Changing ideas about eukaryotic origins

Tom A. Williams and T. Martin Embley

Institute for Cell and Molecular Biosciences, Newcastle University, Newcastle upon Tyne, UK

The origin of eukaryotic cells is one of the most fascinating challenges in biology, and has inspired decades of controversy and debate. Recent work has led to major upheavals in our understanding of eukaryotic origins and has catalysed new debates about the roles of endosymbiosis and gene flow across the tree of life. Improved methods of phylogenetic analysis support scenarios in which the host cell for the mitochondrial endosymbiont was a member of the Archaea, and new technologies for sampling the genomes of environmental prokaryotes have allowed investigators to home in on closer relatives of founding symbiotic partners. The inference and interpretation of phylogenetic trees from genomic data remains at the centre of many of these debates, and there is increasing recognition that trees built using inadequate methods can prove misleading, whether describing the relationship of eukaryotes to other cells or the root of the universal tree. New statistical approaches show promise for addressing these questions but they come with their own computational challenges. The papers in this theme issue discuss recent progress on the origin of eukaryotic cells and genomes, highlight some of the ongoing debates, and suggest possible routes to future progress.

1. What did we think before?

In the rooted ‘three domains’ tree [1], the eukaryotic nuclear lineage is a deep branching sister group to the Archaea, implying that eukaryotes are as old as that group of prokaryotes (figure 1). The species at the base of eukaryotes in the three domains tree are parasites like *Giardia* and *Microsporidia* which lack classical mitochondria, in agreement with the hypothesis that they were descended from lineages (often called Archezoans—[2]) that diverged from other eukaryotes before the mitochondrial endosymbiosis. In the three domains tree, the eukaryotes—cells with a nucleus—existed before the mitochondrial endosymbiosis. The apparent agreement between phylogeny and cell biology made this version of early evolution compelling. Thus, although competing hypotheses were in circulation at the time [3–7], and many genes on eukaryotic genomes were already known to conflict with the three domains tree [8,9], it is the one that appeared in standard textbooks and works of popular science. A tree diagram is the single figure in the ‘Origin of Species’ [10, pp. 160–161] and so it was natural that there is only a single figure in an updated popular science version [11] of Darwin’s classic. The tree chosen was an unrooted version of the three domains tree, depicting Archaea and eukaryotes as separate groups and with the Archezoans clearly labelled at the base of eukaryotes.

The papers in this theme issue describe and discuss how this view of eukaryotic evolution has radically changed over the past few years, and identify major ongoing controversies and challenges. The contributors sometimes offer very different perspectives on these issues, so there is principled disagreement as well as consensus. In part, this reflects not only the rapid and exciting progress being made but also the inherent difficulty of inferring ancient events from small amounts of incomplete data using imperfect methods, and the ambiguity and scale of the scientific questions that are being asked. Some of the most marked changes in thinking are about the nature of the host for the mitochondrial endosymbiont and the recognition that organelles related to mitochondria are ubiquitous among eukaryotes, including former Archezoans. These changes have removed a major line of evidence for the view that the mitochondrial
host was already a eukaryote and, in turn, have led to more serious consideration of hypotheses in which an Archaeon was the host for the mitochondrial endosymbiosis in founding the eukaryotic lineage. Debates about the role of the mitochondrial endosymbiont in eukaryotic genome evolution, and the evolution and diversity of contemporary mitochondrial homologues, including hydrogenosomes and mitosomes, are now major topics of investigation. The origins of genes and the extent of non-endosymbiotic lateral gene transfers in eukaryotic evolution are still controversial, but it is now clear that eukaryotes owe a major genomic debt to Archaea and Bacteria as well as possessing a previously under-appreciated talent for gene invention and innovation. Whether viruses have also played a role in eukaryotic origins and evolution is hotly debated, fuelled to a degree by recent discoveries of unexpectedly large and gene-rich DNA viruses.

Microbial ecologists have long known that cultured and studied microbes comprise only a small fraction of extant unicellular life, so it is to be expected that our understanding of cellular evolution has been limited by incomplete and biased sampling of natural microbial diversity. New metagenomic and single cell genome sequencing methods hold enormous promise to sample the hitherto unstudied majority of microbial life. As discussed in this issue, these methods have already identified new archaeal lineages that are more closely related to eukaryotes than any yet sampled, and that share genes previously thought to define important aspects of the biology of eukaryotic cells. Concerns about the accuracy of trees for inferring deep eukaryotic relationships or gene origins, which are often made using overly simple statistical models and short sequences, occupy a number of our contributors. The need to consider the fit between model and data, and to recognize that poor models will generally make poor trees, is an oft repeated and important cautionary message. Trees and networks of various sorts will continue to play a major role in studies aiming to investigate eukaryotic evolution and to disentangle vertical and horizontal descent, but existing methods are fraught with problems and the search for congruence between independent lines of evidence will always be important.

2. A new host for the mitochondrial endosymbiont

The three domains tree describes eukaryotes and Archaea as separate groups and has a fully formed eukaryotic cell as the host for the mitochondrial endosymbiont [1]. However, at the same time as some analyses were recovering the three domains tree, other analyses (reviewed in [12]) of the same data but often using better methods were supporting another hypothesis called the ‘eocyte tree’ [3]. In the ‘eocyte tree’, eukaryotes originate from within the Archaea as the sister group of species like *Sulfolobus*—which James Lake [13] classified within a separate kingdom called Eocyta or ‘dawn cells’ [3], and which Woese et al. later named the Crenarchaeota [1]. Support for the eocyte tree has continued to accumulate in recent years with improved evolutionary models and wider sampling of environmental Archaea [12,14,15]. Thus, analyses of universal core genes using better-fitting models place eukaryotes within the diversity of Archaea, branching with a group called the ‘TACK’ superphylum which contains the lineages Thaumarchaeota, Aigarchaeota, Crenarchaeota and Korarchaeota [16–19]. As eocytes were originally defined phylogenetically as the sister group of eukaryotes [3], these new trees are consistent with the eocyte hypothesis. Our special issue opens with a personal perspective by James Lake [13] describing the genesis and development of the eocyte hypothesis and other seminal contributions, including his highly original ‘ring of life’ hypothesis that invokes large gene flows as major drivers in eukaryotic evolution. This ‘ring of life’ is also the focus of the paper by Mcinerney et al. [20], who argue that it is the best-supported and most general hypothesis to explain the different types of data that speak to eukaryotic origins.

If the trees that place the origin of the eukaryotic nuclear lineage within the Archaea are correct, then we should expect to find new species that are more similar to eukaryotes at the level of genes and proteins. Eugene Koonin [21] discusses recent data that are consistent with this hypothesis and demonstrates how understanding archaeal genome evolution is important for understanding early eukaryotic evolution. Consistent with the predictions of recent phylogenomic analyses, prokaryotic homologues of key eukaryotic components, including genes involved in the cytoskeleton and ubiquitin-mediated protein degradation, are found only among the TACK Archaea. But Koonin [21] also shows that homologues of other signature eukaryotic genes, including components of the cell division, membrane remodelling, and RNA interference machineries, have a patchy distribution across the sequenced diversity of Archaea, suggesting a complex history of gene loss and potentially horizontal transfer throughout archaeal evolution. As mentioned in §1, limited and potentially biased sampling of...
natural microbial diversity may limit our inferences of early evolution. The paucity of genomes is particularly acute for Archaea because the exploration of this domain has traditionally lagged behind Bacteria and eukaryotes. This situation is rapidly changing because of advances in single cell and metagenomic approaches that now enable the genomes of uncultured microbes to be sequenced directly from the environment [22]. The most spectacular finding to date has been that of the Lokiarchaeota, an archaeal lineage that appears to contain the closest relatives of eukaryotes discovered so far [23,24]. Consistent with its sister group relationship to eukaryotes, Lokiarchaeota have more eukaryotic signature genes than any other Archaea yet described [23]. Saw et al. [25] describe the methods they used to sequence and assemble the genome of *Lokiarchaeum* and other uncultured members of the TACK group, and the implications of the Lokiarchaeota gene repertoires for the origins of key eukaryotic features such as the cytoskeleton, membrane remodelling and phagocytosis. This final trait has often been argued [26] to be a key ability of the ancestral host cell that acquired the mitochondrial endosymbiont. Intriguingly for theories of eukaryogenesis, the ESCRT machinery—found in eukaryotes as well as in Lokiarchaeota and some other TACK Archaea—has recently been shown to regulate the reformation of the nuclear envelope after mitosis [27].

3. Endosymbiosis, mitochondrial homologues and the origins of bacterial genes on eukaryotic genomes

The rejection of the Archezoa hypothesis, and the discovery of mitochondrial homologues in parasites and anaerobes that were previously thought to primitively lack them [28], has stimulated interest in ideas that propose that the mitochondrial endosymbiont. Intriguingly for theories of eukaryogenesis, the ESCRT machinery—found in eukaryotes as well as in Lokiarchaeota and some other TACK Archaea—has recently been shown to regulate the reformation of the nuclear envelope after mitosis [27].

The rejection of the Archezoa hypothesis, and the discovery of mitochondrial homologues in parasites and anaerobes that were previously thought to primitively lack them [28], has stimulated interest in ideas that propose that the mitochondrial endosymbiont. Intriguingly for theories of eukaryogenesis, the ESCRT machinery—found in eukaryotes as well as in Lokiarchaeota and some other TACK Archaea—has recently been shown to regulate the reformation of the nuclear envelope after mitosis [27].

Eukaryotic genome evolution from within

Eukaryotic genomes encode a significant fraction—as much as 63% according to recent analyses of the yeast genome [44]—of eukaryote-specific genes that underpin key aspects of eukaryotic biology. Traditional models for eukaryotic gene origins emphasized the duplication and functional divergence of pre-existing genes [45], but there is increasing evidence that the de novo origin of new genes from noncoding sequence is also important. McIvyagh & Guerzoni [46] provide an overview of these data and provide interesting examples from across the eukaryotic tree, some of which are functionally important and subject to positive selection. Evidence for widespread de novo gene origination in modern eukaryotes provides a plausible mechanism by which eukaryote-specific genes could have evolved in the nascent eukaryotic stem lineage during the origin of eukaryotes.

The results suggest that rates of gene transfer in these groups are broadly similar, providing some support for the idea that the importance and dynamics of HGT may be qualitatively similar among prokaryotes and eukaryotes. This result, if held to hold more generally, would suggest an ongoing flux of bacterial genes into eukaryotic genomes from a variety of sources in addition to the large-scale gains associated with ancestral endosymbioses.
One of the most distinctive features of eukaryotic genomes in comparison to prokaryotes is the preponderance of noncoding sequence, which in many lineages outweighs or even dwarfs the quantity of coding DNA. While much of this excess material is probably selfish or non-functional [47], high-profile debate currently rages over the extent to which noncoding elements contribute to eukaryotic phenotypic complexity by regulating the expression of coding sequences [48–52]. Elliott & Gregory [53] contribute to this debate by providing new insights into the relationships between genome size, coding capacity, repetitive content and other genomic parameters from the largest survey of eukaryotic genome diversity to date. Their data underline striking differences between the streamlined, gene-rich genomes of prokaryotes and the large, highly repetitive genomes of many eukaryotes. These differences may arise from the fundamental changes in the population genetic environment that accompanied the origin of eukaryotes, ranging from increased cell size (and concomitant reduction in population densities) to the evolution of meiosis and sex. The relative contributions of genetic drift [54], mutation [55] and selection [56,57]—perhaps at multiple levels [58]—to the origin and evolution of eukaryotes and their genomes remains a fascinating area of debate, and broad comparative data of the type presented by Elliott & Gregory [53] will continue to play an important role in contrasting the predictions of the leading hypotheses.

5. How good are our methods for inferring the past?

Much of the progress discussed in this volume has been facilitated by the increasing ease with which whole genomes and transcriptomes can be sequenced, even for uncultured organisms. In principle, obtaining representative sampling is no longer a major hurdle, but the increasing rate of data generation has largely outstripped the computational power needed to analyse it. This has created a situation where undesirable trade-offs are made between dataset size and model adequacy, and this is hindering progress. Better phylogenetic models are already available that recognize that the evolutionary process is complex and may change over time and between species, but they come with a cost of increased analysis time and hence cannot be used for large numbers of species. As improved taxonomic sampling is already known to affect the accuracy of phylogenetic reconstructions [59], improving the scalability of complex methods to handle more data is highly desirable. Nicolas Lartillot [60] provides an overview of these issues and highlights potential solutions to some of the outstanding problems. Bayesian approaches provide a natural framework for fitting more complex and biologically motivated models to genome data, but Lartillot [60] argues that future progress may depend on the development of alternatives to standard Markov Chain Monte Carlo (MCMC) algorithms. MCMC has underpinned the successes of Bayesian phylogenetics to date, but the technique is now 50 years old and can struggle to achieve convergence on large-scale genomic datasets, even with continuing advances in computational power.

Probabilistic supertrees [61] synthesize information from a set of input gene trees to infer an overall species tree while allowing for some disagreement between the histories of the individual genes, whether due to horizontal transfer or more prosaic sources of phylogenetic error. They, therefore, represent an interesting and potentially very valuable ‘middle ground’ between the complex, hierarchical models of gene and genome evolution described by Szöllősi et al. [43] and Lartillot [60] and the simpler ‘supermatrix’ or concatenation approaches that have frequently been used to investigate the evolutionary history of genomes and species. Early supertree methods based on parsimony are known to have problems, so Akanni et al. [62] used a recently developed Bayesian probabilistic supertree method in their contribution. Their analysis evaluates the evidence for large-scale gene flows from Bacteria into archaeal genomes. Recently published work has argued that large gene flows have been an important factor in the evolution and ecology of Archaea [63,64]. While the supertree they recover for Archaea suggests a strong vertical signal, composite trees including Archaea and Bacteria were poorly resolved for deeper nodes, which Akanni et al. [62] suggest results from a mixture of vertical and horizontal signals, consistent with published work claiming episodic inter-domain transfer. These are intriguing results that raise interesting questions about the different effects of HGT on Bacteria and Archaea, and why these two prokaryotic groups should behave differently. It also suggests that the archaeal host lineage that merged with the mitochondrial endosymbiont might have been similarly chimaeric in terms of its genome content.

The limited reliability of single gene trees inferred using overly simple methods is at the core of a number of contributions to this issue. Moreira & López-García [65] discuss how better trees have been used to evaluate proposals that viruses have played a key role in eukaryotic origins. These ideas were originally prompted by the discovery of the Mega-viridae, giant amoeba-infecting viruses whose unexpectedly large genomes (1–2.5 Mbp, comparable in size with many cellular genomes) encode homologues of core components of the eukaryotic DNA replication and translation machineries [66–68]. Viral homologues branched outside the eukaryotic clade in early trees, suggesting that an ancient Megavirus, perhaps part of a ‘fourth domain’ of life, might have donated these genes to the ancestral eukaryote [67]. Moreira & López-García [65] note that placing viruses in phylogenetic trees is exceptionally challenging because of their high rates of sequence evolution, which—not unlike the deep divergences between the cellular domains—can induce artefacts such as long-branch attraction, the spurious grouping of fast-evolving sequences due to chance convergences in the substitution process. Their new analyses, in combination with a review of recent work, lead them to suggest that the presence of eukaryotic genes on viral genomes is best explained by horizontal acquisition from their eukaryotic hosts. They conclude that there is no compelling support for a viral contribution to the origin of eukaryotes or for the hypothesis that viruses represent a primaevæal fourth domain of life.

Many of the contributions in the volume favour hypotheses that have prokaryotes first and eukaryotes as a derived group formed through a merger involving Archaea and Bacteria. This prokaryote to eukaryote polarization of cellular evolution is consistent with published data using ancient paralogues and phylogenetic networks to root the universal tree on the bacterial stem [1,69–72]. It is also consistent with—albeit patchy and incomplete—fossil evidence.
for prokaryotes and prokaryotic metabolism more than a billion years before the earliest eukaryotic fossils [12,73,74], and with the observation that all known eukaryotes have a mitochondrial homologue—implying that the origin of alpha-proteobacteria occurred before the radiation of known eukaryotes [28]. Nevertheless, the trees used for paragene rooting were inferred using overly simple phylogenetic methods that are known to be unreliable for reconstructing ancient events [5,75], leaving room for criticism and debate. As a result, hypotheses that eukaryotes, or at least cells carrying much of the complexity that we associate with eukaryotes, might pre-date prokaryotes have persisted in the literature [6,7]. In these ‘eukaryotes first’ or ‘eukaryotes early’ scenarios, all three groups of cellular life are either held to have arisen contemporaneously, or prokaryotes are proposed to have originated through simplification of a complex ancestor that possessed many of the features that persist in modern eukaryotes [76–78]. Mariscal & Doolittle [79] provide a lucid historical overview of ‘eukaryotes first’ scenarios, examining their original motivations and discussing how they have fared as new data have accumulated. Their contribution brings clarity to a confusing and sometimes contradictory literature and, importantly, it attempts to clarify what is meant by ‘eukaryotes first’ and to identify how these ideas might be tested.

Gouy et al. [80] tackle the question of what came first from a methodological perspective, questioning whether alternatives to the bacterial root depicted in universal trees (figure 1) can really be rejected, given the limitations of the models used to recover it [69–71]. They suggest that the use of better models and more careful attention to the properties of data are needed to re-evaluate the root position, and we firmly agree that this is urgently needed. In particular, the inference that eukaryotes branch within Archaea presumes a root outside of those two groups—a tenuous assumption, according to Gouy et al. [80]. They also argue that the preference for the bacterial root is influenced by a persistent bias that favours simple to complex evolutionary scenarios, an unhelpful progressivist attitude that is also criticized by Mariscal & Doolittle [79].

The question of how best to root phylogenetic trees is an outstanding one at all levels of the taxonomic hierarchy, with the recent controversy about the root of the eukaryotic tree providing another important example [81–83]. Most of the published tree-based methods for rooting rely on outgroup rooting. This has well-known problems, because the outgroup is often highly divergent from the ingroup, and this makes analyses susceptible to the well-known long-branch artefact that has bedevilled work on early evolution, as discussed by a number of our contributors. As an alternative to outgroup rooting, Williams et al. [84] evaluate the potential of non-reversible and non-stationary substitution models, which infer the root of the tree as an integral part of the analysis. These are models in which the probability of the tree depends on the starting point of the substitutional process, so that the inferred trees are rooted. These methods have previously shown promise [85,86], but have not been applied more generally because of the additional computational burden of model fitting in comparison to standard models. Two recently described models were applied to infer the root of the universal tree and obtained a root either within the bacterial domain or on the branch separating the Bacteria and Archaea, providing some support for prokaryotes-first hypotheses and suggesting that gene sequences contain a rooting signal that can be extracted. However, as with the methods discussed by Larotit [60] and Gouy et al. [80], current implementations are slow, limiting the size of the datasets that can be analysed—a serious difficulty given the established importance of broad taxonomic sampling for inferring phylogenetic trees [59] and the models, while promising, are by no means consummated.

6. Some concluding remarks

Inferring ancient events from small amounts of data using methods that are not completely up to the job is unlikely to be error-free, and some views will no doubt change again. Nonetheless, the papers in this theme issue—and those in another recent collection [87]—testify to an era of remarkable excitement in the field of eukaryotic origins. The debate about the relative importance of non-endosymbiotic gene transfer, and bulk versus continual transfer hypotheses as a source(s) of prokaryotic genes on eukaryotic and archaeal genomes is particularly vibrant. Some of the discussion is fuelled by the inherent difficulties in trying to infer events from trees that are poorly resolved, because of saturation and other complexities of gene evolution, and also because of still limited sampling of microbial diversity. Nevertheless, it is very clear, and has been for some time [8,9,88], that widespread HGT means that no single tree can depict the history of all genes on prokaryotic or eukaryotic genomes. Trees and non-tree-based methods like networks will continue to be complementary and synergistic approaches for analysing how genomes evolve.

One area that is particularly exciting is the exploration of uncultured microbial diversity, which has the potential to hone in on the closest extant relatives of the mitochondrial endosymbiont [89] and of the proposed archaeal host lineage [23] and provide an experimental framework for testing currently favoured hypotheses. Those partners in early eukaryotic evolution, like all ancestors, are long-extinct—but better sampling of their modern relatives can help to improve trees and to refine inferences about the gene content and cellular features of our prokaryotic ancestors. The discovery of the Lokiarchaeota, with their enhanced content of genes previously thought to be eukaryotic specific, is a particularly exciting discovery and provides evidence that phylogenetic methods, however imperfect, can be used to infer ancient relationships [24]. But sequence data can only take us so far and a major challenge now is to isolate Lokiarchaeota and other relevant environmental lineages into culture so that the cellular manifestation of their genome content—their biology and physiology—can actually be studied in the laboratory.

Competing interests. We have no competing interests.

Funding. T.A.W. and T.M.E. were supported by a European Research Council Advanced Investigator Programme Grant (ERC-2010-AdG-266701) and a Programme Grant from the Wellcome Trust (no. 045404) to T.M.E.

Acknowledgements. We express our sincere thanks to all the contributing authors for their willingness to produce timely articles for this theme issue. We also thank Helen Eaton, Ruth Milne and the rest of the editorial team of Philosophical Transactions of Royal Society B for the enjoyable opportunity to put this volume together.
AUTHOR PROFILES

Martin Embley completed his PhD in Microbiology at Newcastle University in 1984 and then taught Industrial Microbiology and Molecular Biology at North East London Polytechnic. After a sabbatical in Erko Stackebrandt’s laboratory learning how to sequence and analyse ribosomal RNA, he moved to the Department of Zoology at the Natural History Museum in London in 1991 to help set up a new DNA laboratory. At the Museum he investigated the early evolution of eukaryotes and the role of mitochondria in eukaryotic origins as well as the functional diversity of mitochondrial homologues including hydrogenosomes and mitosomes. In 2004, he moved to the Institute of Cell and Molecular Biosciences at Newcastle University to take up the Chair of Molecular Evolution. His group uses phylogenomics and experimental cell biology to investigate the chimaeric origins of eukaryotic cells and to identify the essential functions of the tiny minimal mitochondria of parasitic protists. Genome analyses and functional studies are also used to understand how obligate intracellular eukaryotic parasites (microsporidians) exploit the cells they infect.

Tom Williams obtained a BA in Genetics and a PhD in Molecular Evolution from Trinity College Dublin. His PhD work with Mario Fares focused on protein folding and the evolution of molecular chaperones in Bacteria and Archaea. From 2010 to 2015, he was a Marie Curie Fellow and then a Research Associate in Martin Embley’s group at Newcastle University, working on phylogenetics and eukaryotic genome evolution. In October 2015, he moves to Bristol University as a Royal Society University Research Fellow.

References


