism had a lower (but still relatively high) fraction of shared LECA introns in comparison with the fraction of duplications with introns, for which it was the top category. The fraction of duplications with introns was dominated by mitochondrial carrier proteins (PF00153), which had shared introns traced to its first duplication and many later duplications. Conversely, duplications in amino acid metabolic genes had a higher fraction of shared LECA introns. This number was largely influenced

by aminotransferases class I and II (PF00155) and serine hydroxymethyltransferase (PF00464) with 6 and 4 shared introns, respectively. Most of the OGs with this function in other Pfams had few, if any, LECA introns (**Supplementary Figure 3.5b**).

Whereas differences in duplication fractions between most cellular localisations were rather subtle, shared intron fractions revealed a more variable picture (**Supplementary Figure 3.5c**). For the extracellular region most LECA introns were shared, in contrast with a relatively low fraction of duplications with introns. Ten duplications in leucine-rich repeat 8 (PF13855), which all shared the same three LECA introns, contributed to a large extent to this number, especially since OGs with this localisation had fewer introns in general (**Supplementary Figure 3.5d**). Introns could be traced back to the majority of duplications related to the endosome, whereas a lower fraction of LECA introns were shared between endosomal paralogs. Most of these duplications were in the PX domain (PF00787) and FYVE zinc finger domain (PF01363), with relatively few shared introns. Paralogs that function in the nuclear envelope did not share any LECA introns, which is in sharp contrast with 36% of nuclear envelope duplications sharing introns. Although these introns were shared between nuclear envelope and another (cytosol) and unknown localisation, the corresponding duplications in the RanBP1 (PF00638) and importin-beta N-terminal domain (PF03810) had been annotated as duplications in nuclear envelope genes.

The most notable difference for the separate phylogenetic origins was the higher fraction of shared introns for alphaproteobacteria-related paralogs and the lower fraction of shared introns for Asgard archaea-related paralogs (**Supplementary Figure 3.5e**). OGs from probable endosymbiont origin had few introns in general, whereas OGs that were probably inherited from the host had more introns (**Supplementary Figure 3.5f**). The low number of LECA introns in alphaproteobacteria-related OGs could account for the low fraction of duplications with introns in alphaproteobacterial acquisitions.

Supplementary Note 3.3: The emergence of two different intron types

Two types of spliceosomal introns emerged during eukaryogenesis: U2 and U12. Three different models for the appearance of two types have been proposed (Burge et al., 1998). The first is the codivergence model, which postulates that the snRNA genes and introns diverged into two different sets after duplication of the snRNA genes. According to the fission/fusion model the two intron types evolved in separate proto-eukaryotic lineages that later fused. Whereas in the first two models the two types originated from primordial spliceosomal introns, in the parasitic invasion model the two types represent two temporally separate invasions of self-splicing group II introns.

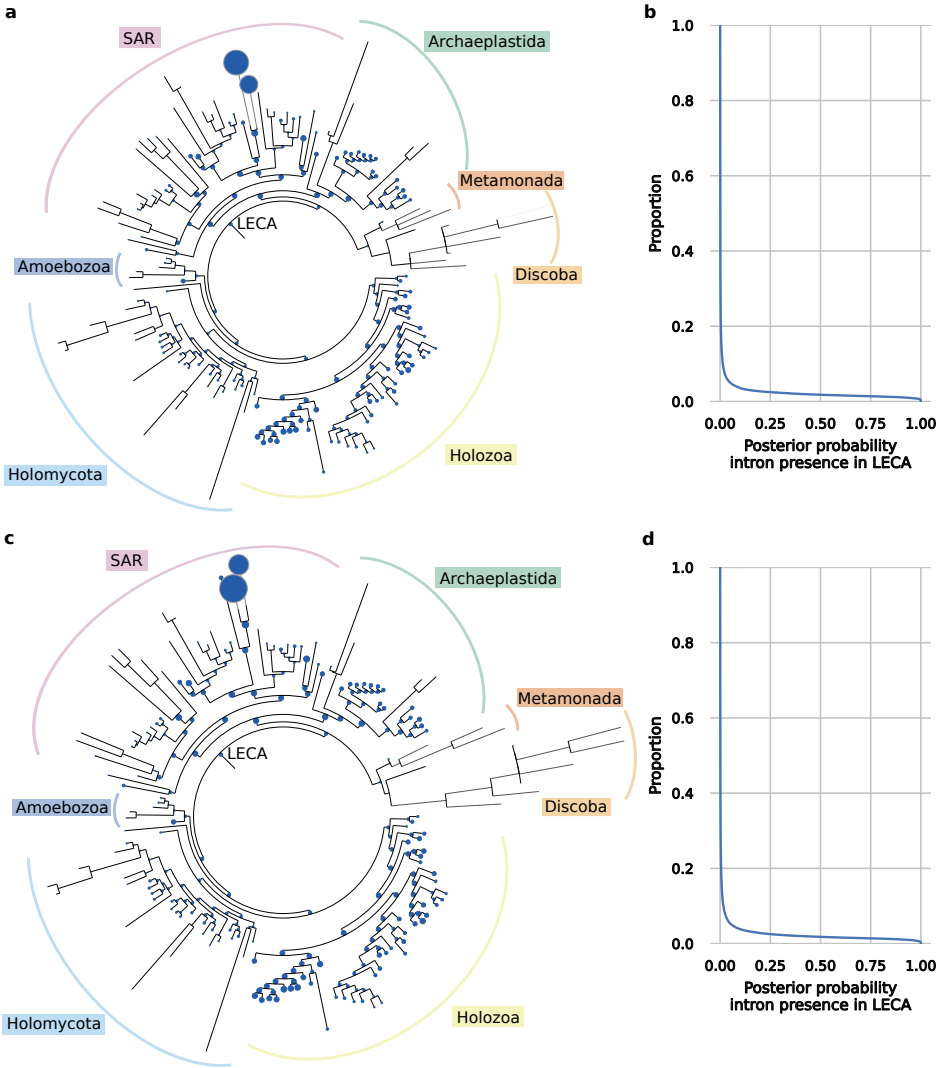
To investigate the origin of the two different types of introns during eukaryogenesis, we predicted the type of all introns. 1.5% (KOGs) and 1.9% (Pfams) of LECA introns were predicted to be of U12-type, which was higher than in any present-day eukaryote in our dataset. 14.8% (KOGs) and 31.1% (Pfams) of these U12-type LECA introns were shared with an intron in a paralog, which is not significantly lower than for U2-type introns (Fisher's exact tests, $P = 0.072$ (KOGs) and $P = 0.53$ (Pfams)). Most of these shared U12-type LECA introns in KOGs were paired to U2-type introns in paralogs (56 U12-U2

pairs and 4 U12-U12 pairs). In contrast, 29 of the 50 shared U12-type LECA introns in Pfams were paired with at least one other U12-type LECA intron in a paralog. The higher numbers of U12-U12 pairs in the Pfams set may result from multiple *bona fide* OGs being combined into a single KOG (see main text).

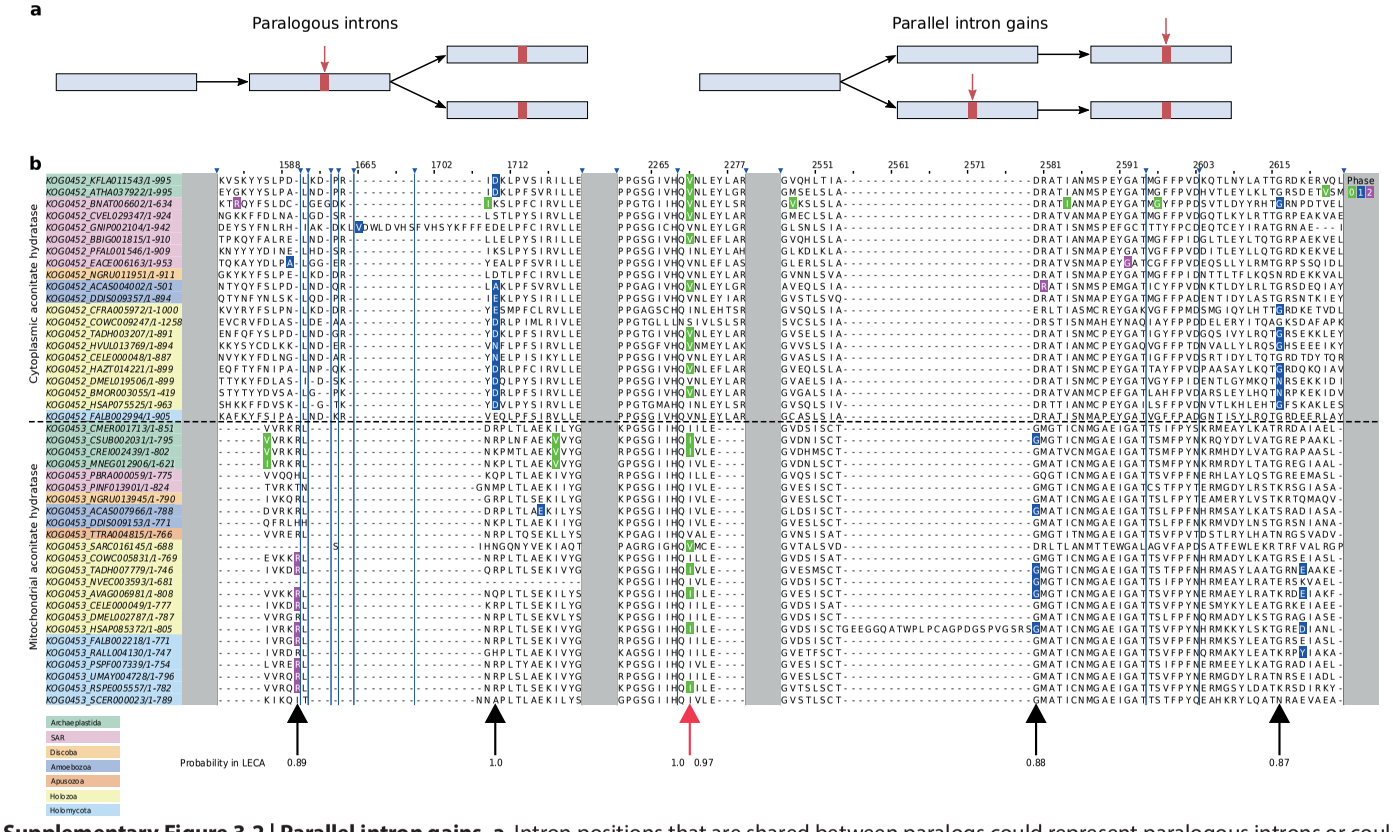
U12-type introns were less often in phase 0 and more in phase 2 than U2-type introns (**Supplementary Figure 3.7a** and **3.8a**) and even more biased towards the 5′ end (**Supplementary Figure 3.7b**). Both observations are consistent with previous studies comparing intron types in present-day eukaryotes (Basu et al., 2008a; Moyer et al., 2020). Assuming that nearly all type conversions were from U12 to U2 conversion, as has been argued based on comparative analyses (Sharp and Burge, 1997), we inferred 32 U12 gains, 9 complete U12 losses and 7 U12-to-U2 conversions before duplications and a further 113 U12 gains, 20 complete U12 losses and 15 U12-to-U2 conversions on the branches that resulted in the LECA families. Paralogs that had a U12-type intron traced back to their pre-duplication lineage were overrepresented in cell cycle and inorganic ion transport and metabolism functions (**Supplementary Figure 3.8b**). Differences in the fraction of U12-type introns among shared introns between functions of KOGs were not significant (**Supplementary Figure 3.7c**) and only a few significant differences between different phylogenetic origins of these paralogs were found (**Supplementary Figure 3.8c**). U12-type introns emerged at least before a large part of the complexification of the cell cycle and comparisons of the branch lengths seemed to suggest that U12-type introns are as old as U2-type introns, if not older (**Supplementary Figure 3.6b**).

The three models have different expectations regarding shared U12-type introns, depending on the timing of minor intron emergence. The occurrence of both types among shared introns and the inferred age of U12-type introns are not consistent with the model of two invasions by group II introns that were clearly separated in time. Divergence from primordial introns in either separate lineages followed by fusion of these lineages or divergence in the same lineage is a more likely scenario based on our findings.

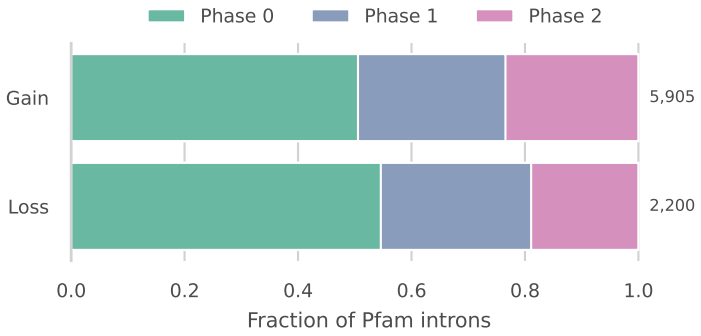
Supplementary Figures



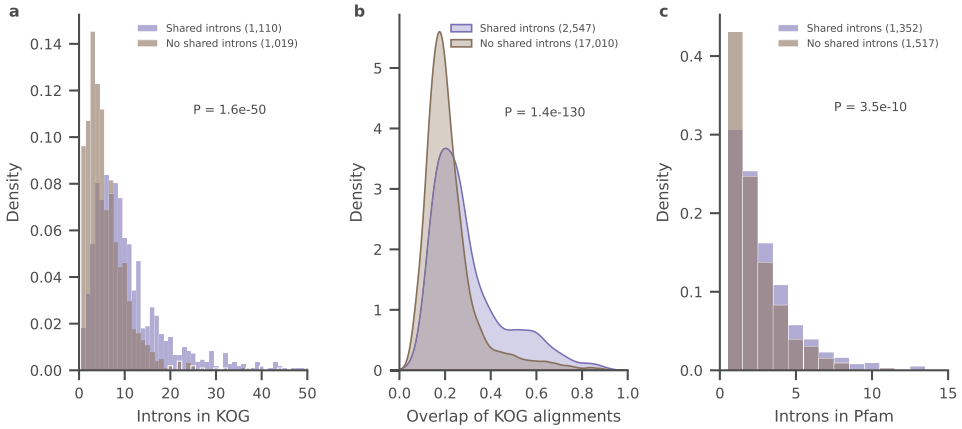
Supplementary Figure 3.1 | Ancestral intron reconstructions. a, Species tree with for each node the estimated number of introns (including missing sites) in KOGs represented as circles. These estimates are based on all used KOGs, including those from separate acquisitions. The size of the LECA node corresponds to 27,108 introns. The two terminal nodes with the highest number of introns correspond to two dinoflagellate species. **b**, Descending empirical cumulative distribution function plot of the posterior probability of a KOG intron to have been present in LECA. **c**, Species tree with for each node the estimated number of introns (including missing sites) in Pfam OGs represented as circles. These estimates are based on all used Pfam OGs, including those from separate acquisitions. The size of the LECA node corresponds to 14,977 introns. **d**, Descending empirical cumulative distribution function plot of the posterior probability of a Pfam OG intron to have been present in LECA.



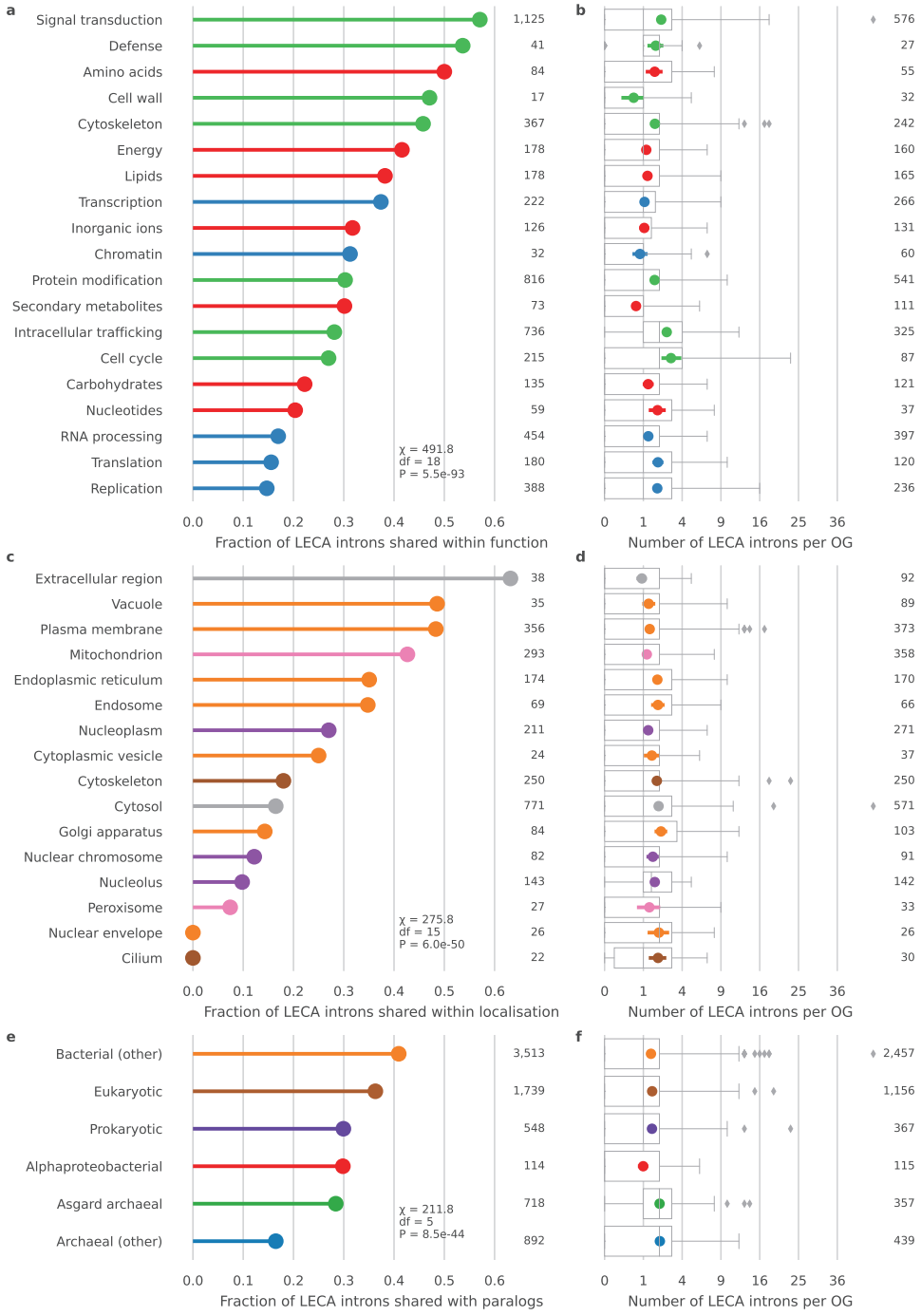
Supplementary Figure 3.2 | Parallel intron gains. a, Intron positions that are shared between paralogs could represent paralogous introns or could be due to parallel intron gains. **b**, Example of parallel intron gains in separate acquisitions. Several aligned sequences from KOG0452 (cytoplasmic aconitate hydratase) and KOG0453 (mitochondrial aconitate hydratase) are shown with their mapped introns. The phase of introns is indicated with colours. Arrows point to introns that were probably present in LECA, with the number corresponding to the posterior probability. The shared intron that has been the result of parallel intron gains is indicated with a red arrow. Different sections of the alignment are separated by grey blocks and the blue stripes correspond to blocks of alignment positions that are only gaps in the sequences shown.



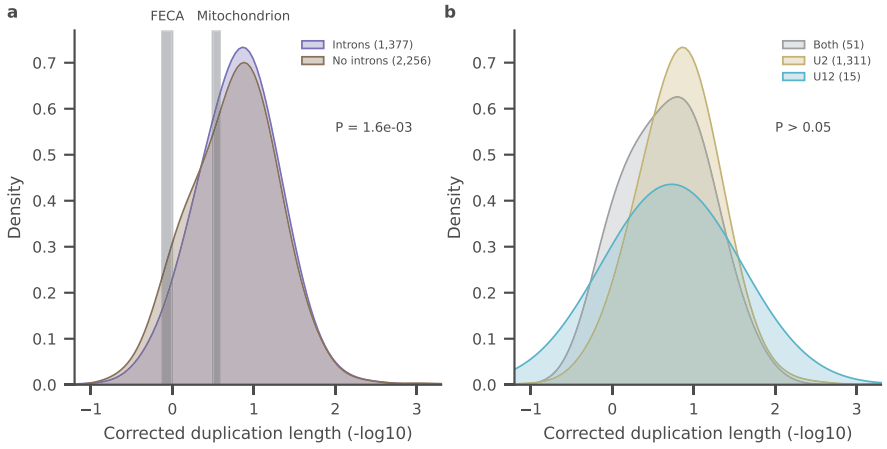
Supplementary Figure 3.3 | Phase distributions of introns in Pfams that were gained or lost before LECA. Numbers indicate the number of inferred intron gains and losses.



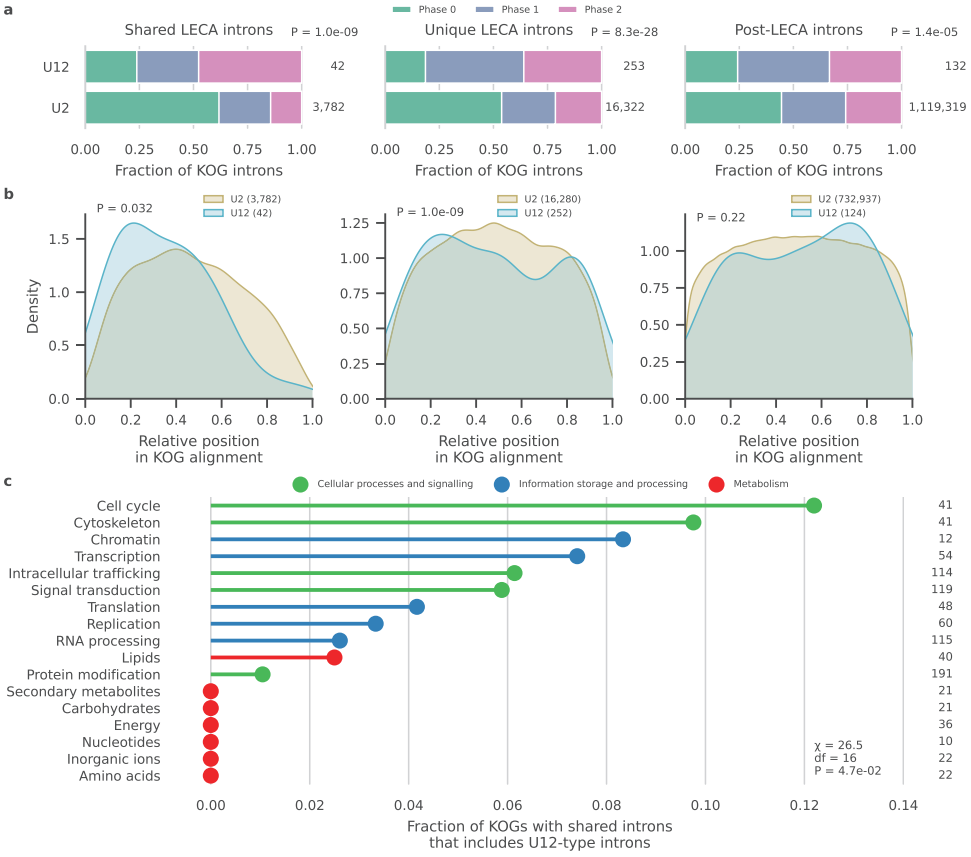
Supplementary Figure 3.4 | Influence of number of LECA introns and overlap of OG alignments on finding shared introns. a, Normalised histograms showing the distribution of the number of LECA introns in a KOG for KOGs with and without shared introns. For clarity, KOGs with more than 50 LECA introns are not shown. **b**, Density plots showing the distribution of the fraction of overlapping positions in the alignment of all pairs of KOGs in the same cluster. Pairs with and without shared introns are depicted separately. Sites with more than 90% gaps were excluded in calculating the overlapping fraction. A lower overlap could be due to domain accretion and loss after duplication. **c**, Normalised histograms showing the distribution of the number of LECA introns in a Pfam OG for Pfam OGs with and without shared introns. For clarity, Pfam OGs with more than 15 LECA introns are not shown. *P* values of Kolmogorov-Smirnov tests are shown. The numbers indicate the number of OGs (**a**, **c**) or pairs of KOGs (**b**). KOGs and Pfam OGs with no LECA introns were not included.


a
Signal transduction
Defense
Amino acids
Cell wall
Cytoskeleton
Energy
Lipids
Transcription
Inorganic ions
Chromatin
Protein modification
Secondary metabolites
Intracellular trafficking
Cell cycle
Carbohydrates
Nucleotides
RNA processing
Translation
Replication
1,125
41
84
17
367
178
178
222
126
32
816
73
736
215
135
59
454
180
388
χ = 491.8
df = 18
P = 5.5e-93
0.0 0.1 0.2 0.3 0.4 0.5 0.6
Fraction of LECA introns shared within function
b
576
27
55
32
242
160
165
266
131
60
541
111
325
87
121
37
397
120
236
0 1 4 9 16 25 36
Number of LECA introns per OG
c
Extracellular region
Vacuole
Plasma membrane
Mitochondrion
Endoplasmic reticulum
Endosome
Nucleoplasm
Cytoplasmic vesicle
Cytoskeleton
Cytosol
Golgi apparatus
Nuclear chromosome
Nucleolus
Peroxisome
Nuclear envelope
Cilium
38
35
356
293
174
69
211
24
250
771
84
82
143
27
26
22
χ = 275.8
df = 15
P = 6.0e-50
0.0 0.1 0.2 0.3 0.4 0.5 0.6
Fraction of LECA introns shared within localisation
d
92
89
373
358
170
66
271
37
250
571
103
91
142
33
26
30
0 1 4 9 16 25 36
Number of LECA introns per OG
e
Bacterial (other)
Eukaryotic
Prokaryotic
Alphaproteobacterial
Asgard archaeal
Archaeal (other)
3,513
1,739
548
114
718
892
χ = 211.8
df = 5
P = 8.5e-44
0.0 0.1 0.2 0.3 0.4 0.5 0.6
Fraction of LECA introns shared with paralogs
f
2,457
1,156
367
115
357
439
0 1 4 9 16 25 36
Number of LECA introns per OG

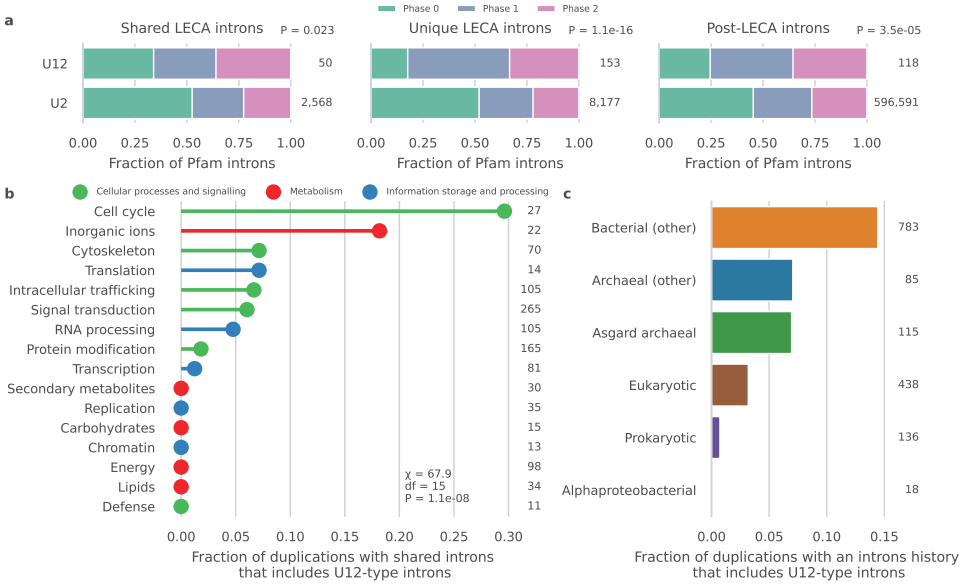
Supplementary Figure 3.5 | Fraction of shared LECA introns and number of LECA introns for Pfam OGs. a, **c**, Fraction of LECA introns shared between pairs of Pfam OGs with the same function (**a**) or localisation (**c**). Only functions and localisations with at least ten LECA introns and ten pairs are shown. Comparisons of the three functional categories were all significant (**Supplementary Table 3.7**), as were 50% of pairwise comparisons (**Supplementary Data 3.6**). 60% of the comparisons between the five localisation categories and 58% of pairwise comparisons were significant (**Supplementary Table 3.8**, **Supplementary Data 3.7**). **e**, Fraction of LECA introns shared between Pfam OGs within an acquisition or invention clade. 73% of pairwise comparisons were significant (**Supplementary Table 3.9**). **b**, **d**, **f**, Number of LECA introns in an OG with a certain function (**b**), localisation (**d**) or phylogenetic origin (**f**) on a square-root scale. Distributions are shown with boxplots and the average number of introns per group with a coloured dot. Coloured bars represent 95% confidence intervals of the mean. Numbers correspond with the number of LECA introns (**a**, **c**, **e**) or number of Pfam OGs (**b**, **d**, **f**).



Supplementary Figure 3.6 | Timing of duplications with introns using branch lengths from phylogenetic trees. a, Density plot showing the duplication lengths of duplications with and without pre-duplication introns. For comparative purposes, the estimated timing (Vosseberg et al., 2021a) of the first eukaryotic common ancestor ('FECA'), which represents the divergence from the Asgard archaeal lineage, and the divergence from the alphaproteobacterial lineage ('Mitochondrion') are depicted in grey. The two distributions are significantly different according to the Kolmogorov-Smirnov test. **b**, Density plot showing the duplication lengths of duplications with only U2-type, only U12-type or both types of pre-duplication introns. All pairwise comparisons with Kolmogorov-Smirnov tests were not significant. Numbers in both panels indicate the number of duplications considered.



Supplementary Figure 3.7 | Comparison of U2- and U12-type introns in KOGs. a, Intron phase distributions for U2- and U12-type post-LECA, unique LECA and shared LECA introns in KOGs. *P* values of χ^2 contingency tests are shown. **b**, Density plots showing the relative positions of introns in the alignment of a KOG, comparing U2- and U12-type introns for the three different groups. *P* values of Kolmogorov-Smirnov tests are shown. **c**, Fraction of KOGs with shared introns that includes U12-type introns in the different functional categories. Only functions with at least ten KOGs with shared introns are shown. Comparisons of the three functional categories (**Supplementary Table 3.10**) and pairwise comparisons of the different functions (**Supplementary Data 3.8**) were not significant. Numbers indicate the number of introns considered (**a**, **b**) or the number of KOGs with shared introns (**c**).



Supplementary Figure 3.8 | Comparison of U2- and U12-type introns in Pfams. a, Intron phase distributions for U2- and U12-type post-LECA, unique LECA and shared LECA introns in Pfam OGs. *P* values of χ^2 contingency tests are shown. **b**, Fraction of Pfam duplications with pre-duplication introns that includes U12-type introns in the different functional categories. Only functions with at least ten duplications with shared introns are shown. Comparisons of the three functional categories were not significant (**Supplementary Table 3.11**). 9.2% of pairwise comparisons were significant, which were only comparisons including the cell cycle and inorganic ions (**Supplementary Data 3.9**). **c**, Fraction of Pfam duplications with pre-duplication introns in either that duplication or a more ancestral duplication that includes U12-type introns according to the different phylogenetic origins. Only the pairwise comparisons of bacterial and eukaryotic and bacterial and prokaryotic duplications were significant (**Supplementary Table 3.12**). Numbers indicate the number of introns (**a**) or the number of duplications (**b**, **c**) considered.

Supplementary Tables

Supplementary Table 3.1 | Pairwise comparisons of intron phases in KOGs.

Introns type 1	Introns type 2	χ^2	d.f.	*P* value	Adjusted *P* value
Unique LECA introns	Shared LECA introns	111	2	6.61×10^{-25}	6.61×10^{-25}
Unique LECA introns	Post-LECA introns	506	2	1.46×10^{-110}	4.37×10^{-110}
Shared LECA introns	Post-LECA introns	467	2	3.66×10^{-102}	5.49×10^{-102}

Supplementary Table 3.2 | Pairwise comparisons of intron phases in Pfams.

Introns type 1	Introns type 2	χ^2	d.f.	*P* value	Adjusted *P* value
Unique LECA introns	Shared LECA introns	2	2	3.29×10^{-1}	3.29×10^{-1}
Unique LECA introns	Post-LECA introns	133	2	1.05×10^{-29}	3.15×10^{-29}
Shared LECA introns	Post-LECA introns	50	2	1.32×10^{-11}	1.98×10^{-11}

Supplementary Table 3.3 | Pairwise comparisons of functional categories of pairs of KOGs with and without introns (unique versus shared introns).

Category 1	Category 2	Odds ratio	*P* value	Adjusted *P* value
Cellular processes and signalling	Metabolism	1.60	1.62×10^{-15}	2.44×10^{-15}
Cellular processes and signalling	Information storage and processing	2.74	3.35×10^{-78}	1.00×10^{-77}
Metabolism	Information storage and processing	1.72	4.29×10^{-13}	4.29×10^{-13}

Supplementary Table 3.4 | Pairwise comparisons of functional categories of Pfam duplications with and without introns (duplications with versus without preduplication introns).

Category 1	Category 2	Odds ratio	*P* value	Adjusted *P* value
Cellular processes and signalling	Metabolism	1.44	4.89×10^{-4}	7.33×10^{-4}
Cellular processes and signalling	Information storage and processing	1.87	2.30×10^{-11}	6.89×10^{-11}
Metabolism	Information storage and processing	1.30	2.49×10^{-2}	2.49×10^{-2}

Supplementary Table 3.5 | Pairwise comparisons of localisation categories of Pfam duplications with and without introns (duplications with versus without preduplication introns).

Category 1	Category 2	Odds ratio	*P* value	Adjusted *P* value
Endomembrane system	Metabolic compartment	1.18	3.34×10^{-1}	4.18×10^{-1}
Endomembrane system	Other	1.19	2.60×10^{-1}	3.71×10^{-1}
Endomembrane system	Nucleus	1.59	6.51×10^{-3}	6.51×10^{-2}
Endomembrane system	Cytoskeleton	1.69	1.68×10^{-2}	8.41×10^{-2}
Metabolic compartment	Other	1.01	1.00×10^{0}	1.00×10^{0}
Metabolic compartment	Nucleus	1.36	1.07×10^{-1}	2.06×10^{-1}
Metabolic compartment	Cytoskeleton	1.44	1.23×10^{-1}	2.06×10^{-1}
Other	Nucleus	1.34	9.17×10^{-2}	2.06×10^{-1}
Other	Cytoskeleton	1.42	1.13×10^{-1}	2.06×10^{-1}
Nucleus	Cytoskeleton	1.06	8.21×10^{-1}	9.12×10^{-1}

Supplementary Table 3.6 | Comparison of phylogenetic origins of Pfam duplications with and without introns (duplications with versus without preduplication introns).

Phylogenetic origin 1	Phylogenetic origin 2	Odds ratio	*P* value	Adjusted *P* value
Eukaryotic	Prokaryotic	1.09	5.32×10^{-1}	5.70×10^{-1}
Eukaryotic	Asgard archaeal	1.17	2.81×10^{-1}	3.38×10^{-1}
Eukaryotic	Bacterial (other)	1.40	7.97×10^{-5}	2.69×10^{-4}
Eukaryotic	Archaeal (other)	2.35	1.09×10^{-9}	1.64×10^{-8}
Eukaryotic	Alphaproteobacterial	3.46	1.26×10^{-5}	6.95×10^{-5}
Prokaryotic	Asgard archaeal	1.07	7.25×10^{-1}	7.25×10^{-1}
Prokaryotic	Bacterial (other)	1.28	6.30×10^{-2}	9.46×10^{-2}
Prokaryotic	Archaeal (other)	2.15	1.39×10^{-5}	6.95×10^{-5}
Prokaryotic	Alphaproteobacterial	3.16	1.47×10^{-4}	3.15×10^{-4}
Asgard archaeal	Bacterial (other)	1.19	2.13×10^{-1}	2.91×10^{-1}
Asgard archaeal	Archaeal (other)	2.01	1.16×10^{-4}	2.89×10^{-4}
Asgard archaeal	Alphaproteobacterial	2.95	5.26×10^{-4}	9.85×10^{-4}
Bacterial (other)	Archaeal (other)	1.69	8.95×10^{-5}	2.69×10^{-4}
Bacterial (other)	Alphaproteobacterial	2.48	1.81×10^{-3}	3.02×10^{-3}
Archaeal (other)	Alphaproteobacterial	1.47	2.93×10^{-1}	3.38×10^{-1}

Supplementary Table 3.7 | Pairwise comparisons of functional categories of pairs of Pfam OGs with and without introns (unique versus shared introns).

Category 1	Category 2	Odds ratio	*P* value	Adjusted *P* value
Cellular processes and signalling	Metabolism	1.30	1.15×10^{-3}	1.15×10^{-3}
Cellular processes and signalling	Information storage and processing	2.75	3.34×10^{-42}	1.00×10^{-41}
Metabolism	Information storage and processing	2.12	1.17×10^{-13}	1.76×10^{-13}

Supplementary Table 3.8 | Pairwise comparisons of localisation categories of pairs of Pfam OGs with and without introns (unique versus shared introns).

Category 1	Category 2	Odds ratio	*P* value	Adjusted *P* value
Metabolic compartment	Endomembrane system	1.07	6.32×10^{-1}	7.03×10^{-1}
Metabolic compartment	Other	2.87	8.85×10^{-13}	3.64×10^{-12}
Metabolic compartment	Nucleus	2.88	1.70×10^{-10}	3.40×10^{-10}
Metabolic compartment	Cytoskeleton	3.32	4.57×10^{-10}	7.61×10^{-10}
Endomembrane system	Other	2.67	9.75×10^{-18}	9.75×10^{-17}
Endomembrane system	Nucleus	2.69	1.09×10^{-12}	3.64×10^{-12}
Endomembrane system	Cytoskeleton	3.09	2.06×10^{-11}	5.15×10^{-11}
Other	Nucleus	1.01	1.00×10^{0}	1.00×10^{0}
Other	Cytoskeleton	1.16	4.67×10^{-1}	6.68×10^{-1}
Nucleus	Cytoskeleton	1.15	5.45×10^{-1}	6.81×10^{-1}

Supplementary Table 3.9 | Comparison of phylogenetic origins of shared introns in Pfam OGs (unique versus shared introns).

Phylogenetic origin 1	Phylogenetic origin 2	Odds ratio	*P* value	Adjusted *P* value
Bacterial (other)	Eukaryotic	1.22	1.19×10^{-3}	1.99×10^{-3}
Bacterial (other)	Prokaryotic	1.62	7.56×10^{-7}	1.89×10^{-6}
Bacterial (other)	Alphaproteobacterial	1.63	1.97×10^{-2}	2.69×10^{-2}
Bacterial (other)	Asgard archaeal	1.75	1.97×10^{-10}	9.86×10^{-10}
Bacterial (other)	Archaeal (other)	3.51	6.32×10^{-46}	9.47×10^{-45}
Eukaryotic	Prokaryotic	1.33	6.41×10^{-3}	9.62×10^{-3}
Eukaryotic	Alphaproteobacterial	1.34	1.90×10^{-1}	2.38×10^{-1}
Eukaryotic	Asgard archaeal	1.43	1.77×10^{-4}	3.80×10^{-4}
Eukaryotic	Archaeal (other)	2.89	2.38×10^{-27}	1.79×10^{-26}
Prokaryotic	Alphaproteobacterial	1.00	1.00×10^{0}	1.00×10^{0}
Prokaryotic	Asgard archaeal	1.08	5.74×10^{-1}	6.63×10^{-1}
Prokaryotic	Archaeal (other)	2.16	3.48×10^{-9}	1.30×10^{-8}
Alphaproteobacterial	Asgard archaeal	1.07	7.39×10^{-1}	7.92×10^{-1}
Alphaproteobacterial	Archaeal (other)	2.15	1.06×10^{-3}	1.99×10^{-3}
Asgard archaeal	Archaeal (other)	2.01	1.00×10^{-8}	3.01×10^{-8}

Supplementary Table 3.10 | Pairwise comparisons of functional categories of KOGs with shared introns regarding intron type.

Category 1	Category 2	Odds ratio	*P* value	Adjusted *P* value
Cellular processes and signalling	Information storage and processing	1.20	7.27×10^{-1}	7.27×10^{-1}
Cellular processes and signalling	Metabolism	8.89	9.42×10^{-3}	2.82×10^{-2}
Information storage and processing	Metabolism	7.41	3.73×10^{-2}	5.60×10^{-2}

Supplementary Table 3.11 | Pairwise comparisons of functional categories of Pfam duplications with introns regarding intron type.

Category 1	Category 2	Odds ratio	*P* value	Adjusted *P* value
Cellular processes and signalling	Information storage and processing	2.19	6.20×10^{-2}	9.30×10^{-2}
Cellular processes and signalling	Metabolism	3.27	1.70×10^{-2}	5.11×10^{-2}
Information storage and processing	Metabolism	1.50	5.60×10^{-1}	5.60×10^{-1}

Supplementary Table 3.12 | Comparisons of phylogenetic origins of Pfam duplications with an intron history regarding intron type.

Phylogenetic origin 1	Phylogenetic origin 2	Odds ratio	*P* value	Adjusted *P* value
Bacterial (other)	Asgard archaeal	2.09	7.31×10^{-2}	1.57×10^{-1}
Bacterial (other)	Archaeal (other)	2.87	4.20×10^{-2}	1.05×10^{-1}
Bacterial (other)	Eukaryotic	3.41	2.37×10^{-10}	3.55×10^{-9}
Bacterial (other)	Alphaproteobacterial	inf	2.35×10^{-1}	4.41×10^{-1}
Bacterial (other)	Prokaryotic	inf	6.05×10^{-6}	4.54×10^{-5}
Asgard archaeal	Archaeal (other)	1.37	7.57×10^{-1}	9.63×10^{-1}
Asgard archaeal	Eukaryotic	1.63	3.10×10^{-1}	5.17×10^{-1}
Asgard archaeal	Alphaproteobacterial	inf	5.89×10^{-1}	8.84×10^{-1}
Asgard archaeal	Prokaryotic	inf	1.44×10^{-2}	7.21×10^{-2}
Archaeal (other)	Eukaryotic	1.19	7.70×10^{-1}	9.63×10^{-1}
Archaeal (other)	Alphaproteobacterial	inf	1.00×10^{0}	1.00×10^{0}
Archaeal (other)	Prokaryotic	inf	4.20×10^{-2}	1.05×10^{-1}
Eukaryotic	Alphaproteobacterial	inf	1.00×10^{0}	1.00×10^{0}
Eukaryotic	Prokaryotic	inf	4.21×10^{-2}	1.05×10^{-1}
Alphaproteobacterial	Prokaryotic	nan	1.00×10^{0}	1.00×10^{0}

Supplementary Data

Supplementary Data files are available online on the website of Communications Biology (https://doi.org/10.1038/s42003-022-03426-5).

Supplementary Data 3.1 | KOG clusters.

Supplementary Data 3.2 | Species list with links to used genome annotation files.

Supplementary Data 3.3 | Pairwise comparisons of functions of pairs of KOGs with and without introns (unique versus shared introns).

Supplementary Data 3.4 | Pairwise comparisons of functions of Pfam duplications with and without introns (duplications with versus without preduplication introns).

Supplementary Data 3.5 | Pairwise comparisons of localisations of Pfam duplications with and without introns (duplications with versus without preduplication introns).

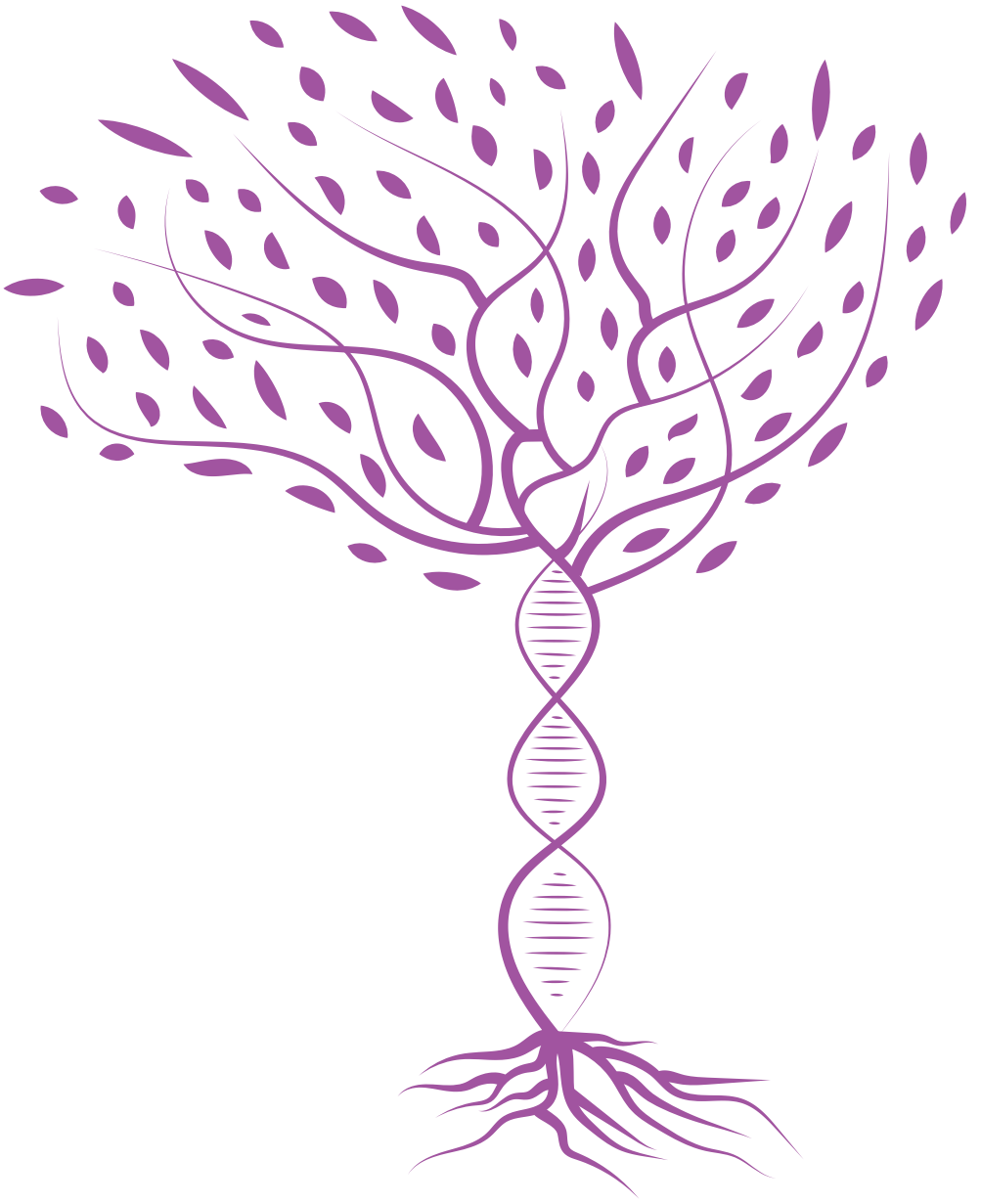
Supplementary Data 3.6 | Pairwise comparisons of functions of pairs of Pfam OGs with and without introns (unique versus shared introns).

Supplementary Data 3.7 | Pairwise comparisons of localisations of pairs of Pfam OGs with and without introns (unique versus shared introns).

Supplementary Data 3.8 | Pairwise comparisons of functions of KOGs with shared introns regarding intron type.

Supplementary Data 3.9 | Pairwise comparisons of functions of Pfam duplications with introns regarding intron type.

Supplementary Data 3.10 | The source data behind the graphs in the paper.



4

Domestication of self-splicing introns during eukaryogenesis: the rise of the complex spliceosomal machinery

Julian Vosseberg and Berend Snel

Biology Direct, 2017

Abstract

The spliceosome is a eukaryote-specific complex that is essential for the removal of introns from pre-mRNA. It consists of five small nuclear RNAs (snRNAs) and over a hundred proteins, making it one of the most complex molecular machineries. Most of this complexity has emerged during eukaryogenesis, a period that is characterised by a drastic increase in cellular and genomic complexity. Although not fully resolved, recent findings have started to shed some light on how and why the spliceosome originated.

In this paper we review how the spliceosome has evolved and discuss its origin and subsequent evolution in light of different general hypotheses on the evolution of complexity. Comparative analyses have established that the catalytic core of this ribonucleoprotein (RNP) complex, as well as the spliceosomal introns, evolved from self-splicing group II introns. Most snRNAs evolved from intron fragments and the essential Prp8 protein originated from the protein that is encoded by group II introns. Proteins that functioned in other RNA processes were added to this core and extensive duplications of these proteins substantially increased the complexity of the spliceosome prior to the eukaryotic diversification. The splicing machinery became even more complex in animals and plants, yet was simplified in eukaryotes with streamlined genomes. Apparently, the spliceosome did not evolve its complexity gradually, but in rapid bursts, followed by stagnation or even simplification. We argue that although both adaptive and neutral evolution have been involved in the evolution of the spliceosome, especially the latter was responsible for the emergence of an enormously complex eukaryotic splicing machinery from simple self-splicing sequences.

Background

Eukaryotic genes are in general composed of coding sequences interspersed by non-coding parts, the introns. Only after removal of these introns and splicing of the exons, a functional protein can be synthesised. The splicing reaction requires one of the most complex machines in the eukaryotic cell, the spliceosome, which consists of five snRNA molecules and over a hundred proteins (Valadkhan and Jaladat, 2010; Wahl et al., 2009). Two types of spliceosomes are present across eukaryotes, namely the major and the minor spliceosome. Each spliceosome splices its own type of introns, the U2-type introns for the major spliceosome and the U12-type introns for the minor counterpart.

The spliceosome is one of the numerous complex characteristics that emerged during eukaryogenesis. Eukaryotes are considered far more complex than prokaryotes, because of these evolved characteristics such as their larger genomes, cell sizes and intracellular compartmentalisation. However, some complex eukaryote-like features, such as large cells and internal membranes, have been observed in certain prokaryotes and some eukaryotes are less complex in organisation, cautioning for a too eukaryocentric view on complexity (Booth and Doolittle, 2015). It has been firmly demonstrated that eukaryotes originated from the merger of two prokaryotes (McInerney et al., 2014), an archaeal host related to the recently discovered Asgard phyla (Spang et al., 2015; Zaremba-Niedzwiedzka et al., 2017) and a bacterial endosymbiont related to the Alphaproteobacteria. Lane and Martin have proposed that the increased complexity of eukaryotes could solely be

enabled by the surplus of energy provided by the mitochondrial endosymbionts (Lane and Martin, 2010), but their reasoning is challenged (Booth and Doolittle, 2015; Lynch and Marinov, 2017). The precise role of the mitochondria in the evolution of eukaryotic complexity remains therefore under lively debate.

The greater complexity of eukaryotes is additionally observed in the complexity of molecular machines, both for machines that are also present in prokaryotes (e.g., the ribosome and respiration chain complexes) (Gray et al., 2010; Lukeš et al., 2011; Speijer, 2011) and eukaryote-specific complexes other than the spliceosome. The evolution of these molecular machines in their cellular context is within the scope of the emerging field of evolutionary cell biology (Brodsky et al., 2012; Lynch et al., 2014; Richardson et al., 2015). One of the questions in this field is how the complexity of these complexes has evolved. For a complete understanding of the evolution of a complex, not only the intermediate steps have to be described, but also the evolutionary forces driving these steps. Multiple models have been proposed, emphasising the adaptive (Speijer, 2011), neutral (Lukeš et al., 2011; Mast et al., 2014; Stoltzfus, 1999, 2012) or maladaptive (Lynch, 2007, 2012) nature of additional components or interactions. Moreover, according to the biphasic model an increase in complexity is followed by a period of reductive evolution (Cuypers and Hogeweg, 2012; Wolf and Koonin, 2013).

Many steps were needed for the emergence of the complex spliceosome in the last eukaryotic common ancestor (LECA). The aim of this review is to reconstruct these steps and the subsequent changes in the complexity of the spliceosome in the distinct eukaryotic lineages, which is important for understanding why and how the complexity of this machine has evolved. This could additionally provide more insight into the evolution of other complex molecular machines. In this review we will focus on the snRNAs and main proteins of the major spliceosome.

LECA's spliceosomes

To separate the evolution of the spliceosome during eukaryogenesis from its evolution after the eukaryotic diversification, the spliceosome of LECA has to be reconstructed. The presence of spliceosomal components in all major eukaryotic lineages has revealed that LECA already had a complex major spliceosome, with five snRNAs and around eighty proteins (Collins and Penny, 2005). Therefore, LECA's spliceosomes would likely not be much unlike typical contemporary spliceosomes.

It has become clear that this complex spliceosome had to remove numerous introns from LECA's transcripts. Multiple approaches have been followed for reconstructing the introns in LECA (reviewed in (Koonin et al., 2013; Rogozin et al., 2012)). The most sophisticated model used to date inferred an intron density of 4.3 introns per kilobase in LECA's genome, which is only a fraction lower than the typical intron density of animal and plant genomes, but much higher than that of most protists (Csuros et al., 2011). Apparently, the complex nature of LECA's spliceosome corresponded with its intron-rich genome.

The probable function of LECA's spliceosomes can be inferred from experimental research on present-day spliceosomes, most of which has been performed in yeast, animals

and plants. The main components of LECA's major (U2-type) spliceosome were the five small nuclear ribonucleoproteins (snRNPs), consisting of snRNAs and dozens of other associated proteins (Collins and Penny, 2005) (composition and function of present-day spliceosomes reviewed in (Matera and Wang, 2014; Valadkhan and Jaladat, 2010; Wahl et al., 2009)). The uridine (U)-rich snRNAs in the spliceosome were U1, U2, U4, U5 and U6, each giving the accompanying snRNP its name. The snRNAs were tightly associated with a ring of either Lsm proteins (U6 snRNA) or Sm proteins (other snRNAs).

Rearrangements during the splicing cycle are crucial for spliceosomal functioning and these conformational changes were in LECA already effected by ATP-dependent RNA helicases. The precise composition of the spliceosome depended on the step in the splicing cycle. For example, U2, U5 and U6 snRNPs were present in the catalytically active spliceosome, whereas U1 and U4 snRNPs dissociated before the splicing reaction, as these were involved in splice site recognition and inhibiting U6 snRNA, respectively. The important regulatory serine/arginine-rich (SR) proteins and heterogeneous nuclear RNPs (hnRNPs), present across eukaryotes (Barbosa-Morais et al., 2006), were involved in exon and intron recognition and thereby splicing out the proper introns and enabling alternative splicing (Matera and Wang, 2014; Shepard and Hertel, 2009; Valadkhan and Jaladat, 2010; Wahl et al., 2009). After recognition of the 5′ and 3′ splice sites and the adenosine branch point nucleotide the first splicing step could be executed. In this transesterification reaction a nucleophilic attack created a covalent bond between the 5′ splice site and the 2′ OH group of the bulged adenosine, resulting in a lariat. In the following second reaction the exon ends were joined together and the lariat intron was released. In essence, LECA's major spliceosome would likely not have been fundamentally different in composition and function from its present-day counterparts.

Minor spliceosome and spliced-leader trans-splicing

Although some earlier studies suggested otherwise (Barbosa-Morais et al., 2006; Collins and Penny, 2005), additional genome data and more sensitive searches revealed that the minor spliceosome evolved early in eukaryotes as well and was probably present in LECA (López et al., 2008; Russell et al., 2006). The minor spliceosome consists of its own specific snRNPs – U11, U12, U4atac and U6atac – which are functionally analogous to their major-spliceosomal counterparts, and U5 snRNP, which is shared between both spliceosomes (Turunen et al., 2013). The associated proteins in the minor spliceosome can be either specific to this complex or shared with the major spliceosome (Turunen et al., 2013). As mentioned before, the minor spliceosome excises a different kind of introns, the U12-type introns. These introns comprise only a small fraction compared with the U2-type introns in the organisms that contain both kinds of introns (Basu et al., 2008b; Lin et al., 2010; Turunen et al., 2013).

Most snRNPs of the major spliceosome are also involved in another related splicing reaction called spliced-leader (SL) *trans*-splicing, in which the SL RNA, which is carried by the SL snRNP, donates the first "exon" to the mRNA. This splicing mechanism is especially prevalent in some protist lineages, where it in some cases may account for all splicing events (Lei et al., 2016). Based on its patchy presence pattern across eukaryotes

it was initially proposed to have been present in LECA and subsequently lost multiple times in many lineages (Collins and Penny, 2005). However, the observed pattern may also result from independent gain events due to horizontal gene transfer (HGT) (Lasda and Blumenthal, 2011) or recurrent mutational acquisition of SL RNA (Douris et al., 2010; Lasda and Blumenthal, 2011; Maeso et al., 2012). Whether the major spliceosome of LECA performed SL *trans*-splicing can therefore not unambiguously be established.

Origin of the spliceosome

LECA likely already possessed two spliceosome types to process two different kinds of introns. These spliceosomes were approximately as complex as the ones typically observed in present-day eukaryotes. This poses the question how the complex spliceosome evolved during eukaryogenesis. Where did the proteins come from, how were they recruited into the spliceosome and what functions did their prokaryotic homologues, if present, execute?

The function of the spliceosome is removing introns from pre-mRNA molecules. The question how the spliceosome originated cannot be decoupled from the origin of the introns they remove. Without introns the spliceosome would be functionless and without the spliceosome the introns would cause the production of aberrant proteins. Different hypotheses have been proposed for the origin of spliceosomal introns. These will shortly be discussed before we turn to the origin of the spliceosome. Both the spliceosomal core and the introns themselves are likely derived from the very same origin, namely self-splicing introns.

Spliceosomal introns

The similarities between spliceosomal introns and group II self-splicing introns have been recognised for a long time. The latter are present in prokaryotes and in eukaryotic organelles. In mitochondria and plastids these introns are *bona fide* introns that lost their mobility potential, whereas in prokaryotes they are more properly regarded as retroelements (Koonin, 2009; Zimmerly and Semper, 2015). Group II introns (reviewed in e.g. (Dayie and Padgett, 2008; Zimmerly and Semper, 2015)) typically have a length of around 2-3 kb and consist of six RNA domains. The large domain I functions as a scaffold and recognises and positions the exons (Costa et al., 2000; Jacquier and Michel, 1990), domains II and III enhance splicing catalysis (Toor et al., 2008) and domain VI contains the adenosine residue that functions as branch point (van der Veen et al., 1986). Domain V is the most conserved domain and contains the catalytic triad, which binds the two catalytic divalent metal ions (Gordon and Piccirilli, 2001; Peebles et al., 1995; Toor et al., 2008). Domain IV is the largest, as it encodes a protein, aptly named intron-encoded protein (IEP). The maturase function of this versatile protein is required for the proper folding of group II introns, promoting RNA recognition and splicing (Matsuura et al., 2001; Wank et al., 1999). Moreover, its reverse transcriptase activity enables reverse splicing, which results in the proliferation of the introns in the host genome (Saldanha et al., 1999; Wank et al., 1999).

There is an overwhelming amount of evidence supporting the homology between

spliceosomal introns and group II self-splicing introns. The splice site recognition, branching mechanism, stereochemical course of the splicing reaction and the presence of similar RNA domain structures and a homologue of the IEP in the spliceosome (see below) demonstrate the similarities between the two intron types (Dayie and Padgett, 2008; Keating et al., 2010; Peters and Toor, 2015; Zimmerly and Semper, 2015). Moreover, there is a known example of a group II intron that was transferred from mitochondria to the nucleus in a plant family and subsequently evolved into a spliceosomal intron (Kudla et al., 2002), which underlines the evolutionary relationship between group II and spliceosomal introns.

Since group II introns are especially abundant in alphaproteobacteria and present in certain mitochondria, the most accepted view, first suggested by Cavalier-Smith in 1991 (Cavalier-Smith, 1991), is that spliceosomal introns originated from the alphaproteobacterial endosymbiont by endosymbiotic gene transfer (EGT) that later evolved into the mitochondria (Zimmerly and Semper, 2015). However, these self-splicing elements are also present in some archaeal lineages, including the Asgardian loki- and heimdallarchaeal lineages (Spang et al., 2015; Zaremba-Niedzwiedzka et al., 2017), suggesting that they also could have been present in the archaeal host. In this context it is noteworthy that many bacterial genes in eukaryotes, proposed to have been acquired upon mitochondrial endosymbiosis (Ku et al., 2015), had more likely been acquired by the archaeal host before (Ettema, 2016; Pittis and Gabaldón, 2016a). Another hypothesis that was put forward but has fallen out of favour stated that the two kinds of introns share a common ancestor in the last universal common ancestor and originated from a kind of 'protospliceosome' in the RNA world (Doolittle, 2013; Vesteg et al., 2012). This hypothesis is related to the introns-early hypothesis, which postulated that protein-coding genes interspersed with introns were the ancestral state (Doolittle, 1978). However, since it has been established that eukaryotes arose from within the Archaea (McInerney et al., 2014; Spang et al., 2015; Zaremba-Niedzwiedzka et al., 2017), it is extremely unlikely that the introns were lost in the bacterial and all non-eukaryotic archaeal lineages, but remained present in the direct line leading to the eukaryotes.

As demonstrated by relatively recent intron gains, not all spliceosomal introns in present-day eukaryotes are derived from group II introns. These introns have an endogenous origin and different sources have been suggested, such as transposable elements, internal gene duplications and intronisation of translatable sequences (Catania et al., 2009). Although it has been proposed, based on these recent intron gains, that spliceosomal introns in general had an endogenous origin (Catania et al., 2009), one should note that the origin of novel introns does not necessarily reflect the origin of the first spliceosomal introns. Given the evidence supporting a relationship with group II introns, an endogenous origin of spliceosomal introns during eukaryogenesis seems very unlikely.

Remnants of group II introns: snRNAs and Prp8

Numerous studies have noted the striking similarities in function and structure between the snRNAs and the group II intron domains and especially U6 snRNA and domain V look very similar (**Figure 4.1**) (Dayie and Padgett, 2008; Keating et al., 2010; Zimmerly

and Semper, 2015). For example, the catalytic triad and bulge are present in both structures, both bind divalent metal ions and they are functionally interchangeable (Dayie and Padgett, 2008; Fica et al., 2013; Keating et al., 2010; Shukla and Padgett, 2002). Parts of U5 snRNA, which is involved in exon recognition, resemble exon-binding sites in domain I and these parts are functionally interchangeable as well (Hetzer et al., 1997; Newman and Norman, 1992; O'Keefe et al., 1996; Peters and Toor, 2015). Also domain VI and U2 snRNA show similarities (Zimmerly and Semper, 2015). The parallels between snRNAs and group II introns have led to the idea that the snRNAs are five pieces of a group II intron (Sharp, 1991). However, since the U1 and U4 snRNAs lack a clear similarity to group II domains, these probably have a different origin (Zimmerly and Semper, 2015). Remarkably, in some organelles group II introns are present in pieces, but splicing occurs normally (Reifschneider et al., 2016; Zimmerly and Semper, 2015). Furthermore, the experimental fragmentation of a group II intron in *Lactococcus lactis* demonstrated the potential for *trans*-splicing (Belhocine et al., 2008). These observations make the hypothesised origin of the snRNAs from group II intron fragments plausible.

As mentioned above, a group II intron usually encodes an IEP. A homologous protein of IEP functions in the spliceosome, namely pre-mRNA processing protein 8 (Prp8), which is present in the U5 snRNP. Prp8 is present in the spliceosomal catalytic core and likely functions as an assembly platform (Galej et al., 2013; Peters and Toor, 2015; Turner et al., 2006). It is the largest and most conserved spliceosomal protein and interacts with the U2 and U6 snRNPs and especially the helicase Brr2 and GTPase Snu114, which are present in the U5 snRNP as well (Galej et al., 2014; Liu et al., 2006; Mozaffari-Jovin et al., 2013; Nguyen et al., 2013; Valadkhan and Jaladat, 2010; Wahl et al., 2009). The first indication for the homology between IEP and Prp8 was the presence of a reverse tran-

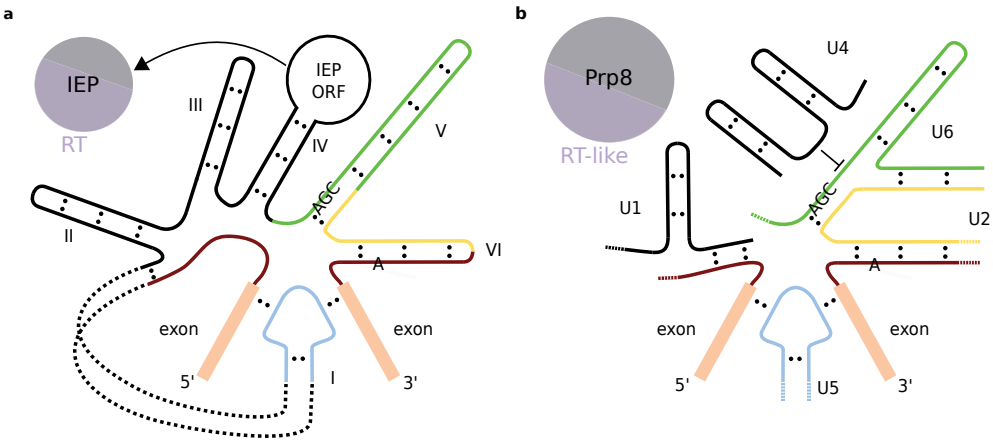


Figure 4.1 | Resemblance between group II introns, and spliceosomal introns and snRNAs. a, Simplified secondary structure of a group II intron (IIA) with its intron-encoded protein (IEP). The largest part of domain I has been omitted. The catalytic triad and adenosine branch point are explicitly depicted. The structures are coloured based on their similarity to spliceosomal structures (**b**). The black RNA domains do not have homologous structures in the spliceosome. **b**, Simplified secondary structure of a spliceosomal intron with the snRNAs and Prp8. U1 and U4 snRNA are not homologous to group II intron domains.

scriptase (RT)-like domain in Prp8, which is similar to the RT domain in IEP (Dlakić and Mushegian, 2011; Qu et al., 2016; Zhao and Pyle, 2016). IEP did not only give rise to Prp8, but also to telomerase and the RT of non-long terminal repeat retrotransposons (Qu et al., 2016). At some point Prp8 must have lost its RT activity (Aravind et al., 2012; Dlakić and Mushegian, 2011), thereby losing the ability for retromobility while maintaining its maturase function, which has occurred frequently for IEPs in organelles as well (Zimmerly and Semper, 2015).

Group II introns can be classified based on RNA structures or phylogenetic groupings of IEP (Simon et al., 2009; Toro and Martínez-Abarca, 2013; Zimmerly and Semper, 2015; Zimmerly et al., 2001). The exon recognition in spliceosomal introns is more similar to the A subtype of group II introns (Zimmerly and Semper, 2015). It is not known how Prp8 and its paralogues relate to the different IEP groups, which could be informative for the source of the group II introns that evolved into the spliceosomal introns.

Sm and Lsm proteins

Each snRNA in the spliceosome is accompanied by a heteroheptamer ring consisting of either Sm or Lsm proteins, which are both members of the Sm family of proteins. For U6 snRNA it is an Lsm ring made up of Lsm2-8, whereas the ring surrounding the other snRNAs consists of SmB, -D3, -G, -E, -F, -D2 and –D1 (Achsel et al., 1999; Collins and Penny, 2005; Lerner and Steitz, 1979; Matera et al., 2007; Mayes et al., 1999; Wahl et al., 2009). The rings function as scaffolds, enabling interactions between the snRNAs and snRNP proteins, and they are specifically involved in snRNP biogenesis (Mura et al., 2013). The central pore of the ring binds to uridine-rich stretches of RNA (Achsel et al., 1999; Matera et al., 2007). The Sm rings remain stably attached to the snRNA, whereas the Lsm ring disassociates from the U6 snRNA, together with the other U6 snRNP proteins (Valadkhan and Jaladat, 2010). This dissociation is essential for the formation of the catalytic core (Chan et al., 2003). U6 snRNA is also unique in the sense that its transcription is performed by RNA polymerase III instead of II, that it receives another 5′ cap, and is not exported to the cytoplasm (Kunkel et al., 1986; Valadkhan and Jaladat, 2010). The import into the nucleus of the other snRNAs is dependent on their interaction with the Sm ring, which is assembled around the snRNA in the cytosol (Matera and Wang, 2014; Matera et al., 2007).

Homologues of Sm and Lsm proteins are present in both bacteria and archaea. The bacterial homologue, Hfq, is encoded by a single-copy gene (Mura et al., 2013). Hfq proteins comprise a homohexameric ring that functions as an RNA chaperone in multiple processes, for instance by mediating inhibiting interactions between non-coding RNAs and target mRNAs (Mura et al., 2013). Archaea have between one and three Sm-like archaeal proteins, making homohexameric or homoheptameric rings, and despite many studies focussing on the structure of these proteins, their function is not well-characterised (Mura et al., 2013).

Although an earlier study was unable to confidently infer the deep phylogenetic relationship between the eukaryotic Sm and Lsm genes (Scofield and Lynch, 2008), a more sophisticated analysis found that each spliceosomal Lsm gene was paired with an Sm gene

(**Figure 4.2a**) (Veretnik et al., 2009). In both studies the relationship with the prokaryotic outgroup was inconclusive. It was suggested that because of the greater divergence of the Sm genes these had acquired a new function in forming the Sm ring, whereas the Lsm ring was the ancestral one. This would be consistent with the observation that Lsm rings are also involved in other processes, whereas Sm rings are specific to the spliceosome (Scofield and Lynch, 2008; Veretnik et al., 2009). Based on this, two waves of duplications were proposed, the first one leading to the seven spliceosomal Lsm genes and then duplication of each Lsm gene to an Lsm–Sm pair. The pairing was confirmed by the observation that several of the pairs have an intron at the same position when intron locations are mapped onto the alignments of these pairs across 22 species. A small number of introns are even shared between certain Lsm-Sm pairs, i.e. Lsm6 and Lsm8 share an identical intron position, as do Lsm3 and SmE. This is not trivial, as it implies that splicing could already take place before the early diversification of the Sm family in eukaryotes. Although the shared introns could reflect independent intron gain events, this is less likely since it is the case for multiple pairs and the inferred shared introns are present in multiple species. Furthermore, given the overall low number of introns (<3%) shared between paralogues originating from gene duplications during eukaryogenesis (Sverdlov et al., 2007) and the high inferred number of introns shared between orthologues in LECA (Rogozin et al., 2003), this would suggest that these duplications occurred relatively late during eukaryogenesis.

Presumably, it started with a homoheptameric flexible Lsm-like ring (**Figure 4.2b**). A first wave of duplications resulted in an Lsm heteroheptamer. Before these duplications splicing already took place and the Lsm ring might already have had a function in splicing. The specific steps from a homomeric to a heteromeric ring are difficult to infer. It has been suggested that once there was a heteromeric ring consisting of two different components, the heptameric nature of the structure accelerated the transition to an entirely heteromeric ring with seven different subunits (Scofield and Lynch, 2008). The reason behind this is that seven is a prime, so the most stable heteromeric ring may be a completely heteromeric one. The resulting heteromeric nature of the ring enabled the steric specificity that is now present in these rings (Scofield and Lynch, 2008). Duplication of the entire ring resulted in the more stable Sm ring, which became associated with all snRNAs but U6. It has been proposed that the origin of the nucleus resulted in this separation between U6 and the other snRNAs, due to the latter's export out and subsequent Sm-mediated import into the nucleus (Veretnik et al., 2009).

Helicases, Snu114 and SR proteins: addition of proteins involved in translation and RNA degradation

The ATP-dependent RNA helicases in the spliceosome are mainly from three families within the SFII superfamily, which is especially predominant in eukaryotes (Anantharaman et al., 2002). One of these is the eIF4A-DeaD family, which has in general only one representative in prokaryotes, DeaD, while in eukaryotes the family has vastly expanded to include around thirty distinct members, most of them functioning in the splicing reaction (Anantharaman et al., 2002). Eukaryotic eIF4A can be regarded as the equivalent

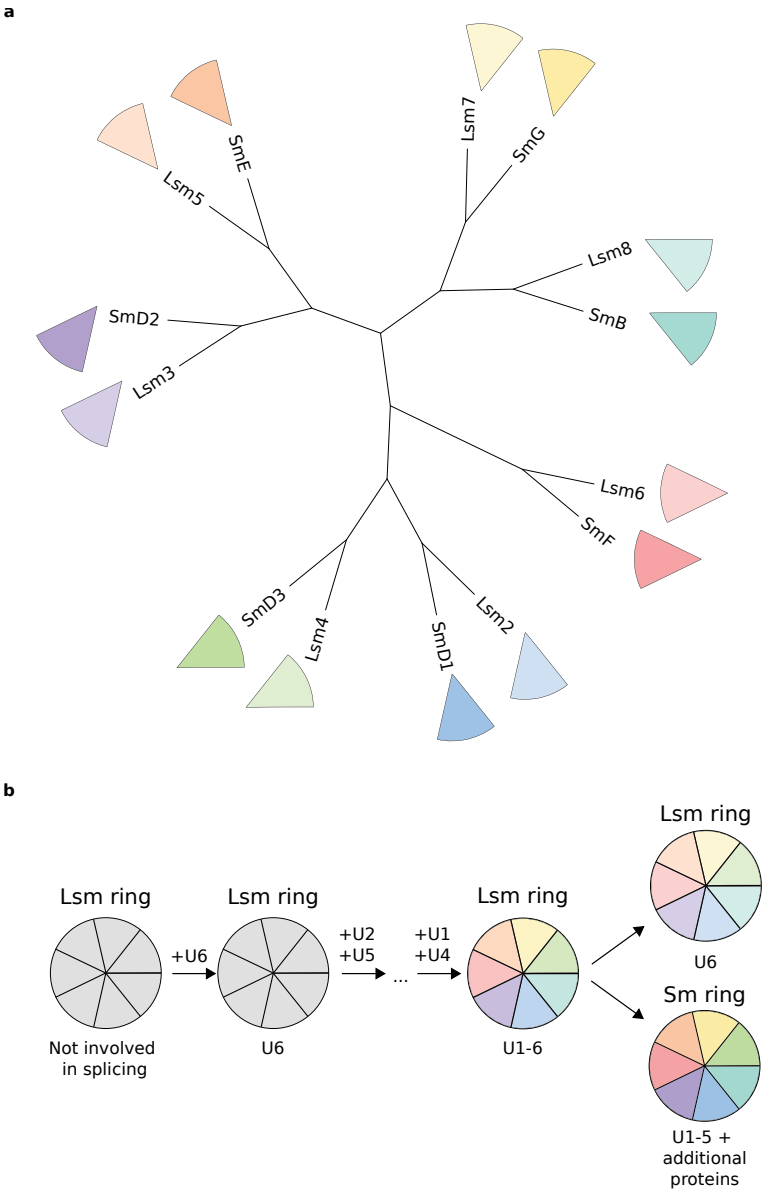


Figure 4.2 | Evolution of the Sm and Lsm rings. a, Tree depicting the scenario on the evolution of the spliceosomal Lsm and Sm proteins, as proposed in Veretnik *et al.* (Veretnik et al., 2009). **b**, Possible scenario for the evolution of the Lsm and Sm rings. A homoheptameric Lsm ring interacted with the *trans*-acting U6 snRNA, thereby facilitating splicing of degenerating self-splicing introns. While the Lsm ring became heteromeric upon duplication and subfunctionalisation of the Lsm protein, the *trans*-acting U2 and U5, which all originated from the introns, and U1 and U4 snRNAs formed RNP complexes with the Lsm ring. Upon duplication of the ring, U6 snRNA was bound by the Lsm ring, whereas the other snRNAs formed a complex with the newly formed Sm ring, followed by the addition of other proteins.

of prokaryotic DeaD, because of their similar function in translation regulation (Anantharaman et al., 2002). The U5 snRNP-specific protein Brr2 is part of the Ski2p-LHR family within the SFII superfamily, whose members typically function in the exosome (Anantharaman et al., 2002).

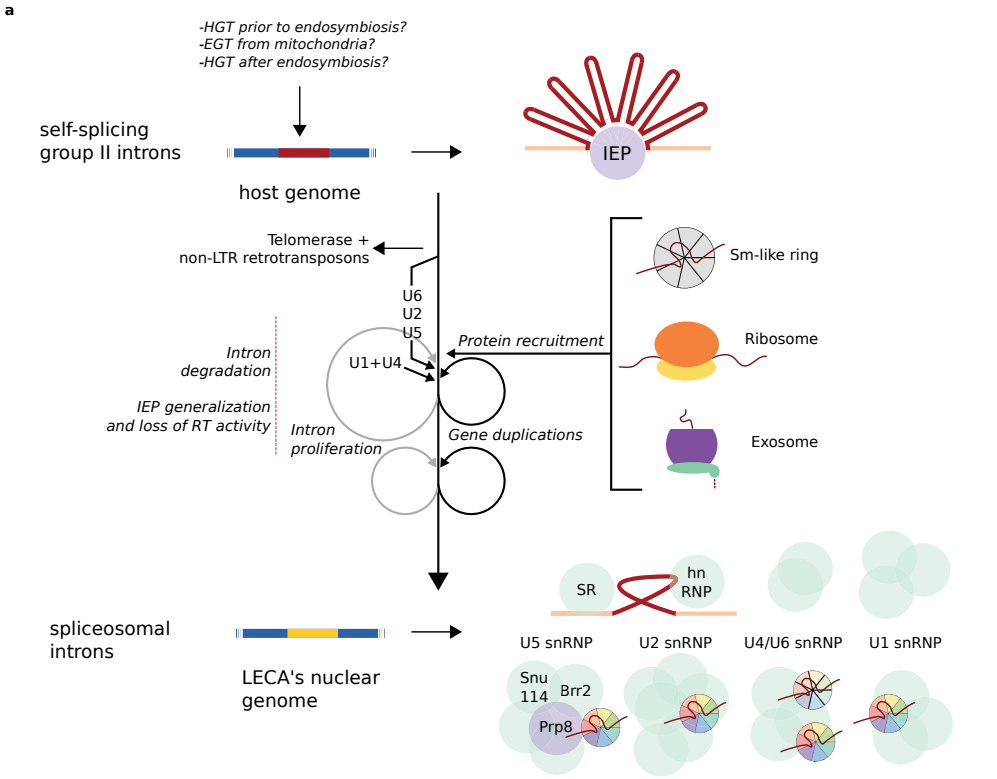
Another protein in U5 snRNP is the aforementioned GTPase Snu114, which interacts with Brr2 and Prp8 and is located near the catalytic site. Snu114 was already present in LECA (Collins and Penny, 2005) and is homologous to the ribosomal translocase EF-2 (Fabrizio et al., 1997). Apparently, multiple proteins involved in RNA degradation and translation were recruited into the spliceosome.

The SR splicing regulator proteins are characterised by an RNA recognition motif, which is also present in multiple prokaryotes, especially cyanobacteria (Califice et al., 2012). A phylogeny based on these motifs pointed to a single origin for SR proteins as a sister group to the SR-like atypical RNPS1/SR45 proteins, albeit with marginal support (Califice et al., 2012). Moreover, the radiation into three SR families and a family comprising the RNPS1/SR45 proteins likely had occurred before LECA. This example emphasises the importance of gene duplications in the origin of the spliceosome, as do the evolutionary histories of Sm proteins and helicases.

The evolutionary history of many other spliceosomal proteins has been clarified to a lesser extent. The exact source of each component, i.e. whether it was present in the archaeal host, the bacterial endosymbiont, was acquired later via HGT or was a unique eukaryotic invention, is obscure as well. The aforementioned examples demonstrate that duplicates of proteins active in other RNA processes in the first stages of eukaryogenesis supplemented the group II intron core in the emerging spliceosome. Subsequent expansions of these protein families resulting in many paralogues within the spliceosome contributed to the vast complexity of the machine (**Figure 4.3a**).

Order of events

Several papers put forward a speculative order of events that led to the emergence of the spliceosome. The starting point for these scenarios is the presence of self-splicing group II introns, including their maturases, in the host genome. For example, Anantharaman *et al.* proposed that the Sm proteins were recruited by the self-splicing introns as protein cofactors, followed by RNA helicases, of which some had an exosomal function (Anantharaman et al., 2002). The subsequent partial degeneration of the introns resulted in the snRNAs that partially replaced the introns themselves in the splicing machinery. On the other hand, the scenario of Martin and Koonin starts with the decay of self-splicing introns, requiring the recruitment of group II-derived RNAs, which evolved into the snRNAs, and associated Sm proteins (Martin and Koonin, 2006). Subsequently, additional proteins were added to this spliceosomal core. The model of Veretnik *et al.* also begins with RNA components, at least U6 snRNA, associated with a homomeric, and later heteromeric Lsm ring (Veretnik et al., 2009). The interaction between U6 snRNA and the Lsm ring could according to this scenario be seen as a 'frozen event'. The addition of other snRNAs, which became later on accompanied by the Sm ring, was the next step. Other components were added to the spliceosome successively. These scenarios differ most in


a
-HGT prior to endosymbiosis?
-EGT from mitochondria?
-HGT after endosymbiosis?
self-splicing
group II introns
IEP
host genome
Telomerase +
non-LTR retrotransposons
U6
U2
U5
U1+U4
Protein recruitment
Sm-like ring
Ribosome
Exosome
Intron
degradation
IEP generalization
and loss of RT activity
Intron
proliferation
Gene duplications
spliceosomal
introns
LECA's nuclear
genome
SR
hn
RNP
U5 snRNP
U2 snRNP
U4/U6 snRNP
U1 snRNP
Snu
114
Brr2
Prp8

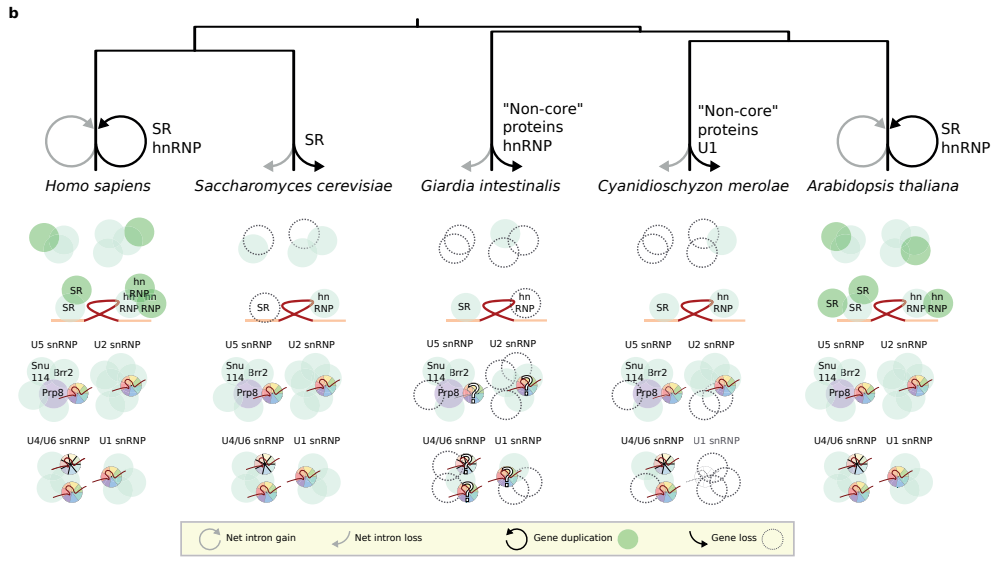

b
SR
hnRNP
SR
"Non-core"
proteins
hnRNP
"Non-core"
proteins
U1
SR
hnRNP
Homo sapiens
Saccharomyces cerevisiae
Giardia intestinalis
Cyanidioschyzon merolae
Arabidopsis thaliana
U5 snRNP
U2 snRNP
U4/U6 snRNP
U1 snRNP
Net intron gain
Net intron loss
Gene duplication
Gene loss

Figure 4.3 | Evolution of the spliceosome. a, Origin of the spliceosome during eukaryogenesis. The major steps resulting in the domestication of self-splicing introns in the early eukaryotes are depicted. **b**, Subsequent evolution after eukaryogenesis resulting in the more complex or simple spliceosomes in five diverse eukaryotes. Besides the gain or loss of notable proteins, the net loss or gain of introns is depicted for each lineage. The internal branches seemed to have experienced no large change of intron density (Csuros et al., 2011). The circles, except Prp8, Snu114 and Brr2, represent an arbitrary number of proteins. The question marks in *Giardia*'s Lsm and Sm rings reflect the ambiguity about their exact composition (Collins and Penny, 2005; Veretnik et al., 2009).

their proposed timing of the origin of snRNAs as distinct units. The models have in common that they regard Sm proteins as early additions to the spliceosome, as they are at the core of the complex.

The timing of the decay of self-splicing introns to spliceosomal ones, on the other hand, differs in these scenarios. Since group II introns have not been detected in nuclear genomes, all introns were apparently converted to spliceosomal introns or completely lost at some point before LECA. Complications with the expression of the targeted gene that arise when a group II intron is integrated in a nuclear gene were suggested to have caused their disappearance (Chalamcharla et al., 2010; Qu et al., 2014). However, their presence in non-coding regions would probably not have posed a challenge, implying that this cannot be a sufficient explanation (Doolittle, 2014). Although the low intracellular Mg^{2+} concentration in eukaryotes may have posed a barrier to group II introns in eukaryotic genomes, including protein-coding genes (Truong et al., 2015), it does not seem an impossible barrier to overcome, especially given that splicing of group II introns can be induced in the cytosol in yeast (Chalamcharla et al., 2010; Qu et al., 2014). Therefore, a more complete and sufficient explanation remains to be postulated.

Spliceosomal diversity after eukaryogenesis

Evidently, much research has focused on the many steps leading to the complex nature of the spliceosome in LECA. Nevertheless, the lack of access to intermediate stages poses a challenge to precisely reconstruct the evolution of the spliceosome. The wide diversity of eukaryotic spliceosomes provides a rich source of complementary data that show both further complexification as well as simplification of the spliceosome (**Figure 4.3b**). The occurrence of these processes has implications for our understanding of the origin of the spliceosome.

Increase in complexity

In at least two lineages the spliceosome has become more complex. The most prominent complexification is the expansion of splicing regulator proteins, which are involved in the recognition of exons and introns, in plants and animals. The SR family has expanded in multicellular eukaryotes, especially in plants (Barbosa-Morais et al., 2006; Richardson et al., 2011). Angiosperms have typically around twenty SR proteins, animals about ten and protists two or three (Richardson et al., 2011). Also the number of hnRNP proteins has increased in multicellular organisms, which is even more pronounced than the SR family expansion (Busch and Hertel, 2012). The greater hnRNP diversity is especially prominent

in vertebrates, whose genomes encode between twenty and forty of these proteins (Barbosa-Morais et al., 2006). Other animals and plants typically have between ten and fifteen hnRNPs, which is much more than the one hnRNP present in yeast (Barbosa-Morais et al., 2006; Busch and Hertel, 2012). Furthermore, other regulatory factors, such as ELAV-like and CELF proteins and kinases that phosphorylate SR proteins, have expanded in vertebrates (Barbosa-Morais et al., 2006; Tang et al., 2012). The diversification of these sets of proteins had already occurred before the last common ancestor of metazoans and the subsequent expansion in vertebrates is proposed to originate from the whole-genome duplications (Barbosa-Morais et al., 2006; Tang et al., 2012). Genome duplications may account for the extensive SR family expansion in plants as well (Richardson et al., 2011).

Although the high number of alternative splicing events in animals relative to other eukaryotes could be related to the expansion of the splicing regulator repertoire in these organisms (Roy and Irimia, 2014), this is not evidently the case in plants (Richardson et al., 2011). It is believed that due to the increased SR and hnRNP repertoire non-optimal splice sites were tolerated, since purifying selection on splice site sequences was relaxed (Busch and Hertel, 2012). The differences in the splicing regulator repertoire might underlie the differential preference for exon skipping in animals, compared to intron retention in plants and other eukaryotes as alternative splicing mechanism (Roy and Irimia, 2014).

Reduction in complexity

In many other eukaryotic lineages the evolution of the spliceosome is characterised by simplification. Both the loss of some subunits and the complete loss of the spliceosome have occurred. Based on draft genomes, introns and spliceosomal genes seem to be completely absent in a few microsporidia species (Akiyoshi et al., 2009; Cuomo et al., 2012) and in a diplomonad species (Andersson et al., 2007). The complete nucleomorph genome of a cryptophyte species also demonstrated a complete loss of introns and spliceosomal RNAs and proteins (Lane et al., 2007).

In contrast with the aforementioned loss of both the major and minor spliceosome and corresponding introns, the loss of only the latter has been more common. The minor spliceosome is present in representatives of all eukaryotic supergroups, but has at least 9 times been lost during eukaryotic evolution (López et al., 2008; Russell et al., 2006). This loss is accompanied by a loss of U12-type introns, which can either be a complete loss of these introns or a conversion to U2-type introns (Bartschat and Samuelsson, 2010). The latter can be accomplished by mutations or a shift of the splice site, which were both found in the lineage leading to *Caenorhabditis elegans* (Bartschat and Samuelsson, 2010). Losses of U12-type introns are more frequently observed than conversions (Lin et al., 2010).

In addition to the complete loss of either the minor spliceosome or both types of spliceosomes, reduced spliceosomes have been observed and more thoroughly analysed in at least three lineages. This loss of spliceosomal subunits is associated with a lower number of introns (Hudson et al., 2012; Korneta et al., 2012; Stark et al., 2015). Numerous proteins can be absent from these spliceosomes. For example, the classical SR proteins appear to

have been lost in some lineages, including *Saccharomyces cerevisiae* (Busch and Hertel, 2012; Collins and Penny, 2005). Even proteins that can be considered to belong to the core of the complex, like the snRNAs, Sm proteins and some other snRNP proteins, are not present in all eukaryotes. For instance, some organisms can perform splicing without a full set of Sm/Lsm proteins (Collins and Penny, 2005; Veretnik et al., 2009). The snRNAs of the diplomonad *Giardia lamblia* have characteristics of both major and minor spliceosomal snRNAs and therefore the reduced spliceosome of this organism is suggested to be a hybrid (Hudson et al., 2012, 2015). Many spliceosomal proteins are missing in this diplomonad, but most U2 snRNP proteins and the core U5 snRNP proteins are still present (Korneta et al., 2012). A similar reduction pattern has been observed in the red alga *Cyanidioschyzon merolae* (Hudson et al., 2015; Stark et al., 2015). The proteins remaining in both organisms correspond with the catalytic core of the spliceosome (Hudson et al., 2015). On top of that, *C. merolae* seems to perform splicing without a U1 snRNP, as both U1 snRNA and U1 snRNP-specific proteins appear to be missing (Stark et al., 2015). This loss is hypothesised to mimic an ancestral state during eukaryogenesis in which U1 had not yet been added to the primordial spliceosome (Hudson et al., 2015). The observations that U1 snRNA does not have a clear analogue in group II introns and that it can be lost in the spliceosome, are arguments for a later addition of the U1 snRNP to the early spliceosome.

Evolutionary models of spliceosomal evolution

Numerous scenarios for the evolution of the spliceosome have been suggested. Usually this concerns a description of what happened, but to truly comprehend the evolution of the spliceosome a transition has to be made from a mere description to addressing the evolutionary forces that shaped this complex machine. A number of hypotheses concerning these forces have been proposed, as mentioned in the introduction. They propose that either the addition or loss of each component of a complex is an adaptation or that solely neutral processes are responsible for the shifts in complexity.

Adaptive model

Since the establishment of the power of natural selection, adaptive explanations for biological observations have been the most prominent and widely accepted. Many biological papers propose an adaptive explanation for their observations, albeit often implicitly. Such explanations can in many cases be criticised as being just-so stories that lack proper evidence (Koonin, 2016). The role of natural selection in reductive evolution is widely established, but this is not the case for its role in the increase in complexity. In that case, each addition should have been selected for. The function of the spliceosome is clear, namely removing spliceosomal introns from pre-mRNAs. The large compositional complexity is believed to have arisen to make splicing more efficient and precise and to stabilise the complex (Koonin, 2016; Martin and Koonin, 2006; Speijer, 2011; Wahl et al., 2009). However, the spliceosome seems to perform worse in these respects compared to the self-splicing capacity of group II introns (Stoltzfus, 1999). Furthermore, in many adaptive scenarios an innovation is necessary to compensate for a detrimental event, which is of

course maladaptive, such as the evolution of snRNAs to compensate for degenerated introns and the higher complexity needed to cope with the expansion of introns into genes and the loss of clearly defined exon-intron boundaries (Anantharaman et al., 2002; Martin and Koonin, 2006; Zimmerly and Semper, 2015). Also a nucleus would be selected for due to the emergence of introns in genes, which resulted in the detrimental synthesis of aberrant proteins (López-García and Moreira, 2006; Martin and Koonin, 2006). Another proposed advantage of a complex spliceosome is that it enables better regulation called fine-tuning, which is especially the case in organisms that have extensive alternative splicing (Speijer, 2011). An issue related to the emergence of the spliceosome is the origin of spliceosomal introns. The main adaptive value of these sequences is proposed to be an expansion of the proteome by facilitating exon shuffling and alternative splicing (Speijer, 2011). This basically means that the increased genomic complexity due to introns is to enable an increase in complexity. Note that in all these adaptive scenarios the present-day function does not necessarily correspond to why the system originated in the first place (Koonin, 2016). In general, many adaptive roles for the spliceosome have been proposed, all giving reasons why splicing could be adaptive once you have it, yet failing to provide a reason for its very origin.

Neutral model

In the constructive neutral evolution model the increase in complexity can be seen as a 'drunkard's walk' into the more complex possibilities of a system (Stoltzfus, 2012). The concept of presuppression is central in this 'walk' (Gray et al., 2010). This means that a certain factor (A) is bound by another factor (B), which does not affect the function of the former. The effects of mutations in factor A that would normally impair its function, are now suppressed by the interaction with factor B. These previously deleterious mutations are therefore now neutral and can become fixed in the population. This results in the dependence of factor A on factor B. In this way other mutations that strengthen this dependence may occur, resulting in a ratchet-like process. Reversal to the ancestral, independent state is possible, but given that there are more possibilities to increase this dependence, this is less likely. Via this mechanism of presuppression neutral evolution could result in a ratchet-like increase in complexity (Lukeš et al., 2011; Mast et al., 2014).

A well-established example of a similar neutral process resulting in increased complexity is subfunctionalisation of paralogues after gene duplication. A combination of constructive neutral evolution and subfunctionalisation may explain the formation of a heteromeric protein complex from a homomeric state. Finnigan *et al.* demonstrated this experimentally for the evolution of the fungal vacuolar H^+-ATPase ring and suggested that this could have been the case in other multi-paralogue complexes as well (Finnigan et al., 2012). As the spliceosome comprises multiple paralogues, such as the Sm proteins and helicases (Anantharaman et al., 2002; Lynch, 2012; Veretnik et al., 2009), a similar mechanism might have been operating in its evolution towards greater complexity as well.

It should be noted that it is difficult to classify an increase in complexity as neutral. As pointed out by Lynch (Lynch, 2007), each embellishment makes a biological system more

susceptible to inactivation by mutations. The additional feature should either provide a direct advantage to become fixed in the population or selection should be inefficient to remove this variant due to the larger effect of genetic drift in case of a small effective population size (Lynch, 2007). The latter is believed to have been the case during eukaryogenesis and this may explain the many complex characteristics of eukaryotes, including complex machineries such as the spliceosome (Koonin, 2009; Lynch, 2007, 2012).

In a neutral scenario the spliceosome would have evolved from the addition of new RNAs and proteins that do not improve the efficacy of the splicing reaction to the catalytic core inherited from the group II intron ancestor. Moreover, at some point the structural RNA elements in the group II introns were replaced by fragments of other group II introns that acted as *trans*-acting RNAs. These primordial snRNAs and an IEP that acts as a general maturase, which does not only assist splicing of its own intron, would have made the RNA domains of the introns and a dedicated IEP redundant. In this way previously deleterious mutations in the introns are now presuppressed by the action of this *trans*-acting RNP complex, resulting in the loss of self-splicing features. This primordial spliceosome would also allow the spread of inactive group II introns and intronised sequences unrelated to group II introns in the genome. The already established nucleus would have prevented aberrant protein synthesis upon invasion of the introns into protein-coding genes. The emergence of introns in genes would have made the eukaryotic lineage dependent on the spliceosome.

Numerous proteins were added to the spliceosomal core during eukaryogenesis. Many of these are clearly derived from proteins that already had an RNA-binding function (Anantharaman et al., 2002; Lukeš et al., 2011; Mura et al., 2013). Coincidental interactions with these proteins could have caused presuppression and subsequent dependence, increasing the complexity of the spliceosome without a clear benefit (Lukeš et al., 2011; Mast et al., 2014; Stoltzfus, 1999). The expanding repertoire of splicing regulatory proteins would have enabled the decay of clearly defined exon-intron boundary features, leading to dependence as well (Lukeš et al., 2011). In these ways, the present-day spliceosome would have been built up "step by unselected step" (Lukeš et al., 2011).

An interplay between neutral and adaptive evolution explains spliceosomal evolution

The lack of clear direct benefits of a complex splicing machinery in the early eukaryotes is a strong argument against an adaptive scenario for its evolution. The only plausible direct benefits are compensations for maladaptive features. In light of the small effective population size inferred to have been present during eukaryogenesis based on the fixation of introns (Koonin, 2009; Rogozin et al., 2012) or paralogous genes (Makarova et al., 2005), or on the proposed early mitochondrial endosymbiosis event (Garg and Martin, 2016; Koonin, 2015; Makarova et al., 2005), this is definitely a possibility. However, a neutral scenario in which these features were tolerated by a more complex spliceosome remains more likely, because a maladaptive intermediate stage does not need to be invoked. Other advantages of spliceosomal introns and concomitantly the spliceosome, like enabling alternative splicing and fine-tuning, work on the long term. These are fully exploited only

in multicellular eukaryotes, making it therefore unlikely that this system has evolved for this particular purpose. The small effective population sizes before LECA, and in animals and plants seem to be largely responsible for the drastic increases in complexity of the spliceosome. A role of adaptive processes is not excluded and likely has played a role in certain interactions, but for each new feature the null-hypothesis of neutral, random evolution should convincingly be disproven (Koonin, 2016). Natural selection has definitely played a role in the simplification of introns and the splicing machinery that can be observed in multiple lineages. The selective pressure for streamlining that characterises organisms like yeast and *Giardia* has resulted in a significant loss of introns and spliceosomal components (Collins and Penny, 2005; Hudson et al., 2015; Korneta et al., 2012). Clearly, complexification in this process is not truly irremediable and can be overcome by natural selection.

The scenario we infer corresponds to a biphasic pattern of evolution, in which a short explosive innovation phase is followed by a much longer reductive phase (Cuypers and Hogeweg, 2012; Wolf and Koonin, 2013). Most of the complexity of the spliceosome emerged during eukaryogenesis. Subsequently, its complexity stabilised or decreased in multiple lineages. However, in the lineages leading to plants and animals, and within the animals the lineage leading to the vertebrates, additional periods of rising complexity took place. Most of the machine's complexity does not seem to evolve gradually at a somewhat constant rate, but instead in rapid bursts. This alternation of periods of increasing and decreasing complexity has also been described for many other processes (Wolf and Koonin, 2013). Although often observed, a biphasic pattern does not offer an explanation *per se*. One potential explanation for these patterns that has been put forward is that complex machines arising through e.g. constructive neutral forces can in subsequent evolutionary time provide an advantage in terms of adaptation in surviving lineages. This explanation has been argued for as a special case of multilevel selection (Doolittle, 1987, 2016) and biphasic genome evolution is one of the most striking outcomes of computational modelling of the interplay between network and genome evolution (Cuypers and Hogeweg, 2012).

Conclusion and future directions

The spliceosome is a complex molecular machine that arose during eukaryogenesis and removes introns from pre-mRNAs, which is required to prevent the production of aberrant proteins. The spliceosome consists of five snRNPs, each comprised of an snRNA and proteins, and additional proteins. There is ample evidence that both the spliceosomal core and the spliceosomal introns originated from self-splicing group II introns, which are widely believed to have been transferred from the mitochondrial endosymbionts to the host DNA. The snRNAs, at least U2, U5 and U6, are likely derived from fragmented group II introns and the U5-snRNP-specific protein Prp8 evolved from the IEP of these introns. Sm proteins, helicases and other proteins were at some point recruited to the spliceosomal core. This addition and the extensive expansion of especially Sm proteins and helicases have drastically increased the complexity of the spliceosome during eukaryogenesis. Apparently, all group II introns in the nucleus were either lost or converted to

spliceosomal introns before LECA. During eukaryotic evolution a pronounced increase in spliceosomal complexity occurred in plants and animals, which mainly involved the regulatory proteins. In other lineages the spliceosome simplified, with U2 and U5 snRNP proteins being the least affected, and concomitantly the number of introns decreased.

The spliceosome-like machineries involved in group II intron splicing in some eukaryotic plastids and mitochondria could be an interesting model for the evolution of *trans*-splicing complexes from self-splicing group II introns, as they are less complex and have evolved more recently. Splicing facilitated by general maturases and other protein factors in plant organelles (Schmitz-Linneweber et al., 2015) and by an RNP complex comprising a *trans*-acting RNA and protein factors in the plastids of the green alga *Chlamydomonas reinhardtii* (Reifschneider et al., 2016), and suggested RNP complexes for excising so-called group III introns in the plastids of the excavate *Euglena* (Zimmerly and Semper, 2015) are interesting examples of recurrent evolution. These might shed more light on the origin of the spliceosome.

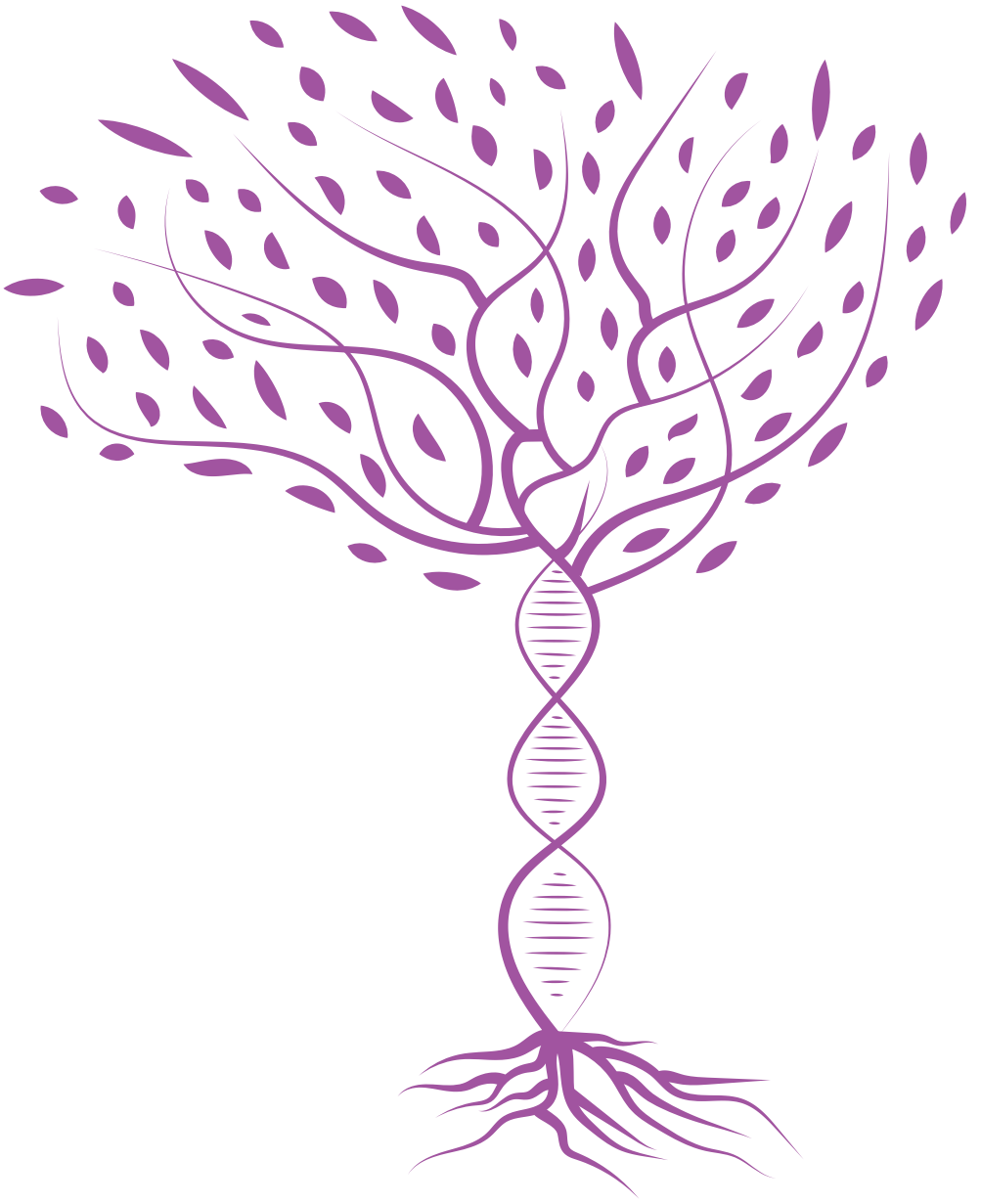
The spliceosome is one of the most complex machines that emerged during eukaryogenesis. Other complex features that originated in the eukaryotic lineage are for example the nuclear pore complex, an elaborate endomembrane system, the RNA interference machinery and the kinetochore (Gould et al., 2016; van Hooff et al., 2017; Rout and Field, 2017; Shabalina and Koonin, 2008), to name a few. Moreover, multiple machineries inherited from the prokaryotic ancestors increased in complexity, like the ribosome, proteasome and exosome (Lukeš et al., 2011; Lynch, 2012; Veretnik et al., 2009). These examples underscore the contribution of gene duplications to increased machine complexity (Lynch, 2012; Veretnik et al., 2009). It is tempting to speculate that the vast expansion of protein families reflects whole-genome duplications or hybridisation events, perhaps in syncytial early eukaryotes (Garg and Martin, 2016). The importance of neutral processes in the evolution of one of the most complex machines suggests that neutral evolution has contributed significantly to the complexity of other less complex machines as well. A profound reconsideration of the evolutionary forces that shaped these complexes is therefore desired, in which neutral processes should be considered as null-hypothesis.

Authors' contributions

J.V. wrote the manuscript, which was read, edited and approved by both authors.

Acknowledgements

We thank Bas Dutilh, other members of the Theoretical Biology and Bioinformatics group and the three reviewers (W. Ford Doolittle, Eugene V. Koonin and Vivek Anantharaman) for useful advice, comments and suggestions on the manuscript and figures. This work is part of the research programme VICI with project number 016.160.638, which is (partly) financed by the Netherlands Organisation for Scientific Research (NWO).



5

Integrating phylogenetics with intron positions illuminates the origin of the complex spliceosome

Julian Vosseberg, Daan Stolker, Samuel H. A. von der Dunk, Berend Snel

Preprint at bioRxiv, 2022

Abstract

Eukaryotic genes are characterised by the presence of introns that are removed from the pre-mRNA by the 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 work has established that these introns and the spliceosomal core originated from self-splicing introns in prokaryotes. Yet it remains largely elusive how the spliceosomal core expanded by recruiting many additional proteins. In this study we use phylogenetic analyses to infer the evolutionary history of the 145 proteins that we could trace back to the spliceosome in the last eukaryotic common ancestor (LECA). We found that an overabundance of proteins derived from ribosome-related processes were 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.

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 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 the 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 recognise 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 machineries 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 recognised by different spliceosome complexes. The vast majority of introns are of U2-type and are recognised by the major spliceosome; U12-type introns comprise a small minority (Moyer et al., 2020). The minor spliceosome specifically recognises U12-type introns and most proteins of the major spliceosome are also part of the minor spliceosome (Bai et al., 2021; Turunen et al., 2013). All snRNAs but U5 have a minor-spliceosome specific 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., 2022b) with a complex spliceosome, prokaryotes lack intragenic introns and a spliceosome, meaning that they 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 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., 2022b). The evolutionary histories of a few gene families in the spliceosome have been described (Anantharaman et al., 2002; Califice et al., 2012; Veretnik et al., 2009) and they suggest 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 machineries 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 human (**Supplementary Table 5.1**) and baker's yeast (**Supplementary Table 5.2**), 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. 145 spliceosomal orthogroups (OGs) could be traced to LECA (**Figure 5.1**, **Supplementary Table 5.3**). 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 human (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 the four minor-spliceosome specific snRNAs were also present in LECA.

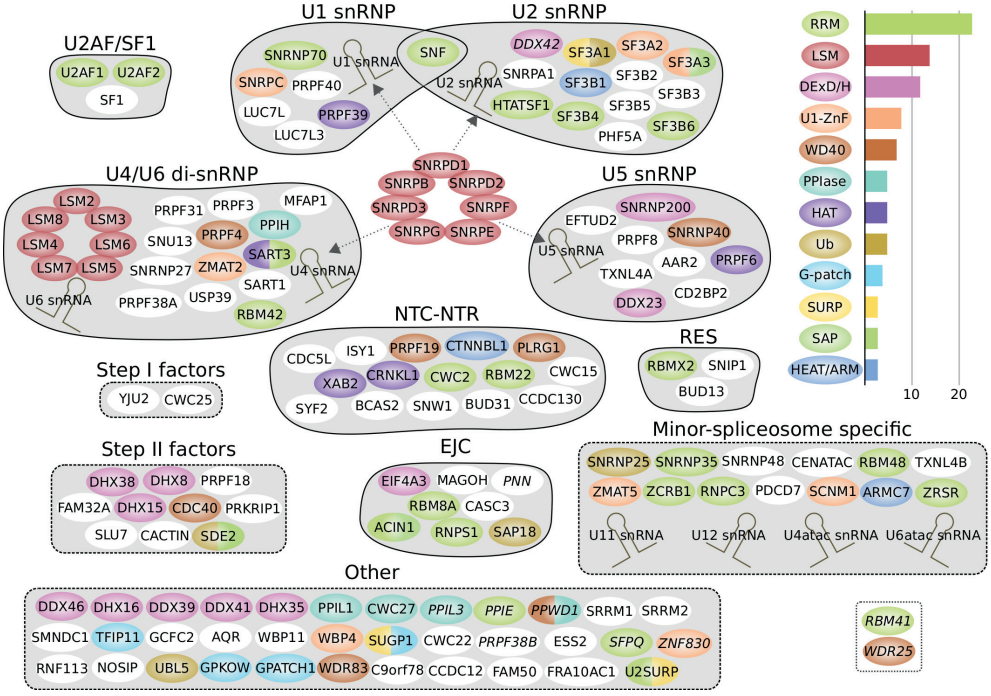


Figure 5.1 | The spliceosome inferred in LECA. The names of OGs with a lower confidence score are in italics (possibly spliceosomal in LECA, see Methods). The OGs are grouped based on the subcomplex they are in or another collection (dashed line), and they are coloured based on their domain composition. 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 spliceosome 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.

From intron-encoded protein to PRPF8

As described above, the U2, U5 and U6 snRNA and the U5-snRNP protein PRPF8, as well as parts of the introns, are remnants of self-splicing group II introns. This means that during eukaryogenesis a system containing only a single RNA (the intron itself) and one protein (IEP) transformed into an enormously complex spliceosome in LECA. In principle, the prokaryotic origins of this system could 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 were not clearly associated with a particular IEP type and their exact phylogenetic position in the IEP tree was unresolved (**Supplementary Figure 5.1a**).

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 were from multiple different IEP types (**Supplementary Figure 5.1b**). 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 5.4**). 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). 56 OGs were most closely related to another spliceosomal OG (**Figure 5.2a**) and therefore their pre-duplication ancestor was probably already part of the spliceosome. By collapsing such close paralogous clades of spliceosomal OGs we identified 102 ancestral spliceosomal units (**Supplementary Figure 5.2**). Duplications of 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 a 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 invention. 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 (**Figure 5.2b**). 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

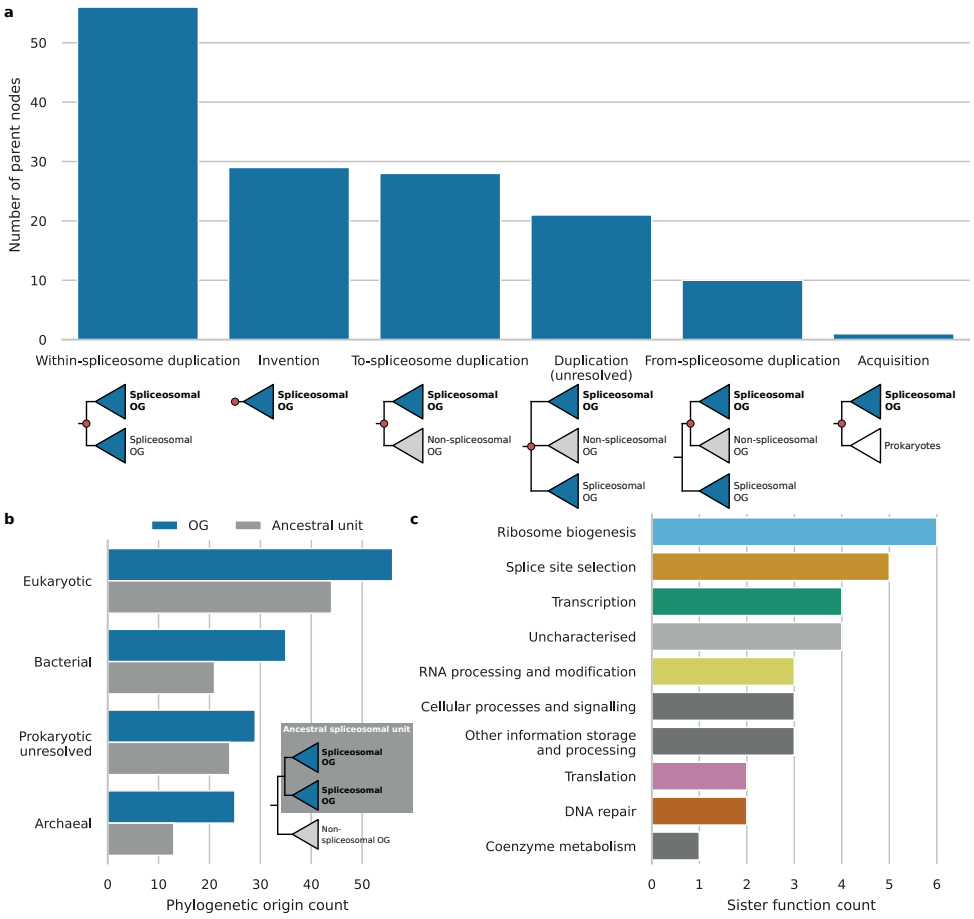


Figure 5.2 | Evolutionary history of spliceosomal proteins before LECA. a, Annotations of the parent nodes of spliceosomal OGs. These parent nodes are shown in red in the example trees below. **b**, Bar plot showing the phylogenetic origins of spliceosomal OGs and ancestral spliceosomal units. **c**, Functions of the sister OGs of ancestral spliceosomal units.

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 (**Figure 5.2c**), especially OGs from archaeal origin. The U5 snRNP protein EFTUD2 is a paralog of elongation factor 2 (**Figure 5.3a**), which catalyses 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). SNU13 and PRPF31 bind to U4 or U4atac snRNA (Nottrott et al., 2002). SNU13 is also part of the C/D-box snoRNP (Watkins et al., 2000) and PRPF31 originated from a C/D-box snoRNP protein (**Figure 5.3b**, **c**). The archaeal orthologs NOP5 and RPL7Ae are part of the functionally equivalent C/D box sRNP (**Figure 5.3d**), which is involved in ribosome biogenesis by modifying rRNA (Aittaleb et al., 2003; Breuer et al., 2021). The eukaryotic DDX helicases, of which six were part of LECA's spliceosome, evolved from prokaryotic DEAD and RHLE proteins, which also function in ribosome assembly (Charollais et al., 2004; Jain, 2008) (**Figure 5.3e**, **f**). A large group of related RNA helicases are the DHX helicases. The ancestral function of DHX helicases was probably related to ribosome biogenesis (**Figure 5.3g**). 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 characterised RNA-binding proteins which 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), emphasising 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 the U6(atac) binding function and was recruited into the spliceosome. Subsequent gene duplications resulted in the two types of heteromeric rings of Lsm/Sm proteins in the spliceosome (**Supplementary Figure 5.3**, Supplementary Information).

A substantial fraction of the LECA spliceosome OGs contains an RNA recognition motif (RRM) (**Figure 5.1**). 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 intra-spliceosome duplications (**Supplementary Figure 5.4a**). 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.

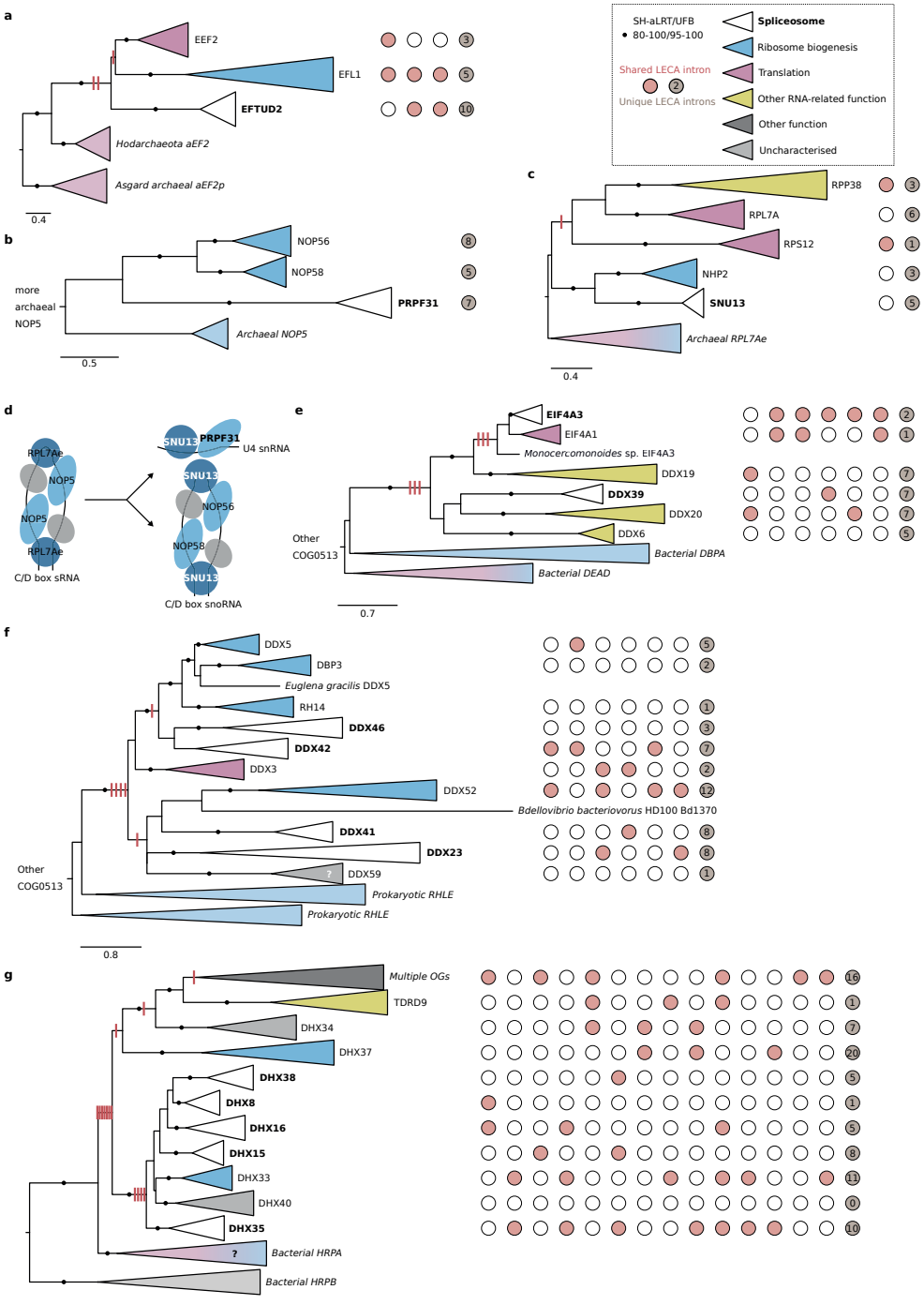

a
EEF2
EFL1
EFTUD2
Hodarchaeota aEF2
Asgard archaeal aEF2p
0.4
b
NOP56
NOP58
PRPF31
more archaeal NOP5
Archaeal NOP5
0.5
SH-aLRT/UFB
80-100/95-100
Shared LECA intron
Unique LECA introns
Spliceosome
Ribosome biogenesis
Translation
Other RNA-related function
Other function
Uncharacterised
c
RPP38
RPL7A
RPS12
NHP2
SNU13
Archaeal RPL7Ae
0.4
d
RPL7Ae
NOP5
NOP5
RPL7Ae
C/D box sRNA
SNU13
PRPF31
U4 snRNA
SNU13
NOP56
NOP58
SNU13
C/D box snoRNA
e
EIF4A3
EIF4A1
Monocercomonoides sp. EIF4A3
DDX19
DDX39
DDX20
DDX6
Other COG0513
Bacterial DBPA
Bacterial DEAD
0.7
f
DDX5
DBP3
Euglena gracilis DDX5
RH14
DDX46
DDX42
DDX3
DDX52
Bdellovibrio bacteriovorus HD100 Bd1370
DDX41
DDX23
DDX59
Other COG0513
Prokaryotic RHLE
Prokaryotic RHLE
0.8
g
Multiple OGs
TDRD9
DHX34
DHX37
DHX38
DHX8
DHX16
DHX15
DHX33
DHX40
DHX35
Bacterial HRPA
Bacterial HRPB

Figure 5.3 | 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 colour. SNU13 was present in both complexes in LECA. The grey 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 characterised but its phylogenetic profile is similar to minor-spliceosome specific proteins (de Wolf et al., 2021) (**Supplementary Figure 5.7b**). **g**, Phylogenetic tree of the DHX helicase family. **a-c**, **e-g**, Eukaryotic LECA OGs are collapsed and coloured 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.

Other large families that contributed substantially to the LECA spliceosome are the U1-type zinc finger and WD40-repeat families. The U1-type zinc finger family contains mainly spliceosomal OGs (**Supplementary Figure 5.5a**). WD40 repeats are present in many eukaryotic proteins with diverse functions. In contrast to the RNA handling functions described before, the proteins that seem to be closely related to the spliceosomal WD40 OGs are mainly involved in intracellular transport, cilia and histone modifications (**Supplementary Figure 5.5b**).

Many minor-spliceosome specific proteins are closely related to a major spliceosome protein

The major and minor spliceosome share many subunits (Bai et al., 2021; Turunen et al., 2013) and this was very likely also the case in LECA. We inferred 13 minor-spliceosome specific proteins in LECA (**Figure 5.1**). 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, SNRNP70 and U2AF1 respectively (**Supplementary Figure 5.4b**, **c**). RNPC3 is closely related to SNF but probably not as sister paralogs (**Supplementary Figure 5.4d**). The sister paralog of RNPC3 is RBM41, which is not well characterised. However, its phylogenetic profile corresponds with minor spliceosome OGs (**Supplementary Figure 5.7**). 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 Figure 5.4a**). ZMAT5 and SCNM1 are part 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, it is likely that these major and minor spliceosome equivalents are sister paralogs. The major spliceosome OG TXNL4A and minor spliceosome OG TXNL4B are clear sister paralogs (**Supplementary Figure 5.5c**). The sister paralog of the WD40-repeat protein CDC40, called WDR25 (**Supplementary Figure 5.5b**), has a presence pattern across eukaryotes that is typical of minor spliceosome OGs (de Wolf et al., 2021), like

RBM41 (**Supplementary Figure 5.7**). This protein has not been characterised 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 the duplication to the minor-spliceosome specific OG is considerably shorter than the one leading to the major spliceosome-specific OG (**Supplementary Figure 5.5d**). 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., 2022b). 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 analysing intron positions in spliceosomal OGs we can relate the duplications in the primordial spliceosome to the spread of the elements that they function on, the introns. We therefore 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 (**Figure 5.3**). Introns shared between spliceosomal paralogs were also found in the Lsm, PPIase and WD40 families (**Supplementary Figure 5.3a**, **Supplementary Table 5.5**). 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 (**Figure 5.3**, **Supplementary Table 5.5**). 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 subfunctionalisation completed multiple times after eukaryogenesis: U1A/U2B"

A notable difference between the LECA spliceosome and the human and yeast spliceosome 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). However, with the availability of more genomes the human and yeast proteins were shown with high confidence to be the result of separate gene duplications (Williams et al., 2013). Additional SNF duplications were identified in other animal lineages (Wil-

liams et al., 2013). We observed even more independent SNF duplications, 22 in total using our set of eukaryotic genomes (**Supplementary Figure 5.6a**). *Guillardia theta* even had an additional third one, probably from the secondary endosymbiont (Supplementary Information).

Drosophila SNF has a dual role in the spliceosome. It is both 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 human and yeast, U1A and U2B" have subfunctionalised and perform the respective functions as indicated by the snRNP in their name. To assess whether a similar subfunctionalisation 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" (**Figure 5.4**). We inferred 16 RRM loss events in U2B"-fate proteins (**Supplementary Figure 5.6b**). 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 proliferation of introns

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 work 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., 2022b). 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

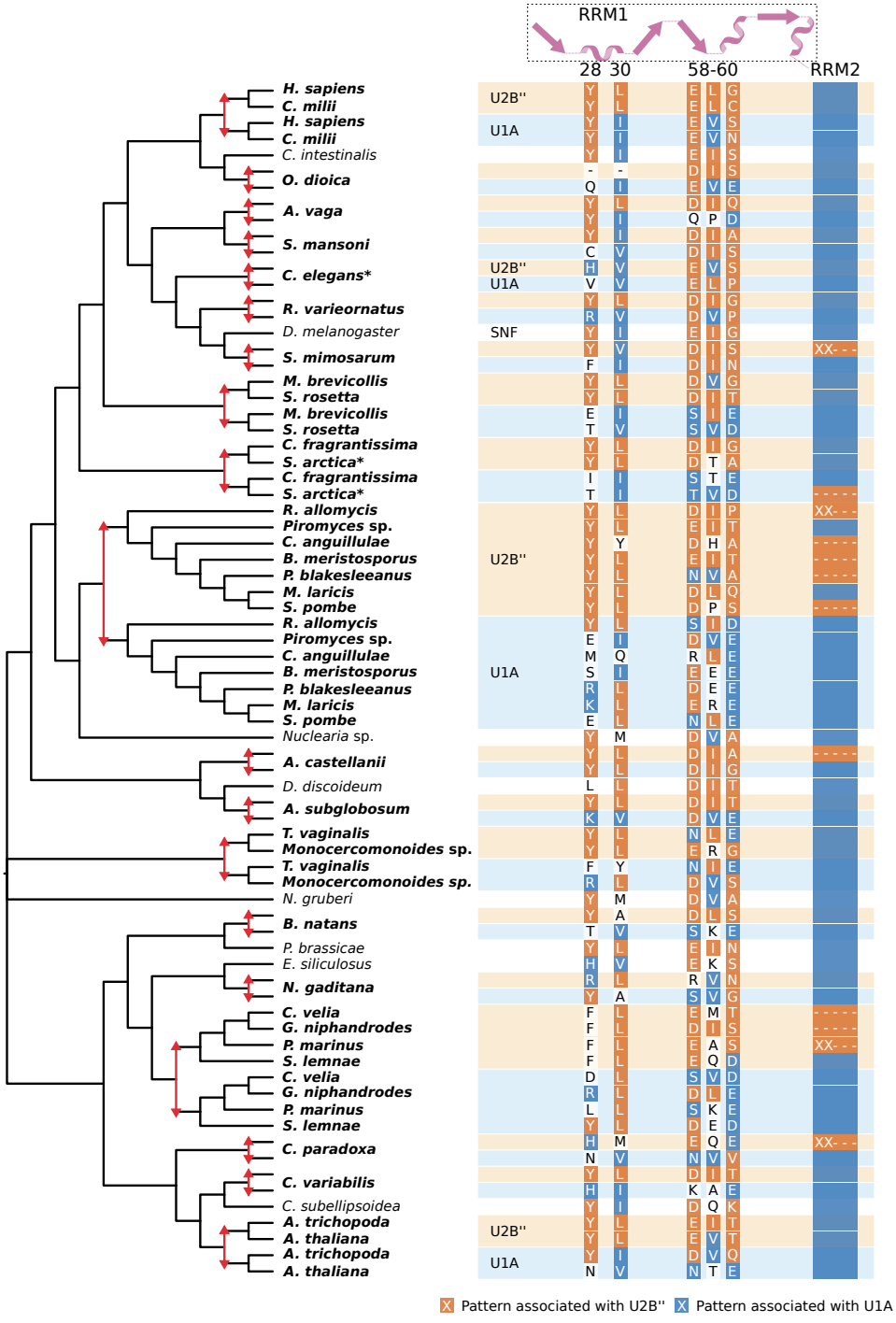

RRM1
28
30
58-60
RRM2
H. sapiens
C. milii
H. sapiens
C. milii
C. intestinalis
O. dioica
A. vaga
S. mansoni
C. elegans*
R. varieornatus
D. melanogaster
S. mimosarum
M. brevicollis
S. rosetta
M. brevicollis
S. rosetta
C. fragrantissima
S. arctica*
C. fragrantissima
S. arctica*
R. allomycis
Piromyces sp.
C. anguillulae
B. meristosporus
P. blakesleeanus
M. laricis
S. pombe
R. allomycis
Piromyces sp.
C. anguillulae
B. meristosporus
P. blakesleeanus
M. laricis
S. pombe
Nuclearia sp.
A. castellanii
D. discoideum
A. subglobosum
T. vaginalis
Monocercomonoides sp.
T. vaginalis
Monocercomonoides sp.
N. gruberi
B. natans
P. brassicae
E. siliculosus
N. gaditana
C. velia
G. niphandrodes
P. marinus
S. lemnae
C. velia
G. niphandrodes
P. marinus
S. lemnae
C. paradoxa
C. variabilis
C. subellipsoidea
A. trichopoda
A. thaliana
A. trichopoda
A. thaliana
U2B''
U1A
SNF
XX- - -
- - - - -
X Pattern associated with U2B''
X Pattern associated with U1A

Figure 5.4 | Independent gene duplications and recurrent sequence evolution in the SNF family. The reconciled tree (see **Supplementary Figure 5.6a** for the full tree) shows the positions of gene duplications (red arrows) and the species names with duplicates are in bold. The coloured rectangles next to the species names correspond with the predicted fate of the duplicates. The most prominent recurrent patterns are depicted with colours 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 is 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.

complexity did not coincide with the increase in intron numbers but followed it instead. We propose a scenario in which intragenic introns emerged early in eukaryogenesis and the complex spliceosome relatively late.

From group II introns to a complex spliceosome

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). Group II introns are present in several mitochondrial and plastid genomes (Zimmerly et al., 2001). Notwithstanding the extent of horizontal gene transfer of these introns among eukaryotes, 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 in which group II introns were present in the archaeal genome before the mitochondrial endosymbiosis also plausible (Vosseberg and Snel, 2017; Vosseberg et al., 2022b).

Some self-splicing introns acquired the capacity to aid the splicing of other introns. Fragments of these introns evolved into the *trans*-acting snRNAs U2/U12, U5 and U6/U6atac. IEP became a general maturase and lost its RT activity. Degeneration of self-splicing introns resulted in primordial spliceosomal introns that required these snRNAs and the general PRPF8 maturase for splicing.

Proteins involved in the assembly and functioning of another large ribonucleoprotein in the cell, the ribosome, became part of the primordial spliceosome, supplemented with other RNA-binding proteins. The evolutionary link with the ribosome emphasises the comparable composition as a ribonucleoprotein with catalysing RNA molecules (ribozymes). In contrast with the aforementioned spliceosomal snRNAs, the U1/U11 snRNA and U4/U4atac 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., 2021a). 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, 2016a; Vosseberg et al., 2021a).

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 U2-type (Vosseberg et al., 2022b). Different scenarios have been postulated for the emergence of two types of introns (Burge et al., 1998). In some scenarios the 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., 2022b). The enormous overlap in composition between the major and minor spliceosome (Bai et al., 2021; Turunen et al., 2013) 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 all snRNAs but U5 have equivalents in the other spliceosome type. This suggests that the divergence between the major and minor spliceosome 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 substantial loss of spliceosomal genes that were part of the LECA spliceosome (**Supplementary Figure 5.7**). 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 different 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 subfunctionalisation 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 subfunctionalisation 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 (Delaney et al., 2014; Williams et al., 2013). 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., 2021a) from different sources (Deutekom et al., 2019; Huerta-Cepas et al., 2016a; Zaremba-Niedzwiedzka et al., 2017). 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 (Farag et al., 2021; Huang et al., 2019; Imachi et al., 2020; Liu et al., 2018, 2021; Seitz et al., 2019; Sun et al., 2021; Tully et al., 2018; Wu et al., 2022; Zhao and Biddle, 2021), 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 26 February 2020, only including manually reviewed proteins (**Supplementary Tables 5.1 and 5.2**). 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; Sales-Lee et al., 2021; de Wolf et al., 2021). Initial evolutionary scenarios of these proteins were inferred based on the approach of Van Hooff *et al.* (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; (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., 2021a) 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 OG were present in both the human and yeast spliceosome, 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 human, 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-characterised, they were classified as possibly spliceosomal. In case of multiple functions, the OG was discarded. The reconstruction of spliceosome OGs in LECA is summarised in **Supplementary Table 5.3**.

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 analysing 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., 2021a). 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 5.6**. 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 (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.

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 (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., 2016a) 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., 2021a), 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 (HMMs) (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 21 April 2022.

EFTUD2

The EF2 family has underwent 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., 2021a; van Wijk and Snel, 2020) 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 euNOGs that were 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 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 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 shortest Opimoda and Diphoda sequence 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 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 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 inferred (LG+R4 model). Per OG the shortest Opimoda and Diphoda sequence was 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 5.5**), 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), the full-length sequences were used for this alignment. Intron positions were mapped onto the alignment using the method described before (Vosseberg et al., 2022b). 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., 2022b). Pre-duplication introns were inferred using Dollo parsimony.

Recurrent duplication and subfunctionalisation 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).

Data availability

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

Acknowledgements

We thank the members of the Theoretical Biology & Bioinformatics group for useful discussions. This work is part of the research programme 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. analysed and interpreted the results. J.V. wrote the manuscript, which was edited and approved by the other authors.

Supplementary Results and Discussion

Gene fusion

Some spliceosomal OGs had a more complex evolutionary history because their domain composition was the result of a gene fusion during eukaryogenesis. Three OGs combine an RRM with another domain that is also part of other spliceosomal OGs: SART3, ACIN1 and U2SURP (**Figure 1**). The statistics in the main text (e.g., sister function and phylogenetic origin) are based on the largest domain, namely the HAT domain for SART3 and the RRM domain for ACIN1 and U2SURP, instead of SAP and SURP, respectively. These SAP and SURP domains are present only in combination with other domains in the spliceosomal OGs. Ubiquitin domains are combined with a SURP domain in SF3A1 and with a SAP domain in SDE2. The main classification for both was based on the ubiquitin domain. In SF3A3, a SAP domain is combined with a U1-type zinc finger and a SURP and G-patch domain are combined in SUGP1. For these proteins, the SURP and SAP domain were respectively used for the statistics. The G-patch domain was also not used for this in case of TFIP11, which combines this domain with a GCFC domain. PPWD1 combines a WD40 and PPIase domain, with the former comprising a larger part of the protein. SNRNP200 has an additional PWI domain. PRPF8 acquired a MPN domain, which is present in proteins involved in deubiquitination, from an Asgard archaea-derived protein.

A specific example of a domain that was acquired after LECA and is worth mentioning is the PPIase domain in PPIE, which was acquired in the opisthokont lineage. The OG name PPIE is due to the PPIase domain in the human protein, despite only the RRM domain probably being present in LECA.

AAR2

The presence of 1-on-1 orthologs of eukaryotic AAR2 in prokaryotes is remarkable since the spliceosome is eukaryote-specific. AAR2 binds to parts of PRPF8 that are not present in IEP, including the RNaseH-like domain (Galej et al., 2013). It has a very sparse presence distribution in prokaryotes. We detected homologs in multiple cyanobacteria, Lokiarchaeota, Gerdarchaeota, one Helarchaeote, one unclassified Asgard archaeon, one planctomycete and one bacterium classified as Ardenticatenales (Chloroflexi). These proteins have not been characterised yet. Their function could provide insight into the transition from AAR2's original prokaryotic function to its spliceosomal function in eukaryotes.

Asgard archaeal EF2

The evolutionary history of EF2 in archaea involved a gene duplication before the last Asgard archaeal common ancestor (Narrowe et al., 2018). The LC3 lineage, which has recently been renamed to Hodarchaeota (Liu et al., 2021), lost one of the paralogs and concomitantly was the only Asgard archaeal lineage to retain the diphthamide biosynthesis genes (Narrowe et al., 2018). In our tree of archaeal EF2 sequences we could distinguish two clear OGs, which corresponded with the "*bona fide*" EF2 (aEF-2) and EF2 "par-

alog" (aEF-2p) described before (Narrowe et al., 2018). Interestingly, the Hodarchaeal sequences were with high support in the aEF-2p clade. The Jordarchaea also encoded solely EF2 sequences from this OG and had retained the HRG motif, which is present in archaea with a single EF2 copy. This motif is the site of diphthamide modification. In fact, the diphthamide biosynthesis COGs COG1736, -1798 and -2102 were present in Hodarchaea, Jordarchaea and the Asgard Lake Cootharaba group (ALCG). For the latter group, no EF2 homologs were detected, probably because of incomplete genomes. These findings corroborate the previously published pattern of losing one of the paralogs and retaining the diphthamide biosynthesis genes in Asgard archaea and Korarchaeota (Narrowe et al., 2018).

Eukaryotic EF2, EFL1 and EFTUD2 are strongly affiliated to the archaeal EF2 from the LC3 lineage. According to our analysis this would suggest that the aEF-2p sequences retained their ribosome translocation function and HRG motifs after duplication in at least the lineages leading to the eukaryotes, Hodarchaea, Jordarchaea and probably ALCG. In these same lineages the diphthamide biosynthesis genes were retained and the "*bona fide*" EF2 was lost. Alternative scenarios with multiple HGT events among Asgard archaea are less likely since it involves multiple genes.

Lsm

The interpretation of the Lsm tree, especially when including a wide range of archaeal sequences, is notoriously difficult due to its unresolvedness. This likely results from a low phylogenetic signal in these relatively short sequences. Our phylogenetic tree from only eukaryotic sequences is not dissimilar to the previously observed Sm/Lsm pairing of proteins with the same position in the Sm and Lsm rings (Veretnik et al., 2009) (**Supplementary Figure 5.3a**). This provides support to the co-duplication of the genes that were part of an ancestral heteroheptameric ring, resulting in separate Sm and Lsm rings. Subsequent duplication of a few single Sm and Lsm genes resulted in two other Lsm/Sm rings in LECA. The Lsm rings are involved in multiple RNA-related processes (Scofield and Lynch, 2008), including U6 snRNA binding, and therefore likely represent the ancestral function of these proteins.

Asgard archaeal SmAP genes can be separated in three OGs that were probably present in the Asgard ancestor (**Supplementary Figure 3b**). The eukaryotic sequences were split in two groups that were related to two separate Asgard archaeal OGs (**Supplementary Figure 5.3c**). Although eukaryotic monophyly could not be rejected (approximately unbiased test, $P = 0.225$), eukaryotic Sm genes seem to originate from two separate host genes.

SNF

The SNF tree is mostly unresolved and proteins that have the same predicted fate cluster together to some extent (**Supplementary Figure 5.6a**). This suggests that there is a conflicting signal between the phylogenetic signal and the fate-specific patterns. The gene duplications that occurred in the ancestors of Tracheophyta, *S. fallax*, *A. castellanii*, *E. huxleyi*, *C. elegans* and *Stegodyphus mimosarum* are clear from the tree. The duplicates

in *Chlorella variabilis*, *Acytostelium subglobosum* and *Nannochloropsis gaditana* are close together in the tree, albeit not monophyletic, making these likely the result of lineage-specific duplications. The previously described duplication in vertebrates can also be seen in the SNF tree (Williams et al., 2013). The paralogs in *Oikopleura dioica*, *Adineta vaga*, *Ramazzottius varieornatus*, *Cyanophora paradoxa* and *Bigelowiella natans* are quite divergent and do not cluster together but probably represent lineage-specific dupliations. The two copies in *Schistosoma mansoni* likely resulted from an ancestral duplication in the Platyhelminthes, either before or after the split with *Schmidtea mediterranea*. Two SNF genes were present in the ancestors of Ichthyosporea, Choanoflagellatea and probably Fungi, likely reflecting separate duplications. An alternative scenario in which a gene duplication took place in an opisthokont ancestor and subsequently one of the copies was lost in the animal, *Capsaspora owczarzaki* and Nuclearia lineages is less likely, as it requires a dual U1 and U2 function to have been maintained in these two paralogs until those lineages separated from the ones that kept both copies. Given the presence of two SNF genes in the Alveolata and Metamonada species that we considered, duplications before the last common ancestors of Alveolata and Metamonada, respectively, is the most parsimonious explanation. In a tree with only Archaeplastida and Cryptista sequences one of the three *G. theta* sequences was together with the red algae, indicating an endosymbiotic origin of the third SNF sequence in this species.

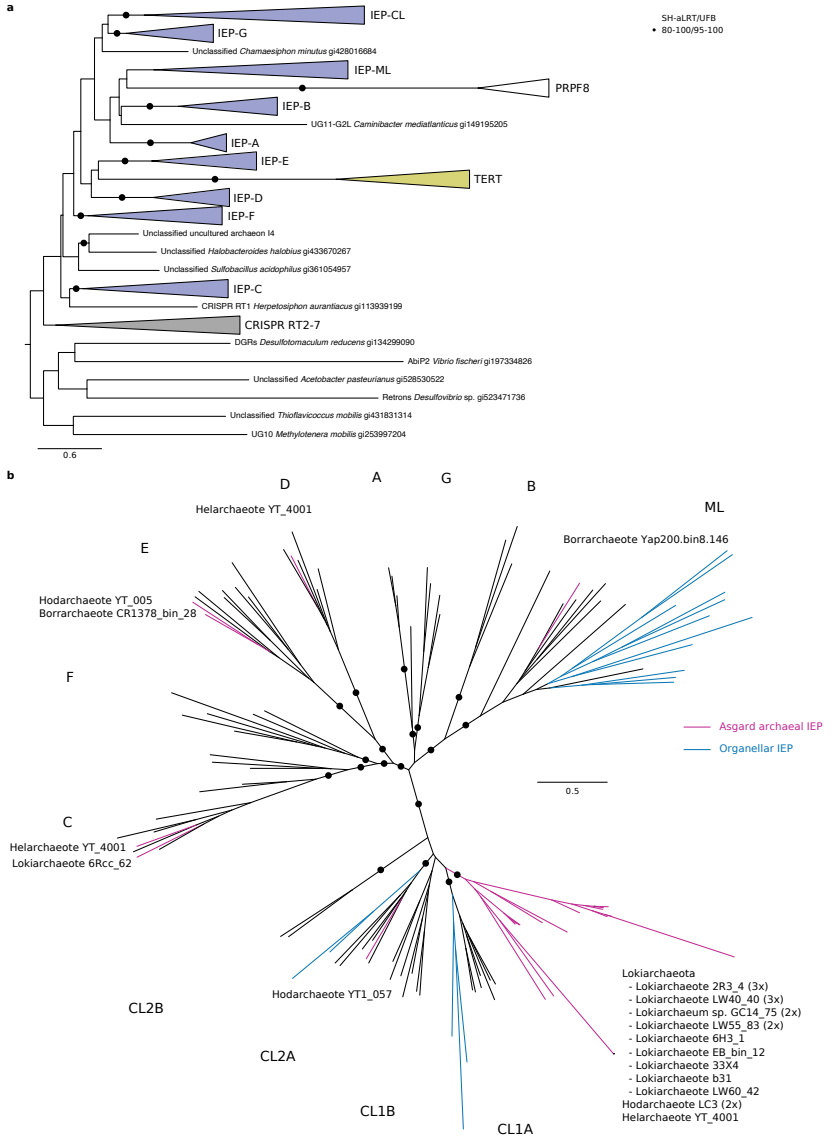
Using the pipeline that we published before (von der Dunk and Snel, 2020), we obtained a high pervasiveness score ($P = 21$, including the excluded *G. theta* would make 22 independent duplication events) and low fate similarity score ($Z_F = 1.867$). The predicted fates of vertebrate, yeast and *A. thaliana* genes were consistent with the U1A and U2B" distinction. However, the *C. elegans* paralogs RNP-3 and RNP-2 were clustered with the U1A and U2B" proteins, respectively, contrary to their functional characterisation (Saldi et al., 2007). This could be due to the functional redundancy observed for these paralogs (Saldi et al., 2007). We adjusted the fate prediction for the *C. elegans* proteins to fit their described function. For three species the predicted fates did not correspond with the fates of orthologs in closely related species. The U2B" prediction of the U1A protein in *Sphaeroforma arctica* was probably caused by its missing second RRM (**Figure 5.4** and **Supplementary Figure 5.6b**). The prediction of the SNF proteins in *Coemansia reversa* and *Ichthyophthirius multifiliis* was also not consistent with close fungal and alveolate orthologs, respectively. The not fully consistent prediction is not unexpected given the low fate similarity.

Besides the recurrent substitutions in U1A and U2B" fate proteins, lineage-specific substitutions could have played a role in the subfunctionalisation. Biochemical analyses of reconstructed ancestral sequences in vertebrates demonstrated how substitutions of five consecutive amino acids in the first RRM effected the changes in snRNA specificity (Delaney et al., 2014). Recurrent substitutions in these positions were not identified by our pipeline.

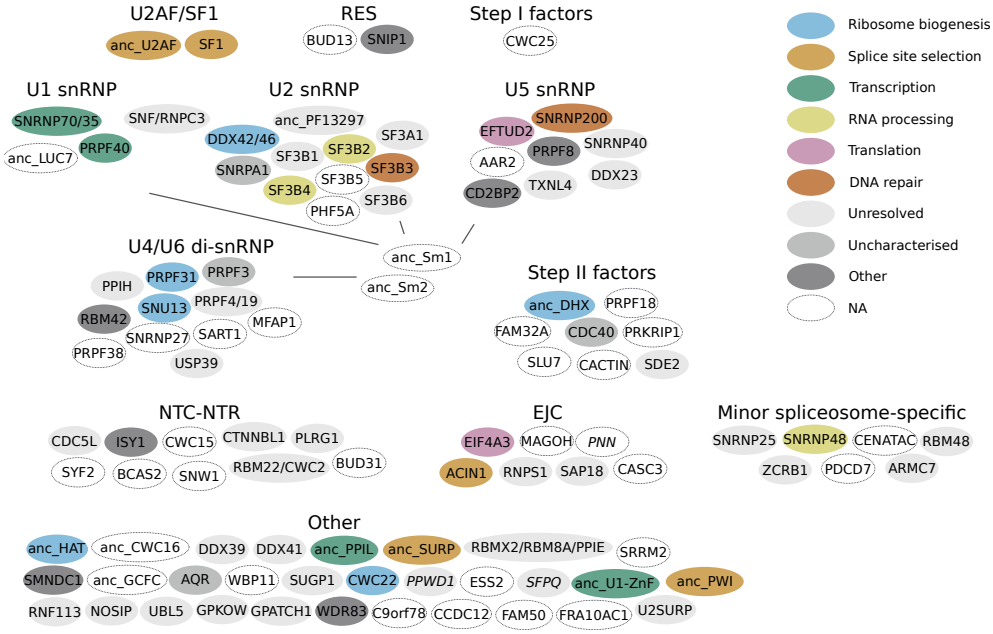
Minor spliceosome

Based on the absence of both minor-spliceosome specific proteins and snRNAs (**Supplementary Figure 5.7b**) we could infer that the minor spliceosome was lost completely 23 times, in the Ichthyophonida, Choanoflagellatea, Chromadorea, *R. varieornatus*, *O. dioica*, *Nuclearia* sp., *Mitosporidium daphniae*, Blastocladiomycota, Kickxellomycotina, *Mortierellomycotina elongata*, Dikarya, Entamoebidae, Haptophyta, Labyrinthulea, Ochrophyta, Myzozoa, Oligohymenophorea, Rhizaria, *G. theta*, Rhodophyta, Chlorophyta, Metamonada and Discoba (if these last two groups are not monophyletic). Additionally, the loss of the minor spliceosome is likely in *D. discoideum*, *A. subglobosum* and *T. trahens* (only RNPC3 detected in those species) and *Encephalitozoon intestinalis* (only ZMAT5 detected). The latter would make the loss in both *M. daphniae* and *E. intestinalis* likely to have happened in a microsporidian ancestor. This parsimonious reconstruction of minor spliceosome loss suggests 10 additional loss events compared with a previous study (López et al., 2008).

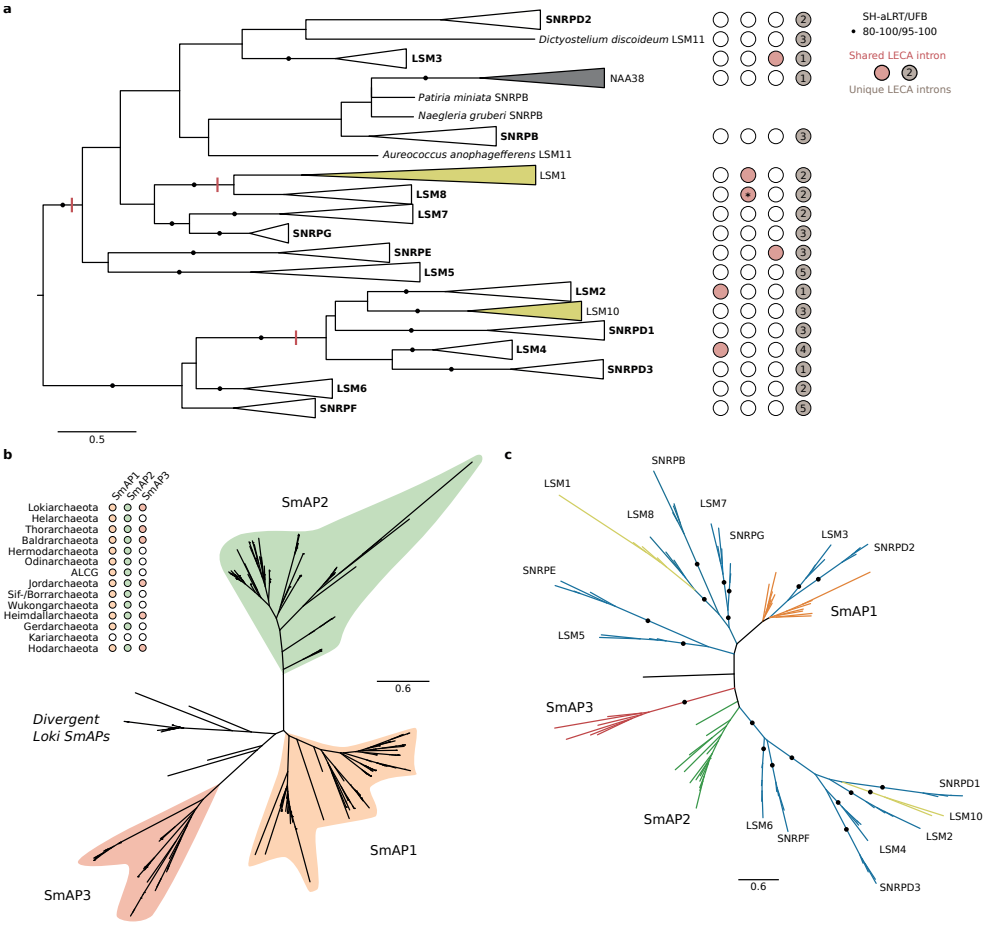
Supplementary Figures



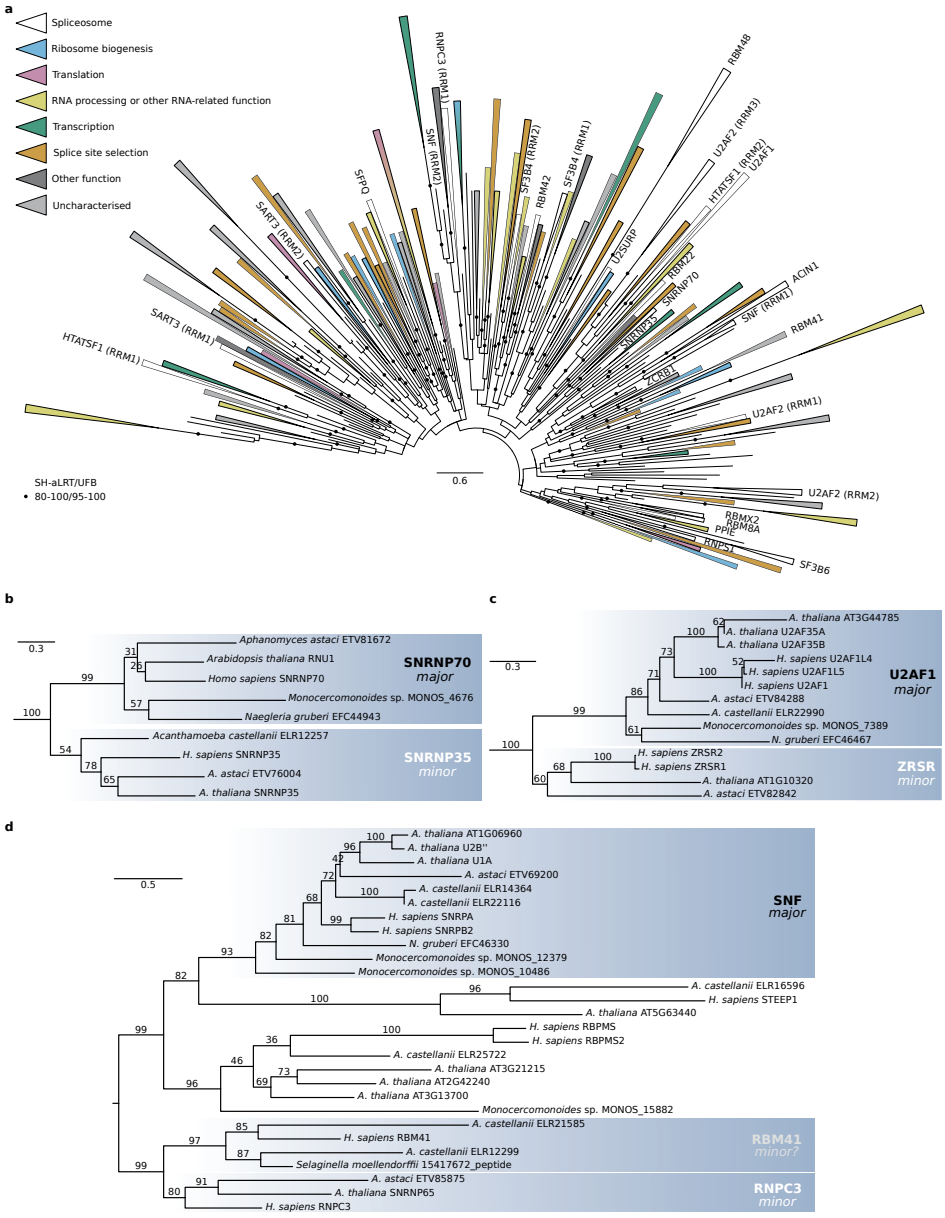
Supplementary Figure 5.1 | Evolutionary history of the intron-encoded protein. a, Phylogenetic position of PRPF8 and TERT in the IEP tree. **b**, Phylogenetic position of Asgard archaeal IEPs. Filled circles correspond with an SH-like approximate likelihood ratio of at least 0.8 and an ultrafast bootstrap value of at least 0.95; scale bars represent the number of substitutions per site. ML: mitochondria-like, CL: chloroplast-like, RT: reverse transcriptase, DGRs: diversity-generating retroelements, UG10 and UG11: RTs of unknown function.



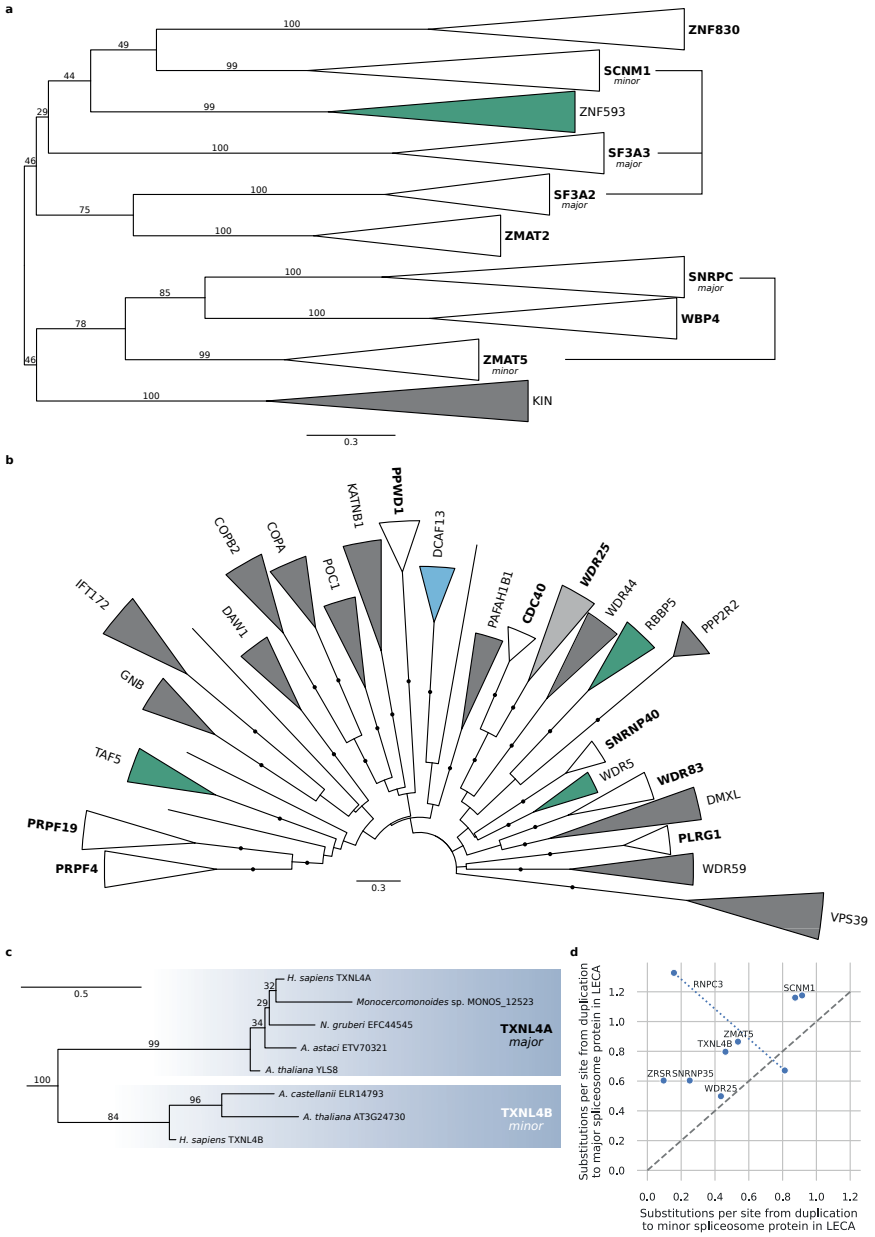
Supplementary Figure 5.2 | Ancestral spliceosomal units. Duplications within the spliceosome are collapsed and the function of the eukaryotic sister OG (if detected) is indicated.



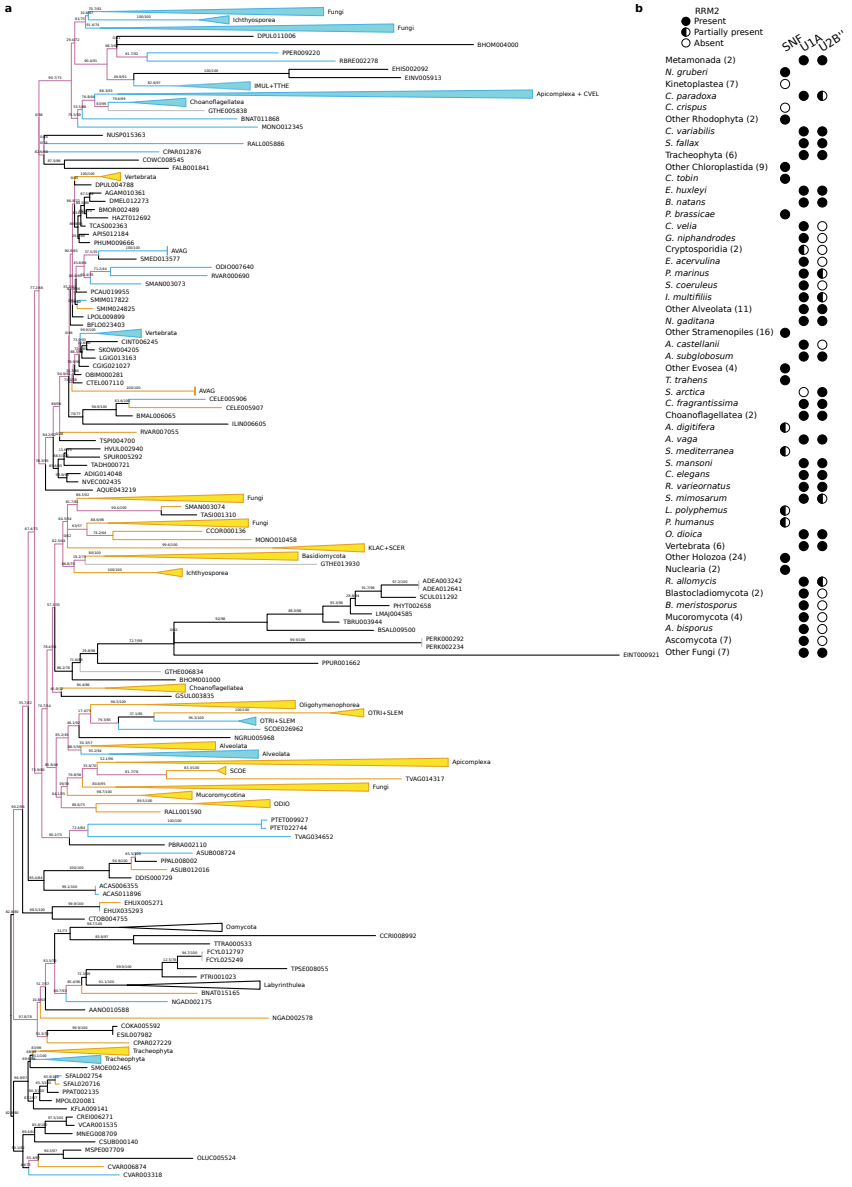
Supplementary Figure 5.3 | Evolutionary history of Lsm/Sm proteins. a, Phylogeny of eukaryotic Lsm/Sm proteins. The tree was rooted using midpoint rooting. The intron positions that are shared between paralogs are indicated. The intron with an asterisk was classified as U12-type intron (Vosseberg et al., 2022b). **b**, Phylogeny of SmAPs in Asgard archaea. The names follow the classification used in previous work (Mura et al., 2003). The presence distribution shows that SmAP3 is restricted to a subset of taxa. **c**, Phylogeny of eukaryotic and Asgard archaeal Sm-family proteins.



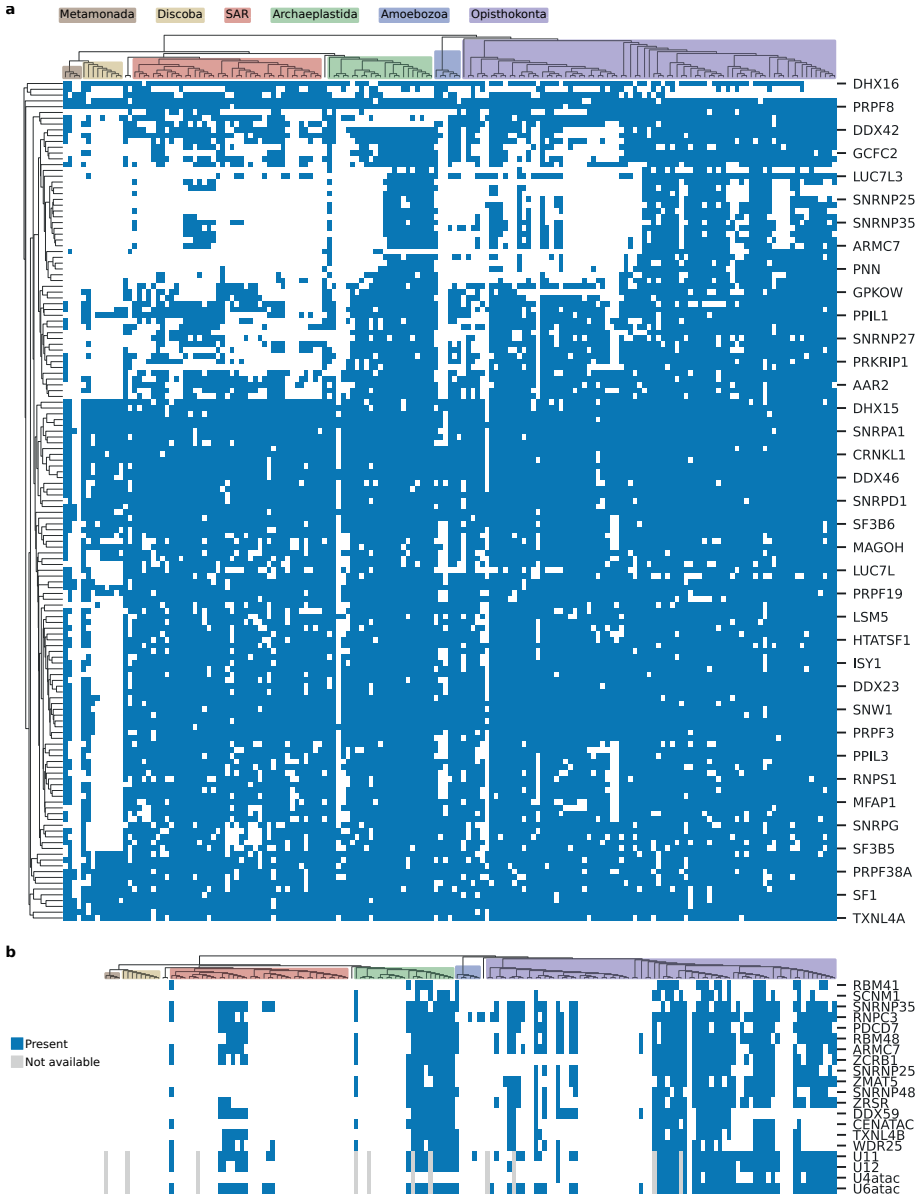
Supplementary Figure 5.4 | Evolutionary history of RRM proteins. a, Phylogeny of RRMs. LECA OGs are collapsed and coloured based on their function. Names are only shown for the spliceosomal OGs. **b**, Phylogeny of SNRNP70 and SNRNP35. The tree was rooted with PPIL4. **c**, Phylogeny of U2AF1 and ZRSR. The tree was rooted with RBM39. **d**, Phylogeny of SNF, STEEP1, RBPMS, RNPC3 and RBM41. The tree was rooted between SNF and RNPC3 based on the topology in the full tree. **b-d**, Branch labels correspond with ultrafast bootstrap values. Major: major spliceosome-specific protein; minor: minor spliceosome-specific protein.



Supplementary Figure 5.5 | Evolutionary history of other large families that contributed to the spliceosome. a, Phylogeny of U1-type zinc finger proteins. **b**, Phylogeny of spliceosomal and closely related WD40 proteins. **c**, Phylogeny of TXNL4A and TXNL4B. The tree was rooted with NXN. **d**, Relation between branch lengths of major- and minor-spliceosome specific proteins. Because no outgroup was used for the SNF/RNPC3 root, the range of possible duplication-to-LECA branch lengths is plotted. **a**, **c**, Branch labels correspond with ultrafast bootstrap values.



Supplementary Figure 5.6 | Evolutionary history of SNF proteins after LECA. a, Phylogenetic tree of SNF. The four letters in the sequence identifiers refer to the species name (see Supplementary Table 1 in (Deutekom et al., 2019)). Paralogs are coloured based on their predicted fate, U1A (blue) or U2B" (yellow), including adjustments. Pink branches connect U1A and U2B" fate proteins of the same species, if necessary. Grey branches reflect duplications without a different fate prediction and the *G. theta* proteins. Branch labels correspond with the SH-like approximate likelihood ratios and ultrafast bootstrap values in percentages. **b**, Presence profile of a second RRM domain in case of a single copy or two paralogs. Numbers in parentheses reflect the number of species.



Supplementary Figure 5.7 | Presence of spliceosomal LECA OGs across eukaryotes. a, Presence of the 145 spliceosomal LECA OGs across 167 eukaryotes. The columns correspond with the species, clustered based on the species tree. Eukaryotic groups are highlighted with different colours. The rows correspond with the spliceosomal LECA OGs; the names of some are indicated. The OGs are clustered with average linkage based on correlation distances. **b**, Presence of minor-spliceosome specific proteins and snRNAs across eukaryotes. The candidate minor-spliceosome specific proteins RBM41, WDR25 and DDX59 are included.

Supplementary Tables

Supplementary Tables 5.1-5.5 are available online on the website of bioRxiv (https://doi.org/10.1101/2022.08.31.505394).

Supplementary Table 5.1 | Spliceosomal proteins in human.

Supplementary Table 5.2 | Spliceosomal proteins in baker's yeast.

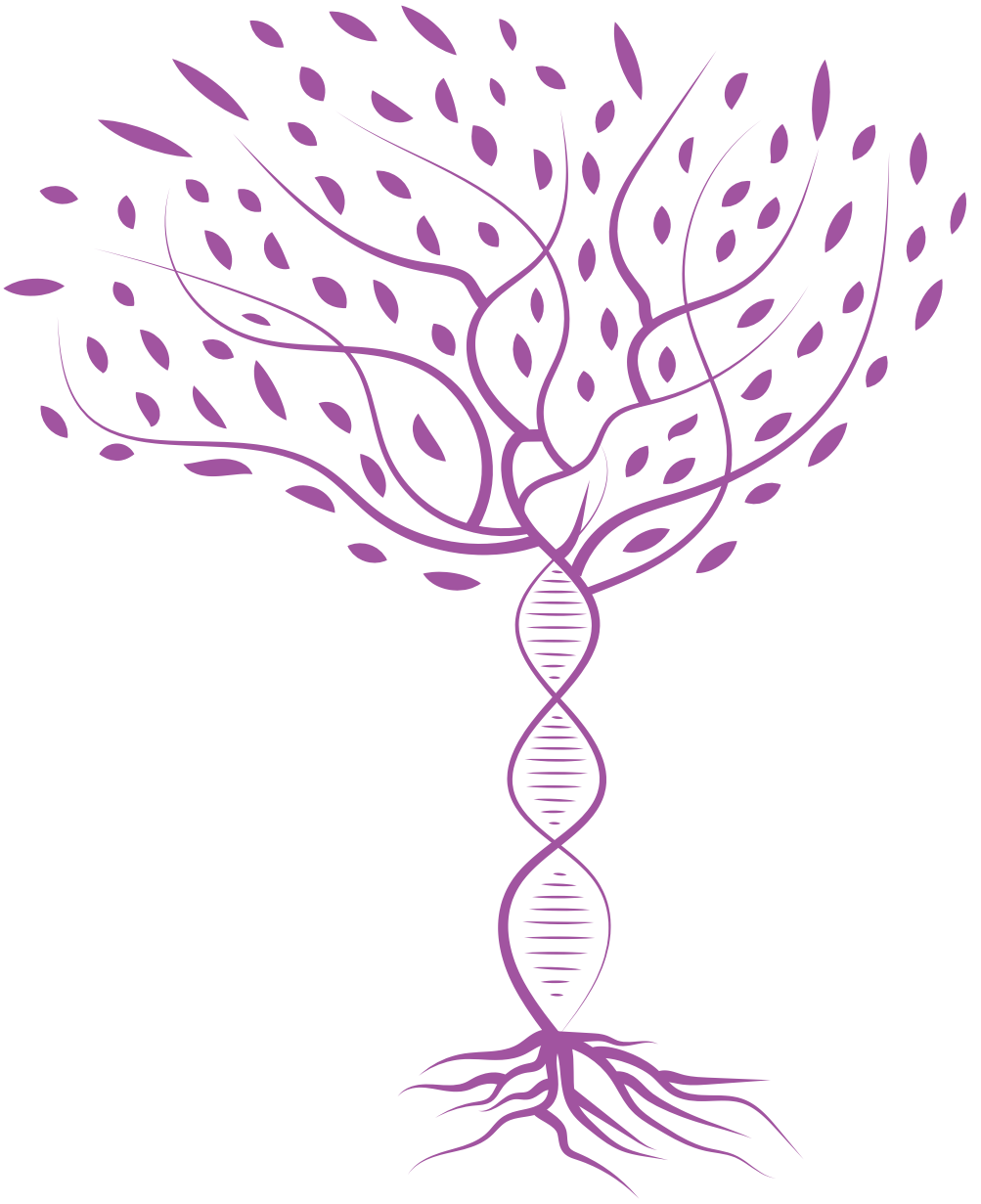
Supplementary Table 5.3 | Spliceosomal LECA OGs.

Supplementary Table 5.4 | Evolutionary histories of spliceosomal LECA OGs.

Supplementary Table 5.5 | Introns inferred in LECA for the spliceosomal LECA OGs and the numbers of intron positions shared with paralogs.

Supplementary Table 5.6 | Programs and settings used for phylogenetic inference.

Family	MAFFT alignment	Alignment trimming	Selected substitution model
IEP-PRPF8-TERT	L-INS-i	trimAl 50%	LG+F+R7
IEP: Asgard archaea	E-INS-i	trimAl 50%	LG+F+R7
EF2 (EFTUD2)	E-INS-i	Divvier	LG+C60+I+G
PRPF31	E-INS-i	trimAl 50%	Q.pfam+F+R5
SNU13	E-INS-i	trimAl 50%	LG+G4
DDX: EIF4A	E-INS-i	Divvier	LG+C60+R7
DDX: DDX3	E-INS-i	Divvier	LG+C60+R6
DHX	E-INS-i	Divvier	LG+C60+R6
LSM: eukaryotes	L-INS-i	trimAl 10%	LG+R4
LSM: Asgard archaea	E-INS-i	trimAl 50%	LG+G4
LSM: eukaryotes and Asgard archaea	E-INS-i	trimAl 50%	LG+R4
RRM	L-INS-i	trimAl 10%	LG+C60+R6
RRM: SNRNP70/35	E-INS-i	trimAl 50%	LG+F+G4
RRM: SNF/RNPC3	E-INS-i	trimAl 50%	LG+I+G4
RRM: U2AF1/ZRSR	E-INS-i	trimAl 60%	LG+G4
U1-type zinc finger	E-INS-i	trimAl 25%	VT+I+G4
WD40	E-INS-i	trimAl gappyout	Q.pfam+R4
TXNL4	E-INS-i	trimAl 50%	LG+G4



6

General discussion

In this thesis we applied phylogenetic analyses to illuminate the order of events in the transition from prokaryotes to eukaryotes. We reconstructed the gene duplications that occurred during eukaryogenesis and the spread of introns in these paralogs. In **chapter 2**, we showed that predominantly genes that were inherited from the host duplicated and relatively few duplications occurred in endosymbiont-derived and metabolic genes. Proteins that build the complex eukaryotic cell, such as the endomembrane system and cytoskeleton, resulted from early duplications, in contrast with the late duplication of regulatory proteins. According to our time estimates, mitochondrial endosymbiosis probably took place between these two duplication waves. In **chapter 3**, we detected many intron positions that were shared between proto-eukaryotic paralogs. We argued that most of these shared introns originated from intron insertions before the duplication event. Our observation that introns were widespread in proto-eukaryotic paralogs strongly suggests that introns were present before most genes duplicated and hence that introns had spread early in eukaryogenesis. These introns, as well as the core of the spliceosomal machinery that removes introns from pre-mRNA molecules, originated from self-splicing group II introns, as reviewed in **chapter 4**. In **chapter 5**, we described that other proteins that were recruited into the spliceosome primarily had a ribosome-related or RNA processing function. Numerous gene duplications shaped the spliceosome into one of the most complex molecular machineries that evolved during eukaryogenesis. The many shared introns between spliceosomal paralogs indicate that introns were widespread through the proto-eukaryotic genome before most spliceosomal complexity originated.

In this final chapter I discuss the implications and future directions of the work described in the previous chapters in relation to recent literature and highlight promising developments that may affect our views on eukaryogenesis.

The ingredients of eukaryogenesis

Gene acquisitions

Eukaryogenesis is a story of at least two microbes, an Asgard archaeon and an alphaproteobacterium, which started a long-lasting, intimate relationship. The serial endosymbiosis theory by Lynn Margulis described multiple endosymbionts that contributed to the eukaryotic cell, including a spirochaete that evolved into the cilium (Sagan, 1967). Additional (endo)symbionts have been proposed by others. The syntrophy hypothesis includes a third microbe, a deltaproteobacterium, that was the host of an archaeal endosymbiont that gave rise to the nucleus (Moreira and López-García, 1998). The diversity of inferred phylogenetic origins of genes that were acquired during eukaryogenesis, more specifically the large numbers of bacterial genes that seemed not to be affiliated with alphaproteobacteria (Pittis and Gabaldón, 2016a), reignited the idea of multiple endosymbioses (Gabaldón, 2018). These additional endosymbiosis are proposed to have been pre-mitochondrial based on the inferred timing of these non-alphaproteobacterial acquisitions before the influx of genes from the mitochondrial endosymbiont (Pittis and Gabaldón, 2016a). Substantial contributions to the proto-eukaryotic lineage could be traced down to Gammaproteobacteria, Firmicutes, Bacteroidetes, Actinobacteria and

Deltaproteobacteria (**Extended Data Fig. 2.4a**). Although the latter could point to a significant role of a deltaproteobacterium in eukaryogenesis, it is more likely that the diffuse phylogenetic signal except for Asgard archaea and alphaproteobacteria primarily reflects horizontal gene transfers among prokaryotes. These transfers have eroded the signal to pinpoint the true phylogenetic donor, if that lineage had not gone extinct. A significant number of genes of bacterial origin are present in Asgard archaeal genomes (Spang et al., 2015). The ratio of archaeal to bacterial genes correlates inversely with their genome size, which is remarkably comparable with what has been observed in eukaryotic genomes (Wu et al., 2022). A larger Asgard archaeal or eukaryotic genome corresponds with a smaller archaeal:bacterial ratio (Wu et al., 2022). The large influx of bacterial genes during eukaryogenesis is in line with the expectations for such large Asgard archaea-derived genomes. Although multiple endosymbioses could have occurred, horizontal gene transfers would be sufficient to explain the gene acquisitions from bacteria.

Like many other studies before, we did not include viral genomes in our analyses. A recent phylogenomics study has identified examples of genetic exchange between nucleocytoplasmic large DNA viruses and proto-eukaryotes (Irwin et al., 2022). This indicates that these large viruses originated before LECA and infected proto-eukaryotes. Molecular dating analyses, combined with the inferred lifestyle of common ancestors, have provided support for an origin of (proto-)eukaryotic host-associated Rickettsiales and Legionellales before LECA (Hugoson et al., 2022; Wang and Luo, 2021). This would mean that proto-eukaryotes lived in association with intracellular gammaproteobacterial Legionellales bacteria (Hugoson et al., 2022) and potentially Chlamydiae (Stairs et al., 2020), and with alphaproteobacterial Rickettsiales bacteria as ectosymbionts or facultative endosymbionts (Schön et al., 2022; Wang and Luo, 2021). Although their genetic contributions to the emerging eukaryotes seem to be limited (**Extended Data Fig. 2.4**), these inferred (endo)symbioses provide insight into the ecological context of eukaryogenesis, with several bacteria in addition to the large viruses exploiting the new complex intracellular structure.

Gene duplications and inventions

The pivotal role of gene duplications in eukaryogenesis has been shown in multiple studies. Expansions in the host-derived histone and E2 ubiquitin-like conjugase families shaped the emerging kinetochore (Tromer et al., 2019). Gene duplications in actin (Stairs and Ettema, 2020), tubulin (Findeisen et al., 2014) and the motor proteins kinesin (Wickstead et al., 2010), dynein (Kollmar, 2016) and myosin (Kollmar and Mühlhausen, 2017) enabled the functioning of a dynamic cytoskeleton in proto-eukaryotes, including a motile cilium (flagellum). Duplications and the subsequent divergence in membrane-trafficking families, such as Rab GTPases, adaptins, protocoatomers, syntaxins and SNAREs, resulted in organelle-specific paralogs of different families that form organelle-specific complexes (Mast et al., 2014; Schlacht et al., 2014). According to the organelle paralogy hypothesis, these paralogs established the identities of the different parts of the endomembrane system (Dacks et al., 2009; Mast et al., 2014; Schlacht et al., 2014).

Because the number of proto-eukaryotic gene duplications per acquisition or inven-

tion is heavily skewed (**Figure 2.1b**), the far majority of these duplications most likely resulted from small-scale duplications of a single gene or a few neighbouring genes. However, the symmetry in parts of the kinetochore complex (Tromer et al., 2019) and the duplication of the entire Lsm ring (**Figure 4.2**) hints at certain larger-scale duplication events. Whole-genome duplications could have played a role in creating paralogs on such a large scale (Makarova et al., 2005; Zhou et al., 2010) and might have been caused by recurrent genome fusions in syncytial proto-eukaryotes (Garg and Martin, 2016). A fusion event is also one of the plausible scenarios for the emergence of two intron and spliceosome types (chapter 3). These large-scale duplication events seem plausible but conclusive data are lacking.

A result from chapter 2 and 5 that we had not put emphasis on, is the contribution of proto-eukaryotic gene inventions. The role of novel folds in eukaryogenesis has been highlighted in previous studies (Aravind et al., 2006; Kauko and Lehto, 2018). However, caution is warranted when inferring a *de novo* gene birth for an OG or a group of paralogous OGs for which no homologs can be detected in prokaryotes. An alternative and likely explanation is that homology could not be inferred because of high rates of sequence evolution. Extreme divergence could have eroded most similarity with homologs. This is especially probable after gene duplication and the gain of a novel function by one of the paralogs (neofunctionalization). While we estimated a near doubling of the proto-eukaryotic genome (chapter 2), the number of duplications is very likely an underestimation when considering these extensively diverged paralogs. Sensitive comparisons between sequences using profile-profile searches and structures, such as performed for the origin of the kinetochore (Tromer et al., 2019), are required to elucidate the extent of contribution of novel folds to eukaryotic complexity.

In sum, two full organisms were the main partners in eukaryogenesis, supplemented with genes from other prokaryotes. *De novo* gene births and the fusion and fission of genes contributed to the increasing genetic repertoire of proto-eukaryotes. Nevertheless, the foremost ingredient that shaped the emerging complex, eukaryotic cell was the large-scale duplication of genes.

Evolution of complexity

Eukaryogenesis resulted in cells that are tremendously more complex than prokaryotic cells. In the previous section I discussed the genetic ingredients to create this complex, new kind of cell. These ingredients, however, do not directly give an explanation *why* the increased genomic and cellular complexity of LECA evolved.

Protein complexes that were already present in prokaryotes became more complex during eukaryogenesis by the addition of new proteins. In addition, large protein complexes originated in proto-eukaryotes, such as the kinetochore (Tromer et al., 2019) and nuclear pore complex (Mast et al., 2014). As discussed in chapter 4, understanding the function of these additional proteins in current-day complexes is not sufficient to explain the evolutionary forces driving the complexification of these systems, because it typically ignores a potential role of neutral evolution. In case of constructive neutral evolution, protein complexes become more complex by the addition of initially superfluous com-

ponents that do not contribute to the function of the complex but also do not impair it. For example, only a minority of proteins in mitochondrial complex I can be found in its bacterial counterpart (Gabaldón et al., 2005). The eukaryote-specific additions do not directly contribute to the NADH:ubiquinone oxidoreductase activity but play a role in the assembly of the complex instead (Stroud et al., 2016). Even though these assembly factors have become essential for its present-day functioning, this additional complexity is most likely dispensable (Schulz et al., 2022). Similarly, supernumerary components are probably present in other complexes. Some kinetochore modules that originated during eukaryogenesis were lost frequently in subsequent eukaryotic evolution, such as the constitutive centromere-associated network (van Hooff et al., 2017). The two different types of introns and the major- and minor-spliceosome specific subunits are likely the result of neutral evolution, as is most of spliceosomal complexity (chapter 4).

Notwithstanding the plausible role of neutral evolution in the complexification of protein complexes, natural selection probably played a large role in large cellular changes such as the establishment of an elaborate endomembrane system and a dynamic cytoskeleton. The ability to perform phagocytosis would have created a distinctive niche for proto-eukaryotes. Moreover, mitochondrial endosymbiosis likely gave the eukaryotic cell a selective advantage. These adaptive traits, however, were probably neutrally embellished with additional components.

The role of neutral processes does not exclude a potential advantage of additional complexity. The complex molecular machines provide a powerful means to fine-tune their regulation in different circumstances. The large genomic and cellular complexity in LECA was probably key to the evolution of morphologically disparate eukaryotes, including multicellular organisms.

These potential advantages raise the question why complex eukaryotic cells evolved only once. What enabled the enormous rise in complexity in proto-eukaryotes? Using theoretical approaches, potential constraints have been explored that could explain the absence of prokaryotes that evolved eukaryote-like complexity. The most debated one is based on bioenergetic constraints. Lane and Martin argued that dedicated internal respiratory membranes controlled by their own genome, as present in mitochondria, released proto-eukaryotes from bioenergetic constraints to become larger (Lane, 2011; Lane and Martin, 2010). The reduction of mitochondrial genomes enabled the nuclear genome to expand by the surplus of energy available (Lane, 2011; Lane and Martin, 2010). This reasoning and the underlying data have been challenged by others and has sparked an ongoing debate (Chiyomaru and Takemoto, 2020; Hampl et al., 2019; Koonin, 2015; Lane, 2017, 2020; Lane and Martin, 2015; Lynch and Marinov, 2015, 2016, 2017). A recent study has provided support for the existence of limits to genome sizes and cell volumes in prokaryotes (Schavemaker and Muñoz-Gómez, 2022). Schavemaker and Muñoz-Gómez also specifically modelled the probable situation at the onset of endosymbiosis and observed that mitochondria could have provided an energetic advantage from the start only in case of a more complex proto-eukaryote that had active cytoplasmic transport with a dynamic cytoskeleton (Schavemaker and Muñoz-Gómez, 2022).

Besides bioenergetic constraints, using horizontal gene transfer in prokaryotes instead of sex places constraints on genome size (Colnaghi et al., 2020), especially in case of

repeat sequences (Colnaghi et al., 2022) like transposable elements (van Dijk et al., 2022). According to these analyses, the invention of meiotic sex would have enabled the increase in proto-eukaryotic genome size (Colnaghi et al., 2020, 2022; van Dijk et al., 2022).

Some complex cellular features, however, are present in certain prokaryotes. Giant bacteria exist that are visible by the naked eye, with a length of even a centimetre in case of the gammaproteobacterium *Candidatus* Thiomargarita magnifica (Volland et al., 2022). It should be noted that these cells are polyploid and contain a vacuole that occupies most of the cell volume (Volland et al., 2022). Organelles that are surrounded by a lipid bilayer are not unique to eukaryotes and are present in multiple prokaryotes. These prokaryotic organelles include thylakoids, magnetosomes and anammoxosomes (Greening and Lithgow, 2020). Other internal membrane structures have been observed in multiple Planctomycetes, *Atribacter laminatus* and *Candidatus* Thiomargarita magnifica (Katayama et al., 2020; Volland et al., 2022; Wiegand et al., 2018). Even phagocytosis does not seem to be unique to eukaryotes. The planctomycete *Candidatus* Uabimicrobium amorphum engulfs other cells by a phagoyctosis-like mechanism (Shiratori et al., 2019). Notwithstanding these examples of prokaryotic cellular complexity, a fundamental gap remains between these prokaryotes and eukaryotes. The constraints described above may have prevented these complex prokaryotes to evolve full eukaryote-like complexity.

Timing proto-eukaryotic events

As mentioned throughout this thesis, many genetic changes occurred during eukaryogenesis. By unravelling the order of these changes during eukaryogenesis we can assess the different hypotheses regarding the origin of eukaryotes. Some information about the order of events can be found in a single gene tree, namely the order of duplications. However, deciphering this order of duplications in a gene family is hindered by low branch support values for deep internal nodes. Although the support for LECA clades is high using the ScrollSaw method, which has also been shown in other studies (Elias et al., 2012; van Wijk and Snel, 2020), most duplication nodes did not receive strong support (**Supplementary Table 2.3**). The use of slightly more groups to select sequences for less expanded families, such as performed for ARF GTPases (Vargová et al., 2021), could increase the phylogenetic signal and provide more resolution. Other events than substitutions, for example the gains and losses of introns (chapter 3), could also aid in resolving the order of duplications.

In chapter 2, we exploited the branch lengths in phylogenetic trees to relatively time the genetic acquisitions from prokaryotes and gene duplications. The statistical analyses and assumptions of our methods have been carefully examined (Susko et al., 2021). The suggestions made in that paper include important considerations for future research. As acknowledged in chapter 2, the inferred timing of acquisition events is dependent on the sampling of prokaryotes. The impact of unsampled donor lineages has been quantified with simulations in a recent study (Tricou et al., 2022). The estimated large impact of these unsampled "ghost" lineages warrants closer examination. Duplications likely represent the latest possible time of acquisition and could therefore provide additional information on the timing of endosymbiosis and HGT events (chapter 2).

Phylogenetic trees contain more time information than we have used for calculating the stem and duplication lengths. For example, the order of duplications and preduplication branches are not used for the duplication lengths, which occasionally cause inconsistencies when older duplications have shorter duplication lengths. By incorporating the order of duplications, preduplication branches and additional information from the prokaryotic sister branch lengths (which also provide estimates for the acquisitions) and simultaneous eukaryotic speciation events in the tree in case of duplications, a higher precision of time estimates may be obtained. Programs specifically developed for dating phylogenies are suited for these analyses, specifically the flexible Bayesian framework used in RevBayes (Höhna et al., 2016). Ideally, one would also like to integrate information across trees for the inference of the same event, including FECA and the first mitochondrial common ancestor, which would require a new, hierarchical framework. In combination with fossil calibrations, these analyses can be used to obtain absolute time estimates instead of relative ones. The linking of the same speciation events ("cross-bracing") has been applied before to the ATPase subunits to time the mitochondrial and plastid endosymbiosis events (Shih and Matzke, 2013).

A timeline of eukaryogenesis

Based on our time estimates from branch lengths and the presence of shared introns, and in the light of recent work and despite all uncertainties, it is possible to draw a rough timeline of eukaryogenesis (**Figure 6.1**). The earliest proto-eukaryotes were already relatively complex prokaryotes with an actin- and tubulin-based cytoskeleton and membrane invaginations, perhaps even a primitive endomembrane system. A protonuclear structure that separated transcription and translation, suggested to be present in current Asgard archaea (Avcı et al., 2022), could have been present in FECA. Introns were in that case not the selective force for a nuclear structure, in contrast with some hypotheses (López-García and Moreira, 2006; Martin and Koonin, 2006). It is also highly unlikely that the membrane structures around the nucleoid in Planctomycetes, *Atribacter* and *Thiomargarita* (Katayama et al., 2020; Volland et al., 2022; Wiegand et al., 2018) are related to introns. Other scenarios for the origin of a nucleus have been proposed (Hendrickson and Poole, 2018). It is plausible that self-splicing group II introns were present in early proto-eukaryotes. The presence of a protonuclear structure facilitated the emergence of the first intragenic introns by preventing the translation of pre-mRNAs with introns into erroneous proteins. The proliferation of introns and the resulting spread through the genome (chapter 3) would have enforced the maintenance of a closed nuclear compartment with a regulated transport machinery.

Our branch length results (chapter 2) indicate that the proto-eukaryotic genome expanded by a first round of gene duplications that enabled the construction of a larger, complex eukaryote-like cell with membrane trafficking between the different parts of the endomembrane system. A considerable influx of bacterial genes transformed the proto-eukaryotic cells in addition to these duplications. This influx could be related to a phagotrophic lifestyle that proto-eukaryotes might have had at this stage (Martijn and Ettema, 2013). These large, slowly dividing cells with active cytoplasmic transport

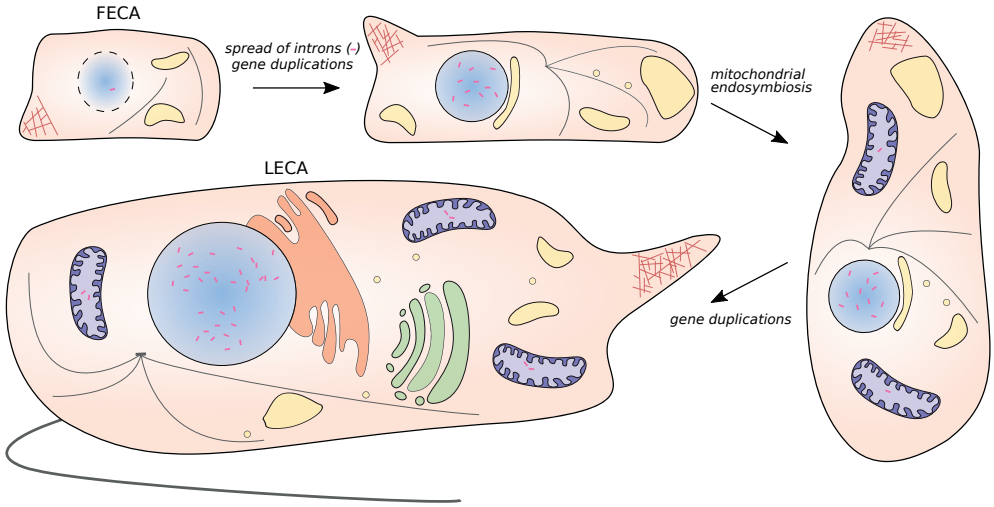


Figure 6.1 | Scenario for eukaryogenesis. FECA is suggested to contain a protonuclear structure around the nucleoid (blue) and a primitive endomembrane system (yellow) and cytoskeleton (red and grey). The potential presence of self-splicing group II introns is indicated in pink. The scenario I propose is a mitochondria-intermediate scenario. Before mitochondrial endosymbiosis, introns had spread through the genome and gene duplications resulted in a more complex endomembrane system including nucleus and dynamic cytoskeleton. After the uptake of the protomitochondrion (purple), additional gene duplications brought about refinements in the endomembrane system, shown as the establishment of the endoplasmic reticulum (orange) and Golgi apparatus (green), and complex regulatory systems. Parts of the figure were created by Max Raas.

would have had an energetic advantage from respiring endosymbionts (Schavemaker and Muñoz-Gómez, 2022). Regardless of the initial circumstances of the symbiosis between proto-eukaryotes and the proteobacterial ancestors of mitochondria, the endosymbionts may have enabled a further increase in cell and genome size (Lane and Martin, 2010; Schavemaker and Muñoz-Gómez, 2022).

A second wave of gene duplications resulted in a refinement of the endomembrane system, intricate signalling systems and complex regulatory systems for gene expression. Most complexification of the spliceosome (chapters 4 and 5) was likely part of this later set of duplications. Concomitantly, the loss and transfer of mitochondrial genes reduced the endosymbiont genomes, which was complemented with a novel protein import system, membrane transporters and further integration of the host and endosymbiont (Roger et al., 2017). Full-blown eukaryotic cells had evolved before the radiation into the extant eukaryotic groups.

Bridging the gap

Many details remain to be elucidated to bridge the gap between FECA and LECA. Novel lineages may reduce this gap by changing the phylogenetic position of either of these ancestors. The discovery of novel eukaryotes that branch more deeply would push LECA further back in time. Similarly, new Asgard archaea that are more closely related to eu-

karyotes than those currently sampled would reduce the gap, as the discovery of these new archaea did in the past (Spang et al., 2015). Further studies on the cell biology of Asgard archaea will illuminate the cellular characteristics of these archaea and hence aid in the reconstruction of FECA (Avcı et al., 2022; Imachi et al., 2020). Ideally, genetic experiments in these archaea would shed light on the functions of Asgard homologs of eukaryotic genes. The expression of Asgard profilin, gelsolin, tubulin, SNARE and ESCRT proteins in other systems has partly elucidated their function (Akıl and Robinson, 2018; Akıl et al., 2020; Hatano et al., 2022; Neveu et al., 2020; Survery et al., 2021). In combination with gene-tree aware ancestral reconstructions (Szöllősi et al., 2013), which can distinguish vertical inheritance scenarios from HGTs during eukaryogenesis from related Asgard archaeal lineages (Wu et al., 2022), these would provide insight into the cellular nature of FECA.

Intermediate stages of eukaryogenesis can be more confidently inferred with the discovery of additional proto-eukaryotic fossils (Porter, 2020). The latest developments in more sensitive homology searches, aided with large-scale protein structure predictions, are likely to reveal deep homologies and facilitate tracing the shared ancestry of protein families (Jumper et al., 2021; Monzon et al., 2022). Ancestral sequence reconstructions and experiments on these reconstructed sequences (Harms and Thornton, 2010, 2013) could inform about the function of proto-eukaryotic, especially preduplication, proteins. These preduplication proteins can be seen as snapshots of intermediate stages of eukaryogenesis and their function could elucidate the appearance and functioning of proto-eukaryotic cells. Moreover, it could reveal the role of neutral versus adaptive evolution in the increased complexity of protein complexes.

Furthermore, studies on analogous systems could provide insight into the intermediate steps during the transition. Evident examples include the further characterisation of the complex prokaryotes mentioned earlier. The spliceosome-like complexes that remove group II introns in certain mitochondria and plastids (chapter 4) could illuminate the rise of the complex spliceosome during eukaryogenesis. Experiments like the introduction of group II introns and retrotransposons in bacteria with or without non-homologous end joining (Lee et al., 2018) and of group II introns in yeast (Chalamcharla et al., 2010; Qu et al., 2014) sharpen ideas on early intron evolution. Endosymbioses, either naturally occuring (Graf et al., 2021) or engineered (Mehta et al., 2018), may shed light on the early stages of mitochondrial endosymbiosis.

Final remarks

As part of all three research chapters, we performed a large-scale analysis similar to a previous study from around fifteen years ago. The use of more advanced methods and the enormously expanded set of sequenced genomes revealed different findings. This was the case for the characterisation of proto-eukaryotic duplications (chapter 2, c.f. Makarova *et al.* (Makarova et al., 2005)), which had been investigated again five years later (Zhou et al., 2010), the extent of shared introns between proto-eukaryotic paralogs (chapter 3, c.f. Sverdlov *et al.* (Sverdlov et al., 2007)) and the composition of the spliceosome in LECA (chapter 5, c.f. Collins and Penny (Collins and Penny, 2005)). To the best of our knowl-

edge, no repetition of these analyses with the increasing amount of genome data had been performed in the intervening years. The findings in these papers, although no longer up-to-date, were widely being referred to, largely without cautionary remarks highlighting potential shortcomings. For the proto-eukaryotic duplications and the LECA spliceosome, our findings were in line with the previously published results. In case of the extent of shared introns, however, we made the complete opposite observation compared with the previous analysis. Possibly, the main findings of some other old analyses regarding eukaryogenesis are also no longer supported by state-of-the-art data and methods.

Most of the functional inference throughout the thesis is derived from a small set of model organisms, especially opisthokonts. Experimental studies in other organisms are highly needed to make more reliable inferences on ancestral functions. The composition of spliceosomes in non-model organisms, for instance, could reveal spliceosomal "jotnarlogs", LECA proteins that were lost in well-characterised organisms (More et al., 2020). Closer examination of the wide diversity of life is key to unravelling the intriguing enigmas in evolutionary biology.

References

Achsel, T., Brahms, H., Kastner, B., Bachi, A., Wilm, M., and Lührmann, R. (1999). A doughnut-shaped heteromer of human Sm-like proteins binds to the 3′-end of U6 snRNA, thereby facilitating U4/U6 duplex formation in vitro. EMBO J. *18*, 5789–5802.

Adl, S.M., Bass, D., Lane, C.E., Lukeš, J., Schoch, C.L., Smirnov, A., Agatha, S., Berney, C., Brown, M.W., Burki, F., et al. (2019). Revisions to the Classification, Nomenclature, and Diversity of Eukaryotes. J. Eukaryot. Microbiol. *66*, 4–119.

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

Akiyoshi, D.E., Morrison, H.G., Lei, S., Feng, X., Zhang, Q., Corradi, N., Mayanja, H., Tumwine, J.K., Keeling, P.J., Weiss, L.M., et al. (2009). Genomic Survey of the Non-Cultivatable Opportunistic Human Pathogen, *Enterocytozoon bieneusi*. PLoS Pathog. *5*, e1000261.

Akıl, C., and Robinson, R.C. (2018). Genomes of Asgard archaea encode profilins that regulate actin. Nature *562*, 439–443.

Akıl, C., Tran, L.T., Orhant-Prioux, M., Baskaran, Y., Manser, E., Blanchoin, L., and Robinson, R.C. (2020). Insights into the evolution of regulated actin dynamics via characterization of primitive gelsolin/cofilin proteins from Asgard archaea. Proc. Natl. Acad. Sci. U. S. A. *117*, 19904–19913.

Akıl, C., Ali, S., Tran, L.T., Gaillard, J., Li, W., Hayashida, K., Hirose, M., Kato, T., Oshima, A., Fujishima, K., et al. (2022). Structure and dynamics of Odinarchaeota tubulin and the implications for eukaryotic microtubule evolution. Sci. Adv. *8*, eabm2225.

Al Jewari, C., and Baldauf, S.L. (2022). Conflict over the Eukaryote Root Resides in Strong Outliers, Mosaics and Missing Data Sensitivity of Site-Specific (CAT) Mixture Models. Syst. Biol., syac029.

Ali, R.H., Bogusz, M., and Whelan, S. (2019). Identifying clusters of high confidence homologies in multiple sequence alignments. Mol. Biol. Evol. *36*, 2340–2351.

Altschul, S.F., Gish, W., Miller, W., Myers, E.W., and Lipman, D.J. (1990). Basic local alignment search tool. J. Mol. Biol. *215*, 403–410.

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

Andersson, J.O., Sjögren, Å.M., Horner, D.S., Murphy, C.A., Dyal, P.L., Svärd, S.G., Logsdon, J.M., Ragan, M.A., Hirt, R.P., and Roger, A.J. (2007). A genomic survey of the fish parasite *Spironucleus salmonicida* indicates genomic plasticity among diplomonads and significant lateral gene transfer in eukaryote genome evolution. BMC Genom. *8*, 51.

Aravind, L., Iyer, L.M., and Koonin, E.V. (2006). Comparative genomics and structural biology of the molecular innovations of eukaryotes. Curr. Opin. Struct. Biol. *16*, 409–419.

Aravind, L., Anantharaman, V., Zhang, D., Souza, D., Francisco, R., and Iyer, L.M. (2012). Gene flow and biological conflict systems in the origin and evolution of eukaryotes. Front. Cell. Infect. Microbiol. *2*, 89.

Avcı, B., Brandt, J., Nachmias, D., Elia, N., Albertsen, M., Ettema, T.J.G., Schramm, A., and Kjeldsen, K.U. (2022). Spatial separation of ribosomes and DNA in Asgard archaeal cells. ISME J. *16*, 606–610.

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

Baños, H., Susko, E., and Roger, A.J. (2022). Are profile mixture models over-parameterized?

Preprint at bioRxiv, 10.1101/2022.02.18.481053.
Barbosa-Morais, N.L., Carmo-Fonseca, M., and Aparício, S. (2006). Systematic genome-wide annotation of spliceosomal proteins reveals differential gene family expansion. Genome Res. *16*, 66–77.
Bartschat, S., and Samuelsson, T. (2010). U12 type introns were lost at multiple occasions during evolution. BMC Genom. *11*, 106.
Basu, M.K., Makalowski, W., Rogozin, I.B., and Koonin, E.V. (2008a). U12 intron positions are more strongly conserved between animals and plants than U2 intron positions. Biol. Direct *3*, 19.
Basu, M.K., Rogozin, I.B., and Koonin, E.V. (2008b). Primordial spliceosomal introns were probably U2-type. Trends Genet. *24*, 525–528.
Belhocine, K., Mak, A.B., and Cousineau, B. (2008). *Trans*-splicing versatility of the Ll.LtrB group II intron. RNA *14*, 1782–1790.
Betts, H.C., Puttick, M.N., Clark, J.W., Williams, T.A., Donoghue, P.C.J., and Pisani, D. (2018). Integrated genomic and fossil evidence illuminates life's early evolution and eukaryote origin. Nat. Ecol. Evol. *2*, 1556–1562.
Booth, A., and Doolittle, W.F. (2015). Eukaryogenesis, how special really? Proc. Natl. Acad. Sci. U. S. A. *112*, 10278–10285.
Breuer, R., Gomes-Filho, J.-V., and Randau, L. (2021). Conservation of archaeal C/D box sRNA-guided RNA modifications. Front. Microbiol. *12*, 654029.
Brodsky, F.M., Thattai, M., and Mayor, S. (2012). Evolutionary cell biology: Lessons from diversity. Nat. Cell Biol. *14*, 651–651.
Brown, M.W., Heiss, A.A., Kamikawa, R., Inagaki, Y., Yabuki, A., Tice, A.K., Shiratori, T., Ishida, K.-I., Hashimoto, T., Simpson, A.G.B., et al. (2018). Phylogenomics Places Orphan Protistan Lineages in a Novel Eukaryotic Super-Group. Genome Biol. Evol. *10*, 427–433.
Buchfink, B., Xie, C., and Huson, D.H. (2015). Fast and sensitive protein alignment using DIAMOND. Nat. Methods *12*, 59–60.
Burge, C.B., Padgett, R.A., and Sharp, P.A. (1998). Evolutionary fates and origins of U12-type introns. Mol. Cell *2*, 773–785.
Burki, F., Roger, A.J., Brown, M.W., and Simpson, A.G.B. (2020). The New Tree of Eukaryotes. Trends Ecol. Evol. *35*, 43–55.
Burki, F., Sandin, M.M., and Jamy, M. (2021). Diversity and ecology of protists revealed by metabarcoding. Curr. Biol. *31*, R1267–R1280.
Busch, A., and Hertel, K.J. (2012). Evolution of SR protein and hnRNP splicing regulatory factors. Wiley Interdisciplinary Reviews: RNA *3*, 1–12.
Califice, S., Baurain, D., Hanikenne, M., and Motte, P. (2012). A Single Ancient Origin for Prototypical Serine/Arginine-Rich Splicing Factors. Plant Physiol. *158*, 546–560.
Candales, M.A., Duong, A., Hood, K.S., Li, T., Neufeld, R.A.E., Sun, R., McNeil, B.A., Wu, L., Jarding, A.M., and Zimmerly, S. (2012). Database for bacterial group II introns. Nucleic Acids Res. *40*, D187–D190.
Capella-Gutiérrez, S., Silla-Martínez, J.M., and Gabaldón, T. (2009). trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. Bioinformatics *25*, 1972–1973.
Carmel, L., Wolf, Y.I., Rogozin, I.B., and Koonin, E.V. (2007). Three distinct modes of intron

dynamics in the evolution of eukaryotes. Genome Res. *17*, 1034–1044.
Catania, F., Gao, X., and Scofield, D.G. (2009). Endogenous Mechanisms for the Origins of Spliceosomal Introns. J. Hered. *100*, 591–596.
Cavalier-Smith, T. (1991). Intron phylogeny: a new hypothesis. Trends Genet. *7*, 145–148.
Chalamcharla, V.R., Curcio, M.J., and Belfort, M. (2010). Nuclear expression of a group II intron is consistent with spliceosomal intron ancestry. Genes Dev. *24*, 827–836.
Chan, S.-P., Kao, D.-I., Tsai, W.-Y., and Cheng, S.-C. (2003). The Prp19p-Associated Complex in Spliceosome Activation. Science *302*, 279–282.
Charollais, J., Dreyfus, M., and 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.
Chiyomaru, K., and Takemoto, K. (2020). Revisiting the hypothesis of an energetic barrier to genome complexity between eukaryotes and prokaryotes. R. Soc. Open Sci. *7*, 191859.
Collins, L., and Penny, D. (2005). Complex Spliceosomal Organization Ancestral to Extant Eukaryotes. Mol. Biol. Evol. *22*, 1053–1066.
Colnaghi, M., Lane, N., and Pomiankowski, A. (2020). Genome expansion in early eukaryotes drove the transition from lateral gene transfer to meiotic sex. eLife *9*, e58873.
Colnaghi, M., Lane, N., and Pomiankowski, A. (2022). Repeat sequences limit the effectiveness of lateral gene transfer and favored the evolution of meiotic sex in early eukaryotes. Proc. Natl. Acad. Sci. U. S. A. *119*, e2205041119.
Costa, M., Michel, F., and Westhof, E. (2000). A three-dimensional perspective on exon binding by a group II self-splicing intron. EMBO J. *19*, 5007–5018.
Csűrös, M. (2008). Malin: maximum likelihood analysis of intron evolution in eukaryotes. Bioinformatics *24*, 1538–1539.
Csuros, M., Rogozin, I.B., and Koonin, E.V. (2011). A Detailed History of Intron-rich Eukaryotic Ancestors Inferred from a Global Survey of 100 Complete Genomes. PLOS Comput. Biol. *7*, e1002150.
Cuomo, C.A., Desjardins, C.A., Bakowski, M.A., Goldberg, J., Ma, A.T., Becnel, J.J., Didier, E.S., Fan, L., Heiman, D.I., Levin, J.Z., et al. (2012). Microsporidian genome analysis reveals evolutionary strategies for obligate intracellular growth. Genome Res. *22*, 2478–2488.
Cuypers, T.D., and Hogeweg, P. (2012). Virtual Genomes in Flux: An Interplay of Neutrality and Adaptability Explains Genome Expansion and Streamlining. Genome Biol. Evol. *4*, 212–229.
Dacks, J.B., and Field, M.C. (2018). Evolutionary origins and specialisation of membrane transport. Curr. Opin. Cell Biol. *53*, 70–76.
Dacks, J.B., Peden, A.A., and Field, M.C. (2009). Evolution of specificity in the eukaryotic endomembrane system. Int. J. Biochem. Cell Biol. *41*, 330–340.
Dacks, J.B., Field, M.C., Buick, R., Eme, L., Gribaldo, S., Roger, A.J., Brochier-Armanet, C., and Devos, D.P. (2016). The changing view of eukaryogenesis – fossils, cells, lineages and how they all come together. J. Cell Sci. *129*, 3695–3703.
Darwin, C. (1859). On the Origin of Species by Means of Natural Selection, or the Preservation of Favoured Races in the Struggle for Life (London: John Murray).
Dayie, K.T., and Padgett, R.A. (2008). A glimpse into the active site of a group II intron and maybe the spliceosome, too. RNA *14*, 1697–1703.

Delaney, K.J., Williams, S.G., Lawler, M., and Hall, K.B. (2014). Climbing the vertebrate branch of U1A/U2B" protein evolution. RNA *20*, 1035–1045.

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

Derelle, R., Philippe, H., and Colbourne, J.K. (2020). Broccoli: Combining Phylogenetic and Network Analyses for Orthology Assignment. Mol. Biol. Evol. *37*, 3389–3396.

Deutekom, E.S., Vosseberg, J., van Dam, T.J.P., and 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.

Deutekom, E.S., Snel, B., and van Dam, T.J.P. (2021). Benchmarking orthology methods using phylogenetic patterns defined at the base of Eukaryotes. Brief. Bioinform. *22*, bbaa206.

Dibb, N.J., and Newman, A.J. (1989). Evidence that introns arose at proto-splice sites. EMBO J. *8*, 2015–2021.

van Dijk, B., Bertels, F., Stolk, L., Takeuchi, N., and Rainey, P.B. (2022). Transposable elements promote the evolution of genome streamlining. Phil. Trans. R. Soc. B *377*, 20200477.

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

Doolittle, W.F. (1978). Genes in pieces: were they ever together? Nature *272*, 581–582.

Doolittle, W.F. (1987). What Introns Have to Tell Us: Hierarchy in Genome Evolution. Cold Spring Harb. Symp. Quant. Biol. *52*, 907–913.

Doolittle, W.F. (2013). The spliceosomal catalytic core arose in the RNA world… or did it? Genome Biol. *14*, 141.

Doolittle, W.F. (2014). The trouble with (group II) introns. Proc. Natl. Acad. Sci. U. S. A. *111*, 6536–6537.

Doolittle, W.F. (2016). Making the Most of Clade Selection. Philos. Sci. *84*, 275–295.

Douris, V., Telford, M.J., and Averof, M. (2010). Evidence for Multiple Independent Origins of *trans*-Splicing in Metazoa. Mol. Biol. Evol. *27*, 684–693.

von der Dunk, S.H.A., and Snel, B. (2020). Recurrent sequence evolution after independent gene duplication. BMC Evol. Biol. *20*, 98.

Eddy, S.R. (2011). Accelerated Profile HMM Searches. PLOS Comput. Biol. *7*, e1002195.

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

Embley, T.M., and Hirt, R.P. (1998). Early branching eukaryotes? Curr. Opin. Genet. Dev. *8*, 624–629.

Esser, C., Ahmadinejad, N., Wiegand, C., Rotte, C., Sebastiani, F., Gelius-Dietrich, G., Henze, K., Kretschmann, E., Richly, E., Leister, D., et al. (2004). A genome phylogeny for mitochondria among α-proteobacteria and a predominantly eubacterial ancestry of yeast nuclear genes. Mol. Biol. Evol. *21*, 1643–1660.

Ettema, T.J.G. (2016). Evolution: Mitochondria in the second act. Nature *531*, 39–40.

Fabrizio, P., Laggerbauer, B., Lauber, J., Lane, W.S., and Lührmann, R. (1997). An evolutionarily conserved U5 snRNP-specific protein is a GTP-binding factor closely related to the ribosomal translocase EF-2. EMBO J. *16*, 4092–4106.

Farag, I.F., Zhao, R., and Biddle, J.F. (2021). "Sifarchaeota," a Novel Asgard Phylum from Costa Rican Sediment Capable of Polysaccharide Degradation and Anaerobic Methylotrophy. Appl. Environ. Microbiol. *87*, e02584-20.

Fica, S.M., Tuttle, N., Novak, T., Li, N.-S., Lu, J., Koodathingal, P., Dai, Q., Staley, J.P., and Piccirilli, J.A. (2013). RNA catalyses nuclear pre-mRNA splicing. Nature *503*, 229–234.

Findeisen, P., Mühlhausen, S., Dempewolf, S., Hertzog, J., Zietlow, A., Carlomagno, T., and Kollmar, M. (2014). Six Subgroups and Extensive Recent Duplications Characterize the Evolution of the Eukaryotic Tubulin Protein Family. Genome Biol. Evol. *6*, 2274–2288.

Finn, R.D., Coggill, P., Eberhardt, R.Y., Eddy, S.R., Mistry, J., Mitchell, A.L., Potter, S.C., 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, G.C., Hanson-Smith, V., Stevens, T.H., and Thornton, J.W. (2012). Evolution of increased complexity in a molecular machine. Nature *481*, 360–364.

Fitch, W.M. (1970). Distinguishing Homologous from Analogous Proteins. Syst. Biol. *19*, 99–113.

Force, A., Lynch, M., Pickett, F.B., Amores, A., Yan, Y., and Postlethwait, J. (1999). Preservation of Duplicate Genes by Complementary, Degenerative Mutations. Genetics *151*, 1531–1545.

Fritz-Laylin, L.K., Prochnik, S.E., Ginger, M.L., Dacks, J.B., Carpenter, M.L., Field, M.C., Kuo, A., Paredez, A., Chapman, J., Pham, J., et al. (2010). The genome of *Naegleria gruberi* illuminates early eukaryotic versatility. Cell *140*, 631–642.

Gabaldón, T. (2018). Relative timing of mitochondrial endosymbiosis and the "pre-mitochondrial symbioses" hypothesis. IUBMB Life *70*, 1188–1196.

Gabaldón, T., and Koonin, E.V. (2013). Functional and evolutionary implications of gene orthology. Nat. Rev. Genet. *14*, 360–366.

Gabaldón, T., Rainey, D., and Huynen, M.A. (2005). Tracing the Evolution of a Large Protein Complex in the Eukaryotes, NADH:Ubiquinone Oxidoreductase (Complex I). J. Mol. Biol. *348*, 857–870.

Galej, W.P., Oubridge, C., Newman, A.J., and Nagai, K. (2013). Crystal structure of Prp8 reveals active site cavity of the spliceosome. Nature *493*, 638–643.

Galej, W.P., Nguyen, T.H.D., Newman, A.J., and Nagai, K. (2014). Structural studies of the spliceosome: zooming into the heart of the machine. Curr. Opin. Struct. Biol. *25*, 57–66.

Galindo, L.J., López-García, P., Torruella, G., Karpov, S., and Moreira, D. (2021). Phylogenomics of a new fungal phylum reveals multiple waves of reductive evolution across Holomycota. Nat. Commun. *12*, 4973.

Garg, S.G., and Martin, W.F. (2016). Mitochondria, the Cell Cycle, and the Origin of Sex via a Syncytial Eukaryote Common Ancestor. Genome Biol. Evol. *8*, 1950–1970.

Gilbert, W. (1978). Why genes in pieces? Nature *271*, 501–501.

González-Pech, R.A., Stephens, T.G., and Chan, C.X. (2019). Commonly misunderstood parameters of NCBI BLAST and important considerations for users. Bioinformatics *35*, 2697–2698.

Gordon, P.M., and Piccirilli, J.A. (2001). Metal ion coordination by the AGC triad in domain 5 contributes to group II intron catalysis. Nat. Struct. Mol. Biol. *8*, 893–898.

Gould, S.B., Garg, S.G., and Martin, W.F. (2016). Bacterial Vesicle Secretion and the Evolu-

tionary Origin of the Eukaryotic Endomembrane System. Trends Microbiol. *24*, 525–534.
Graf, J.S., Schorn, S., Kitzinger, K., Ahmerkamp, S., Woehle, C., Huettel, B., Schubert, C.J., Kuypers, M.M.M., and Milucka, J. (2021). Anaerobic endosymbiont generates energy for ciliate host by denitrification. Nature *591*, 445–450.
Grau-Bové, X., Navarrete, C., Chiva, C., Pribasnig, T., Antó, M., Torruella, G., Galindo, L.J., Lang, B.F., Moreira, D., López-Garcia, P., et al. (2022). A phylogenetic and proteomic reconstruction of eukaryotic chromatin evolution. Nat. Ecol. Evol. *6*, 1007–1023.
Gray, M.W., Lukeš, J., Archibald, J.M., Keeling, P.J., and Doolittle, W.F. (2010). Irremediable Complexity? Science *330*, 920–921.
Greening, C., and Lithgow, T. (2020). Formation and function of bacterial organelles. Nat. Rev. Microbiol. *18*, 677–689.
Guindon, S., Dufayard, J.-F., Lefort, V., Anisimova, M., Hordijk, W., and Gascuel, O. (2010). New Algorithms and Methods to Estimate Maximum-Likelihood Phylogenies: Assessing the Performance of PhyML 3.0. Syst. Biol. *59*, 307–321.
Guy, L., and Ettema, T.J.G. (2011). The archaeal 'TACK' superphylum and the origin of eukaryotes. Trends Microbiol. *19*, 580–587.
Hampl, V., Čepička, I., and Eliáš, M. (2019). Was the Mitochondrion Necessary to Start Eukaryogenesis? Trends Microbiol. *27*, 96–104
Hankeln, T., Friedl, H., Ebersberger, I., Martin, J., and Schmidt, E.R. (1997). A variable intron distribution in globin genes of *Chironomus*: evidence for recent intron gain. Gene *205*, 151–160.
Harms, M.J., and Thornton, J.W. (2010). Analyzing protein structure and function using ancestral gene reconstruction. Curr. Opin. Struct. Biol. *20*, 360–366.
Harms, M.J., and Thornton, J.W. (2013). Evolutionary biochemistry: revealing the historical and physical causes of protein properties. Nat. Rev. Genet. *14*, 559–571.
Harris, C.R., Millman, K.J., van der Walt, S.J., Gommers, R., Virtanen, P., Cournapeau, D., Wieser, E., Taylor, J., Berg, S., Smith, N.J., et al. (2020). Array programming with NumPy. Nature *585*, 357–362.
Hatano, T., Palani, S., Papatziamou, D., Salzer, R., Souza, D.P., Tamarit, D., Makwana, M., Potter, A., Haig, A., Xu, W., et al. (2022). Asgard archaea shed light on the evolutionary origins of the eukaryotic ubiquitin-ESCRT machinery. Nat. Commun. *13*, 3398.
Hauser, M., Mayer, C.E., and Söding, J. (2013). kClust: fast and sensitive clustering of large protein sequence databases. BMC Bioinform. *14*, 248.
Hendrickson, H.L., and Poole, A.M. (2018). Manifold Routes to a Nucleus. Front. Microbiol. *9*, 2604.
Hetzer, M., Wurzer, G., Schweyen, R.J., and Mueller, M.W. (1997). *Trans*-activation of group II intron splicing by nuclear U5 snRNA. Nature *386*, 417–420.
Hoang, D.T., Chernomor, O., von Haeseler, A., Minh, B.Q., and Vinh, L.S. (2018). UFBoot2: Improving the Ultrafast Bootstrap Approximation. Mol. Biol. Evol. *35*, 518–522.
Hogeweg, P., and Konings, D.A.M. (1985). U1 snRNA: The evolution of its primary and secondary structure. J. Mol. Evol. *21*, 323–333.
Höhna, S., Landis, M.J., Heath, T.A., Boussau, B., Lartillot, N., Moore, B.R., Huelsenbeck, J.P., and Ronquist, F. (2016). RevBayes: Bayesian Phylogenetic Inference Using Graphical Models and an Interactive Model-Specification Language. Syst. Biol. *65*, 726–736.

van Hooff, J.J.E., Tromer, E., van Wijk, L.M., Snel, B., and Kops, G.J.P.L. (2017). Evolutionary dynamics of the kinetochore network in eukaryotes as revealed by comparative genomics. EMBO Rep. *18*, 1559–1571.

van Hooff, J.J.E., Tromer, E., van Dam, T.J.P., Kops, G.J.P.L., and Snel, B. (2019). Inferring the Evolutionary History of Your Favorite Protein: A Guide for Molecular Biologists. BioEssays *41*, 1900006.

Huang, J.-M., Baker, B.J., Li, J.-T., and 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, A.J., Moore, A.N., Elniski, D., Joseph, J., Yee, J., and Russell, A.G. (2012). Evolutionarily divergent spliceosomal snRNAs and a conserved non-coding RNA processing motif in *Giardia lamblia*. Nucleic Acids Res. *40*, 10995–11008.

Hudson, A.J., Stark, M.R., Fast, N.M., Russell, A.G., and Rader, S.D. (2015). Splicing diversity revealed by reduced spliceosomes in *C. merolae* and other organisms. RNA Biol. *12*, 1–8.

Hudson, A.J., McWatters, D.C., Bowser, B.A., Moore, A.N., Larue, G.E., Roy, S.W., and Russell, A.G. (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., Capella-Gutiérrez, S., Pryszcz, L.P., Marcet-Houben, M., and Gabaldón, T. (2014). PhylomeDB v4: zooming into the plurality of evolutionary histories of a genome. Nucleic Acids Res. *42*, D897–D902.

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

Huerta-Cepas, J., Serra, F., and Bork, P. (2016b). ETE 3: reconstruction, analysis, and visualization of phylogenomic data. Mol. Biol. Evol. *33*, 1635–1638.

Huerta-Cepas, J., Forslund, K., Coelho, L.P., Szklarczyk, D., Jensen, L.J., von Mering, C., and 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., Heller, D., Hernández-Plaza, A., Forslund, S.K., Cook, H., Mende, D.R., Letunic, I., Rattei, T., Jensen, L.J., et al. (2019). eggNOG 5.0: a hierarchical, functionally and phylogenetically annotated orthology resource based on 5090 organisms and 2502 viruses. Nucleic Acids Res. *47*, D309–D314.

Hugoson, E., Guliaev, A., Ammunét, T., and Guy, L. (2022). Host Adaptation in Legionellales Is 1.9 Ga, Coincident with Eukaryogenesis. Mol. Biol. Evol. *39*, msac037.

Hunter, J.D. (2007). Matplotlib: A 2D Graphics Environment. Comput. Sci. Eng. *9*, 90–95.

Imachi, H., Nobu, M.K., Nakahara, N., Morono, Y., Ogawara, M., Takaki, Y., Takano, Y., Uematsu, K., Ikuta, T., Ito, M., et al. (2020). Isolation of an archaeon at the prokaryote–eukaryote interface. Nature *577*, 519–525.

Irwin, N.A.T., Pittis, A.A., Richards, T.A., and Keeling, P.J. (2022). Systematic evaluation of horizontal gene transfer between eukaryotes and viruses. Nat. Microbiol. *7*, 327–336.

Jacquier, A., and Michel, F. (1990). Base-pairing interactions involving the 5′and 3′-terminal nucleotides of group II self-splicing introns. J. Mol. Biol. *213*, 437–447.

Jain, C. (2008). The *E. coli* RhlE RNA helicase regulates the function of related RNA helicases during ribosome assembly. RNA *14*, 381–389.

James, R.H., Caceres, E.F., Escasinas, A., Alhasan, H., Howard, J.A., Deery, M.J., Ettema, T.J.G., and Robinson, N.P. (2017). Functional reconstruction of a eukaryotic-like E1/E2/(RING) E3 ubiquitylation cascade from an uncultured archaeon. Nat. Commun. *8*, 1120.

Jamy, M., Biwer, C., Vaulot, D., Obiol, A., Jing, H., Peura, S., Massana, R., and Burki, F. (2022). Global patterns and rates of habitat transitions across the eukaryotic tree of life. Nat. Ecol. Evol. *6*, 1458–1470.

Jékely, G. (2003). Small GTPases and the evolution of the eukaryotic cell. BioEssays *25*, 1129–1138.

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.

Kalvari, I., Nawrocki, E.P., Ontiveros-Palacios, N., Argasinska, J., Lamkiewicz, K., Marz, M., Griffiths-Jones, S., Toffano-Nioche, C., Gautheret, D., Weinberg, Z., et al. (2021). Rfam 14: expanded coverage of metagenomic, viral and microRNA families. Nucleic Acids Res. *49*, D192–D200.

Kalyaanamoorthy, S., Minh, B.Q., Wong, T.K.F., von Haeseler, A., and Jermiin, L.S. (2017). ModelFinder: fast model selection for accurate phylogenetic estimates. Nat. Methods *14*, 587–589.

Katayama, T., Nobu, M.K., Kusada, H., Meng, X.-Y., Hosogi, N., Uematsu, K., Yoshioka, H., Kamagata, Y., and Tamaki, H. (2020). Isolation of a member of the candidate phylum 'Atribacteria' reveals a unique cell membrane structure. Nat. Commun. *11*, 6381.

Katoh, K., and Standley, D.M. (2013). MAFFT multiple sequence alignment software version 7: improvements in performance and usability. Mol. Biol. Evol. *30*, 772–780.

Katoh, K., and Toh, H. (2008). Recent developments in the MAFFT multiple sequence alignment program. Brief. Bioinform. *9*, 286–298.

Kauko, A., and Lehto, K. (2018). Eukaryote specific folds: Part of the whole. Proteins *86*, 868–881.

Keating, K.S., Toor, N., Perlman, P.S., and Pyle, A.M. (2010). A structural analysis of the group II intron active site and implications for the spliceosome. RNA *16*, 1–9.

Kim, D., Lee, J., Cho, C.H., Kim, E.J., Bhattacharya, D., and Yoon, H.S. (2022). Group II intron and repeat-rich red algal mitochondrial genomes demonstrate the dynamic recent history of autocatalytic RNAs. BMC Biol. *20*, 2.

Klinger, C.M., Spang, A., Dacks, J.B., and Ettema, T.J.G. (2016). Tracing the archaeal origins of eukaryotic membrane-trafficking system building blocks. Mol. Biol. Evol. *33*, 1528–1541.

Kollmar, M. (2016). Fine-Tuning Motile Cilia and Flagella: Evolution of the Dynein Motor Proteins from Plants to Humans at High Resolution. Mol. Biol. Evol. *33*, 3249–3267.

Kollmar, M., and Mühlhausen, S. (2017). Myosin repertoire expansion coincides with eukaryotic diversification in the Mesoproterozoic era. BMC Evol. Biol. *17*, 211.

Koonin, E.V. (2005). Orthologs, Paralogs, and Evolutionary Genomics. Annu. Rev. Genet. *39*, 309–338.

Koonin, E.V. (2006). The origin of introns and their role in eukaryogenesis: a compromise solution to the introns-early versus introns-late debate? Biol. Direct *1*, 22.

Koonin, E.V. (2009). Intron-Dominated Genomes of Early Ancestors of Eukaryotes. J. Hered. *100*, 618–623.

Koonin, E.V. (2015). Energetics and population genetics at the root of eukaryotic cellular and genomic complexity. Proc. Natl. Acad. Sci. U. S. A. *112*, 15777–15778.

Koonin, E.V. (2016). Splendor and misery of adaptation, or the importance of neutral null for understanding evolution. BMC Biol. *14*, 114.

Koonin, E.V., Csuros, M., and Rogozin, I.B. (2013). Whence genes in pieces: reconstruction of the exon–intron gene structures of the last eukaryotic common ancestor and other ancestral eukaryotes. Wiley Interdiscip. Rev. RNA *4*, 93–105.

Korneta, I., Magnus, M., and Bujnicki, J.M. (2012). Structural bioinformatics of the human spliceosomal proteome. Nucleic Acids Res. *40*, 7046–7065.

Koumandou, V.L., Wickstead, B., Ginger, M.L., Giezen, M. van der, Dacks, J.B., and Field, M.C. (2013). Molecular paleontology and complexity in the last eukaryotic common ancestor. Crit. Rev. Biochem. Mol. Biol. *48*, 373–396.

Ku, C., Nelson-Sathi, S., Roettger, M., Garg, S., Hazkani-Covo, E., and Martin, W.F. (2015). Endosymbiotic gene transfer from prokaryotic pangenomes: Inherited chimerism in eukaryotes. Proc. Natl. Acad. Sci. U. S. A. *112*, 10139–10146.

Kudla, J., Albertazzi, F., Blazević, D., Hermann, M., and Bock, R. (2002). Loss of the mitochondrial *cox2* intron 1 in a family of monocotyledonous plants and utilization of mitochondrial intron sequences for the construction of a nuclear intron. Mol. Gen. Genom. *267*, 223–230.

Kufel, J., Allmang, C., Petfalski, E., Beggs, J., and Tollervey, D. (2003). Lsm Proteins Are Required for Normal Processing and Stability of Ribosomal RNAs*. J. Biol. Chem. *278*, 2147–2156.

Kunkel, G.R., Maser, R.L., Calvet, J.P., and Pederson, T. (1986). U6 small nuclear RNA is transcribed by RNA polymerase III. Proc. Natl. Acad. Sci. U. S. A. *83*, 8575–8579.

Lambowitz, A.M., and Belfort, M. (2015). Mobile Bacterial Group II Introns at the Crux of Eukaryotic Evolution. Microbiol. Spectr. *3*, 3.1.04.

Lane, N. (2011). Energetics and genetics across the prokaryote-eukaryote divide. Biol. Direct *6*, 35.

Lane, N. (2014). Bioenergetic constraints on the evolution of complex life. Cold Spring Harb. Perspect. Biol. *6*, a015982.

Lane, N. (2017). Serial endosymbiosis or singular event at the origin of eukaryotes? J. Theor. Biol. *434*, 58–67.

Lane, N. (2020). How energy flow shapes cell evolution. Curr. Biol. *30*, R471–R476.

Lane, N., and Martin, W. (2010). The energetics of genome complexity. Nature *467*, 929–934.

Lane, N., and Martin, W.F. (2015). Eukaryotes really are special, and mitochondria are why. Proc. Natl. Acad. Sci. U. S. A. *112*, E4823–E4823.

Lane, C.E., Heuvel, K. van den, Kozera, C., Curtis, B.A., Parsons, B.J., Bowman, S., and Archibald, J.M. (2007). Nucleomorph genome of *Hemiselmis andersenii* reveals complete intron loss and compaction as a driver of protein structure and function. Proc. Natl. Acad. Sci. U. S. A. *104*, 19908–19913.

Larue, G.E., Eliáš, M., and Roy, S.W. (2021). Expansion and transformation of the minor spliceosomal system in the slime mold *Physarum polycephalum*. Curr. Biol. *31*, 3125-3131.e4.

Lasda, E.L., and Blumenthal, T. (2011). *Trans*-splicing. Wiley Interdiscip. Rev. RNA *2*, 417–434.

Lax, G., Eglit, Y., Eme, L., Bertrand, E.M., Roger, A.J., and Simpson, A.G.B. (2018). Hemimastigophora is a novel supra-kingdom-level lineage of eukaryotes. Nature *564*, 410–414.

Le, S.Q., Gascuel, O., and Lartillot, N. (2008). Empirical profile mixture models for phylogenetic reconstruction. Bioinformatics *24*, 2317–2323.

Le, S.Q., Dang, C.C., and Gascuel, O. (2012). Modeling protein evolution with several amino acid replacement matrices depending on site rates. Mol. Biol. Evol. *29*, 2921–2936.

Lee, G., Sherer, N.A., Kim, N.H., Rajic, E., Kaur, D., Urriola, N., Martini, K.M., Xue, C., Goldenfeld, N., and Kuhlman, T.E. (2018). Testing the retroelement invasion hypothesis for the emergence of the ancestral eukaryotic cell. Proc. Natl. Acad. Sci. U. S. A. *115*, 12465–12470.

Lei, Q., Li, C., Zuo, Z., Huang, C., Cheng, H., and Zhou, R. (2016). Evolutionary Insights into RNA trans-Splicing in Vertebrates. Genome Biol. Evol. 8, 562–577.

Lekontseva, N.V., Stolboushkina, E.A., and Nikulin, A.D. (2021). Diversity of LSM Family Proteins: Similarities and Differences. Biochem. (Mosc.) *86*, S38–S49.

Lerner, M.R., and Steitz, J.A. (1979). Antibodies to small nuclear RNAs complexed with proteins are produced by patients with systemic lupus erythematosus. Proc. Natl. Acad. Sci. U. S. A. *76*, 5495–5499.

Lin, C.-F., Mount, S.M., Jarmołowski, A., and Makałowski, W. (2010). Evolutionary dynamics of U12-type spliceosomal introns. BMC Evol. Biol. *10*, 47.

Liu, S., Rauhut, R., Vornlocher, H.-P., and Lührmann, R. (2006). The network of protein–protein interactions within the human U4/U6.U5 tri-snRNP. RNA *12*, 1418–1430.

Liu, Y., Zhou, Z., Pan, J., Baker, B.J., Gu, J.-D., and Li, M. (2018). Comparative genomic inference suggests mixotrophic lifestyle for Thorarchaeota. ISME J. *12*, 1021–1031.

Liu, Y., Makarova, K.S., Huang, W.-C., Wolf, Y.I., Nikolskaya, A.N., 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.

Lo Gullo, G., De Santis, M.L., Paiardini, A., Rosignoli, S., Romagnoli, A., La Teana, A., Londei, P., and Benelli, D. (2021). The Archaeal Elongation Factor EF-2 Induces the Release of aIF6 From 50S Ribosomal Subunit. Front. Microbiol. *12*, 631297.

Long, M., and Deutsch, M. (1999). Association of intron phases with conservation at splice site sequences and evolution of spliceosomal introns. Mol. Biol. Evol. *16*, 1528–1534.

López, M.D., Rosenblad, M.A., and Samuelsson, T. (2008). Computational screen for spliceosomal RNA genes aids in defining the phylogenetic distribution of major and minor spliceosomal components. Nucleic Acids Res. *36*, 3001–3010.

López-García, P., and Moreira, D. (2006). Selective forces for the origin of the eukaryotic nucleus. BioEssays *28*, 525–533.

López-García, P., and Moreira, D. (2020). The Syntrophy hypothesis for the origin of eukaryotes revisited. Nat. Microbiol. *5*, 655–667.

Lukeš, J., Archibald, J.M., Keeling, P.J., Doolittle, W.F., and Gray, M.W. (2011). How a neutral evolutionary ratchet can build cellular complexity. IUBMB Life *63*, 528–537.

Lynch, M. (2007). The frailty of adaptive hypotheses for the origins of organismal complexity. Proc. Natl. Acad. Sci. U. S. A. *104*, 8597–8604.

Lynch, M. (2012). The Evolution of Multimeric Protein Assemblages. Mol. Biol. Evol. *29*,

1353–1366.
Lynch, M., and Marinov, G.K. (2015). The bioenergetic costs of a gene. Proc. Natl. Acad. Sci. U. S. A. *112*, 15690–15695.
Lynch, M., and Marinov, G.K. (2016). Reply to Lane and Martin: Mitochondria do not boost the bioenergetic capacity of eukaryotic cells. Proc. Natl. Acad. Sci. U. S. A. *113*, E667–E668.
Lynch, M., and Marinov, G.K. (2017). Membranes, energetics, and evolution across the prokaryote-eukaryote divide. eLife *6*, e20437.
Lynch, M., Field, M.C., Goodson, H.V., Malik, H.S., Pereira-Leal, J.B., Roos, D.S., Turkewitz, A.P., and Sazer, S. (2014). Evolutionary cell biology: Two origins, one objective. Proc. Natl. Acad. Sci. U. S. A. *111*, 16990–16994.
Maeso, I., Roy, S.W., and Irimia, M. (2012). Widespread Recurrent Evolution of Genomic Features. Genome Biol. Evol. *4*, 486–500.
Makarova, K.S., Wolf, Y.I., Mekhedov, S.L., Mirkin, B.G., and Koonin, E.V. (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., and Koonin, E.V. (2004). Comparative Genomics, Evolution and Origins of the Nuclear Envelope and Nuclear Pore Complex. Cell Cycle *3*, 1625–1650.
Maris, C., Dominguez, C., and Allain, F.H.-T. (2005). The RNA recognition motif, a plastic RNA-binding platform to regulate post-transcriptional gene expression. FEBS J. *272*, 2118–2131.
Martijn, J., and Ettema, T.J.G. (2013). From archaeon to eukaryote: the evolutionary dark ages of the eukaryotic cell. Biochem. Soc. Trans. *41*, 451–457.
Martijn, J., Vosseberg, J., Guy, L., Offre, P., and Ettema, T.J.G. (2018). Deep mitochondrial origin outside the sampled alphaproteobacteria. Nature *557*, 101–105.
Martin, W., and Koonin, E.V. (2006). Introns and the origin of nucleus–cytosol compartmentalization. Nature *440*, 41–45.
Martin, W., and Müller, M. (1998). The hydrogen hypothesis for the first eukaryote. Nature *392*, 37–41.
Martin, W.F., Roettger, M., Ku, C., Garg, S.G., Nelson-Sathi, S., and Landan, G. (2017). Late mitochondrial origin is an artifact. Genome Biol. Evol. *9*, 373–379.
Mast, F.D., Barlow, L.D., Rachubinski, R.A., and Dacks, J.B. (2014). Evolutionary mechanisms for establishing eukaryotic cellular complexity. Trends Cell Biol. *24*, 435–442.
Matera, A.G., and Wang, Z. (2014). A day in the life of the spliceosome. Nat. Rev. Mol. Cell Biol. *15*, 108–121.
Matera, A.G., Terns, R.M., and Terns, M.P. (2007). Non-coding RNAs: lessons from the small nuclear and small nucleolar RNAs. Nat. Rev. Mol. Cell Biol. *8*, 209–220.
Matsuura, M., Noah, J.W., and Lambowitz, A.M. (2001). Mechanism of maturase-promoted group II intron splicing. EMBO J. *20*, 7259–7270.
Mayes, A.E., Verdone, L., Legrain, P., and Beggs, J.D. (1999). Characterization of Sm-like proteins in yeast and their association with U6 snRNA. EMBO J. *18*, 4321–4331.
Maynard Smith, J., and Szathmáry, E. (1995). The Major Transitions in Evolution (Oxford: Freeman).
McInerney, J.O., O'Connell, M.J., and Pisani, D. (2014). The hybrid nature of the Eukaryota

and a consilient view of life on Earth. Nat. Rev. Microbiol. *12*, 449–455.
McKinney, W. (2010). Data Structures for Statistical Computing in Python. In Proceedings of the 9th Python in Science Conference (SciPy 2010) (Austin, Texas), pp. 56–61.
Méheust, R., Bhattacharya, D., Pathmanathan, J.S., McInerney, J.O., Lopez, P., and Bapteste, E. (2018). Formation of chimeric genes with essential functions at the origin of eukaryotes. BMC Biol. *16*, 30.
Mehta, A.P., Supekova, L., Chen, J.-H., Pestonjamasp, K., Webster, P., Ko, Y., Henderson, S.C., McDermott, G., Supek, F., and Schultz, P.G. (2018). Engineering yeast endosymbionts as a step toward the evolution of mitochondria. Proc. Natl. Acad. Sci. U. S. A. *115*, 11796–11801.
Minh, B.Q., Nguyen, M.A.T., and von Haeseler, A. (2013). Ultrafast approximation for phylogenetic bootstrap. Mol. Biol. Evol. *30*, 1188–1195.
Minh, B.Q., Schmidt, H.A., Chernomor, O., Schrempf, D., Woodhams, M.D., von Haeseler, A., and 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, M.C., Nagata, S., Tamaki, S., Tomita, M., and 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., and Bateman, A. (2022). Reciprocal Best Structure Hits: Using AlphaFold models to discover distant homologues. Bioinform. Adv. *2*, vbac072.
More, K., Klinger, C.M., Barlow, L.D., and Dacks, J.B. (2020). Evolution and Natural History of Membrane Trafficking in Eukaryotes. Curr. Biol. *30*, R553–R564.
Moreira, D., and López-García, P. (1998). Symbiosis Between Methanogenic Archaea and δ-Proteobacteria as the Origin of Eukaryotes: The Syntrophic Hypothesis. J. Mol. Evol. *47*, 517–530.
Moyer, D.C., Larue, G.E., Hershberger, C.E., Roy, S.W., and Padgett, R.A. (2020). Comprehensive database and evolutionary dynamics of U12-type introns. Nucleic Acids Res. *48*, 7066–7078.
Mozaffari-Jovin, S., Wandersleben, T., Santos, K.F., Will, C.L., Lührmann, R., and Wahl, M.C. (2013). Inhibition of RNA Helicase Brr2 by the C-Terminal Tail of the Spliceosomal Protein Prp8. Science *341*, 80–84.
Muñoz-Gómez, S.A., Susko, E., Williamson, K., Eme, L., Slamovits, C.H., Moreira, D., López-García, P., and Roger, A.J. (2022). Site-and-branch-heterogeneous analyses of an expanded dataset favour mitochondria as sister to known Alphaproteobacteria. Nat. Ecol. Evol. *6*, 253–262.
Mura, C., Phillips, M., Kozhukhovsky, A., and Eisenberg, D. (2003). Structure and assembly of an augmented Sm-like archaeal protein 14-mer. Proc. Natl. Acad. Sci. U. S. A. *100*, 4539–4544.
Mura, C., Randolph, P.S., Patterson, J., and Cozen, A.E. (2013). Archaeal and eukaryotic homologs of Hfq. RNA Biol. *10*, 636–651.
Narrowe, A.B., Spang, A., Stairs, C.W., Caceres, E.F., Baker, B.J., Miller, C.S., and Ettema, T.J.G. (2018). Complex evolutionary history of translation elongation factor 2 and diphthamide biosynthesis in archaea and parabasalids. Genome Biol. Evol. *10*, 2380–2393.

Nawrocki, E.P., and Eddy, S.R. (2013). Infernal 1.1: 100-fold faster RNA homology searches. Bioinformatics *29*, 2933–2935.

Neveu, E., Khalifeh, D., Salamin, N., and Fasshauer, D. (2020). Prototypic SNARE Proteins Are Encoded in the Genomes of Heimdallarchaeota, Potentially Bridging the Gap between the Prokaryotes and Eukaryotes. Curr. Biol. *30*, 2468–2480.

Newman, A.J., and Norman, C. (1992). U5 snRNA interacts with exon sequences at 5′ and 3′ splice sites. Cell *68*, 743–754.

Nguyen, L.-T., Schmidt, H.A., von Haeseler, A., and Minh, B.Q. (2015). IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. Mol. Biol. Evol. *32*, 268–274.

Nguyen, T.H.D., Li, J., Galej, W.P., Oshikane, H., Newman, A.J., and Nagai, K. (2013). Structural Basis of Brr2-Prp8 Interactions and Implications for U5 snRNP Biogenesis and the Spliceosome Active Site. Structure *21*, 910–919.

Nottrott, S., Urlaub, H., and Lührmann, R. (2002). Hierarchical, clustered protein interactions with U4/U6 snRNA: a biochemical role for U4/U6 proteins. EMBO J. *21*, 5527–5538.

Ohno, S. (1970). Evolution by Gene Duplication (New York: Springer-Verlag).

O'Keefe, R.T., Norman, C., and Newman, A.J. (1996). The Invariant U5 snRNA Loop 1 Sequence Is Dispensable for the First Catalytic Step of pre-mRNA Splicing in Yeast. Cell *86*, 679–689.

O'Malley, M.A., Leger, M.M., Wideman, J.G., and Ruiz-Trillo, I. (2019). Concepts of the last eukaryotic common ancestor. Nat. Ecol. Evol. *3*, 338–344.

Peebles, C.L., Zhang, M., Perlman, P.S., and Franzen, J.S. (1995). Catalytically critical nucleotide in domain 5 of a group II intron. Proc. Natl. Acad. Sci. U. S. A. *92*, 4422–4426.

Peters, J.K., and Toor, N. (2015). Group II intron lariat: Structural insights into the spliceosome. RNA Biol. *12*, 913–917.

Pisani, D., Cotton, J.A., and McInerney, J.O. (2007). Supertrees disentangle the chimerical origin of eukaryotic genomes. Mol. Biol. Evol. *24*, 1752–1760.

Pittis, A.A., and Gabaldón, T. (2016a). Late acquisition of mitochondria by a host with chimaeric prokaryotic ancestry. Nature *531*, 101–104.

Pittis, A.A., and Gabaldón, T. (2016b). On phylogenetic branch lengths distribution and the late acquistion of mitochondria. Preprint at bioRxiv, 10.1101/064873.

Polycarpou-Schwarz, M., Gunderson, S.I., Kandels-Lewis, S., Seraphin, B., and Mattaj, I.W. (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.

Poole, A.M., and Gribaldo, S. (2014). Eukaryotic origins: how and when was the mitochondrion acquired? Cold Spring Harb. Perspect. Biol. *6*, a015990.

Porter, S.M. (2020). Insights into eukaryogenesis from the fossil record. Interface Focus *10*, 20190105.

Price, M.N., Dehal, P.S., and Arkin, A.P. (2010). FastTree 2 – approximately maximum-likelihood trees for large alignments. PLOS ONE *5*, e9490.

Qu, G., Dong, X., Piazza, C.L., Chalamcharla, V.R., Lutz, S., Curcio, M.J., and Belfort, M. (2014). RNA–RNA interactions and pre-mRNA mislocalization as drivers of group II intron loss from nuclear genomes. Proc. Natl. Acad. Sci. U. S. A. *111*, 6612–6617.

Qu, G., Kaushal, P.S., Wang, J., Shigematsu, H., Piazza, C.L., Agrawal, R.K., Belfort, M., and

Wang, H.-W. (2016). Structure of a group II intron in complex with its reverse transcriptase. Nat. Struct. Mol. Biol. *23*, 549–557.

Reifschneider, O., Marx, C., Jacobs, J., Kollipara, L., Sickmann, A., Wolters, D., and Kück, U. (2016). A Ribonucleoprotein Supercomplex Involved in *trans*-Splicing of Organelle Group II Introns. J. Biol. Chem. *291*, 23330–23342.

Richardson, D.N., Rogers, M.F., Labadorf, A., Ben-Hur, A., Guo, H., Paterson, A.H., and Reddy, A.S.N. (2011). Comparative Analysis of Serine/Arginine-Rich Proteins across 27 Eukaryotes: Insights into Sub-Family Classification and Extent of Alternative Splicing. PLOS ONE *6*, e24542.

Richardson, E., Zerr, K., Tsaousis, A., Dorrell, R.G., and Dacks, J.B. (2015). Evolutionary cell biology: functional insight from "endless forms most beautiful." Mol. Biol. Cell *26*, 4532–4538.

Roger, A.J., Muñoz-Gómez, S.A., and Kamikawa, R. (2017). The Origin and Diversification of Mitochondria. Curr. Biol. *27*, R1177–R1192.

Rogozin, I.B., Wolf, Y.I., Sorokin, A.V., Mirkin, B.G., and Koonin, E.V. (2003). Remarkable Interkingdom Conservation of Intron Positions and Massive, Lineage-Specific Intron Loss and Gain in Eukaryotic Evolution. Curr. Biol. *13*, 1512–1517.

Rogozin, I.B., Carmel, L., Csuros, M., and Koonin, E.V. (2012). Origin and evolution of spliceosomal introns. Biol. Direct *7*, 11.

Rout, M.P., and Field, M.C. (2017). The Evolution of Organellar Coat Complexes and Organization of the Eukaryotic Cell. Annu. Rev. Biochem. *86*, 637–657.

Roy, S.W., and Gilbert, W. (2005). The pattern of intron loss. Proc. Natl. Acad. Sci. U. S. A. *102*, 713–718.

Roy, S.W., and Irimia, M. (2014). Diversity and evolution of spliceosomal systems. Methods Mol. Biol. *1126*, 13–33.

Roy, S.W., Gozashti, L., Bowser, B.A., Weinstein, B.N., and Larue, G.E. (2020). Massive intron gain in the most intron-rich eukaryotes is driven by intron-like transposable elements of unprecedented diversity and flexibility. Preprint at bioRxiv, 10.1101/2020.10.14.339549.

Russell, A.G., Charette, J.M., Spencer, D.F., and Gray, M.W. (2006). An early evolutionary origin for the minor spliceosome. Nature *443*, 863–866.

Sagan, L. (1967). On the origin of mitosing cells. J. Theor. Biol. *14*, 225-IN6.

Saldanha, R., Chen, B., Wank, H., Matsuura, M., Edwards, J., and Lambowitz, A.M. (1999). RNA and Protein Catalysis in Group II Intron Splicing and Mobility Reactions Using Purified Components. Biochemistry *38*, 9069–9083.

Saldi, T., Wilusz, C., MacMorris, M., and Blumenthal, T. (2007). Functional redundancy of worm spliceosomal proteins U1A and U2B". Proc. Natl. Acad. Sci. U. S. A. *104*, 9753–9757.

Sales-Lee, J., Perry, D.S., Bowser, B.A., Diedrich, J.K., Rao, B., Beusch, I., Yates, J.R., Roy, S.W., and Madhani, H.D. (2021). Coupling of spliceosome complexity to intron diversity. Curr. Biol. *31*, 4898–4910.

Schavemaker, P.E., and Muñoz-Gómez, S.A. (2022). The role of mitochondrial energetics in the origin and diversification of eukaryotes. Nat. Ecol. Evol. *6*, 1307–1317.

Schlacht, A., Herman, E.K., Klute, M.J., Field, M.C., and Dacks, J.B. (2014). Missing Pieces of an Ancient Puzzle: Evolution of the Eukaryotic Membrane-Trafficking System. Cold Spring Harb. Perspect. Biol. *6*, a016048.

Schmitz-Linneweber, C., Lampe, M.-K., Sultan, L.D., and Ostersetzer-Biran, O. (2015). Organellar maturases: A window into the evolution of the spliceosome. Biochim. Biophys. Acta Bioenerg. *1847*, 798–808.

Schön, M.E., Zlatogursky, V.V., Singh, R.P., Poirier, C., Wilken, S., Mathur, V., Strassert, J.F.H., Pinhassi, J., Worden, A.Z., Keeling, P.J., et al. (2021). Single cell genomics reveals plastid-lacking Picozoa are close relatives of red algae. Nat. Commun. *12*, 6651.

Schön, M.E., Martijn, J., Vosseberg, J., Köstlbacher, S., and Ettema, T.J.G. (2022). The evolutionary origin of host association in the Rickettsiales. Nat. Microbiol. *7*, 1189–1199.

Schulz, L., Sendker, F.L., and Hochberg, G.K.A. (2022). Non-adaptive complexity and biochemical function. Curr. Opin. Struct. Biol. *73*, 102339.

Scofield, D.G., and Lynch, M. (2008). Evolutionary Diversification of the Sm Family of RNA-Associated Proteins. Mol. Biol. Evol. *25*, 2255–2267.

Seabold, S., and Perktold, J. (2010). Statsmodels: Econometric and Statistical Modeling with Python. In Proceedings of the 9th Python in Science Conference (SciPy 2010) (Austin, Texas), pp. 92–96.

Seemann, T. (2014). Prokka: rapid prokaryotic genome annotation. Bioinformatics *30*, 2068–2069.

Seitz, K.W., Dombrowski, N., Eme, L., Spang, A., Lombard, J., Sieber, J.R., Teske, A.P., Ettema, T.J.G., and Baker, B.J. (2019). Asgard archaea capable of anaerobic hydrocarbon cycling. Nat. Commun. *10*, 1822.

Shabalina, S.A., and Koonin, E.V. (2008). Origins and evolution of eukaryotic RNA interference. Trends Ecol. Evol. *23*, 578–587.

Shah, N., Nute, M.G., Warnow, T., and Pop, M. (2019). Misunderstood parameter of NCBI BLAST impacts the correctness of bioinformatics workflows. Bioinformatics *35*, 1613–1614.

Sharp, P.A. (1991). Five easy pieces. Science *254*, 663–663.

Sharp, P.A., and Burge, C.B. (1997). Classification of Introns: U2-Type or U12-Type. Cell *91*, 875–879.

Shepard, P.J., and Hertel, K.J. (2009). The SR protein family. Genome Biol. *10*, 242.

Shih, P.M., and Matzke, N.J. (2013). Primary endosymbiosis events date to the later Proterozoic with cross-calibrated phylogenetic dating of duplicated ATPase proteins. Proc. Natl. Acad. Sci. U. S. A. *110*, 12355–12360.

Shimodaira, H. (2002). An Approximately Unbiased Test of Phylogenetic Tree Selection. Syst. Biol. *51*, 492–508.

Shiratori, T., Suzuki, S., Kakizawa, Y., and Ishida, K. (2019). Phagocytosis-like cell engulfment by a planctomycete bacterium. Nat. Commun. *10*, 5529.

Shukla, G.C., and Padgett, R.A. (2002). A Catalytically Active Group II Intron Domain 5 Can Function in the U12-Dependent Spliceosome. Mol. Cell *9*, 1145–1150.

Simon, D.M., Kelchner, S.A., and Zimmerly, S. (2009). A Broadscale Phylogenetic Analysis of Group II Intron RNAs and Intron-Encoded Reverse Transcriptases. Mol. Biol. Evol. *26*, 2795–2808.

Söding, J. (2005). Protein homology detection by HMM–HMM comparison. Bioinformatics *21*, 951–960.

Sonnhammer, E.L.L., and Koonin, E.V. (2002). Orthology, paralogy and proposed classifica-

tion for paralog subtypes. Trends Genet. *18*, 619–620.
Sousa, F.L., Neukirchen, S., Allen, J.F., Lane, N., and Martin, W.F. (2016). Lokiarchaeon is hydrogen dependent. Nat. Microbiol. *1*, 16034.
Spang, A., Saw, J.H., Jørgensen, S.L., Zaremba-Niedzwiedzka, K., Martijn, J., Lind, A.E., van Eijk, R., Schleper, C., Guy, L., and Ettema, T.J.G. (2015). Complex archaea that bridge the gap between prokaryotes and eukaryotes. Nature *521*, 173–179.
Spang, A., Stairs, C.W., Dombrowski, N., Eme, L., Lombard, J., Caceres, E.F., Greening, C., Baker, B.J., and Ettema, T.J.G. (2019). Proposal of the reverse flow model for the origin of the eukaryotic cell based on comparative analyses of Asgard archaeal metabolism. Nat. Microbiol. *4*, 1138–1148.
Speijer, D. (2011). Does constructive neutral evolution play an important role in the origin of cellular complexity? BioEssays *33*, 344–349.
Stairs, C.W., and Ettema, T.J.G. (2020). The Archaeal Roots of the Eukaryotic Dynamic Actin Cytoskeleton. Curr. Biol. *30*, R521–R526.
Stairs, C.W., Dharamshi, J.E., Tamarit, D., Eme, L., Jørgensen, S.L., Spang, A., and Ettema, T.J.G. (2020). Chlamydial contribution to anaerobic metabolism during eukaryotic evolution. Sci. Adv. *6*, eabb7258.
Stark, M.R., Dunn, E.A., Dunn, W.S.C., Grisdale, C.J., Daniele, A.R., Halstead, M.R.G., Fast, N.M., and Rader, S.D. (2015). Dramatically reduced spliceosome in *Cyanidioschyzon merolae*. Proc. Natl. Acad. Sci. U. S. A. *112*, E1191–E1200.
Steinegger, M., Meier, M., Mirdita, M., Vöhringer, H., Haunsberger, S.J., and Söding, J. (2019). HH-suite3 for fast remote homology detection and deep protein annotation. BMC Bioinform. *20*, 473.
Stoltzfus, A. (1999). On the Possibility of Constructive Neutral Evolution. J. Mol. Evol. *49*, 169–181.
Stoltzfus, A. (2012). Constructive neutral evolution: exploring evolutionary theory's curious disconnect. Biol. Direct *7*, 35.
Strassert, J.F.H., Irisarri, I., Williams, T.A., and Burki, F. (2021). A molecular timescale for eukaryote evolution with implications for the origin of red algal-derived plastids. Nat. Commun. *12*, 1879.
Stroud, D.A., Surgenor, E.E., Formosa, L.E., Reljic, B., Frazier, A.E., Dibley, M.G., Osellame, L.D., Stait, T., Beilharz, T.H., Thorburn, D.R., et al. (2016). Accessory subunits are integral for assembly and function of human mitochondrial complex I. Nature *538*, 123–126.
Sun, J., Evans, P.N., Gagen, E.J., Woodcroft, B.J., Hedlund, B.P., Woyke, T., Hugenholtz, P., and Rinke, C. (2021). Recoding of stop codons expands the metabolic potential of two novel Asgardarchaeota lineages. ISME COMMUN. *1*, 30.
Survery, S., Hurtig, F., Haq, S.R., Eriksson, J., Guy, L., Rosengren, K.J., Lindås, A.-C., and Chi, C.N. (2021). Heimdallarchaea encodes profilin with eukaryotic-like actin regulation and polyproline binding. Commun. Biol. *4*, 1–14.
Susko, E., Steel, M., and Roger, A.J. (2021). Conditions under which distributions of edge length ratios on phylogenetic trees can be used to order evolutionary events. J. Theor. Biol. *526*, 110788.
Sverdlov, A.V., Csuros, M., Rogozin, I.B., and Koonin, E.V. (2007). A glimpse of a putative pre-intron phase of eukaryotic evolution. Trends Genet. *23*, 105–108.

Szathmáry, E. (2015). Toward major evolutionary transitions theory 2.0. Proc. Natl. Acad. Sci. U. S. A. *112*, 10104–10111.

Szöllősi, G.J., Rosikiewicz, W., Boussau, B., Tannier, E., and Daubin, V. (2013). Efficient Exploration of the Space of Reconciled Gene Trees. Syst. Biol. *62*, 901–912.

Tamarit, D., Caceres, E.F., Krupovic, M., Nijland, R., Eme, L., Robinson, N.P., and Ettema, T.J.G. (2022). A closed *Candidatus* Odinarchaeum chromosome exposes Asgard archaeal viruses. Nat. Microbiol. *7*, 948–952.

Tang, Y.H., Han, S.P., Kassahn, K.S., Skarshewski, A., Rothnagel, J.A., and Smith, R. (2012). Complex Evolutionary Relationships Among Four Classes of Modular RNA-Binding Splicing Regulators in Eukaryotes: The hnRNP, SR, ELAV-Like and CELF Proteins. J. Mol. Evol. *75*, 214–228.

The UniProt Consortium (2019). UniProt: a worldwide hub of protein knowledge. Nucleic Acids Res. *47*, D506–D515.

Toor, N., Keating, K.S., Taylor, S.D., and Pyle, A.M. (2008). Crystal Structure of a Self-Spliced Group II Intron. Science *320*, 77–82.

Toro, N., and Martínez-Abarca, F. (2013). Comprehensive Phylogenetic Analysis of Bacterial Group II Intron-Encoded ORFs Lacking the DNA Endonuclease Domain Reveals New Varieties. PLOS ONE *8*, e55102.

Toro, N., and Nisa-Martínez, R. (2014). Comprehensive Phylogenetic Analysis of Bacterial Reverse Transcriptases. PLOS ONE *9*, e114083.

Torriani, S.F.F., Stukenbrock, E.H., Brunner, P.C., McDonald, B.A., and Croll, D. (2011). Evidence for Extensive Recent Intron Transposition in Closely Related Fungi. Curr. Biol. *21*, 2017–2022.

Tria, F.D.K., Brückner, J., Skejo, J., Xavier, J.C., Zimorski, V., Gould, S.B., Garg, S.G., and Martin, W.F. (2019). Gene duplications trace mitochondria to the onset of eukaryote complexity. Preprint at bioRxiv, 10.1101/781211.

Tria, F.D.K., Brueckner, J., Skejo, J., Xavier, J.C., Kapust, N., Knopp, M., Wimmer, J.L.E., Nagies, F.S.P., Zimorski, V., Gould, S.B., et al. (2021). Gene Duplications Trace Mitochondria to the Onset of Eukaryote Complexity. Genome Biol. Evol. *13*, evab055.

Tricou, T., Tannier, E., and de Vienne, D.M. (2022). Ghost lineages can invalidate or even reverse findings regarding gene flow. PLOS Biol. *20*, e3001776.

Tromer, E., Bade, D., Snel, B., and Kops, G.J.P.L. (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, E.C., Hooff, J.J.E. van, Kops, G.J.P.L., and Snel, B. (2019). Mosaic origin of the eukaryotic kinetochore. Proc. Natl. Acad. Sci. U. S. A. *116*, 12873–12882.

Truong, D.M., Hewitt, F.C., Hanson, J.H., Cui, X., and Lambowitz, A.M. (2015). Retrohoming of a Mobile Group II Intron in Human Cells Suggests How Eukaryotes Limit Group II Intron Proliferation. PLOS Genet. *11*, e1005422.

Tully, B.J., Graham, E.D., and Heidelberg, J.F. (2018). The reconstruction of 2,631 draft metagenome-assembled genomes from the global oceans. Sci. Data *5*, 170203.

Turner, I.A., Norman, C.M., Churcher, M.J., and Newman, A.J. (2006). Dissection of Prp8 protein defines multiple interactions with crucial RNA sequences in the catalytic core of the spliceosome. RNA *12*, 375–386.

Turunen, J.J., Niemelä, E.H., Verma, B., and Frilander, M.J. (2013). The significant other: splicing by the minor spliceosome. Wiley Interdiscip. Rev. RNA *4*, 61–76.

Valadkhan, S., and Jaladat, Y. (2010). The spliceosomal proteome: At the heart of the largest cellular ribonucleoprotein machine. Proteomics *10*, 4128–4141.

Vargová, R., Wideman, J.G., Derelle, R., Klimeš, V., Kahn, R.A., Dacks, J.B., and Eliáš, M. (2021). A Eukaryote-Wide Perspective on the Diversity and Evolution of the ARF GTPase Protein Family. Genome Biol. Evol. *13*, evab157.

van der Veen, R., Arnberg, A.C., van der Horst, G., Bonen, L., Tabak, H.F., and Grivell, L.A. (1986). Excised group II introns in yeast mitochondria are lariats and can be formed by self-splicing in vitro. Cell *44*, 225–234.

Veretnik, S., Wills, C., Youkharibache, P., Valas, R.E., and Bourne, P.E. (2009). Sm/Lsm Genes Provide a Glimpse into the Early Evolution of the Spliceosome. PLOS Comput. Biol. *5*, e1000315.

Vesteg, M., Šándorová, Z., and Krajčovič, J. (2012). Selective forces for the origin of spliceosomes. J. Mol. Evol. *74*, 226–231.

Virtanen, P., Gommers, R., Oliphant, T.E., Haberland, M., Reddy, T., Cournapeau, D., Burovski, E., Peterson, P., Weckesser, W., Bright, J., et al. (2020). SciPy 1.0: fundamental algorithms for scientific computing in Python. Nat. Methods *17*, 261–272.

Volland, J.-M., Gonzalez-Rizzo, S., Gros, O., Tyml, T., Ivanova, N., Schulz, F., Goudeau, D., Elisabeth, N.H., Nath, N., Udwary, D., et al. (2022). A centimeter-long bacterium with DNA contained in metabolically active, membrane-bound organelles. Science *376*, 1453–1458.

Vosseberg, J., and 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, J.J.E., Marcet-Houben, M., van Vlimmeren, A., van Wijk, L.M., Gabaldón, T., and Snel, B. (2020). Data for: Timing the origin of eukaryotic cellular complexity with ancient duplications. Dataset at figshare, 10.6084/m9.figshare.10069985.v3.

Vosseberg, J., van Hooff, J.J.E., Marcet-Houben, M., van Vlimmeren, A., van Wijk, L.M., Gabaldón, T., and Snel, B. (2021a). Timing the origin of eukaryotic cellular complexity with ancient duplications. Nat. Ecol. Evol. *5*, 92–100.

Vosseberg, J., Schinkel, M., Gremmen, S., and Snel, B. (2021b). Data for: The spread of the first introns in proto-eukaryotic paralogs. Dataset at figshare, 10.6084/m9.figshare.16601744.v2.

Vosseberg, J., Gremmen, S., Schinkel, M., and Snel, B. (2022a). Code for: The spread of the first introns in proto-eukaryotic paralogs. Code at figshare, 10.6084/m9.figshare.19411820.v1.

Vosseberg, J., Schinkel, M., Gremmen, S., and Snel, B. (2022b). The spread of the first introns in proto-eukaryotic paralogs. Commun. Biol. *5*, 476.

Wahl, M.C., Will, C.L., and Lührmann, R. (2009). The Spliceosome: Design Principles of a Dynamic RNP Machine. Cell *136*, 701–718.

Wang, S., and Luo, H. (2021). Dating Alphaproteobacteria evolution with eukaryotic fossils. Nat. Commun. *12*, 3324.

Wang, Z., and Wu, M. (2014). Phylogenomic Reconstruction Indicates Mitochondrial Ancestor Was an Energy Parasite. PLOS ONE *9*, e110685.

Wank, H., SanFilippo, J., Singh, R.N., Matsuura, M., and Lambowitz, A.M. (1999). A Reverse

Transcriptase/Maturase Promotes Splicing by Binding at Its Own Coding Segment in a Group II Intron RNA. Mol. Cell *4*, 239–250.

Waskom, M.L. (2021). seaborn: statistical data visualization. J. Open Source Softw. *6*, 3021.

Waterhouse, A.M., Procter, J.B., Martin, D.M.A., Clamp, M., and Barton, G.J. (2009). Jalview Version 2—a multiple sequence alignment editor and analysis workbench. Bioinformatics *25*, 1189–1191.

Watkins, N.J., Ségault, V., Charpentier, B., Nottrott, S., Fabrizio, P., Bachi, A., Wilm, M., Rosbash, M., Branlant, C., and 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, G.T., Holton, N., Hall, K.B., and Wahl, M.C. (2018). Molecular principles underlying dual RNA specificity in the *Drosophila* SNF protein. Nat. Commun. *9*, 2220.

Wickstead, B., Gull, K., and Richards, T.A. (2010). Patterns of kinesin evolution reveal a complex ancestral eukaryote with a multifunctional cytoskeleton. BMC Evol. Biol. *10*, 110.

Wiegand, S., Jogler, M., and Jogler, C. (2018). On the maverick Planctomycetes. FEMS Microbiol. Rev. *42*, 739–760.

van Wijk, L.M., and Snel, B. (2020). The first eukaryotic kinome tree illuminates the dynamic history of present-day kinases. Preprint at bioRxiv, 10.1101/2020.01.27.920793.

Wilkinson, M.E., Charenton, C., and Nagai, K. (2020). RNA Splicing by the Spliceosome. Annu. Rev. Biochem. *89*, 359–388.

Will, C.L., Schneider, C., Hossbach, M., Urlaub, H., Rauhut, R., Elbashir, S., Tuschl, T., and 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, S.G., and Hall, K.B. (2010). Coevolution of *Drosophila snf* Protein and Its snRNA Targets. Biochemistry *49*, 4571–4582.

Williams, S.G., Harms, M.J., and Hall, K.B. (2013). Resurrection of an Urbilaterian U1A/U2B"/SNF Protein. J. Mol. Biol. *425*, 3846–3862.

Williams, T.A., Foster, P.G., Nye, T.M.W., Cox, C.J., and Embley, T.M. (2012). A congruent phylogenomic signal places eukaryotes within the Archaea. Proc. R. Soc. B *279*, 4870–4879.

Williams, T.A., Cox, C.J., Foster, P.G., Szöllősi, G.J., and Embley, T.M. (2020). Phylogenomics provides robust support for a two-domains tree of life. Nat. Ecol. Evol. *4*, 138–147.

Woese, C.R. (1987). Bacterial evolution. Microbiol. Rev. *51*, 221–271.

Woese, C.R., Kandler, O., and Wheelis, M.L. (1990). Towards a natural system of organisms: proposal for the domains Archaea, Bacteria, and Eucarya. Proc. Natl. Acad. Sci. U. S. A. *87*, 4576–4579.

Wolf, Y.I., and Koonin, E.V. (2013). Genome reduction as the dominant mode of evolution. BioEssays *35*, 829–837.

de Wolf, B., Oghabian, A., Akinyi, M.V., Hanks, S., Tromer, E.C., van Hooff, J.J.E., van Voorthuijsen, L., van Rooijen, L.E., Verbeeren, J., Uijttewaal, E.C.H., et al. (2021). Chromosomal instability by mutations in the novel minor spliceosome component CENATAC. EMBO J. *40*, e106536.

Wong, D.K., Grisdale, C.J., Slat, V.A., Rader, S.D., and Fast, N.M. (2022). The evolution of pre-mRNA splicing and its machinery revealed by reduced extremophilic red algae. J.

Eukaryot. Microbiol., e12927.
Wu, F., Speth, D.R., Philosof, A., Crémière, A., Narayanan, A., Barco, R.A., Connon, S.A., Amend, J.P., Antoshechkin, I.A., and Orphan, V.J. (2022). Unique mobile elements and scalable gene flow at the prokaryote–eukaryote boundary revealed by circularized Asgard archaea genomes. Nat. Microbiol. *7*, 200–212.
Yang, D., Oyaizu, Y., Oyaizu, H., Olsen, G.J., and Woese, C.R. (1985). Mitochondrial origins. Proc. Natl. Acad. Sci. U. S. A. *82*, 4443–4447.
Yenerall, P., and Zhou, L. (2012). Identifying the mechanisms of intron gain: progress and trends. Biol. Direct *7*, 1–10.
Yoshihama, M., Nakao, A., Nguyen, H.D., and Kenmochi, N. (2006). Analysis of Ribosomal Protein Gene Structures: Implications for Intron Evolution. PLOS Genet. *2*, e25.
Zachar, I., Szilágyi, A., Számadó, S., and Szathmáry, E. (2018). Farming the mitochondrial ancestor as a model of endosymbiotic establishment by natural selection. Proc. Natl. Acad. Sci. U. S. A. *115*, E1504–E1510.
Zaremba-Niedzwiedzka, K., Caceres, E.F., Saw, J.H., Bäckström, D., Juzokaite, L., Vancaester, E., Seitz, K.W., Anantharaman, K., Starnawski, P., Kjeldsen, K.U., et al. (2017). Asgard archaea illuminate the origin of eukaryotic cellular complexity. Nature *541*, 353–358.
Zhang, L.-Y., Yang, Y.-F., and Niu, D.-K. (2010). Evaluation of Models of the Mechanisms Underlying Intron Loss and Gain in *Aspergillus* Fungi. J. Mol. Evol. *71*, 364–373.
Zhao, C., and Pyle, A.M. (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., and Biddle, J.F. (2021). Helarchaeota and co-occurring sulfate-reducing bacteria in subseafloor sediments from the Costa Rica Margin. ISME COMMUN. *1*, 25.
Zhou, X., Lin, Z., and Ma, H. (2010). Phylogenetic detection of numerous gene duplications shared by animals, fungi and plants. Genome Biol. *11*, R38.
Zimmerly, S., and Semper, C. (2015). Evolution of group II introns. Mob. DNA *6*, 7.
Zimmerly, S., Hausner, G., and Wu, X. (2001). Phylogenetic relationships among group II intron ORFs. Nucleic Acids Res. *29*, 1238–1250.

Appendix

Abbreviations

Numbers in parentheses refer to the page where the abbreviation is introduced

BBH bidirectional best hit (36)
BLAST basic local alignment tool (19)
Tool to quickly find similar sequences
COG cluster of orthologous groups (35)
Originally based on one eukaryotic, one archaeal and five bacterial genomes; updated since and included in the eggNOG database
DNA deoxyribonucleic acid (10)
ENOG eggNOG non-supervised orthologous group (48)
euNOG eukaryotic eggNOG non-supervised orthologous group (76)
FECA first eukaryotic common ancestor (17)
First proto-eukaryotic descendant of the last common ancestor of eukaryotes and their closest prokaryotic relatives
HGT horizontal gene transfer (35)
Gene exchange between two organisms that is not due to vertical inheritance
HMM hidden Markov model (20)
Profile based on a multiple sequence alignment; used for sensitive homology searches
hnRNP heterogeneous nuclear ribonucleoprotein (100)
IEP intron-encoded protein (101)
KOG eukaryotic orthologous group (35)
Originally based on seven eukaryotic genomes; included in the eggNOG database
LECA last eukaryotic common ancestor (14)
Most recent common ancestor of all extant eukaryotes
mRNA messenger RNA (10)
OG orthogroup (18)
The set of genes descended from a single gene in the last common ancestor of a specified set of species
PPIase peptidylprolyl isomerase (120)
RNA ribonucleic acid (10)
rRNA ribosomal RNA (123)
RNP ribonucleoprotein (98)
RRM RNA recognition motif (120)
RT reverse transcriptase (104)
SAR Stramenopiles, Alveolata and Rhizaria (15)
SmAP Sm-like archaeal protein (123)
snoRNP small nucleolar RNP (123)
The nucleolus is a large structure in the nucleus where ribosome biogenesis takes place
snRNA small nuclear RNA (16)
snRNP small nuclear RNP (16)
SR serine/arginine-rich (protein) (100)
sRNP small RNP (123)
tRNA transfer RNA (123)

Samenvatting

Een van de belangrijkste gebeurtenissen in de evolutie is het ontstaan van de eerste complexe cellen uit simpele cellen. Ongeveer de eerste helft van de geschiedenis van het leven op aarde waren er alleen de relatief simpele cellen van bacteriën en **archaea**. Archaea lijken op het eerste gezicht erg op bacteriën, maar zijn op moleculair niveau fundamenteel anders. Uit een specifieke groep archaea zijn zo'n twee miljard jaar geleden de eerste **eukaryoten** ontstaan. Vrijwel alle levende wezens die we zonder microscoop kunnen zien, zoals algen, planten, dieren en schimmels, zijn eukaryoten, die zijn opgebouwd uit complexe cellen. Eukaryote cellen zijn groter en hebben veel meer DNA dan bacteriën en archaea. Bovendien bevatten ze compartimenten die elk hun eigen taak hebben. Qua grootte en organisatie kun je een bacteriële of archaeale cel vergelijken met een tent, terwijl een eukaryote cel meer lijkt op een huis met verschillende kamers.

Over hoe sommige organismen de tent voor een huis hebben ingeruild, is veel onduidelijk. Er is dan ook een levendig en soms verhit debat gaande tussen onderzoekers over het ontstaan van complex leven. Het ontbreken van levende tussenvormen en het beperkte aantal duidelijke fossielen van deze periode maken het ontstaan van eukaryoten (**eukaryogenese**) tot een intrigerend vraagstuk. Voor het reconstrueren van evolutionaire gebeurtenissen kunnen we echter ook het DNA van hedendaagse soorten gebruiken. Onze genen zijn gevormd door miljoenen jaren van evolutie. Hoewel ze behoorlijk zijn veranderd, zijn de echo's van een ver verleden er nog in terug te horen. Er is momenteel een enorme hoeveelheid genetische materiaal beschikbaar van diverse organismen. Met deze data en programma's om genetische echo's te detecteren en modelleren hebben we de evolutie van duizenden genen kunnen reconstrueren.

Een drijvende kracht achter de toename in de complexiteit van de cel waren de talloze **genduplicaties** die hebben plaatsgevonden tijdens eukaryogenese. Zo'n duplicatie van een gen kan ertoe leiden dat de twee kopieën, **paralogen** genaamd, elk een deel van de taken van het vooroudergen overnemen of dat een van beide een volledig nieuwe rol gaat vervullen. In het eerste deel van het proefschrift hebben we de genduplicaties tijdens eukaryogenesis gereconstrueerd en gekarakteriseerd. Met name genen die betrokken zijn bij het bouwen en reguleren van de complexe cel blijken het resultaat van duplicaties. Relatief weinig duplicaties hebben plaatsgevonden in metabole genen.

Een belangrijk moment is het ontstaan van **mitochondriën** geweest. Mitochondriën zijn een van de onderdelen van eukaryote cellen en functioneren als energiecentrales. Ooit waren dit vrijlevende bacteriën, die ergens tijdens de evolutie zijn opgenomen door toekomstige eukaryote cellen. Er is veel discussie over of het verkrijgen van het mitochondrium de eerste, cruciale stap of de laatste stap was. Met de evolutionaire reconstructies hebben we ook relatieve tijdschattingen verkregen voor de genduplicaties en voor de opname van de mitochondriale voorouder in de toekomstige eukaryote cel, waardoor we vroege en late genduplicaties kunnen onderscheiden.

Naast de duplicatie van genen is de structuur van de genen ook veranderd tijdens eukaryogenese. Binnen de genen zijn stukken DNA gekomen, **introns** genaamd, die als advertenties de codering in het DNA onderbreken. Als een gen actief is, wordt de nucleotidevolgorde overgeschreven in RNA, dat erg lijkt op DNA. Voordat dit RNA vertaald kan worden naar een eiwit, moeten de introns uit het RNA verwijderd worden. We hebben

de positie van introns in genen van de eukaryote voorouder gereconstrueerd en deze informatie gecombineerd met de eerdergenoemde duplicatie-informatie. Als paralogen een intron delen op "dezelfde" (homologe) plek in het gen, betekent dit waarschijnlijk dat het intron al aanwezig was in het vooroudergen voor de duplicatie. We hebben veel intronposities gedetecteerd die gedeeld werden tussen paralogen van genduplicaties tijdens eukaryogenese. Dit impliceert dat introns zich al vroeg tijdens eukaryogenese verspreid hadden.

Introns worden verwijderd door het **spliceosoom**, een van de meest complexe moleculaire machines in eukaryote cellen. Zowel de introns als de kern van de spliceosomale machinerie ontstonden uit zelfsplicende introns die voorkomen in bacteriën en archaea, maar daar de genen niet onderbreken. Om de oorsprong van het zeer complexe spliceosoom tijdens eukaryogenese op te helderen hebben we de evolutionaire geschiedenissen van de verschillende spliceosomale eiwitten gereconstrueerd. Hieruit bleek dat aan de spliceosomale kern voornamelijk eiwitten zijn toegevoegd die voorheen een functie hadden die gerelateerd is aan het ribosoom. Dit grote complex vertaalt RNA naar eiwit. Talloze genduplicaties vormden het spliceosoom tot een zeer complexe machinerie in de eukaryote voorouder. De vele gedeelde introns tussen spliceosomale paralogen die we hebben gevonden, wijzen erop dat introns wijdverspreid waren voordat de meeste spliceosomale complexiteit was ontstaan.

Met behulp van de analyses uit de in het proefschrift beschreven studies hebben we een tijdlijn kunnen schetsen van de evolutionaire gebeurtenissen in de opkomende eukaryoten. We leidden uit de vele gedeelde intronposities tussen paralogen bijvoorbeeld af dat een fysieke scheiding tussen het verwijderen van de introns en het vertalen van RNA naar eiwit al in een vroeg stadium nodig was. Dit maakt het zeer aannemelijk dat de celkern, het compartiment dat het DNA opslaat en voor deze scheiding zorgt, vroeg is ontstaan. Eiwitten die betrokken zijn bij de bouw van de complexe eukaryote cel met de verschillende compartimenten en bij het actief transport hiertussen, zijn volgens onze tijdsschattingen vroeg ontstaan. De opname van de vrijlevende bacterie die later een mitochondrium werd, was waarschijnlijk een tussentijdse gebeurtenis. Daarna is de cel nog complexer geworden, met name in het reguleren van processen.

De oorsprong van eukaryoten is het onderwerp van een levendig debat tussen biologen. Wanneer de energiecentrale is verkregen, staan hierin meestal centraal. Met het werk in dit proefschrift kunnen meer tussentijdse cellulaire en genetische veranderingen worden onderscheiden om op te helderen hoe complex leven is opgekomen. Een beter begrip van het ontstaan van complex leven kan ons meer leren over de fundamenten van evolutie en celbiologie. Misschien kan het ons zelfs meer vertellen over de kans dat iets vergelijkbaars elders in het universum gebeurt.

Acknowledgements

Eleven years ago, I started my journey at Utrecht University when I enrolled in the bachelor's programme Biology. I was the first in my family to attend university and I did not have a clue what being a scientist really meant. During my studies I explored multiple paths, from cardiac development to metagenomics. I had the opportunity to go abroad for an internship, which had a large impact on my personal development. The research projects that I carried out during my masters enthused me about becoming a scientist. The colleagues in both the Bakkers and Ettema groups played a major role in this. I would like to thank especially **Jeroen**, **Sonja**, **Thijs** and **Joran** for putting me on the trajectory towards a PhD.

Berend, first of all I want to thank you for offering me a PhD position when you still barely knew me. Our first meetings felt a bit awkward and there was some miscommunication in the beginning. It took some time to get used to your supervision style but quite soon I embraced it. I will remember the long silences (for those unfamiliar, it's Berend's thinking time, do not disturb), sighs (these are not directed to anyone) and random walks into my office (there is nothing wrong if Berend leaves again without saying anything). You gave me the freedom and support to explore what I like, for example in the projects that I carried out and in teaching. Your knowledge is impressive. I greatly valued that you appreciated my opinion and advice and explicitly asked for it. Thank you for your supervision.

Next, I would like to thank my independent advisors, **Geert** and **Robin**, for their advice. You provided Berend and me with a good external perspective and our progress meetings were of added value. I am also grateful to the **assessment committee** for taking the time to read and assess the thesis.

Peter and **Sam**, you were my targets when asking for assistants for my course and now again as my paranymphs. Peter, I cannot sufficiently stress the importance of your contribution to the development and further improvement of the data science course. Thanks for the long chats that we could have, which were slightly distracting from work sometimes, but your down-to-earth views, to-the-point questions and helpful advice resonated with me. Sam, my laid-back office mate, who is always in for a social activity. Thanks for the valuable suggestions for the general introduction and discussion. It happened multiple times over the years that after we had had a nice discussion about our projects, you came up with good ideas after some additional thinking. It was much easier letting go of the data science course last year knowing that you were still involved.

I have changed offices twice and shared the office with multiple people. In strict lockdown times I missed you! Of course, social chatter with office mates is essential for a joyful PhD but they can also provide valuable feedback on your work, especially on figures. **Dajo**, it was fun sharing an office with you during my first months. Our casual chats made it easier for me to settle into the group. **Eva**, we were not always synchronised and I remember your wondering face when I was still talking to you after you had already put your noise-cancelling headset back on. Your creativity and love for corals, hedgehogs, sea sparkle and many other eukaryotes is inspiring. **Juliane**, my fellow cool office mate! Thanks for listening to my frustrations and excitements. I cannot stop smiling when thinking back to our discussions about your greyscale figures and my lollipop plots. I

enjoyed our talks about private life, also outside office hours during dinners, mud runs and in-line skate tours.

Not only my office mates, but all colleagues of **Theoretical Biology and Bioinformatics** have contributed to my PhD. **Jan Kees**, I want to thank you for making sure the computational infrastructure worked perfectly and helping me with silly questions or requests. **Jolien**, your sharp, scrutinising mind has laid the foundation for the work in this thesis. I am looking forward to being your colleague again in Wageningen, where we can continue our political discussions. Moving on to the other members of the Snel group, I start with thanking the former group members **Leny**, **John** and **Eelco**. I learnt a lot from you. **Laura**, thanks for the nice conversations at the coffee machine and good luck with wrapping up your thesis! With your perseverance I have no doubt that you will succeed. **Max** and **Bastiaan**, you are the newbies with contagious enthusiasm. Max, I am grateful that you shared your impressive cover art which helped me with creating the figure in the discussion. I hope you will continue drawing these cartoons. Bastiaan, I wish you all the best with your PhD! A large part of my PhD years consisted of teaching. **Michael**, **Joséphine** and **Julia**, thank you for everything you have done for the data science course. **Can**, thanks for supervising my teaching qualification trajectory and for your valuable feedback on my portfolio.

Besides being a supervisee, I acted as the daily supervisor of a few bachelor and master students for their research projects and writing assignment. **Sjoerd**, **Michelle**, **Daan**, **Tony** and **Martijn**, I enjoyed supervising you. It was nice to see the diversity in working styles that you had and maybe unwittingly you taught me a lot too. You have all contributed to the content of this thesis and to future project ideas.

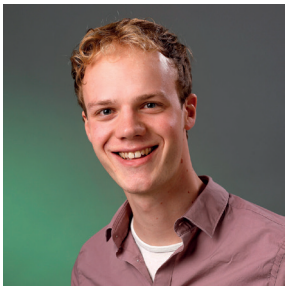
Voor het laatste deel van het dankwoord schakel ik over naar het Nederlands. **Davi** en **Marloes**, jullie speelden een belangrijke rol in de tijd die ik buiten het Kruyt heb doorgebracht de afgelopen jaren. Davi, zonder jouw creativiteit was de cover maar saai geweest. Ondanks mijn vage ideeën heb je een prachtig ontwerp gemaakt. Ik bewonder je geduld en pogingen tot doorvragen als je mij weer eens vroeg om in simpele taal uit te leggen wat ik precies deed. Ik heb meerdere metaforen op jou uitgeprobeerd, maar met name de "oerblender" is blijven hangen. Marloes, bij jou vind ik altijd een luisterend oor. Geinteresseerd vroeg je mij vaak hoe mijn werkdag eruit had gezien. Het was fijn om een drukke dag op het werk af te sluiten met jou op de tennisbaan of door lekker samen te eten, hoewel ik je geen bobotie meer ga voorschotelen…

Mam, **pap** en **Michiel**, jullie hebben me altijd gesteund en gestimuleerd om te doen wat ik leuk vind en om door te leren. Daarmee hebben jullie het mogelijk gemaakt dat dit boekje er is.

Michael, je bent mijn belangrijkste steun, ook in het laatste jaar dat voor jou behoorlijk stressvol is geweest. Bedankt dat je in me gelooft en me weet af te remmen als dat nodig is. Ik wil je ook graag bedanken voor het lezen van een eerste versie van de inleiding en het meedenken over de vormgeving van dit boek. Ik ben trots op jou en op wat we samen al hebben bereikt. Ik heb zin in alle avonturen die we samen nog gaan beleven!

Curriculum vitae

Julian Vosseberg was born in Harderwijk, the Netherlands, on 9 August 1993. After finishing his pre-university education (vwo) (*cum laude*) at the Christelijk College Nassau-Veluwe in Harderwijk, he registered for the bachelor's programme Biology at Utrecht University. He graduated in 2014 (*cum laude*) and continued his studies at Utrecht University as master's student of the Cancer, Stem Cells and Developmental Biology programme. He did his major research project in the lab of Jeroen Bakkers at the Hubrecht Institute to study cardiac development in zebrafish. For a second internship he went to the lab of Thijs Ettema at Uppsala University, Sweden, to find novel alphaproteobacteria in metagenomes. In 2017 he started as a PhD candidate and teacher in the Theoretical Biology and Bioinformatics group at Utrecht University under the supervision of Berend Snel. In March 2022 he was elected as member of the municipal council in Nieuwegein. He currently works as a postdoctoral researcher with Thijs Ettema at Wageningen University & Research.



List of publications

Vosseberg, J., Stolker, D., von der Dunk, S.H.A., and Snel, B. Integrating phylogenetics with intron positions illuminates the origin of the complex spliceosome. *Preprint at bioRxiv*, 10.1101/2022.08.31.505394.

Martijn, J., **Vosseberg, J.**, Guy, L., Offre, P., and Ettema, T.J.G. (2022). Phylogenetic affiliation of mitochondria with Alpha-II and Rickettsiales is an artefact. *Nature Ecology and Evolution*, 10.1038/s41559-022-01871-3.

Schön, M.E.*, Martijn, J.*, **Vosseberg, J.**, Köstlbacher, S., and Ettema, T.J.G. (2022). The evolutionary origin of host association in the Rickettsiales. *Nature Microbiology*, 7:1189-1199.

Vosseberg, J., Schinkel, M., Gremmen, S., and Snel, B. (2022). The spread of the first introns in proto-eukaryotic paralogs. *Communications Biology*, 5:476.

van Gerven, M.R., Taschner-Mandl, S., Matser, Y.A.H., **Vosseberg, J.**, Bozsaky, E., Koster, J., Tytgat, G.A.M., Molenaar, J.J., and van den Boogaard, M. (2022). Mutational spectrum of ATRX aberrations in neuroblastoma and associated patient and tumor characteristics. *Cancer Science*, 113:2167-2178.

Vosseberg, J.*, van Hooff, J.J.E.*, Marcet-Houben, M., van Vlimmeren, A., van Wijk, L.M., Gabaldón, T., and Snel, B. (2021). Timing the origin of eukaryotic cellular complexity with ancient duplications. *Nature Ecology and Evolution*, 5:92-100.

Martijn, J.*, Schön, M.E.*, Lind, A.E., **Vosseberg, J.**, Williams, T.A., Spang, A., and Ettema, T.J.G. (2020). Hikarchaeia demonstrate an intermediate stage in the methanogen-to-halophile transition. *Nature Communications*, 11:5490.

Deutekom, E.S., **Vosseberg, J.**, van Dam, T.J.P., and Snel, B. (2019). Measuring the impact of gene prediction on gene loss estimates in Eukaryotes by quantifying falsely inferred absences. *PLOS Computational Biology*, 15:e1007301.

Vosseberg, J., Martijn, J., and Ettema, T.J.G. (2018). Draft genome sequence of "*Candidatus* Moanabacter tarae," representing a novel marine verrucomicrobial lineage. *Microbiology Resource Announcements*, 7:e00951-18.

Martijn, J., **Vosseberg, J.**, Guy, L., Offre, P., and Ettema, T.J.G. (2018). Deep mitochondrial origin outside the sampled alphaproteobacteria. *Nature*, 557:101-105.

Vosseberg, J., and Snel, B. (2017). Domestication of self-splicing introns during eukaryogenesis: The rise of the complex spliceosomal machinery. *Biology Direct*, 12:30.

*These authors contributed equally

PhD portfolio

The result of a PhD trajectory is much more than the content of the thesis. How I developed as a researcher can only partly be inferred from the chapters. All setbacks and unsuccessful attempts have been smoothened out in the stories that we present. Yet this search and re-search is what science is at its core. By teaching classes and representing my PhD programme in the PhD council I developed in more areas than just research. By listing the activities that I did during my PhD trajectory, I aim to represent a complementary picture on my personal and scientific development.

Courses and events

Competence area	Name	ECTS*
Research skills and knowledge	Analytics and algorithms for omics data	3.0
	Advanced omics for life sciences	1.5
	Bioinformatics and evolutionary genomics	1.5
Responsible conduct of science	This thing called science	2.0
	PhD day 2019: Transparent science	0.3
Communication	Breaking science	2.0
	Analytic storytelling	0.7
	PhD day 2021: When creativity meets science	0.2
	PhD day 2018: Talkin' science	0.3
Teaching	Supervising research of MSc students	1.2
	Supervision of high school students	2.0
	Start to teach	0.6
Professional development	Honest networking	0.1
	PhD activating career event (PhACE)	NA
	Career services workshop 'What is your team role?'	NA

*ECTS: European credit transfer and accumulation system

Presentations at conferences

- Oral presentation at the EMBO workshop "Comparative genomics of unicellular eukaryotes: Interactions and symbioses" in Sant Feliu de Guíxols, Spain (September 2022)
- Oral presentation at the online SMBE meeting (July 2021)
- Oral presentation at the online BioSB conference (June 2021)
- Oral presentation at the online NLSEB meeting (April 2021)
- Poster presentation at the online Origins conference (January 2021)
- Research pitch at the online Science4Life Symposium (November 2020)
- Oral presentation at the EMBO workshop "Comparative genomics of eukaryotic microbes: Genomes in flux, and flux between genomes" in Sant Feliu de Guíxols, Spain (October 2019)

- Oral presentation at the SMBE meeting in Manchester, UK (July 2019)
- Poster presentation at NWO Life2019 in Bunnik, NL (May 2019)
- Poster presentation at the Science4Life Symposium in Utrecht (November 2018)
- Poster presentation at the Evolution conference in Montpellier, France (August 2018)

Peer-reviewing

I reviewed three manuscripts for *Genome Biology and Evolution* (with Berend Snel), *Frontiers in Bioinformatics* and *Database: The Journal of Biological Databases and Curation*.

Outreach

- "Understanding the origin of the eukaryotic cell: gene duplications to the rescue" (behind the paper blog post written with Jolien van Hooff; https://natureecoevocommunity.nature.com/posts/understanding-the-origin-of-the-eukaryotic-cell-gene-duplications-to-the-rescue)
- "Timeline of early eukaryotic evolution" (Utrecht University news article; https://www.uu.nl/en/news/timeline-of-early-eukaryotic-evolution)
- 3-minute pitch for a lay audience as part of the Breaking Science competition: https://youtu.be/V5APTzA89Sw

Teaching

I created, coordinated and taught a new bachelor course, called *Data science en biologie*. In 2022 I obtained my teaching qualification (BKO).

- Course development and coordination
  - *Data science en biologie* (novel bachelor biology course, level 2; 2019, 2020 and 2021)
- Teaching
  - Giving lectures, tutorials and practicals and supervising project groups for *Data science en biologie* (level 2 bachelor biology, 2019-2021)
  - Giving a lecture for *Evolutie en biodiversiteit* (level 1 bachelor Biology, 2020)
  - Giving a guest lecture to high school students (U-talent, 2020)
  - Supervising poster assignment groups for *Bioinformatica* (level 1 bachelor Biomedical Sciences, 2020)
  - Giving a guest lecture and guiding a paper discussion for *Systems biology* (master Molecular and Cellular Life Sciences, 2019-2020)
  - Assisting tutorials for *Systeembiologie* (level 1 bachelor Biology, 2018)
  - Assisting a practical for *Genoombiologie* (level 3 bachelor Biology, 2017)
- Supervising students
  - Two master students for their major research project (9 months)
  - One master student for her minor research project (6 months)
  - One bachelor student for his research project and thesis (10 weeks)
  - One master student for his writing assignment (5 weeks)
  - Three high school students for their thesis project (3 days)

Other activities

- Member of the PhD council of the Graduate School of Life Sciences (2018-2021): representing the Computational Life Sciences programme, member of the PhD survey 2020 committee, member of the Supervisor of the Year 2019 committee
- Member of the hiring committee of a PhD confidential advisor at the Science faculty
- Member of the hiring committee of a tenure track researcher in the Theoretical Biology and Bioinformatics group

Did you take a walk through a park lately? Summer is coming; flowers have come up and bees are zooming around. Endless forms most beautiful that are all visible by the naked eye, live on this Earth. But this has not always been the case.

It all started around four billion years ago; there it was: life. For roughly the first half of life's history, there were only small and compact bacterial cells. And then something astonishing happened. Complex cells emerged: larger cells, with much more DNA. And when you zoom in, you see compartments dedicated to specific tasks. While bacterial cells are in that sense like a tent, the complex cells are like a house with different rooms.

How did the tent become a house two billion years ago? We weren't there and, unfortunately, I don't have a time machine. There are also no living intermediates and no clear fossils either. How could we possibly figure out? Well, we are living evidence; we are descendants of these first complex cells. Inside us we have DNA that has been shaped by years and years of evolution. It has changed a lot along the way, but it still carries echoes of aeons ago.

At the moment, there is an enormous amount of DNA available from diverse organisms. With computer programs I used these data to reconstruct the evolution of thousands of genes and obtain relative time estimates. Based on that I was able to draw a timeline of the evolution of the first complex cells. I found that, for example, the compartment that contains the DNA was very likely established early and that the uptake of the powerhouse of the cell, which used to be a free-living bacterium itself, was likely an intermediate event.

There is a lively debate on the origin of complex life and much remains to be discovered. Understanding how complex cells originated can teach us more about the fundaments of life. And maybe it can tell us how likely it is that something similar would happen to life somewhere else in the Universe.

The next time you go to the park, enjoy the plants, animals and fungi. The seed for those beautiful living creatures was planted roughly 2 billion years ago in the enormous transition from simple, bacteria-like cells to large, complex cells.