# Reconstructing the Last Common Ancestor: Epistemological and Empirical Challenges

Amadeo Estrada[1], Edna Suárez-Díaz[2], Arturo Becerra[2*]

1. Posgrado en Ciencias Biológicas, Universidad Nacional Autónoma de México

2. Facultad de Ciencias, Universidad Nacional Autónoma de México, Circuito Exterior Ciudad Universitaria, Coyoacán, DF 04510, México

*Corresponding author email: abb@ciencias.unam.mx

AB, ORCID ID:0000-0002-7076-0342

ESD, ORCID ID:0000-0003-2259-3110

## Abstract

Reconstructing the genetic traits of the Last Common Ancestor (LCA) and the Tree of Life (TOL) are two examples of the reaches of contemporary molecular phylogenetics. Nevertheless, the whole enterprise has led to paradoxical results. The presence of Lateral Gene Transfer poses epistemic and empirical challenges to meet these goals; the discussion around this subject has been enriched by arguments from philosophers and historians of science. At the same time, a few but influential research groups have aimed to reconstruct the LCA with rich-in-detail hypotheses and high-resolution gene catalogs and metabolic traits. We argue that LGT poses insurmountable challenges for detailed and rich in details reconstructions and propose, instead, a middle-ground position with the reconstruction of a *slim LCA* based on traits under strong pressures of Negative Natural Selection, and for the need of consilience with evidence from organismal biology and geochemistry. We defend a cautionary perspective that goes beyond the statistical analysis of gene similarities and assumes the broader consequences of evolving empirical data and epistemic pluralism in the reconstruction of early life.

**Keywords:** Last Common Ancestor (LCA), Horizontal Gene Transfer (HGT), Lateral Gene Transfer (LGT), Tree of Life (TOL), Network of Life (NOL), Reconstruction of Ancestral States.

**Declarations:**

Funding: Scholarship No. 460604 to A.E.  (CONACYT)

Conflicts of interest: The authors declare that there is no conflict of interest.

## Introduction

The reconstruction of the Last Common Ancestor (LCA) and the Tree of Life (TOL) are two burgeoning research areas of molecular phylogenetics that have led to paradoxical results, and important attention from philosophers and historians alike. On one hand, the presence of Lateral (or Horizontal) Gene Transfer (LGT, HGT) poses insurmountable epistemic and empirical challenges to meet these goals, since the heredity trace is blurred, as pointed out by Ford Doolittle (Doolittle 2000). On the other, a few prominent research groups have offered high-resolution reconstructions of the LCA -full of meticulous details of metabolic traits and gene catalogs- based on a statistical analysis of genomic data. In this paper, we offer a middle ground position: we argue that the reconstructions of a *slim* LCA and a *background* TOL are possible, and we highlight a number of underappreciated -both scientific and philosophical- arguments to support this view. Nevertheless, such a *slim* LCA is forcefully succinct and incomplete. Although incomplete, we believe these slim reconstructions are worthy and informative.

As early as 1993, Elena Hilario and Peter Gogarten argued that LGT, involved in ATPase genes, posed a serious challenge for the reconstruction of phylogenies, suggesting that a Network of Life (NOL) should replace the traditional representation of the Tree of Life. A few years later, Doolittle (1999a, 1999b, 2000) became one of the first molecular evolutionists to systematize the methodological challenges in the reconstruction of the three Domains of Life (Archaea, Bacteria and Eukarya), and the LCA. This makes an implicit point, in his early papers about the reconstruction of the LCA, and explicitly in 2000, even if the LCA were a concrete entity, from which afterward the three major groups, Domains, have shared by LGT an important part of their genes. He argued that the evidence that a large amount of LGT had taken place during the history of life made it methodologically impossible to retrieve a unique organismal phylogeny, or TOL, from the analysis of genetic/genomic data sets, and "that lateral transfer of genes between species, phyla or domains is so frequent that all reconstruction attempts are doomed to failure" (Doolittle 2000), concerning the

LCA. Indeed, even in a small rate's scenario, biologists agree that LGT has important phylogenetic consequences: its constancy, its presence even between Domains, the variety of its causes, and the lack of major impediments for it to take place, make it a pervasive evolutionary phenomenon that violates the assumption of vertical genetic inheritance at the basis of phylogenetic inferences. Although the LCA could be a well-defined entity (Velasco 2018), the amount of LGT that occurred since early life poses relevant methodological challenges for any kind of reconstruction. Further discussions on the subject of the LCA as a common ancestor and the common ancestry question, problematizations on the TOL, the repercussions of LGT upon the reconstruction of life history, and the separation of this problem from that of the LCA, are relevant distinctions (Velasco 2018). For Doolittle, the history of life is better represented by a network or a web of life. This idea has gained traction in the collaboration between biologists, historians, and philosophers of science, who have defended a pluralist view concerning the representation of life's history (Bapteste et al 2009, O'Malley 2010, O'Malley, Martin and Dupré 2010, O'Malley 2018).

Well-grounded concerns about the possibility of reconstructing the LCA (Doolittle 2000), or even about its existence (Woese 1998), however, have a counterpart from other practitioners in the field that have consistently attempted to do so. Based on statistical analyses of genomic sequences, a few well-known research groups have published richly detailed reconstructions of biochemical and other phenotypic traits ascribed to the LCA (Weiss et al. 2016a). In practice, these scientists have downplayed the role of LGT in challenging the preeminence of vertical heredity in biological evolution. The strictly statistical approach and a bold research strategy has resulted in divergent and even contradictory evolutionary hypotheses unsupported by independent evidence, between different research groups, and at times in single research groups. The LCA has been characterized 1. as close to the origin of life (Koonin 2003, Weiss et al. 2016a), or as being far away from the origin of life ( (Mirkin et al. 2003, Delaye et al. 2005)); 2. as having an small genome (Koonin 2003), or as having a genome similar in size to many free living bacterias today (Kyrpides et al. 1999, Harris et al. 2003, Mirkin et al. 2003, Delaye et al. 2005, Yang et al. 2005, Ouzounis et al. 2006, Ranea et al. 2006, Becerra et al. 2014); 3. as being autotrophic (Martin et al. 2008, Koonin and Martin 2005, Weiss et al. 2016a), or as being heterotrophic (Delaye et al. 2005, Becerra et al 2014, Muñoz et al 2018); 4. as being hyperthermophilic (Woese 1987, Weiss et al. 2016a); or as being Mesophilic (Galtier et al. 1999, Groussin et al. 2013, Cantine and Fournier 2018); 5. as constituted by a RNA genome (Mushegian and Koonin 1996, Koonin 2003), or as having a DNA genome (Ouzounis et al. 2006, Delaye et al. 2005, Becerra et al. 2014); 6. as being a simple cell (Koonin 2003), or as having a complex cell, similar to today's bacterias (Kyrpides et al. 1999, Harris et al. 2003, Mirkin et al. 2003, Delaye et al. 2005, Yang et al. 2005, Ouzounis et al. 2006, Ranea et al. 2006, Becerra et al. 2014). New findings and changes in what we think about certain subjects are common in science. Nevertheless, we think that these extreme divergences between and even inside some researchers' characterizations of the LCA are linked to the fact of relying on statistical approaches only without other kinds of data outside the sequence comparisons methods. In doing so, researchers can become subjects of contradictory algorithm results.

A middle-ground position is not only possible but also suggestive of future research. Robust reconstructions of the LCA require consilience between genetic data -without which there is no possibility of phylogenetic reconstruction-, organismal or cellular, and geochemical evidence. In this approach, Negative Natural Selection (NNS) plays a relevant role in identifying the few preserved genetic characters which, as we will argue below, should be compatible with a *background* TOL representing *cell lineages*. Our argument begins with a review of the twin consequences of gene-centrism for phylogenetic inferences: the epistemic fatalism surrounding the retrieval of the TOL and the LCA, and the optimist obliteration of non-genetic evidence in the search for biological ancestors (Section 1).[1] Section 2 introduces the research strategy adopted by those who propose rich-in-

[1] Eric Bapteste and John Dupré (2013) have called attention to the *standard model or entities ontology* and have argued in favor of a *processual microbial ontology*. Although their arguments are relevant for our proposal and for our emphasis on cellular processes, we have restricted ourselves to a tangential reference to their ideas.

phenotypic-details hypotheses for the LCA; we argue that this approach has serious epistemic disadvantages for the reconstruction of early forms of life, despite being rewarded in scientific practice. Our methodological approach is presented in Section 3, where we emphasize the role of NNS to identify highly preserved characters that provide a scarcely detailed but more robust reconstruction of the LCA. By requiring that evolutionary hypotheses be supported by independent non-genetic evidence (organismal, geochemical) we suggest that a robust LCA must be compatible with a *background* TOL that represents cell lineages. We thus arrive at a less epistemically pessimistic conclusion, far from the position of the impossibility of altogether reconstructing the LCA. Finally, we have added some reflections on the strengths and weaknesses of our approach in the concluding remarks, proposing that it provides a pluralistic research strategy that brings together advances in molecular, organismal (cell) biology and the earth sciences in the study of early life.

## 1. The gene-centric perspective and Doolittle's fatalism

Molecular databases and bioinformatics have revolutionized the practice of phylogenetics and the search for ancestral stages. Molecular phylogenetics has challenged the methods of taxonomy as well as some previously accepted phylogenies, and/or allowed inquiring into relatively unknown groups before such techniques were available (Suárez-Díaz and Anaya-Muñoz 2008, Suárez-Díaz 2014, Strasser 2019). One of the first major results of such developments in early evolution studies was the proposal by George E. Fox and Carl Woese of the three Domains of Life. By comparing sequences of the small subunit of ribosomal RNA of organisms over different clades, they concluded a monophyletic origin for the three major groups of life, which they defined as three lineages (Woese and Fox 1977). Since their publication, the automation of DNA sequencing, the growth in computability, and the development of statistical comparative methods have transformed the phylogenetic analysis of genes, entire genomes, and protein sequences. More recently, molecular evolutionists have proposed the existence of two -instead of three- ancestral main groups, Archaea and Bacteria (Zhou et al. 2018), and have argued that the LCA had a well-defined differentiation between genotype and phenotype (Delaye et al. 2005, Becerra et al. 2007).

Many challenges, however, remain in the study of early life. In 1999 Ford Doolittle introduced the formal discussion on the methodological and empirical obstacles plaguing the quest of molecular phylogeneticists for the true TOL. Mutational saturation and artifacts related to within-molecule and between-lineages differences in evolutionary rates could be misleading about deep branching and the rooting of the tree (Doolittle 1999a, p. 2125). Perhaps more importantly, Doolittle pointed out two consequences of the observation that many genes give believably different phylogenies for the same organisms, almost certainly because they have been 'laterally transferred', namely: (i) that it is not logical to equate gene phylogeny and organismal or cellular phylogenies, and (ii) that unless organisms are construed as either less or more than the sum of their genes, there is no unique organismal phylogeny. Thus, there is a problem with the very conceptual basis of phylogenetic classification (Doolittle 1999a, p. 2125).

Doolittle's claims are compelling and justified for a gene-centric view of organismal evolution, and they translate into a strong argument against the representation of phylogenetic relations as a unique TOL. Moreover, Doolittle has argued that the LGT also obscures the retrieval of ancestral characters present in the LCA (Doolittle 2000). Thus, Doolittle's alternative was to suggest an entangled representation of the three main Domains, with gene exchanges showing rich interrelation at several stages. This view of an entangled web has justifiably captured the imagination of philosophers and students of biology alike, who have received it as a transformative representation of evolution that builds upon and goes beyond the Darwinian tree of life (Bapteste et al. 2009, Bapteste and Dupré 2013, O'Malley 2018). Doolittle, indeed, has joined the call for the scientific community to participate in philosophical inquiries and debates, and argued for a pluralistic approach that takes into consideration the different scales at which evolution takes place:

> This is not to say that similarities and differences between organisms are not to be accounted for by evolutionary mechanisms, but descent with modification is only one of these mechanisms, and a single treelike pattern is not the necessary (or expected) result of their collective operation. Pattern pluralism (the recognition that different evolutionary models and representations of relationships will be appropriate, and true, for different taxa or at different scales or for different purposes) is an attractive alternative to the quixotic pursuit of a single true TOL (Doolittle and Bapteste 2007, p. 2043).[2]

Doolittle's fatalism concerning the TOL clearly targets a gene-based tree. In a rarely cited statement, he argues that the TOL does not represent the history of life, which is better portrayed by a network that represents *cellular*, not *genomic*, phylogenies:

> That is, the donors for many LGT events would have been *cellular lineages* that have since gone extinct, because LUCA [Last Universal Common Ancestor] is by this and the previous conceptualization the last universal common cellular (not necessarily genomic) ancestor (Doolittle and Brunet 2016, our emphasis).

Doolittle has warned about the high importance of LGT and its repercussions in TOL and LCA reconstructions (Doolittle 2000), and although his works have been widely cited and discussed, many practitioners in the field have downplayed the importance of his claims, in order to continue with such reconstructions; the ensuing argumentations have been centered on the amount of LGT that has

[2] On the question of pluralism, also, "[a]universal Tree of Life (TOL) has long been a goal of molecular phylogeneticists, but reticulation at the level of genes and possibly at the levels of cells and species renders any simple interpretation of such a TOL, especially as applied to prokaryotes, problematic." (Doolittle and Brunet 2016, p. 1).

taken place through the history of life. The reconstruction optimism has posed LGT as a much rarer phenomenon than what Doolittle discusses (for example, Theobald 2010 and Puigbò et al. 2013). In the day-to-day practice of molecular phylogeneticists, many have continued with the publication of detailed genetic reconstructions, an approach doomed to discern between cellular and genetic lineages. For Doolittle, it is the cellular lineages that provide the basis to infer the properties of the LCA and to delineate the pattern of evolution -a network, on his account. More pointedly, as philosophers and biologists have argued, cell and organismal lineages are populations characterized by a diversity of reproductive forms that go beyond sexual reproduction (Sterner 2017).

In the context of all these conceptual nuances, the last decade has seen molecular evolutionists producing growing evidence of LGT as a common event in prokaryotic evolution (McInerney et al. 2008; Sapp 2009; see references in O'Malley 2018). On the other hand, attempts to quantify LGT have been made to assess the feasibility of characterizing past life (Puigbò et al. 2013; Gogarten et al. 2002; Huang and Gogarten, 2006). The problem is that there are major difficulties to measure LGT, not the least because the statistical criteria and bioinformatic tools used to estimate it share the same methodological constraints that plague phylogenetic reconstructions (Cortez et al. 2009).

Less attention has been given to the implications of LGT beyond molecular phylogenetics. Often, the scientific literature includes statements that assume that only genes provide data for reconstructing the TOL (Koonin 2003; Weiss et al. 2016a; Kim and Caetano-Anollés 2011; Kim and Caetano-Anollés 2012). Though Doolittle has called for an in-depth knowledge and understanding of biological processes, his warning -as we will show in the next section-, has not been generally attended to. Beneath the molecular and statistical methods of molecular phylogenetics, the complicated phenotypic aspects of organismal biology and the geochemical constraints acting upon cellular and organismal populations are overlooked by a gene-centric perspective. This approach, however, is nowadays dominant in scientific practice (Hagen 1999, 2003, Morgan 1998, Suárez-Díaz 2014). Phenotypic traits are nevertheless diagnostic to each of the main Domains at the cellular level, and those characters may not be traceable with genetics or genomics. Thus, for instance, the presence of the nucleus, of membrane differences, multicellularity, and the presence of endosymbionts, as well as differences in the tertiary structure of proteins, are all well-established taxonomic traits where the genetic sequence signal has been either lost or still scarcely known in its full genetic complexity (Russell et al. 1997, Yang and Honig 2000).

## 2. The gene-centric approach and reconstruction's optimism

As we mentioned before, LGT cannot be reliably measured with precision -because of the same methodological challenges that plague phylogenetic reconstructions. Nevertheless, we know that this is a pervasively common phenomenon in the history of life, taking place even between Domains, of which some estimates have been given. LGT has been assessed as a common event in Bacteria and in Archaea (for instance, it has been evaluated in experimental settings), more abundant in these than in Eukarya (Husnik and McCutcheon, 2018). In Eukarya, also, LGT is a constant event produced mainly by viruses and by bacteria that has been going on for enough time -2,5 billion years- to produce important evolutionary results (Keeling and Palmer 2008).

Doolittle's methodological concerns regarding the limits of the gene-centric approach pose a serious warning to research in molecular phylogenetics. Several research teams, however, have concluded that the relevance and profusion of LGT does not preclude attempts to reconstruct the LCA (Huang and Gogarten 2006, Puigbò et al. 2013), and still others have been more extreme in their disregard of warnings about the gene-centric view. In practice, this has meant the publication of several rich-in-detail hypotheses on the phenotypic traits of a theoretical construct, the LCA (a *cell population* that is postulated as the common ancestor to all living beings).

Possibly one of the most detailed gene-centric characterizations of the LCA is the one advanced by Madeline C. Weiss and her colleagues at William F. Martin's lab at the Institute of Molecular Evolution in Düsseldorf (Weiss et al 2016a). Peter Gogarten and David Deamer (2016) wrote a response to their original publication, and an ensuing debate took place in the pages of *Nature* (Weiss et al. 2016b). Gogarten and Deamer's critique went from pointing out basic conceptual misunderstandings -conflating the origin of life with the LCA-, to questioning their methodological approach, all the way up to a lack of empirical evidence for the proposed phenotypic traits of the LCA. They pointed out that there is relevant genetic data that supports the notion that the LCA lived during a period as late as one hundred to two hundred million years after the origin of life. As the theoretical ancestor of current biodiversity, the LCA must have had both a transcriptional and a replicational apparatus; it must have had also the transmembrane proteins which are present today in all the major Domains; this set of properties, Gogarten and Deamer claim, hardly counts as a "half-alive" -as characterized by Weiss et al.- *progenote*, and it is not consistent with what biologists nowadays know about the origin of life (Delaye et al. 2005, Becerra et al. 2007, Goldman et al. 2013).[3] The point we want to make from this important debate is the scarce reliability of detailed reconstructions based on a strictly genetic analysis.[4]

Methodologically, Gogarten and Deamer claim that a computational approach is subject to false positive and false negative errors. Indeed, all ancestral state reconstructions run the risk of producing one of those two types of errors.[5] The inference of traits for the LCA is not an exception. False

[3] A progenote is "a hypothetical biological entity in which phenotype and genotype had an imprecise, rudimentary linkage relationship" (Woese & Fox 1977).

[4] By locating 355 protein families, Weiss et al. inferred a set of traits, including that the LCA (which equivocally for the authors is the same as the progenote) was anaerobic, $CO_2$ fixing, $H_2$ dependent with a Wood–Ljungdahl pathway, $N_2$-fixing and thermophilic, among many other biochemical details. What drew the most pointed critique, however, was the speculation that this organism consisted of a cell-sized compartment bound by a mineral membrane, a hypothesis that revealed the lack of familiarity with current advances in the biochemistry of lipids and on the geochemistry of hydrothermal vents (the supposed environment of this organism). This membrane had an "underlying lipid bilayer added to provide a permeability barrier to ions such as $Na^+$ and $H^+$". But, as Gogarten and Deamer point out, an obvious question was the source of the lipid forming the membrane, and the process by which it would adhere to the mineral surface.

To this, they added a critique of the supposed ATP-ase and sodium–proton exchange mechanism embedded in the bilayer, finally concluding that "these biochemical systems and catalysts are characteristic of an advanced form of life having ribosomes, translation, genes and a genetic code, far beyond what most would imagine as the first form of life" (Gogarten and Deamer 2016).

[5] The complete methodological critique goes like this: "Weiss et al. propose a computational scheme that provides a shortcut to identify genes that may have been present in LUCA. This approach is subject to two types of error: False positives. The stated criterion for inclusion in the LUCA gene set is that the gene needs to be

positives take place when a broad distribution of a given gene or genes is the result of LGT, and not of vertical heredity from the ancestor. Statistical analyses might indicate that a character was also present in the ancestor. To discern among those false positives, scientists need to check them against independent sorts of evidence: cellular (organismal), biochemical, and/or geochemical data. All of those are data produced in fields of knowledge which are independent of molecular phylogenetics and of the gene-centric perspective. Metabolic pathways, for instance, are known to be prone to LGT, and so the more a research team relies on the statistical analysis of functionally related genes, the more false positives will result (Cohen et al. 2011).

By contrast, false negatives result when a trait was present in the ancestor, but was lost in the descendants, erasing the actual genetic signal. This may lead scientists astray to conclude that a certain character was absent in the ancestor, when in fact it was present and, as in Gogarten and Deamer's (2016) critique, this type of error results in a gene set that is strongly biased toward genes that have a limited distribution in today's organisms. Again, this is an outcome that can only be discernible and corrected with the use of independent empirical evidence. To emphasize: gene-centered strategies do not produce false positives and false negatives *per se*, but the statistical tools of gene analyses cannot provide scientists the necessary information to distinguish between hypotheses that are, for instance, incompatible with the primitive Earth's conditions, the fossil record, or other biological data.

present in two archaeal and two bacterial groups. From the presented tables, it is clear that orders are considered as distinct groups. The criterion identifies a gene as present in LUCA if a single transdomain transfer occurred before the two 'groups' (that is, orders) in the receiving domain split, or if the transferred gene was subsequently transferred between the two groups. Given that gene transfer within domains occurs more frequently than transfer between domains, a large number of false positives are expected under the implemented scheme. False negatives, which are likely to be an even bigger problem. The authors correctly assume that gene transfer between the domains has occurred. Consequently, many genes that were present in LUCA will not be inferred as present. ATP synthases and aminoacyl-transfer RNA (tRNA) synthetases illustrate this point, because only one out of at least five ATP-synthase subunits was part of the inferred LUCA set, and only eight aminoacyl-tRNA synthetases, whereas LUCA appears to have used the full complement of today's genetically encoded amino acids, and had a functional ATPase/ATP synthase. The consequence of these errors is that the inferred LUCA gene set is strongly biased toward genes that have a limited distribution and utility in today's organisms" (Gogarten and Deamer 2016, p.1).

Some highly productive research groups in the buoyant field of molecular phylogenetics, have aimed to reconstruct the phylogeny of organisms exclusively based on quantitative analyses of sequences. This kind of research has gained visibility beyond biology, reaching general audiences and luring students of science. Their conclusions have been proved to be unwarranted and have often led to *contradictory hypotheses* even within the same team and in consecutive publications, with no recognition of their divergent conclusions. In our view, the exclusively gene-centric approach is oblivious to the diversity of evolutionary mechanisms and to the power of independent evidence from independent data sources. More examples of this reductionist approach -some recent, and some older ones- will help to further illustrate our point.

In 1996, Arcady R. Mushegian and Eugene V. Koonin, from the National Center for Biotechnology Information (NCBI), argued that the LCA should have had a very small genome. They supported their claims on the first sequenced genomes from *Haemophilus influenzae* and *Mycoplasma genitalium*, two pathogenic but widely divergent bacteria. Their conclusions obviated a well-studied biological phenomenon: because of genetic redundancy with their hosts, pathogens, and endosymbionts are characterized by gene loss. Instead, the authors argued that if these extant living forms could survive with such a small genome, then the LCA would have had a similar size genome (Mushegian and Koonin 1996). As others soon noticed, from a strictly biological perspective, those minimal genomes are exclusively found in pathogenic bacteria (Becerra et al. 1997). Organismal evidence, thus, promptly helps to illuminate the challenges faced by hypotheses resulting from a strictly gene-centric statistical approach.

The amount of phenotypic detail included in such hypotheses also provides plenty of specific and speculative LCA phenotypic traits, including a high definition of metabolic pathways. This prolific detail, retrieved from quantitative analysis of sequences and scarcely supported by other types of evidence, has been qualified as “cherry-picking” by other scholars (Gogarten and Deamer 2016).

Examples abound of the recurrent practice of molecular phylogeneticists to ascribe metabolic traits and gene catalogs unsupported by independent data to their reconstruction of the LCA. Mirkin et al. 2003, Koonin 2003, Yang et al. 2005, Sobolevsky and Trifonov 2006, Ouzounis et al. 2006, Ranea et al. 2006, Kim and Caetano-Anollés 2011, and Weiss et al. 2016a are only a few of the many publications that attempt to provide highly specific phenotypic details based exclusively on computational genetic analyses.

The appeal of the statistical frame of mind nevertheless persists. In 2003, Mirkin et al. with Koonin as coauthor, argued for a LCA of about 600 COG's (Clusters of Orthologous Groups); this would constitute a large genome and would have been located in a complex cellular structure. A few months later, Koonin (2003), again defended the view that the LCA might have had a minimum genome of about 63 gene sequences and should have been a simple cell form as they suggested in 1996 (Mushegian and Koonin 1996). Our aim here is not to point out that contradictory results -or inconsistencies- may justifiably come from any team's results, but to underline a problem and advance a suggestion. Producing new hypotheses, even contradictory to previous ones, is a legitimate occurrence in scientific practice, and it may respond to changes in technological capacities, or in new available procedures. But our focus here is on the lack of independent data to discriminate between plausible hypotheses, some of them frankly incompatible with geological or cellular evidence.

In brief, overlooking empirical restrictions beyond the analyses of genetic information has led to unwarranted conclusions on the nature of the LCA. In fact, *there are many things the LCA could not have been*. Any hypothesis of the LCA must be confronted with current empirical knowledge from the Earth sciences, as well as what scientists know about biochemistry and metabolic pathways; conversely, the scarcity of biochemical and geochemical knowledge surrounding the early stages of life poses a severe epistemic constraint upon the understanding of the LCA. In this scenario, phylogenetic reconstructions aiming at a lower resolution -a LCA *grosso modo*, which is, paradoxically, a *slimmer* LCA- but compatible with independent kinds of data, have important epistemic advantages. The most relevant of these is the robustness of findings from consilience, which significantly reduces the possibility of false negatives and false positives. The seeming "completeness" of reconstructions offered by practitioners of the gene-centric approach is attractive. More often than not, however, these reconstructions lack the robustness required of confirmed scientific hypotheses. Speculation in ancestral reconstruction is expected and valued; it is also something achievable with today's computability capacities. In the next section, we argue in favor of a pluralistic toolkit that can reduce the number of speculative hypotheses and can produce more robust results.

## 3. Negative Natural Selection and robustness in a *slimmer* LCA and a background TOL

Negative Natural Selection (NNS) fixates and maintains adaptive biological traits, provides long-term stability, and removes deleterious mutations -more common than beneficial ones. Therefore, NNS has also been called *purifying selection* or *background selection*, as it eliminates less-adapted variants, allowing better-adapted variants to increase their population frequencies. L. Loewe in 2008 expresses this idea succinctly:

> Because more DNA changes are harmful than beneficial, negative selection plays an essential role in maintaining the long-term stability of biological structures by removing deleterious mutations. Thus, negative selection is sometimes also called purifying selection or background selection. One key reason why this form of selection is so prevalent is the success of evolution in optimizing biological structures. [...] The main consequence of negative selection is the extinction of less-adapted variants. If the best-adapted variant does not change because it is at a

> stable local optimum, then negative selection will remove all new variants for that optimal trait (Loewe 2008).

As we argued in section two, extremely detailed reconstructions lead to non-robust hypotheses, but we also find that some reconstructions are possible and useful. For the purpose of reconstructing a slim LCA as well as a background TOL, the practice of locating and choosing highly preserved characters under NNS is still the most useful and relevant strategy. Traits linked to basic cellular functions and reproduction, with which scientists could hypothetically infer the gene catalogs that describe the properties of the LCA, are subject to strong NNS (Delaye et al 2005, Theobald 2010, Goldman et al 2013). Rates of NNS vary among different traits because the loss in fitness depends on the negative repercussions on the functionality of the trait in question. Consequently, a reduction in metabolic tolerance to a substance can convey a reduction in fitness in a living form, while another trait related to a basal function for life, like a mutation in a ribosomal RNA, can easily result in the impossibility to perform cell translation altogether. Therefore, NNS rates on genes codifying basic biological (reproductive) functions are stronger in general than on many other metabolic functions. It follows that NNS would be a strong pressure in a one copy gene scenario if such gene is of survival or adaptive relevance. Therefore the choice of ribosomal RNA in Woese and Fox's original work in 1977 was so informative; this molecule is very similar among different Domains, which explains its success as a reliable marker for phylogenetic inferences, of which two are central to our interests in this paper: an ancestor to today's biodiversity and a basic TOL. Although this paper does not aim to produce a LCA characterization and genetic catalog but to problematize the characterization process, we want to point out that we conceptualize the LCA as a population of cells. The number of basal traits of life that are common to all biodiversity indicates that the LCA could have been made of a complex cell and, therefore, distant from the origin of life. Based on that, it would be difficult to sustain a small population or, even more difficult, a single cell that originated modern biodiversity at this stage in life history.

Biological information is lost with increased temporal distance. The older the organisms, the more difficult it becomes to build hypotheses as to their traits and phylogenetic relations. Elliott Sober first described this problem as an *information destroying* process (Sober 1988), which is mainly due to the effect of weak NNS over genes. Metabolic pathways, by contrast, are broadly diversified along evolutionary patterns, largely because of LGT (Husnik and McCutcheon 2018). At the same time, there are many different pathways to arrive at similar metabolic results (Hügler and Sievert, 2011; Kleiner et al. 2012). Gaining, or even losing metabolic pathways, is always possible by means of LGT or other evolutionary processes (Cohen et al. 2011). Moreover, gaining new genes and metabolic pathways can provide adaptive characters without putting survival at risk. This is the process that allows biologists to explore how LGT-acquired-genes were established in a different clade, and even among different Domains. The presence or absence of NNS provides, thus, the mechanism to explain why some traits vary rapidly and broadly, as it happens with neutral sequences where NNS is absent, and it also explains that sequences linked to highly adaptive traits are more consistent and similar between different clades -in the presence of strong NNS. Because of LGT, Doolittle has argued on the impossibility to retrieve a genetic LCA to root the TOL, and on the impediments to reconstruct a TOL (he argues in favor of a Network of Life instead). This is relevant for most metabolic processes -which also account for most of the LGT (Cohen et al. 2011). Metabolism-linked genes have been established to be frequently and abundantly exchanged even between different Domains, even though not every gene transferred by LGT is fixed (Doolittle 1999a, with Bapteste 2007, with Brunet 2016; Cohen et al. 2011).

NNS fixes and preserves traits, some of which are widely distributed because of vertical inheritance, although LGT can also produce a wide distribution of characters. In fact, NNS is a powerful tool for identifying conserved traits even if the genetic sequence similitude has been lost. This happens with many proteins, which conserve their tertiary structure but have lost the “original” genetic signal, showing differences in sequences (Russell et al. 1997). Recognition of this phenomenon has grown in the past few years and is being complemented with studies of the effect of NNS on these proteins (Jácome et al. 2015). So, NNS is not only useful for preserving gene sequences, but also phenotypic traits. In some other cases, there are different functions related to the same genetic sequence. LGT can produce a wide distribution of a given gene among different clades and, in a short amount of time, even new functions. A high rate of LGT, for instance, can produce a wide distribution of a gene among different Domains very rapidly and lead scientists to conclude it as a conserved trait (a false positive), when in fact the gene may be of recent acquisition but widely spread. This is a scenario where consilience with other data is welcomed. NNS produces conservation of adaptive traits in vertical inheritance, even if these could have been acquired initially by LGT. Hence the difficulties of inferring traits based only on sequence comparisons, and the utility of NNS.

NNS has been proved a very useful tool in phylogenetic reconstructions, nevertheless the choice of highly conserved characters -as a research strategy- has its own weaknesses. These include:

1. There is substantially big data of molecular genetics for inferring early life but, by contrast, scientists still have scarce information on phenotypic traits of early life.
2. It is methodologically challenging to distinguish between genes conserved by NNS, from those which were widely distributed by LGT, a circumstance that, as we mentioned before, is highly conducive to false positives.
3. There are few genetic sequences that point to the presence of strong NNS, namely, informational traits and basic metabolic pathways. This leaves biologists with a much smaller data set than what is available in international big data banks.
4. Our methodological suggestion limits the details biologists aim to retrieve about ancient organisms such as the LCA.
5. Conserved traits may fail to distinguish between the last common ancestor and older common ancestors; especially when there are long stretches of time with few extant phylogenetic branches to sample, as is the case for the LCA of today's three domains.

Nevertheless, the advantage of the proposed strategy surpasses its limitations. Although there is scarce phenotypic information, scientists have enough data today to produce a few well supported hypotheses. This is a growing research field, including for instance the study of the evolution of tertiary protein structures, where analyses that consider NNS seem to be a promising field (DeepMind, News in Focus, Nature 2020:588). On point two, above, phylogenies are also an important tool to identify LGT and to distinguish those traits from others which are under the effect of NNS. A slimmer reconstruction of the LCA provides the advantage of more robust hypotheses. Although it limits the details of the scientists' reconstructions, NNS is a parameter that helps to choose between relevant traits and counterbalances the data from sequence comparisons.

The conclusion is that a *slim* characterization of the LCA, although less detailed and based on traits under strong pressures of NNS, is, in general, more robust and testable than a high-resolution catalog of metabolic genes. It is not free, however, of methodological challenges, as we pointed out above. Characterizing the LCA as a cell population provides a fruitful counterpoint to gene-centrism and exclusive statistical analyses for the better understanding of processes and context of early life evolution and the possibilities of diversification through evolutionary mechanisms. This is true even if all phylogenetic inferences originate from genomic data, which is necessary but not sufficient in this endeavor. Genomic data point towards both, a TOL and a NOL, depending on which data are sampled, although cellular data clearly indicate a TOL. The "complete" reconstruction of the LCA, in terms of its metabolism and environment, does not need to be a solvable question once and for all. The LCA, as all ancestral reconstructions and phylogenies (including the TOL and the NOL), is a working hypothesis for representing the history of life and it is mostly useful as a research pathway or heuristic. Perhaps counterintuitively, in this view, a *slimmer* (*grosso modo*) characterization of the LCA becomes more useful and robust, as other sources of independent evidence contribute to its reconstruction. Moreover, the information of a background TOL may be useful in reconstructions of different stages of ancestral life, and becomes even more instrumental at this point than the NOL.

So far, the study of the major taxonomic groups of life reveals that -LGT notwithstanding- main clades such as Mammalia or Amphibia, just as Gymnosperms and Angiosperms, for example, are very well differentiated, internally coherent and cohesive branches. A similar process happens with the main Domains of life: they do share genes, but these two/three groups are clearly diagnosable and differentiable from a cellular or organismic point of view. Some few traits that are under the strongest pressures of NNS, and therefore can be used for reconstructions, consistent with Woese's original three Domains, are the cell membrane, the transcription and translation apparatuses, glycolysis, glutamic acid, lipids, pyrimidines and purines catabolism and anabolism (Rivas et al. 2018). In translation, for example, ribosomal proteins, elongation factors, aminoacyl-tRNA synthetases among others, are strongly conserved (Dealye et al 2005). In transcription, RNA polymerase Beta (Delaye et al 2005) is also highly conserved, as well as the palm domain of the DNA and RNA polymerases (Jácome et al 2015). This list can be enlarged as new molecular procedures and data sources indicate relevant information about homologies in protein tertiary structure, although the amino acid sequence has been lost. All of these sets of traits have a distinct pattern of extremely conserved sequence or structure consistent with NNS. Other research groups working on the reconstruction of the LCA hypothesis have reached different results, though there is a major consensus that genes involved in translation and transcription, in DNA replication and repair, in proteins associated to membrane structure, as well as nucleotide and sugar metabolism (Becerra et al. 2007) are all basal traits which are useful to infer a *slim* LCA. All those traits are consistently similar among all forms of life, although they might be insufficient to infer the detailed metabolic pathways of the LCA offered.

We certainly agree with Doolittle that the reconstruction of both the LCA and the TOL face major problems because of LGT, simply because LGT is a major impediment for the reconstruction of vertical or phylogenetic relations (Doolittle 1999a, 1999b, 2009, Fournier et al. 2015). With a gene-centric approach the TOL is blurred and becomes an intricate web. This happens because the TOL represents gene distribution, not life history at the level of cell (or organismal) populations. But also, there is a growing body of evidence that suggests the feasibility of a *genetic background tree* standing behind the NOL, and the desirability of doing so. Cohen et al. (2011, p. 1485) arrive at similar conclusions:

> Others and we have previously shown that the biological function of a gene family is important in determining its propensity to undergo HGT. In agreement with previous studies, the lowest HGT levels are observed for the informational genes (involved in transcription and translation), where the most pronounced trend was found in genes associated with the ribosome and related with translation (COG functional category: ‘‘translation, ribosomal structure, and biogenesis’’).

This means that it is plausible that a superposed web of life obscures statistically robust Trees of Life that interconnect the three main Domains. The three major groups of life are consistent from the perspective of organisms and cells, a result which does not seem to be reachable if we base our reconstructions on only genetic, genomic, or molecular comparisons. What we call a *background tree* is a TOL distinguishable through the web of genes because it can be reconstructed from a limited number of genetic traits (just like our *slim* LCA). It means that a TOL is represented with a lower resolution than a NOL, but it is historically informative. This option must not be seen as a desperate “tree-rescue strategy”, an attempt to separate the “wheat” of vertical inheritance, from the “chaff” of LGT. Core genes, as scientists and philosophers are ready to argue, will only deliver *a history of the genes, not of the cells -or species, or organisms* (Bapteste et al. 2009, p.8):

> To maintain that the history of the core genes “represents” the species history requires some argument that the history of these parts is somehow “essential” to a species’ genealogy. But post-Darwinian biologists are generally loath to attribute any special essentialist status to either genes or species. If they fail to essentialize (which should be expected), then any such core-gene tree, which might well be an

> interesting and at times scientifically fruitful representation, cannot be considered to represent the species history (ibid p. 9).

Thus, the background tree, in their conception, does not aim to retrieve the history of bacterial *species.* In our view, this tree is to be understood as a heuristic tool which provides crucial but not complete information about the LCA.

Due to NNS, the divergence between traits involved in informational and other functions is less than among clades (Cohen et al 2011, Rivas et al. 2018), and so they provide valuable information to reconstruct a background TOL. Some of these traits also include the translation system present in all life forms -although there are some punctual differences between Archaea and Bacteria-, and a genetic code that is almost identical in all major groups. These few differences provide valuable evidence in favor of a LCA population, with the help of a background TOL, which could be represented as standing "behind" the genetic NOL. The number of traits that might be used for those reconstructions may grow with new empirical evidence and genetic data, but the important point here is to emphasize the need for independent evidence, besides gene genealogies, to support hypotheses about the processes that trace back to the LCA population.

## Concluding remarks

LGT poses an unsurpassable constraint in the reconstruction of a genetic TOL and of the LCA; somehow contradictorily, several research groups have attempted to reconstruct the last common ancestor in great detail. We have argued that sequence comparisons provide necessary but not sufficient evidence for the hypothetical traits that made up a background or statistically relevant TOL, nor for the LCA's genetic catalog. The importance of genes as historically informative is a central tenet in evolutionary biology and in the practice of phylogenetic inferences, but the gene-centric view of biological processes has its limits in the reconstruction of cell lineages of the LCA. The background or statistically relevant TOL, and its use as a tool to reconstruct a slim LCA, are research programs in need of an organismal and processual perspective that complements molecular studies; as we have argued, lack of independent evidence from other disciplines in favor of statistical data analysis not only leads to false negative and false positive results, but also to the never-ending search for better statistical methods, a methodological anxiety that has characterized the field since its early days (Suárez-Díaz and Anaya-Muñoz 2008).

The incorporation of non-genetic empirical evidence in phylogenetic reconstructions results in a slimmer characterization of the LCA based on traits under strong pressures of NNS. Such reconstruction is potentially more robust, although much less detailed and would provide less information than what the merely genetic analyses produce. The LCA does not need to be an empirically solvable question to increase our knowledge of early life, but an investigative pathway (Holmes 2004), or heuristic from which we can postulate hypotheses in order to solve other biological inquiries. Our position is not only theoretical, but it suggests a research pathway. A number of research questions that would be productive for a research program on the background tree we are proposing are: a) a multi- and interdisciplinary effort to look for consilience with macroevolutionary trends and clades, with which NNS is expected to be compatible if our proposal holds true; and b) if our hypothesis of a background tree is correct, phylogenetic and comparative genome research on the traits now included in section 3 –cell membrane, the transcription and translation apparatuses, glycolysis, glutamic acid, lipids, pyrimidines and purines catabolism and anabolism (Rivas et al. 2018), ribosomal proteins, elongation factors, aminoacyl-tRNA synthetases, RNA polymerase Beta (Delaye et al 2005), as well as the palm domain of the DNA and RNA polymerases (Jácome et al 2015)–, should be inconsistent with a NOL, but compatible with a number of possible background trees.

The scientific debate regarding which model best represents the history of life, a tree or a network, has benefited greatly from the deserved interest of philosophers and historians of biology; simultaneously, the biological consensus that the evolutionary principle of descent with modification is explained by several mechanisms that result in a diversity of historical patterns (trees, networks, webs) has provided empirical stimulus to philosophical pluralism in many ways. John Dupré is among those whose work on the contrast between monistic and pluralistic approaches to evolutionary patterns and mechanisms has influenced the current biological debate, and rightly so (Dupré 2003). Others, such as Maureen O'Malley, have looked to recent biological research to argue for a more diversified view of life, one that seeks to correct the bias of Eukarya in the philosophy of science (O'Malley et al. 2019) and include Bacteria as more than exceptions (O'Malley 2013). Eric Bapteste and Dupré (2013) have taken a step further, arguing for a processual ontology that is not committed to a hierarchical ontology that depends on vertical inheritance and thus privileges genealogical relations. It is in the context of hierarchical ontology that biologists may wrongly assume that by reconstructing genetic genealogies they are describing the history of species or of cell lineages.

In this paper, we have arrived at similar conclusions, coming from a different place. Genetic genealogies cannot provide the history of species or cell lineages, even if they -among other types of evidence- are indispensable for understanding the historical processes through which entities are stabilized (Bapteste and Dupré 2013, p. 399). The evolution of the last ancestral population of microbial cells is then the updated version of a fundamental question: what can be said of the last common ancestor to all living organisms? Although the question in itself might not be solved because of the lack of substantial empirical evidence, characterizing the LCA is a legitimate endeavor. This remains a productive research pathway for evolutionary biology, with which we might better understand the entangled tree of life.

**Acknowledgements:** A. Estrada thanks the Posgrado en Ciencias Biológicas at the Universidad Nacional Autónoma de México, as well as Consejo Nacional de Ciencia y Tecnología (CONACYT) for their support with fellowship No. 460604. I also thank Professor Daniel Piñero for his advice. We also acknowledge the invaluable help of Vivette García Deister and Luis Delaye for their insight on this paper.

## References

Bapteste E. Dupré J (2013) Towards a processual microbial ontology. Biology and Philosophy, 28, 379–404 https://doi.org/10.1007/s10539-012-9350-2

Bapteste E, O'Malley MA, Beiko R et al. (2009) Prokaryotic evolution and the tree of life are two different things. Biology Direct 4, 34. doi:10.1186/1745-6150-4-34

Becerra A, Islas S, Leguina JI et al. (1997) Polyphyletic gene losses can bias backtrack characterizations of the cenancestor. Journal of Molecular Evolution, 45, 115-117.

Becerra A, Delaye L, Islas S et al. (2007) The Very Early Stages of Biological Evolution and the Nature of the Last Common Ancestor of the Three Major Cell Domains. Annual Review of Ecology, Evolution, and Systematics, 38, 361-379.

Becerra A, Rivas M, García-Ferris C et al (2014) A phylogenetic approach to the early evolution of autotrophy: the case of the reverse TCA and the reductive acetyl-CoA pathways. Int Microbiol. doi: 10.2436/20.1501.01.211.

Cantine MD, Fournier GP (2018). Environmental Adaptation from the Origin of Life to the Last Universal Common Ancestor. Orig Life Evol Biosph. doi: 10.1007/s11084-017-9542-5.

Cohen O, Gophna U, Pupko T (2011) The Complexity Hypothesis Revisited:

Connectivity Rather Than Function Constitutes a Barrier to Horizontal Gene Transfer.

Molecular Biology and Evolution, 28, 1481-1489.

Cortez D, Delaye L, Lazcano A et al. (2009) Composition-Based Methods to Identify Horizontal Gene Transfer. Methods in Molecular Biology, vol 532. Humana Press

Delaye L, Becerra A, Lazcano A (2005) The Last Common Ancestor: What's in a name? Origins of Life and Evolution of Biospheres, 35, 6, 537-554.

Doolittle WF (1999a) Phylogenetic Classification and the Universal Tree. Science, 284, 5423, 2124-2128.

Doolittle WF (1999b) Lateral genomics. Trends in Cell Biology, 9:M5-M8.

Doolittle, W.F. (2000) Searching for the common ancestor. Res. Microbiol. 151, 85–89

Doolittle WF, Bapteste E (2007) Pattern pluralism and the Tree of Life hypothesis. Proceedings of the National Academy of Sciences of the United States of America, 104, 2043-2049.

Doolittle WF (2009) The practice of classification and the theory of evolution, and what the demise of Charles Darwin's tree of life hypothesis means for both of them. Philosophical Transactions Royal Society of London Series B Biological Sciences, 1527, 2221-2228.

Doolittle WF, Brunet T (2016) What Is the Tree of Life? PLoS Genetic 12, : e1005912. https://doi.org/10.1371/journal.pgen.1005912

Dupré J (2003) Human Nature and the Limits of Science. Clarendon Press, Oxford
2003, ISBN: 9780199248063

Fournier GP, Andam CP, Gogarten JP (2015) Ancient horizontal gene transfer and the last common ancestors. BMC Evol Biol 15, 70. https://doi.org/10.1186/s12862-015-0350-0

Galtier N, Tourasse N, Gouy M (1999) A nonhyperthermophilic common ancestor to extant life forms. Science. doi: 10.1126/science.283.5399.220.

Gogarten JP, Doolittle WF, Lawrence JG (2002) Prokaryotic evolution in light of gene transfer. Molecular Biology and Evolution. 19, 12, 2226-38.

Gogarten JP, Deamer D (2016) Is LUCA a thermophilic progenote? Nature Microbiology, 1.

Goldman AD, Bernhard TM, Dolzhenko E et al. (2013) LUCApedia: a database for the study of ancient life. Nucleic Acids Res. doi: 10.1093/nar/gks1217.

Groussin M, Boussau B, Charles S et al. (2013) The molecular signal for the adaptation to cold temperature during early life on EarthBiol. Lett. http://doi.org/10.1098/rsbl.2013.0608

Harris JK, Kelley ST, Spiegelman GB et al. (2003) The genetic core of the universal ancestor. Genome Res. 13:407–12

Hagen JB (1999) Naturalists, Molecular Biologists, and the Challenges of Molecular
Evolution. Journal of the History of Biology 32: 321–341

Hagen JB (2003) The Statistical Frame of Mind in Systematic Biology from Quantitative Zoology to Biometry. Journal of the History of Biology 36: 353–384.

Hilario E, Gogarten JP (1993) Horizontal transfer of ATPase genes--the tree of life becomes a net of life. Biosystems. 31(2-3):111-119. doi:10.1016/0303-2647(93)90038-e

Holmes FL (2004) Investigative Pathways. Patterns and Stages in the Careers of Experimental Scientists. Yale University Press. ISBN: 9780300100754

Huang J, Gogarten JP (2006) Ancient horizontal gene transfer can benefit phylogenetic reconstruction. Trends in Genetics, 22, 7, 361-366.

Hügler M, Sievert SM (2011) Beyond the Calvin cycle: autotrophic carbon fixation in the ocean. Annu Rev Mar Sci 3:261–289. https://doi.org/10.1146/annurev-marine-120709-142712

Husnik F, McCutcheon JP (2018) Functional horizontal gene transfer from bacteria to eukaryotes. Nature Reviews Microbiology, 16, 67-79.

Jácome R, Becerra A, Ponce de León S et al (2015) Structural Analysis of Monomeric RNA-Dependent Polymerases: Evolutionary and Therapeutic Implications. PLoS One.

Keeling PJ, Palmer JD (2008) Horizontal gene transfer in eukaryotic evolution. Nat Rev Genet. 605-18. doi: 10.1038/nrg2386.

Kim KM, Caetano-Anollés G (2011) The proteomic complexity and rise of the primordial ancestor of diversified life. BMC Evolutionary Biology, 11.

Kim KM, Caetano-Anollés G (2012) The evolutionary history of protein fold families and proteomes confirms that the archaeal ancestor is more ancient than the ancestors of other superkingdoms. BMC Evolutionary Biology. doi: 10.1186/1471-2148-12-13.

Kleiner M, Petersen JM, Dubilier N (2012) Convergent and divergent evolution of metabolism in sulfur-oxidizing symbionts and the role of horizontal gene transfer. Curr Opin Microbiol. doi: 10.1016/j.mib.2012.09.003.

Koonin EV (2003) Comparative genomics, minimal gene-sets and the last universal common ancestor. Nature Reviews. Microbiology, 1, 2, 127-36.

Kyrpides N, Overbeek R, Ouzounis C. (1999) Universal protein families and the functional content of the last universal common ancestor. J. Mol. Evol. 49:413–23

Loewe L, (2008) Negative selection. Nature Education 1(1):59

Martin W, Baross J, Kelley D et al. (2008) Hydrothermal vents and the origin of life. Nat Rev Microbiol 6:805-814

McInerney JO, Cotton JA, Pisani D (2008) The prokaryotic tree of life: past, present and future? Trends in Ecology and Evolution, 23, 276-81.

Mirkin BG, Fenner TI, Galperin MY et al. (2003) Algorithms for computing parsimonious evolutionary scenarios for genome evolution, the last universal common ancestor and dominance of horizontal gene transfer in the evolution of prokaryotes. BMC Evolutionary Biology, 3, 2.

Morgan G (1998) Emile Zuckerkandl, Linus Pauling, and the molecular evolutionary clock, 1959-1965. Journal of the History of Biology, 1, 155-178.

Muñoz-Velasco I, García-Ferris C, Hernandez-Morales R et al. (2018) Methanogenesis on Early Stages of Life: Ancient but Not Primordial. Orig Life Evol Biosph. doi: 10.1007/s11084-018-9570-9.

Mushegian A, Koonin EV (1996) A minimal gene set for cellular life derived by comparison of complete bacterial genomes. Proceedings of the National Academy of Sciences, 93, 1026810273.

O'Malley MA (2010) Ernst Mayr, the tree of life, and philosophy of biology. Biology and Philosophy 25:529–552. doi 10.1007/s10539-010-9214-6

O'Malley MA, Martin W, Dupré J (2010) The tree of life: introduction to an evolutionary debate. Biol Philos25, 441–453. https://doi.org/10.1007/s10539-010-9208-4

O'Malley MA (2013) Philosophy and the microbe: a balancing act. Biology & Philosophy, 28(2), 153-159.

O'Malley MA (2018) W. Ford Doolittle: evolutionary provocations and a pluralistic vision. Dreamers, Visionaries, and Revolutionaries in the Life Sciences, eds. Oren Harman and Michael r. dietrich (eds). Dreamers, Visionaries, and Revolutionaries in the Life Sciences. Chicago: University of Chicago Press.

O'Malley MA, Leger MM, Wideman JG et al (2019) Concepts of the last eukaryotic common ancestor. Nature ecology & evolution, 3(3), 338-344

Ouzounis CA, Kunin V, Darzentas N, et al (2006) A minimal estimate for the gene content of the last universal common ancestor--exobiology from a terrestrial perspective. Research in Microbiology, 157, 1, 57-68.

Puigbò P, Wolf Y, Koonin EV (2013) Seeing the Tree of Life behind the phylogenetic forest. BMC Biology 11, 46.

Ranea J, Sillero A, Thornton J, et al. (2006) Protein Superfamily Evolution and the Last Universal Common Ancestor (LUCA). Journal of Molecular Evolution, 63, 513-525.

Rivas M, Becerra A, Lazcano A (2018) On the Early Evolution of Catabolic Pathways: A Comparative Genomics Approach. I The Cases of Glucose, Ribose, and the Nucleobases Catabolic Routes. Journal of Molecular Evolution, 86, 27-46.

Russell R, Saqi M, Sayle R, et al. (1997) Recognition of analogous and homologous protein folds: analysis of sequence and structure conservation. Journal of Molecular Biology, 269.

Sapp J (2009) The new foundations of evolution: On the tree of life. Oxford: Oxford University Press.

Sober E (1988) Reconstructing the Past: Parsimony, evolution and inference. Bradford
MIT Press.

Sobolevsky Y, Trifonov E (2006) Protein Modules Conserved Since LUCA. Journal of Molecular Evolution, 63, 5, 622-634.

Sterner B (2017) Individuating population lineages: a new genealogical criterion. Biology & Philosophy 32(5), 683-703.

Strasser B (2019) Collecting Experiments: Making Big Data Biology. Chicago: University of Chicago Press.

Suárez-Díaz E, Anaya-Muñoz V (2008) History, objectivity, and the construction of molecular phylogenies. Studies in History and Philosophy of Biological and Biomedical Sciences. 39, 451-468 doi.org/10.1016/j.shpsc.2008.09.002.

Suárez-Díaz E (2014) The long and winding road of molecular data in phylogenetic analysis. Journal of the History of Biology, 47: 443-478.

Theobald DL (2010) A formal test of the theory of universal common ancestry. Nature. doi: 10.1038/nature09014.

Velasco J (2018) Universal Common Ancestry, LUCA, and the Tree of Life: Three Distinct Hypotheses about the Evolution of Life. Biology and Philosophy (2018) 33: 31

Watson T (2019) The trickster microbes that are shaking up the tree of life. Nature. doi: 10.1038/d41586-019-01496-w.

Weiss M, Sousa F, Mrnjavac N et al (2016a) The physiology and habitat of the last universal common ancestor. Nature Microbiology, 1, 9.

Weiss M, Neukirchen S, Roettger M et al (2016b) Reply to ‘Is LUCA a thermophilic progenote? Nature Microbiology 1, 16230 doi:10.1038/nmicrobiol.2016.230

Woese C, Fox G (1977) Phylogenetic structure of the prokaryotic domain: The primary kingdoms. Proceedings of the National Academy of Sciences, 74, 5088–5090. doi: 10.1073/pnas.74.11.5088

Woese (1987) Bacterial evolution. Microbiol Rev 51(2):21

Woese C (1998) The universal ancestor. Proc Natl Acad Sci USA, 95:6854-6859.

Yang AS, Honig B (2000) An integrated approach to the analysis and modeling of protein sequences and structures. III. A comparative study of sequence conservation in protein structural families using multiple structural alignments. J Mol Biol. 691-711. doi: 10.1006/jmbi.2000.3975.

Yang S, Doolittle RF, Bourne P (2005) Phylogeny determined by protein domain content. Proceedings of the National Academy of Sciences, 102, 373–378.

Zhou Z, Liu Y, Li M, Gu JD (2018) Two or three domains: a new view of tree of life in the genomics era. Applied Microbiology and Biotechnology, 102, 3049–3058. doi.org/10.1007/s00253-018-8831-x