PHILOSOPHICAL TRANSACTIONS B

rstb.royalsocietypublishing.org

Review

Cite this article: Mariscal C, Doolittle WF. 2015 Eukaryotes first: how could that be? *Phil. Trans. R. Soc. B* **370**: 20140322. http://dx.doi.org/10.1098/rstb.2014.0322

Accepted: 24 May 2015

One contribution of 17 to a theme issue 'Eukaryotic origins: progress and challenges'.

Subject Areas:
evolution, genomics

Keywords:
eukaryotes, LUCA, LECA, streamlining, convergence

Author for correspondence:
W. Ford Doolittle
e-mail: w.ford.doolittle@dal.ca

Eukaryotes first: how could that be?

Carlos Mariscal[1,2] and W. Ford Doolittle[2]

[1]Departments of Philosophy, and [2]Biochemistry and Molecular Biology, Dalhousie University, PO Box 15000, Halifax, Nova Scotia, Canada B3H 4R2

In the half century since the formulation of the prokaryote : eukaryote dichotomy, many authors have proposed that the former evolved from something resembling the latter, in defiance of common (and possibly common sense) views. In such 'eukaryotes first' (EF) scenarios, the last universal common ancestor is imagined to have possessed significantly many of the complex characteristics of contemporary eukaryotes, as relics of an earlier 'progenotic' period or RNA world. Bacteria and Archaea thus must have lost these complex features secondarily, through 'streamlining'. If the canonical three-domain tree in which Archaea and Eukarya are sisters is accepted, EF entails that Bacteria and Archaea are *convergently* prokaryotic. We ask what this means and how it might be tested.

1. Introduction

Our aim in this review is to typologize and critique a class of evolutionary scenarios we call 'eukaryotes first' (EF), and provide some criteria by which they might be rejected. EF is the claim that importantly many of the features that now distinguish members of the clade Eukarya from prokaryotes were already present in the last universal common ancestor (LUCA) of all life. In a convenient terminology, such features can be called eukaryotic 'cellular signature structures' (CSSs) [1] and are made up at least in part by 'eukaryotic signature proteins' (ESPs) [2]. CSSs might include elements of the endomembrane system (such as the nuclear membrane), the cytoskeleton, spliceosomal introns and some of the complex RNA-based regulatory networks now being described in eukaryotes. To the extent that LUCA had ESPs and CSSs, it was *phenotypically* a eukaryotic cell even if not, cladistically speaking, a member of Eukarya. We must navigate through the common conflation in biology between systematics and evolutionary history, or between cladistics and phenetics. EF is not a claim that members of the clade designated Eukarya or Eukaryota—which comprises the last eukaryotic common ancestor (LECA) and all its descendants—gave rise to either of the prokaryotic clades, Bacteria and Archaea. Nor does EF, in any version of which we are aware, imagine that alpha-protoeobacterial or cyanobacterial cells are escaped mitochondria or plastids. EF for us means 'eukaryotes first' but not 'Eukarya first'.

EF challenges prevailing beliefs in two ways. First, it goes against what we think is the majority view about ESPs and CSSs, that their origins or acquisition of modern function represent advances achieved in Eukarya since its divergence from prokaryotes—that is, between the first eukaryotic common ancestor (FECA), all of whose descendants other than LECA are extinct, and LECA [3]. Even those who question that evolutionary complexification is intrinsically progressive overall often understand the history of the eukaryotic lineage in this way [4]. EF denies this view in whole or part, and puts at least some eukaryote-typical ESPs or CSSs in LUCA, generally as relics from an earlier 'progenote' stage or 'RNA world'.

Second, when coupled with prevailing versions of the universal tree of life, EF has important implications for prokaryotic evolution. The three-domain tree now seen in most textbooks [5] has its deepest branching separating Bacteria from a clade subsequently giving rise to a monophyletic Archaea and a monophyletic Eukarya (figure 1-3 and 1-4). If this tree is accepted and EF is to be defended, then ESPs and CSSs present in LUCA must have been lost twice (once in the line leading to Bacteria and once in the line leading to Archaea). To the extent that Bacteria and Archaea show similar structures or processes that can be seen

THE ROYAL SOCIETY PUBLISHING

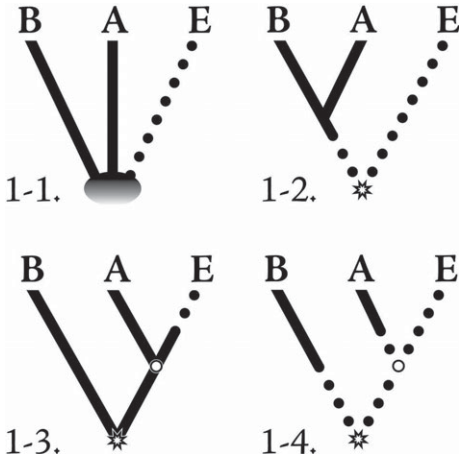

Figure 1. Four options for the evolution of eukaryote-like cellular complexity, represented by the dotted line. Where it exists, the last common ancestor exclusive to Archaea and Eukarya, LAECA, is shown as an open circle, and LUCA is shown as a starburst. In the first scenario, all three branches share a common ancestor in the form of a heterogeneous community of organisms [6–9]. It is unclear that any form of comparative genomic analyses can test this. With the second, Bacteria and Archaea are sisters, and simplification from a eukaryote-like ancestral state began after their divergence from the eukaryotic lineage, which remains primitively complex [10]. With the third, which we take as having been the consensus or 'received' view for the last several decades, the tree is rooted on the line leading to Bacteria, and most complexification develops after Eukarya and Archaea diverge from each other (after LAECA). The fourth possibility differs from the third in that LAECA already possessed important complex, eukaryote-typical features which it inherited from LUCA. Thus, Bacteria and Archaea are independently 'streamlined' and the features that make them similar as prokaryotes are convergent. This interpretation unites the canonical three-domain phylogeny with EF thinking. How the likely possibility that Eukarya branch within—rather than as sister to—Archaea affects this interpretation is discussed in the text (§4). 1-2 and 1-4 are EF scenarios, as we define the concept.

as *streamlined versions* of or replacements for the lost ESPs and CSSs, they exhibit *evolutionary convergence* as 'prokaryotes'.

An increasingly well-supported modification of the three-domain tree has eukaryotes emerging from *within* a paraphyletic Archaea, however [11–13]. If this new tree is accepted and several deeper archaeal lineages all share with each other and with Bacteria such similar 'prokaryotic' structures or processes then—if EF is *still* to be defended—convergence must have occurred multiple times. There would have to have been multiple instances of eukaryote → prokaryote 'streamlining' versus only one of gain (complexification in the prokaryote → eukaryote transition) as envisioned by standard evolutionary progressivist views. So advancing phylogenetic and phylogenomic research and an adherence to principles of parsimony might be expected to pull the rug out from under EF theorizing, if evolutionary model problems and lateral gene transfer (LGT) do not ultimately confound us.

However, there are four higher-order reasons why EF thinking may not (and possibly should not) go away. First, EF views are various: some are as non-committal as the inference that LUCA was 'more complex' than most of its descendants and are silent about the possession of specific contemporary ESPs or CSSs. Woese's notion of genomic 'annealing' is not without intuitive appeal [6]. Second, what it is to be a 'prokaryote' is unclear [14,15]. If genomic reduction is all that is entailed, Bacteria and the one or more archaeal lineages that have, according to EF, separately achieved that status may be convergent only in a trivial sense. Third, diverse instances of genomic reduction are increasingly well documented and understood at the cellular and population levels [16,17]. Finally, evolutionary progressivism and the equation of complexity with progress are persistent biases even among biologists, and we should always guard against them [18].

In §2 of this review, we provide a chronology of some EF scenarios. No doubt we have missed many, and in any case, the line between principled and off-the-wall evolutionary speculation is often hard to draw. In §3, we parse the parsimonious possibilities for such EF schemes as a function of three-domain tree topologies. In §4, we discuss the emerging and more complex eukaryogenesis story and its relevance to EF. Finally, in §5, we explore the feasibility of an ancestrally eukaryotic lineage converging on prokaryotic attributes and consider five eukaryotic cellular systems as examples. Our goal is not to decide on the truth of any of the EF hypotheses, but to set out conditions for their more thorough consideration and possible rejection.

2. Some eukaryotes-first scenarios

The 1962 paper of Stanier & van Niel [19], which set the stage for a subsequent 'prokaryote : eukaryote' dichotomy, was agnostic as to which came first, emphasizing instead their parallel evolutionary radiation.

> . . . if we look at the microbial world in its entirety, we can now see that evolutionary diversification through time has taken place on two distinct levels of cellular organization, each of which embodied, within certain limits, the same kinds of evolutionary potentialities [19, p. 33].

But combining this dichotomy with the then dominant 'autogenous origin' hypothesis—that cyanobacteria ('blue-green algae') are the progenitors of all algae and higher plants, and indeed (by loss of photosynthesis), all eukaryotes—polarizes the relationship [20]. Margulis, while presenting the endosymbiont hypothesis as an alternative to autogenous origin, nevertheless also held that eukaryotes emerged from *within* the prokaryotes [21]. Most pertinently, such a view still underwrites the current phylogenetic consensus, in which Eukarya arise as sisters to or within Archaea. The standard view of eukaryogenesis as a 'major transition', ratcheting up organismal and cellular complexity, also seems to speak against EF scenarios [4]. Nevertheless, some authors have imagined such scenarios, often boasting to be the first to have done so. Our purpose in this section is to recount, without initial critical comment, several such claims.

(a) Reanney's genome reduction scenario (1974)

A notably early EF proposal would be Reanney's [22]. Based on 'the fact that many stages of evolution appear to have been accompanied by physical loss of superfluous DNA', he 'postulated that the genomes of prokaryotes—where almost every gene is represented by one copy only—represent the results of this process carried to its extreme'. Moreover, he mused, 'certain features of very early evolution which have been eliminated from prokaryotes may survive in eukaryotes' [22].

(b) Woese and Fox's progenote (1977–1982)

For Carl Woese, a motivating belief was that major structural and functional differences between the translational

machineries of Bacteria, Archaea and eukaryotes reflect different solutions to problems of translational accuracy and efficiency as yet only partially solved in their common ancestor. That ancestor was thus more primitive than any contemporary pro- or eukaryote: Woese & Fox [7] called it 'the progenote'. Moreover, they imagined LGT to have been so common early on that one might think of the progenote as a state of 'genetic communion', not an identifiable single common ancestral cell or species (figure 1-1). They did accept as fact the endosymbiont hypothesis for the origin of organelles and its logical consequence that the engulfing host was organizationally prokaryotic at the time of engulfment, insofar as it lacked compartmentalized organelles for respiration and photosynthesis. But, they suggested that we 'revise those attitudes concerning eukaryotic evolution that are based on the preconception that the cytoplasmic component arose relatively late in the evolutionary scheme from the bacterial tree' [7]. Although his focus was on translation, Woese, in 1982, would argue that other characteristically eukaryotic cellular attributes also testify to an early divergence, and that eukaryotes are in fact more 'primitive'—more progenote-like—than prokaryotes:

> When one looks at the eukaryotic cell in molecular detail, at its nuclear organization—control mechanisms, introns and so on—it feels in a way less, not more, advanced than its prokaryotic counterparts—less streamlined, less straightforwardly controlled. In fact, one might profitably consider that the urcaryote [the earliest eukaryote] in ways resembles the progenote more than do prokaryotes. I would like to suggest that the eukaryotic cell evolved from the progenote at a somewhat later time than did its prokaryotic counterparts, and that this is in fact responsible for the origin of the nucleus [8, p. 14].

(c) 'Introns early' (1978)

Three year earlier, one of us (W.F.D.) and James Darnell independently advanced this same idea [23,24]. Stunned (as were most geneticists) by the discovery of introns, dubious of Gilbert's [25] notion that introns' 'function' was to facilitate eukaryote evolution, and cognizant of Woese and Fox's three-domain phylogeny, we independently proposed that introns were present at a pre-cellular (or 'progenote') evolutionary stage and had participated in the assembly of the first genes. They were, we suggested, subsequently eliminated by genomic 'streamlining' from Bacteria and Archaea, a process that similarly simplified many inefficient and 'primitive' processes and structures that modern eukaryotes remain saddled with. In our (W.F.D.'s) words . . .

> The assumption that the tightly organized prokaryotic genome is 'primitive' and represents the sort of organization found in the common ancestor of the prokaryote and eukaryote genomes is so common that it is usually only implicitly stated. However, there is no direct evidence in favour of this assumption and some indirect evidence against it. I would like to argue that the eukaryotic genome, at least in that aspect of its structure manifested as 'genes in pieces' is in fact the primitive original form [23, p. 581].

(d) Hartman's kronocyte (1984)

Hartman [26], also impressed by how much nuclear biochemistry is tied up in the making and processing of RNA, both coding and non-coding, ventured that the nucleus, such as mitochondria and plastids, derived from an endosymbiont. The host for all three (and the original possessor of many typically eukaryotic cellular features) was to have been a primitive RNA-genomed creature he called the kronocyte (after Zeus's father). The notion of some sort of RNA-world-generated RNA-genomed ancestor, converting to DNA independently at the base of one or more of the three domains, remained popular for some time. For instance, Mushegian & Koonin [27] inferred the existence of such an entity from the absence of several key proteins of bacterial DNA replication from Archaea and eukaryotes.

(e) Sogin's fourth domain (1991)

Sogin [28] presented a more fleshed-out version of the RNA-genomed, cytoskeleton-equipped host scenario. From a progenote still enmeshed in the 'RNA world' emerged a DNA-based lineage which branched into Bacteria and Archaea and a primitive RNA-based lineage with an 'RNA-dominated infrastructure', in which the 'major innovation was the cytoskeleton'. This allowed it to engulf bacterial and archaeal symbionts, one of the latter becoming the nucleus. This archaeal genome 'contributed DNA and the majority of protein-coding regions found in contemporary eukaryotes, while the proto-eukaryotic [RNA-genomed] lineage contributed coding information for synthesis of both the translation apparatus and the cytoskeleton' [28]. One abiding mystery this scenario claimed to resolve was the origin of protein families that are characteristic of eukaryotic cellular structures and absent or present only in very distant homologues among prokaryotes. The former were first called ESPs by Hartman & Fedorov [2,29], who claimed that there are 300–400 of them, those with identifiable functions being (in the main) 'components of the cytoskeleton, inner membranes, RNA-modification machinery, and the major elements of intracellular control systems such as ubiquitin, inositol phosphates, cyclins, and the GTP binding proteins' [2]. Kurland *et al.* [1] later introduced the term CSS for eukaryote-specific structures (nucleus, nucleoli, Golgi apparatus and so forth) in which many ESPs play a role.

(f) Forterre's thermoreduction hypotheses (1995)

Forterre [30] imagined a cellular LUCA characterized by 'a higher gene content with multicopy genes, a bigger cell volume and diverse molecular mechanisms or structures (possibly a nuclear membrane)'. Although this ancestor's genome was already DNA, it remained (like modern eukaryotes with their introns and non-coding RNAs) heavily invested in RNA metabolism. Because RNA (mRNA in particular) is heat-labile, it must have been a mesophile. Adaptation to thermophily (then thought to have occurred at the base of both Bacteria and Archaea) entailed a loss of any nuclear membrane (so that mRNA processing times were reduced) and of extra DNA and extra gene copies. Such streamlining meant to Forterre that prokaryotes, though simpler than eukaryotes, are more highly evolved: the 'pro' in prokaryotes is a misnomer. In a recent and more richly elaborated formulation of this hypothesis [31], Forterre accepts the 'classical' rooted three-domain tree in which Archaea and Eukarya are sisters, and takes on board the findings of homologues of several ESPs in one or another newly discovered archaeal lineage [32,33]. He proposes that the last archael/eukaryal common ancestor (LAECA) was more complex than any contemporary archaeon and that Archaea, independently from Bacteria, underwent 'thermoreduction'. But, the eukaryal lineage still had many complexities yet to acquire, so this LAECA was a 'bug-in-between'. Forterre thus approaches what we wager will be the emerging consensus

as more and more ESPs are found in deeply diverging archaeal genomes (see §4).

(g) The ribonucleoprotein worldview (1999–2009)

Penny and co-workers added r-selection to thermoreduction as a driver of prokaryotic genomic streamlining and elaborated arguments for a universal common ancestor more eukaryote-like than prokaryotic [9,34–36]. RNA world relics passed on through LUCA to contemporary eukaryotes include diplo(poly)ploidy, telomerase RNA, linear genomes, rRNA processing by snoRNA and splicing with snRNAs. All these would have been lost either once or twice (depending on where the three domain tree is rooted, see below) in prokaryotes, which were thus 'the new kids on the block'. Latter articulations of their scenario re-emphasize the 'continuity hypothesis', in which the ribonucleoprotein (RNP) complexes involved in so many contemporary eukaryotic cellular processes are directly descended from RNPs that facilitated the transition from an RNA to a DNA–protein world, and bring to bear arguments about 'molecular crowding'. The latter phenomenon, more acute in larger cells, it is claimed, drove subcellular compartmentalization (before any endosymbiosis) and secured a continuing role for RNPs.

(h) Glansdorff and Labedan's community LUCA (2008)

In a sort of mash-up of Forterre's thermoreduction, Woese and Fox's progenote and Hartman's kronocyte hypotheses, these authors infer that 'LUCA was a protoeukaryote, with a RNA genome inherited from its progenote ancestor', and that the 'RNA LUCA was a metabolically and morphologically heterogenous community, constantly shuffling around genetic material'. Some of this community's members were phagotrophs, whereas others were 'thermoreduced' into Archaea or Bacteria [37]. Such views are stubbornly recurrent and although it seems reasonable that LGT ('shuffling around') has always been important, to view communities as ancestors is to conflate the notions of common ancestor and common ancestry. As suggested in figure 1-1, community hypotheses are not amenable to analysis by parsimony or logic.

(i) The Planctomycetes–Verrucomicrobia–Chlamydiae-first hypothesis (2010)

Forterre and Gribaldo, in synthesizing work of several groups over the previous two decades, proposed (tentatively) an EF scenario in which the Planctomycetes–Verrucomicrobia–Chlamydiae (PVC) superphylum—shown by some analyses to be the earliest diverging bacteria—inherited from LUCA several ESPs and even versions of CSSs that were subsequently lost in the lines leading to non-PVC bacteria (i.e. *all* other bacteria) and in Archaea, after its divergence from the eukaryotic lineage [38]. Among such ESPs and CSSs were membrane coat proteins that allow formation of an intracytoplasmic membrane system, a nucleus-like enclosure for the DNA, and phagocytosis-like particle engulfment, structures and processes unknown among bacteria outside the PVC clade. That these and other eukaryote-like PVC characteristics are homologous to their eukaryotic counterparts (as needed for any EF claim) was vigorously denounced in 2011 in a paper authored by an astonishingly broad collection of the field's luminaries, who themselves otherwise hold warring views about eukaryogenesis [39]. They assert that 'all of the PVC traits that are currently cited as evidence for aspiring eukaryoticity (sic) are either analogous (the result of convergent evolution), not homologous, to eukaryotic traits, or else they are the result of horizontal gene transfers' [39]. This may well be the most thoroughgoing published rebuttal of any EF hypothesis, and aims at the kind of care in distinguishing convergence (analogy) from retained similarity (homology) we recommend below. Of course, such concern is not immune to biases, for instance taking the non-exclusivity of ESPs among the PVC group as evidence against the EF notion.

(j) Kurland's new root (2013)

Harish *et al.* [10] have recently produced a new and very EF-friendly universal tree. They use an elaborate genome-content approach to obtain an 'akaryote' (Bacteria and Archaea) branch and a eukaryote branch, with an inferred most recent universal common ancestor (MRUCA) that was in gene content more complex than its descendants. The authors admit that 'though we cannot draw the conclusion that MRUCA was morphologically similar to a eukaryote, elements and cohorts of [protein superfamilies] from its proteome are recognizable in the proteomes of modern eukaryotes' [10]. Moreover, they infer that MRUCA (cell, species or population) was the survivor of a mass extinction event much more recent than the 3.5–4.0 billion years ago usually assumed for LUCA. Re-rooting the tree so that Bacteria and Archaea are sisters means that we need not consider their shared prokaryotic traits to be convergent. Eukaryote-like features shared by Archaea and eukaryotes to the exclusion of Bacteria may parsimoniously be ancestral to all three domains (present in LUCA, lost in Bacteria). But such a rooting is not widely accepted and seems increasingly unlikely, as more and more and more evidence situating eukaryotes *within* Archaea accumulates.

3. Parsing possibilities and reasoning by parsimony

As we have seen, the radical element these scenarios share is the notion that many ESPs and CSSs are primitive features retained from LUCA rather than evidence of the advanced state of eukaryotes *vis-à-vis* prokaryotes. Although claims that LUCA was a population (figure 1-1) are difficult to analyse and may be logically incoherent, most of the EF hypotheses discussed in §2 take the form shown in figure 1-2 or 1-4, and are to be contrasted with what we think to be the standard or received view (figure 1-3). The latter three trees can be subjected to simple parsimony reasoning, which we consider the first tool for evaluating evolutionary hypotheses [40]. All else being equal, trees 1-2 or 1-3 would be preferred to 1-4, because either requires only one event (simplification after the Bacteria + Archaea clade diverged from Eukarya, or complexification after Eukarya diverged from Archaea) rather than two (independent or convergent simplification in both bacterial and archaeal branches). But, 1-2 is inconsistent with most phylogenies and applying simple parsimony to 1-3 and 1-4 (to which we will devote the rest of our attention) is reasonable only if loss and gain of eukaryote-typical complexity are equally probable.

Some authors argue that genome reduction is, in fact, a more likely or common evolutionary outcome than genome growth. Indeed, Wolf and Koonin recently proposed:

> a general model composed of two distinct evolutionary phases: the short, explosive, innovation phase that leads to an abrupt increase in genome complexity, followed by a much longer reductive phase, which encompasses either a neutral ratchet of genetic material loss or adaptive genome streamlining [17, p. 829].

If along both bacterial and archaeal branches there were great expansions in typical effective populations sizes, as many believe characterize prokaryotes *vis-à-vis* eukaryotes, selection for streamlining might well have independently effected convergence to simpler cellular organization. By this reasoning, similar streamlining has not occurred in the eukaryotes either, because complexity confers fitness in that lineage, or because smaller eukaryotic population sizes allowed retention of the mildly deleterious products of selfish and constructive neutral evolutionary forces [41,42].

Thus, the relative credibility of the EF scenario 1-4 depends not so much on parsimony as on a convincing demonstration that the prokaryotic character of Bacteria and Archaea reflects convergent streamlining rather than retained simplicity (as in 1-3). How might that be shown, and what is a prokaryote other than 'not a eukaryote'? Woese, and more recently and emphatically Norman Pace [14], would answer 'little or nothing'. We deal with this attitude and its implications in §5 of this review. But first, we must address the further complication created by recent revisions to the universal tree.

4. The return of the eocytes

In the 1980s, James Lake proposed, on the basis of structural studies on ribosomes, that eukaryotes were specifically related to the Crenarchaeota, one of the two then accepted divisions of Archaea [43]. This notion suffered 30 years of neglect, during which it was widely believed that Archaea is monophyletic and Eukarya is its sister (as in figure 1-3 and 1-4). New data and analyses, some from newly discovered archaeal phyla, have revived Lake's 'eocyte hypothesis' insofar as something resembling a crenarchaeal–eukaryotic grouping is concerned [11–13], and several of the papers in this issue argue effectively for this. Topologically, these new trees should make EF even harder to defend, because convergent streamlining to prokaryote status would have to have occurred as many more times as there are archaeal lineages branching off before LECA. Thus, much hinges on just how many such branches there are—on the precise structure of the archaeal–eukaryotic tree [13]. Until that is settled, we can expect continued arguments between the more traditionally minded who would envision a progressive accumulation in Archaea of precursors to the complex feature that eukaryotes then went on to perfect, and those who see the ESPs exhibited by Archaea as remnants left after streamlining from a relatively complex LAECA [31,33,44,45].

Ettema and co-workers [13,32] would be in the former camp. Their 'phagocytosing archaean theory' (PhAT), seemingly now well supported by the discovery of the Lokiarchaeota [13] is of the former type, and envisions that an archaean of the TACK superphylum lost its cell wall, 'allowing for the evolution of a more flexible actin-based cytoskeleton', this 'matured into a primitive phagocytosis machinery' which encouraged 'rampant' LGT, to protect against which 'a protective membrane [was] formed via invagination events, giving rise to a primitive karyotic cell type' [32]. Koonin & Yutin [33], endorse the opposing position, writing that . . .

> The complexity of the archaeal ancestor was apparently fixed in the emerging eukaryotes thanks to endosymbiosis. In contrast, the proto-eukaryotic features were differentially lost in archaeal lineages in the course of reductive evolution, resulting in the currently observed dispersed eukaryome [33, p. 13].

Parsimony analysis can, in principle, identify the ESPs and CSSs present in the genome of the 'bug-in-between' (LAECA). This will require a taxon-rich, diverse, rooted and uncontroversial deep archaeal tree, in which Eukarya are securely located [46]. Consensus will be difficult to achieve, however. One can predict continuing controversies over evolutionary models for deep phylogeny and ancestral state reconstruction, and LGT will always be a wild card: where ESPs currently exist may say little about lineage histories. Moreover, even if such analyses reliably show LAECA to have been complex, we cannot tell, using only archaeal and eukaryotic data, what happened between LUCA and LAECA—gradual acquisition of the building blocks of eukaryotic CSSs that come to full modern function in LECA, or loss of much of an even richer heritage of complexity present in LUCA. For this, we must still look at what it is that unites Bacteria and Archaea as prokaryotes, and ask if we think these features are *plausibly* products of convergent reductive streamlining of eukaryotic antecedents. If not, then they must have been retained from their common ancestor, LUCA, which was to that extent a prokaryote. In §5, we question whether the possibilities open to streamlining lineages are so limited (by physical or genetic and developmental constraints) that similar solutions are expected to be achieved by independent streamlining episodes. We also investigate whether the onerousness of such proposed episodes renders the conversion from eukaryotes to prokaryotes mechanistically implausible.

5. A principled approach to convergence and common descent

To reiterate, the received view (figure 1-3) considers typically 'prokaryotic' traits shared by Bacteria and Archaea to the exclusion of eukaryotes to be due largely to common descent (retained similarity, or homology), whereas EF (figure 1-4) interprets them as due to convergent streamlining evolution. *A priori*, the received view will be more plausible if it turns out that genomic and organismal streamlining is a difficult or rare process in evolution or if independent methods indicate that shared 'prokaryotic' features of Bacteria and Archaea are retained from LUCA. If streamlining is a dominant evolutionary process, as some now maintain [16,17], the tree in figure 1-4 is perhaps as plausible as the tree in figure 1-3. But even if we accepted that, comparing trees this way is too coarse-grained an analysis; a full answer must deal with the murky specifics of the history of life on Earth. We must have independent, principled ways to distinguish convergent from retained similarity.

Convergent evolution, the repetition of form or function in evolution, can occur for a variety of reasons, some of which might initially appear to be quite implausible [47]. Key to diagnosing whether two taxa are similar due to convergence or common descent is a precise statement of what one of us (C.M.) has called *specificity*. Specificity can be defined as the ratio of the number of identical characters to total possible characters within a given biological context.

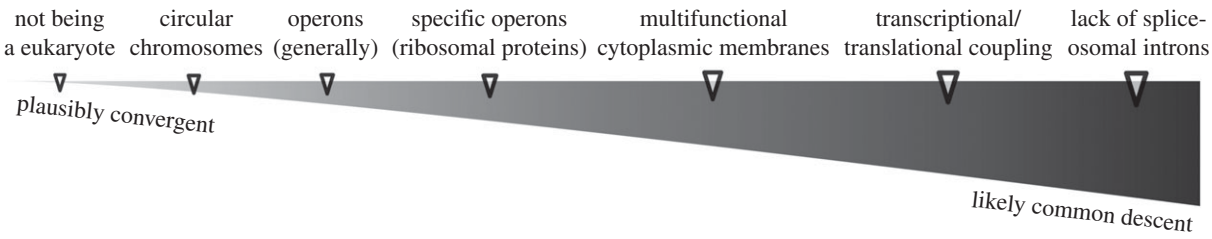

Figure 2. A continuum of specificity. Labelled are features common to Archaea and Bacteria that would be instances of convergent evolution if tree 1-4 is correct, though we stress these labels are early approximations. See §5 for a discussion of these features and their implications for the plausibility of EF views.

As an example, consider the genetic code. Crick [48] once argued that the only reason the genetic code was universal was because all life shares common ancestry; the code is a 'frozen accident'. His reasoning was based on two factors: the specificity of the code and (discussed later) the difficulty (indeed 'lethality') of evolving alternative codes once one had been established.

Specificity, in this case, is the fraction of all mathematically or physically possible codes represented by existing codes—one or a few divided by a very large number. But it may be that some codes are superior to others, because of nucleic acid chemistry (for instance, a stereochemical fit between anticodon and amino acid, whose generality Crick was at pains to discount) or because of their buffering of the effects of mutation. When accounting for such features as hydrophobicity and production of physicochemically similar amino acids, it has been argued that 99.97% of possible codes are mutationally less robust than our current code [49]. But even if we suppose the vast majority of codes are worse than our current code, it might still follow that there are tens of thousands of equivalent or better codes. So if all of life exhibits one of a few of these possible options, it follows that the genetic code has a high degree of specificity [50–52]. Moreover, we must distinguish *selected specificity*, in which natural selection constrains choice among alternatives and might plausibly produce convergence of independently evolving entities, and *neutral specificity*, in which alternatives have equal fitness and similar outcomes bespeak common descent more likely than convergence. The argument for the genetic code being a frozen accident can now be understood as a claim about the code's high neutral specificity and low or moderate selective specificity.

Specificity also applies to structures or processes not so easily enumerated as are possible codes. As an example, consider blubber, the subcutaneous layer of fat used by many endothermic animals to regulate their body heat. Relative to the sheer number of possible ways to regulate body heat, blubber is actually a fairly specific trait. An alternative way to describe this feature, 'insulation', would be less specific, but cover more cases, including fur, feathers and so on. An even broader description, 'thermoregulation', would include not only blubber, fur and feathers, but even radically different features, such as behaviour. The less specific a trait, the more likely it is to convergently evolve, all else being equal. But very unspecific traits, such as thermoregulation, are weak examples of either convergence or retained similarity, because any number of evolved features might be included as convergent, even if the evolutionary pressures and underlying structures were quite different!

Crick's argument also rests on an analysis of the difficulty of evolving different codes, which depends on the selective advantage of any alternative code relative to the cost of new variations. If the evolutionary cost of modifying the code is low or if the selective advantage of an alternative code is very high, we would expect natural selection to move towards the more optimal code—even from very distant starting points. Convergence on an ideal code would not be difficult in such a scenario. In fact, there are many instances on Earth of variations in the genetic code [53] but all are, as Crick surmised, minor. The vast majority of likely possible codes are not explored because of the advantage of moving to an entirely new code, even if superior, cannot overcome the cost of doing so. Each of three popular models for codon reassignment ('codon capture', 'ambiguous intermediate' and 'genome streamlining') is sufficiently codon-or tRNA-specific and onerous that it is difficult to imagine reassignment of many codons at once, even in small genomes [50]. So even when the number of selectively equivalent alternatives is high, transitions between them may be very difficult (highly constrained), rendering convergence inherently unlikely.

We have seen the EF view can be understood as the notion that LUCA was complex and contained many ESPs. To reject it requires that we argue that those prokaryotic structures and processes shared by Archaea and Bacteria and thought to be the result of convergent streamlining are instead more likely to be retained characteristics of LUCA, later replaced in Eukarya or the archaeal lineage immediately ancestral to FECA. Such an argument entails showing

(1) that sharing of these traits is not a consequence of LGT, insofar as this is a means to effect convergence, and
(2) that these traits are highly specific (representing but a few of many physical or chemical possibilities) and also neutral (not constrained to just these few possibilities by general properties of biological systems, such as the availability of amino acids), so that similarities would be neither trivial nor inevitable because of selection for optimality.

We have given a first-pass approximation of what some distinguishing prokaryotic traits may be and positioned them on a spectrum from possibly convergent to likely common descent (figure 2). We offer preliminary comments about a few of them.

(a) That Archaea and Bacteria are both 'prokaryotes'

This is a similarity of the lowest possible specificity, of little value in distinguishing common descent from convergence

and considered by Pace [14] to be little more than the recognition that Bacteria and Archaea are not eukaryotes. While Doolittle and Zhaxybayeva [15] argue for the continued utility of the term 'prokaryote', its traditional definition is indeed largely negative.

(b) Circular chromosomes and operons

We consider these traits only somewhat more specific and of limited onerousness in terms of independent achievement (convergence). Circles and lines represent two of two logical configurations for information encoded as it was from stages prior to LUCA, and in fact both can be found in prokaryotes [54]. Similarly, within Bacteria and Archaea, operons are conserved in gene order only over short evolutionary distances, although certain clusters of genes do recur [55]. Some clustering is expected as the result of 'selfish operon' selection [56], and various regulatory regimens may favour the proximity of genes for related functions (sometimes even in what have been called super- or 'uber-operons' [57]). The existence of such forces makes the existence of operons a fairly low specificity trait, as likely the result of convergence as of retention from a common ancestor. That there is significant conservation in content and sometimes gene order between clusters encoding translation-related (especially ribosomal) proteins between Bacteria and Archaea [58] might be thought to speak to their homology (presence in LUCA). But presumably any explanation for differential conservation and retention over such a long time must invoke strong selection on such genes, both in terms of coordinated high expression and product interaction/coevolution. There is a general conundrum here: any claim as to the antiquity and homology of similar structures and processes in diverse lineages must invoke selection to maintain them, and yet—if invoked—this same selection should be adequate to create them. It is only if there are many equally good ways to satisfy such selection (high neutral specificity) that similarity implies homology, and we know too little about this for complex systems such as the ribosome, even yet [59].

(c) Multifunctional cytoplasmic membranes

There have been recent attempts to make the definition of 'prokaryote' more positive. For example, Whitman [60] considers one positive shared character, a multifunctional cytoplasmic membrane, on which a 'proton motive force is generated . . . by respiration, photosynthesis, or ATP hydrolysis to empower key cellular process such as ATP biosynthesis, NAD^+ reduction by reverse electron transport, nutrient uptake, motility and secretion'. In eukaryotes, such chemiosmotic processes occur on organellar membranes and as one referee for this manuscript has pointed out, it is difficult to see how a transition to the prokaryote condition could be effected twice, convergently. Thus, much will depend on a detailed comparison of membrane-associated processes in Bacteria and Archaea. Furthermore, LGT may often be a confounding factor, as in the acquisition of bacterial respiratory functions by Haloarchaea [61]. More to the point, though, EF advocates have not focused on such traits, and in our view, an adequately complex entity between FECA and LECA that had yet to acquire organelles might still comprise a starting point for EF. EF does not mean Eukarya-first.

(d) Transcriptional/translational coupling

Whitman [60] also considers the coupling of transcription and translation. What precludes this in eukaryotes (and thus permits infestation by introns) is the nuclear membrane, which *is* sometimes claimed to be characteristic of LUCA in EF scenarios. Martin & Koonin [62] argued, *contra* Pace [14], that such coupling (which is only *assumed* for the majority of Bacteria and Archaea) is a positive 'prokaryotic' trait. Coupling *per se* seems a low specificity trait, but similarities between translational mechanisms in Bacteria and Archaea are clearly suitable subjects for discriminating convergence from homology or common descent (and thus whether LUCA had a nucleus). Canonical eukaryote translation initiation requires capped mRNAs and 'scanning' to the first AUG, whereas prokaryotes use base pairing between a Shine–Dalgarno (SD) sequence in leader regions upstream of the start codon and an anti-SD sequence at the 3′-end of 16S rRNA. The SD interaction is conserved across Bacteria and Archaea [63] and unless mandated by some physical constraints unknown to us, might be a high specificity trait bespeaking common ancestry and against the notion that nuclear membranes and contemporary eukaryote-type translation initiation were present from LUCA to LECA, and abandoned in Bacteria and archaeal lineages branching off before LECA. But Archaea and Bacteria also have leader-independent mechanisms for initiation and there are mechanistic commonalities as well as strong similarities otherwise between archaeal and eukaryotic translation (and transcription) systems [64]. Compelling arguments for the ancestral character of prokaryotic translational components (and thus against convergence) will have to be made on a gene-by-gene basis, and it is hard to see an overall system-level consensus emerging about the homologous nature of transcriptional/translational coupling in Bacteria and Archaea. Interestingly, recent theorizing would have the nuclear envelope arise as a defence against those group II introns brought in by the pre-mitochondrial proteobacterium [65].

(e) Spliceosomal introns

The claim for eukaryotic primitivity based on the presence of spliceosomal introns—the 'introns early' hypothesis—was advanced before anything was known about eukaryotic pre-mRNA splicing mechanisms and when figure 1-1 was a reasonable understanding of domain phylogeny [23,24]. Parallel or convergent loss of introns by streamlining in Bacteria and Archaea did not seem less probable than gain in eukaryotes. Even if the tree topology of figure 1-3 and 1-4 were accepted, convergence was arguably not wildly improbable, given some additional motivating force like increase in effective population size, as suggested in §3. But phylogenetic analyses now place the eukaryotes within Archaea, entailing not two events of loss (as in figure 1-4) but one (for Bacteria) and as many more as there are independently diverging archaeal lineages below LECA. (We think it unlikely that any Archaea will have spliceosomal introns.)

Moreover, refashioning the spliceosomes that service pre-mRNA introns back into their original group II form (as found in both Bacteria and [rarely] Archaea) should be nearly impossible. The widely accepted story about spliceosomal intron origins derives them from group II introns by the fragmentation of these potentially self-splicing mobile elements into 'five easy pieces' [66]—the canonical snRNAs—and the

stepwise acquisition of as many as 200 proteins, so many that the spliceosome is often said to resemble the ribosome in the complexity of its structure and assembly. There is variation of course: *Saccharomyces cerevisiae* has half as many splicing proteins as humans and the thermoacidophilic red alga *Cyanidioschyzon merolae* has only half that many again—43 identifiable 'core splicing proteins' [67] and apparently no U1 associated proteins (or snRNA). There is a consensus now that LECA was intron-rich, and Irimia and Roy [68] recently concluded that 'by the time of LECA, a very complex, modern-looking spliceosome, composed of at least 78 proteins and all the snRNAs, had already been established'. So spliceosome streamlining has happened, but it is rare and never complete, as long as introns are retained. Complete intron removal has happened only once, in the tiny (0/572 Mb) nucleomorph genome of the cryptophyte alga *Hemiselmis andersenii* [69]. So, *contra* Penny and co-workers [9,34–36], we think it unlikely that LAECA or any of its predecessors (LUCA and earlier) had a genome with snRNA-dependent introns. Indeed, the prevailing hypothesis is that these entered the eukaryotic lineage as group II introns in the genome of the pre-mitochondrial symbiont [66,70]. Notably, much of the non-coding RNA of eukaryotes now thought to play vital regulatory roles is intronic, so alternative splicing, as primitive and appealingly progenotic as it might seem [9], would not have been available to LAECA or LUCA.

6. Conclusion

We have explored several heterodox views that imagine prokaryotes as evolving from eukaryotes, rather than the other way around. None of these posits Archaea and Bacteria evolving within the monophyletic clade of Eukarya, but we hold that it is not irrational or illogical to question whether LUCA could have been more similar to modern eukaryotes than prokaryotes in terms of cellular and genomic complexity, however those might be measured. It seems certain that early in cell evolution the machineries of replication, transcription and translation underwent selection for increased accuracy and efficiency, probable that in some systems this entailed reduction in the number of components and complexities of their interaction, and possible that the process was not complete before the separation of cellular lineages leading to Bacteria and Archaea [71]. There is no reason to believe that complexity is always and of itself adaptive and destined to increase, and in fact, there are many well-documented examples of genomic reduction and streamlining [16,17,41,47]. Nor is it proven that the diversification and high level of differentiated cell types exhibited by eukaryotes was only possible because of the more complex nature of their cells [4].

More problematic but also more testable would be claims about the survival from LUCA of specific identifiable ESPs or CSSs—that LUCA had spliceosomal introns or a nuclear membrane for instance. Such claims must be examined on a case-by-case basis, and it is almost certain that such examination will be confounded by (i) disagreements over evolutionary models in phylogenetics, (ii) differing histories of the many components of any complex CSS, (iii) arguments over the role of LGT in those histories, (iv) diversity among eukaryotes and (especially prokaryotes), making nonsense of any generalizations about what is typical for either and (v) problematic attributions of function to proteins known only from genomic or metagenomic DNA sequence data.

Still, if archaeal phylogenomics continues to advance as spectacularly as it has in the past few years, we may some day achieve consensus about which ESPs and CSSs appeared when in the archaeal radiation—in particular which of them LAECA had. For such LAECAn components—if there are any—when they appeared along the line from LUCA to LAECA and whether they were present in a single or multiple lineages will then be the remaining points of contention. Detailed argumentation of the sort we could only allude to in §5 will be required to reach a consensus on which—because of their extraordinarily high degree of 'neutral specificity'—must have been present in LUCA. A consensus on the meaning of words will be required if we are to decide whether their presence made LUCA a 'eukaryote'.

Competing interests. We declare we have no competing interests.

Acknowledgments. We thank the Natural Sciences and Engineering Research Council of Canada (grant no. GLDSU/447989) for support and Austin Booth, Tyler Brunet, Letitia Meynell and Gordon McOuat for helpful comments.

References

1. Kurland CG, Collins LJ, Penny D. 2006 Genomics and the irreducible nature of eukaryote cells. *Science* **312**, 1011–1014. (doi:10.1126/science.1121674)
2. Hartman H, Fedorov A. 2002 The origin of the eukaryotic cell: a genomic investigation. *Proc. Natl Acad. Sci. USA* **99**, 1420–1425. (doi:10.1073/pnas.032658599)
3. Koumandou VL, Wickstead B, Ginger ML, van der Giezen M, Dacks JB, Field MC. 2013 Molecular paleontology and complexity in the last eukaryotic common ancestor. *Crit. Rev. Biochem. Mol. Biol.* **48**, 373–396. (doi:10.3109/10409238.2013.821444)
4. Booth A, Doolittle WF. In press. Eukaryogenesis, how special really? *Proc. Natl Acad. Sci. USA*, 20142376. (doi:10.1073/pnas.1421376112)
5. Woese CR, Kandler O, Wheelis ML. 1990 Towards a natural system of organisms: proposal for the domains Archaea, Bacteria, and Eucarya. *Proc. Natl Acad. Sci. USA* **87**, 4576–4579. (doi:10.1073/pnas.87.12.4576)
6. Woese C. 1998 The universal ancestor. *Proc Natl Acad Sci USA* **95**, 6854–6859. (doi:10.1073/pnas.95.12.6854)
7. Woese CR, Fox GE. 1977 The concept of cellular evolution. *J. Mol. Evol.* **10**, 1–6. (doi:10.1007/BF01796132)
8. Woese CR. 1982 Archaebacteria and cellular origins: an overview. *Z. B. L. Bakt. Mik. Hyg. I C* **3**, 1–17.
9. Penny D, Collins LJ, Daly TK, Cox SJ. 2014 The relative ages of eukaryotes and akaryotes. *J. Mol. Evol.* **79**, 228–239. (doi:10.1007/s00239-014-9643-y)
10. Harish A, Tunlid A, Kurland CG. 2013 Rooted phylogeny of the three superkingdoms. *Biochimie* **95**, 1593–1604. (doi:10.1016/j.biochi.2013.04.016)
11. Williams TA, Foster PG, Cox CJ, Embley TM. 2013 An archaeal origin of eukaryotes supports only two primary domains of life. *Nature* **504**, 231–236. (doi:10.1038/nature12779)
12. Williams TA, Embley TM. 2014 Archaeal 'dark matter' and the origin of eukaryotes. *Genome Biol. Evol.* **6**, 474–481. (doi:10.1093/gbe/evu031)
13. Spang A *et al.* 2015 Complex Archaea that bridge the gap between prokaryotes and eukaryotes. *Nature* **521**, 173–179. (doi:10.1038/nature14447)

14. Pace NR. 2006 Time for a change. *Nature* **441**, 289. (doi:10.1038/441289a)
15. Doolittle WF, Zhaxybayeva O. 2013 What is a prokaryote? In *The prokaryotes—prokaryotic biology and symbiotic associations* (eds E Rosenberg, EF DeLong, E Stackebrandt, S Lory, F Thompson), pp. 21–37. Berlin, Germany: Springer.
16. Giovannoni SJ, Thrash JC, Temperton B. 2014 Implications of streamlining theory for microbial ecology. *ISME J.* **8**, 1553–1565. (doi:10.1038/ismej.2014.60)
17. Wolf YI, Koonin EV. 2013 Genome reduction as the dominant mode of evolution. *Bioessays* **35**, 829–837. (doi:10.1002/bies.201300037)
18. Gould SJ. 1989 *Wonderful life*. New York, NY: WW Norton.
19. Stanier RY, van Niel CB. 1962 The concept of a bacterium. *Arch. Mikrobiol.* **42**, 17–35. (doi:10.1007/BF00425185)
20. Klein R, Cronquist A. 1967 A consideration of the evolutionary and taxonomic significance of some biochemical, micromorphological and physiological characters in the Thallophytes. *Quart. Rev. Biol.* **42**, 105–296. (doi:10.1086/405346)
21. Margulis L. 1970 *Origin of eukaryotic cells*. New Haven, CT: Yale University Press.
22. Reanney DC. 1974 On the origin of prokaryotes. *J. Theor. Biol.* **48**, 243–251. (doi:10.1016/0022-5193(74)90194-5)
23. Doolittle WF. 1978 Genes in pieces: were they ever together? *Nature* **272**, 581–582. (doi:10.1038/272581a0)
24. Darnell JE. 1978 Implications of RNA–RNA splicing in evolution of eukaryotic cells. *Science* **202**, 1257–1260. (doi:10.1126/science.364651)
25. Gilbert W. 1978 Why genes in pieces? *Nature* **271**, 501. (doi:10.1038/271501a0)
26. Hartman H. 1984 The origin of the eukaryotic cell. *Speculat. Sci. Technol.* **7**, 77–81.
27. Mushegian AR, Koonin EV. 1996 A minimal gene set for cellular life derived by comparison of complete bacterial genomes. *Proc. Natl Acad. Sci. USA* **93**, 10 268–10 273. (doi:10.1073/pnas.93.19.10268)
28. Sogin ML. 1991 Early evolution and the origin of eukaryotes. *Curr. Opin. Genet. Dev.* **1**, 457–463. (doi:10.1016/S0959-437X(05)80192-3)
29. Fedorov A, Hartman H. 2004 What does the microsporidian *E. cuniculi* tell is about the origin of the eukaryotic cell? *J. Mol. Evol.* **59**, 695–702. (doi:10.1007/s00239-003-0085-1)
30. Forterre P. 1995 Thermoreduction, a hypothesis for the origin of prokaryotes. *C. R. Acad. Sci. II* **318**, 415–422.
31. Forterre P. 2013 The common ancestor of Archaea and Eukarya was not an archaeon. *Archaea* **2013**, 1–18. (doi:10.1155/2013/372396)
32. Martijn J, Ettema TJG. 2013 From archaeon to eukaryote: the evolutionary dark ages of the eukaryotic cell. *Biochem. Soc. Trans.* **41**, 451–457. (doi:10.1042/BST20120292)
33. Koonin EV, Yutin N. 2014 The dispersed archaeal eukaryome and the complex archaeal ancestor of eukaryotes. *Cold Spring Harbor Perspect. Biol.* **6**, a01688. (doi:10.1101/cshperspect.a016188)
34. Penny D, Poole A. 1999 The nature of the last universal common ancestor. *Curr. Opin. Genet. Dev.* **9**, 672–677. (doi:10.1016/S0959-437X(99)00020-9)
35. Poole A, Jeffares D, Penny D. 1999 Early evolution: prokaryotes, the new kids on the block. *Bioessays* **21**, 880–889. (doi:10.1002/(SICI)1521-1878(199910)21:10<880::AID-BIES11>3.0.CO;2-P)
36. Collins LJ, Kurland CG, Biggs P, Penny D. 2009 The modern RNP world of eukaryotes. *J. Hered.* **100**, 597–604. (doi:10.1093/jhered/esp064)
37. Glansdorff N, Ying X, Labedan B. 2008 The last universal common ancestor: emergence, constitution and genetic legacy of an elusive forerunner. *Biol. Direct.* **3**, 29. (doi:10.1186/1745-6150-3-29)
38. Forterre P, Gribaldo S. 2010 Bacteria with a eukaryotic touch: a glimpse of ancient evolution? *Proc. Natl Acad. Sci. USA* **107**, 12 739–12 740. (doi:10.1073/pnas.1007720107)
39. McInerney JO, Martin WF, Koonin EV, Allen JF, Galperin MY, Lane N, Archibald JM, Embley TM. 2011 Planctomycetes and eukaryotes: a case of analogy not homology. *Bioessays* **33**, 810–817. (doi:10.1002/bies.201100045)
40. Harvey PH, Pagel M. 1991 *The comparative method in evolutionary biology*. Oxford, UK: Oxford University Press.
41. Lynch M. 2007 *The origins of genome architecture*. Sunderland, MA: Sinauer Associates.
42. Gray MW, Lukes J, Archibald JM, Keeling PJ, Doolittle WF. 2010 Cell biology. Irremediable complexity? *Science* **330**, 920–921. (doi:10.1126/science.1198594)
43. Lake JA, Henderson E, Oakes M, Clark MW. 1984 Eocytes: a new ribosome structure indicates a kingdom with a close relationship to eukaryotes. *Proc. Natl Acad. Sci. USA* **81**, 3786–3790. (doi:10.1073/pnas.81.12.3786)
44. Wolf YI, Makarova KS, Yutin N, Koonin EV. 2011 Updated clusters of orthologous gene for Archaea: a complex ancestor of the Archaea and the byways of horizontal gene transfer. *Biol. Direct.* **7**, 46. (doi:10.1186/1745-6150-7-46)
45. Csuros M, Miklos I. 2009 Streamlining and large ancestral genomes in Archaea inferred with a phylogenetic birth-and-death model. *Mol. Biol. Evol.* **26**, 2087–2095. (doi:10.1093/molbev/msp123)
46. Raymann K, Brochier-Armanet C, Gribaldo S. 2015 The 2 domains tree of life is linked to a new root for the Archaea. *Proc. Natl Acad. Sci. USA* **112**, 6670–6675. (doi:10.1073/pnas.1420858112)
47. McShea DW, Hordijk W. 2013 Complexity by subtraction. *Evol. Biol.* **40**, 504–520. (doi:10.1007/s11692-013-9227-6)
48. Crick FHC. 1968 The origin of the genetic code. *J. Mol. Biol.* **38**, 367–379. (doi:10.1016/0022-2836(68)90392-6)
49. Wagner A. 2005 *Robustness and evolvability in living systems*. Princeton, NJ: Princeton University Press.
50. Knight RD, Freeland SJ, Landweber LF. 2001 Rewiring the keyboard: evolvability of the genetic code. *Nat. Rev. Genet.* **2**, 49–58. (doi:10.1038/35047500)
51. Itzkovitz S, Alon U. 2007 The genetic code is nearly optimal for allowing additional information within protein-coding sequences. *Genome Res.* **17**, 405–412. (doi:10.1101/gr.5987307)
52. Philip GK, Freeland SJ. 2011 Did evolution select a nonrandom 'alphabet' of amino acids? *Astrobiology* **11**, 235–240. (doi:10.1089/ast.2010.0567)
53. Elzanowski A, Ostell J, Leipe D, Soussov V. 2000 *The genetic codes*. Bethesda, MD: National Center for Biotechnology Information (NCBI). See http://www.ncbi.nlm.nih.gov/Taxonomy/Utils/wprintgc.cgi (accessed 25 April 2015).
54. Casjens S. 1998 The diverse and dynamic structure of bacterial genomes. *Annu. Rev. Genet.* **32**, 339–377. (doi:10.1146/annurev.genet.32.1.339)
55. Koonin EV, Wolf YI. 2008 Genomics of Bacteria and Archaea: the emerging dynamic view of the prokaryotic world. *Nucleic Acids Res* **36**, 6688–6719. (doi:10.1093/nar/gkn668)
56. Lawrence JG, Roth JR. 1996 Selfish operons: horizontal transfer may drive the evolution of gene clusters. *Genetics* **143**, 1842–1860.
57. Lathe III, WC, Snel B, Bork P. 2000 Gene context conservation of a higher order than operons. *Trends Biochem. Sci.* **25**, 474–479. (doi:10.1016/S0968-0004(00)01663-7)
58. Wang J, Dasgupta I, Fox GE. 2009 Many nonuniversal archaeal ribosomal proteins are found in conserved gene clusters. *Archaea* **2**, 241–251. (doi:10.1155/2009/971494)
59. Shoji S, Dambacher CM, Shajani Z, Williamson JR, Schultz PG. 2011 Systematic chromosomal deletion of bacterial ribosomal protein genes. *J. Mol. Biol.* **413**, 751–761. (doi:10.1016/j.jmb.2011.09.004)
60. Whitman WB. 2009 The modern concept of the procaryote. *J. Bacteriol.* **191**, 2000–2005. (doi:10.1128/JB.00962-08)
61. Baymann F, Schoepp-Cothenet B, Lebrun E, van Lis R, Nitschke W. 2012 Phylogeny of Rieske/cytb complexes with a special focus on haloarchaeal enzymes. *Genome Biol. Evol.* **4**, 832–841. (doi:10.1093/gbe/evs056)
62. Martin W, Koonin EV. 2006 A positive definition of prokaryotes. *Nature* **442**, 868. (doi:10.1038/442868c)
63. Nakagawa S, Niimura Y, Miura K, Gojobori T. 2010 Dynamic evolution of translation initiation mechanisms in prokaryotes. *Proc. Natl Acad. Sci. USA* **107**, 6382–6387. (doi:10.1073/pnas.1002036107)
64. Nakamoto T. 2009 Evolution and the universality of the mechanism of initiation of protein synthesis. *Gene* **432**, 1–6. (doi:10.1016/j.gene.2008.11.001)
65. Martin W, Koonin EV. 2006 Introns and the origin of nucleus-cytosol compartmentalization. *Nature* **440**, 41–45. (doi:10.1038/nature04531)

66. Doolittle WF. 2014 The trouble with group II introns. *Proc. Natl Acad. Sci. USA* **111**, 6536–6537. (doi:10.1073/pnas.1405174111)
67. Stark MR, Dunn EA, Dunn WS, Grisdale CJ, Daniele AR, Halstead MR, Fast NM, Rader SD. 2015 Dramatically reduced spliceosome in *Cyanidioschyzon merolae*. *Proc. Natl Acad. Sci. USA* **112**, E1191–E1200. (doi:10.1073/pnas.1416879112)
68. Irimia M, Roy SW. 2014 Origin of spliceosomal introns and alternative splicing. *Cold Spring Harbor Perspect. Biol.* **6**, a016071. (doi:10.1101/cshperspect.a016071)
69. Lane CE, van den Heuvek K, Kozera C, Curtis BA, Parsons BJ, Bowman S, Archilbald JM. 2007 Nucleomorph genome of *Hemiselmis andersenii* reveals complete intron loss and compaction as a driver of protein structure and function. *Proc. Natl Acad. Sci. USA* **104**, 19 908–19 913. (doi:10.1073/pnas.0707419104)
70. Cavalier-Smith T. 1991 Intron phylogeny: a new hypothesis. *Trends Genet.* **7**, 145–148. (doi:10.1016/0168-9525(91)90102-V)
71. Doolittle WF, Brown JR. 1994 Tempo, mode, the progenote and the universal root. *Proc. Natl Acad. Sci. USA* **91**, 6721–6728. (doi:10.1073/pnas.91.15.6721)