naturereseach

Peer Review Information

Journal: Nature Ecology & Evolution

Manuscript Title: Timing the origin of eukaryotic cellular complexity with ancient duplications

Corresponding author name(s): Toni Gabaldón, Berend Snel

Editorial Notes:

Reviewer Comments & Decisions:

Decision Letter, initial version:

12th February 2020

*Please ensure you delete the link to your author homepage in this e-mail if you wish to forward it to your co-authors.

Dear Toni,

Your Article, "Timing the origin of eukaryotic cellular complexity with ancient duplications" has now been seen by three reviewers. You will see from their comments copied below that while they find your work of considerable potential interest, they have raised quite substantial concerns that must be addressed. In light of these comments, we cannot accept the manuscript for publication, but would be very interested in considering a revised version that addresses these serious concerns.

We hope you will find the reviewers' comments useful as you decide how to proceed. If you wish to submit a substantially revised manuscript, please bear in mind that we will be reluctant to approach the reviewers again in the absence of major revisions.

I should stress that we will be very reluctant to send a revision back to reviewers unless we see extensive new analyses to address the major problems of the study, including the flaws pointed out by the three referees relating to the use of normalized branch lengths to estimate the order in which gene duplications occurred.

If you choose to revise your manuscript taking into account all reviewer and editor comments, please highlight all changes in the manuscript text file.

We are committed to providing a fair and constructive peer-review process. Please do not hesitate to contact us if there are specific requests from the reviewers that you believe are technically impossible or unlikely to yield a meaningful outcome.

If revising your manuscript:

* Include a “Response to reviewers” document detailing, point-by-point, how you addressed each referee comment. If no action was taken to address a point, you must provide a compelling argument. This response will be sent back to the referees along with the revised manuscript.

* If you have not done so already we suggest that you begin to revise your manuscript so that it conforms to our Article format instructions at http://www.nature.com/natecolevol/info/final-submission. Refer also to any guidelines provided in this letter.

* Include a revised version of any required reporting checklist. It will be available to referees (and, potentially, statisticians) to aid in their evaluation if the manuscript goes back for peer review. A revised checklist is essential for re-review of the paper.

Please use the link below to submit a revised paper:

[REDACTED]

<strong>Note:</strong> This URL links to your confidential home page and associated information about manuscripts you may have submitted, or that you are reviewing for us. If you wish to forward this email to co-authors, please delete the link to your homepage.

If you wish to submit a suitably revised manuscript we would hope to receive it within 6 months. If you cannot send it within this time, please let us know. We will be happy to consider your revision so long as nothing similar has been accepted for publication at Nature Ecology & Evolution or published elsewhere. Should your manuscript be substantially delayed without notifying us in advance and your article is eventually published, the received date would be that of the revised, not the original, version.

Nature Ecology & Evolution is committed to improving transparency in authorship. As part of our efforts in this direction, we are now requesting that all authors identified as ‘corresponding author’ on published papers create and link their Open Researcher and Contributor Identifier (ORCID) with their account on the Manuscript Tracking System (MTS), prior to acceptance. This applies to primary research papers only. ORCID helps the scientific community achieve unambiguous attribution of all scholarly contributions. You can create and link your ORCID from the home page of the MTS by clicking on ‘Modify my Springer Nature account’. For more information please visit please visit <a href="http://www.springernature.com/orcid">www.springernature.com/orcid</a>.

Please do not hesitate to contact me if you have any questions or would like to discuss the required revisions further.

Thank you for the opportunity to review your work.

[REDACTED]

Reviewer expertise:

Reviewer #1: eukaryogenesis, phylogenomics

Reviewer #2: eukaryogenesis, phylogenomics

Reviewer #3: eukaryogenesis, phylogenomics

Reviewers' comments:

Reviewer #1 (Remarks to the Author):

In this manuscript, Vosseberg et al. investigate gene duplications on the eukaryote stem as a way to get a handle on the origin of eukaryotic complexity. Their analyses lead to several interesting and novel findings:

(i) duplications caused the pre-LECA genome to almost double, but most duplications were restricted to a subset of families, primarily genes involved in information processing, storage, cell processes and signalling. Consistent with this function, archaea-origin genes duplicated more frequently than bacteria-origin genes (contra the recent preprint of Tria et al).
(ii) Alphaproteobacteria and Asgards are the largest contributors to the pre-LECA stem (this has long been known for alphas, but the finding of a sister group relationship between Heimdallarchaeota and eukaryotes in many families is new, interesting, and bears on the 2D/3D tree of life debate).
(iii) Based on comparisons of normalized branch lengths, some of the archaeal origin duplications seemed to occur before the mitochondrial endosymbiosis, consistent with the "mito intermediate" view of eukaryote origins.

These findings are important, if they are correct. Findings (i) and (ii) seem well supported by the analyses, finding (iii) perhaps less so. Their validity hinges on the methods used to analyse the genomes, and I have several questions/concerns about those analyses. Before outlining those, I want to commend the authors for clearly laying out the methods used in what was clearly an elaborate analysis in sufficient detail that they could be followed (and perhaps critiqued). I also want to emphasize that this is a novel and thought-provoking study and that the purpose of the comments below is to raise issues that might conceivably be addressed.

(1) Pfam domains were used to trace the origins of eukaryotic gene families. But Pfams are biased towards characterized genes from model organisms. The authors use the relationship between Pfam number and gene number in a range of modern eukaryotes to extrapolate the gene content number of LECA. But I wonder, why not use "unbiased" gene families directly (e.g. using mcl clustering or a related method?) It is not clear that the evolutionary dynamics of uncharacterised gene families will be the same as those of families for which a well-annotated Pfam exists.

(2) The ScrollSaw method, and a criterion of presence on either side of a putative eukaryote root between Opimoda and Diphoda, was used to infer LECA families. But the position of the eukaryote root is not clear. It would be beyond the scope of this work to resolve that issue as well, but it would be worth investigating whether the conclusions change if competing root hypotheses are considered, in particular the excavate root (He et al. 2014). The "competing" Tria et al. preprint mentioned above

recovers excavates at the root based on patterns of shared genes among eukaryote groups.

(3) The use of normalized branch lengths to determine relative gene age rests on strong assumptions and has been criticized in the literature (https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5516564/). Generally, branch lengths are a product of evolutionary rate and time. Normalizing the stem length by the post-LECA lengths works if the post-LECA evolutionary rate is the same as the stem rate, but it is difficult to determine when/if this is the case. In the reference linked above, it was argued that mitochondrial genes might have shorter stems because they had to adapt less during eukaryogenesis (remaining within the same compartment). More broadly, it seems clear that many genes underwent functional divergence during eukaryotic origins: consider genes found patchily in archaea but conserved in eukaryotes. A rate slowdown post-LECA would make stems appear longer when normalized, and faster evolutionary rates post-LECA would have the opposite effect.

I do not know how these effects can be controlled for, but they seem to make the branch length distributions difficult to interpret. Perhaps focusing on distributions of normalized stem lengths for the subset of genes that do not show functional divergence on the eukaryote stem would help to clarify the situation.

Finally, another issue with measuring stem lengths is that the stem is measured back to the branching point with the closest prokaryotic relative. They are therefore dependent on sampling of the prokaryotic sister groups (that is, differences in evolutionary time and differences in sampling rate of archaeal and bacterial diversity are conflated). Again, I do not see how this can be easily addressed but it should be acknowledged as another potential explanation for the branch length results. That is, better sampling of alphaproteobacterial relatives of the mitochondrion relative to Asgards would explain the differences in stem lengths.

(4) "Removal of interspersed prokaryotes" in the analysis: when prokaryote sequences were interspersed between eukaryote clades in the trees, they were removed unless they were Asgard archaea. This would seem to straightforwardly bias the results to those observed, that Asgards are the most frequent archaeal donors of eukaryote genes.

(5) The comparison with the Tria et al. preprint is interesting, and is an important paragraph because the two analyses, while in principle investigating very similar questions, come to almost diametrically opposed conclusions about the kinds of genes duplicated, and their evolutionary origin. A little more detail would be useful here to demonstrate to the reader that the Tria et al results arise from poor sequence clustering. One possibility would be to show gene trees for the families mentioned, labelling the subset of sequences obtained in the clusters of Tria et al.

Reviewer #2 (Remarks to the Author):

I very much like the conclusion that many evolutionary events occurred on the lineage from FECA to LECA, and that the emergence of eukaryotic features was a drawn-out process.

While this is a step up from Tria et al, I am concerned that the chosen approach to time estimation moved the times estimated for duplications events closer to FECA, artificially compressing the time between FECA and the duplication events. Following a gene duplication substitution rates can increase several fold. Tubulins, histones and chaperons provide examples for dramatic rate increases following duplications. The chosen estimation approach thus is likely to place these events further into the past. In addition, I suspect that the extent to which substitution rates are accelerated differs for different types of proteins – the transition from homodimer to heterodimer weakens selection pressures on the protein interface. The rate increase following duplication will be different (and smaller) for proteins that function as monomers and that were duplicated via divergence followed gene transfer (the most common way gene families expand in prokaryotes). I wonder if the different profiles for different groups of proteins might reflect differences in rate acceleration following duplications, and not differences in timing.

Another concern is that the analysis appears to be contingent on a particular placement of LECA relative to the genomes analyzed. Currently two placements of LECA are supported by different groups: at the split between the Amorphea/Unikonts and the Diaphoretickes, or at the base or inside the Excavates. Given that the Excavates are not thoroughly sampled, if the true placement of LECA were within the Excavates , this would imply that some of the events inferred as pre-LECA might actually have been post LECA. A more detailed discussion of the impact of LECA's phylogenetic placement on the performed analyses appears warranted. A discussion of the possible LECA nodes should make use of recent manuscripts on Eukaryotic taxonomy (e.g., Adl SM et al. Revisions to the Classification, Nomenclature, and Diversity of Eukaryotes. J Eukaryot Microbiol. 2019;66(1):4–119).

Line 429: Given that the Excavata are possibly paraphyletic, sampling only two representatives is problematic.

Line 466ff: Presence in Opimoda and Diphoda seems a necessary but not a sufficient condition. If LECA were placed among the Excavate, some Diphoda may have the gene, without it being present in LECA – see above.

Line 558: "Therefore, … " This is rather strange reasoning, resembling the streetlight effect <https://en.wikipedia.org/wiki/Streetlight_effect> . BBHs between the five (or three) supergroups would have been more likely to actually have been present in LECA. Declaring something a LECA node that actually is a post LECA node may lead to more and better resolved "LECA" nodes, but might not reflect reality.

Supplementary Materials line 75: "Such a pattern would fit into a 'Big Bang' hypothesis for eukaryogenesis". This does not appear logical. If an accelerated rate following a duplication occurs, it makes is difficult to place the duplication, but why would one conclude that these happened all at the same time? In many gene families with more than one duplication the different substitution events are clearly separated from one another.

Line 36: Given that prokaryotes of enormous size (Thiomargarita namibiensis and Epulopiscium fishelsoni) and with endomembrane system (Gemmata obscuriglobus) are known, some qualifier should be added.

The ridgeline plots represent the data after some smoothing of the curves was applied. More details on how this was done and on how the original data look as histograms should be provided.

The given links to additional data and code did not work.

Reviewer #3 (Remarks to the Author):

Many open questions remain regarding the origin of eukaryotes from prokaryotic ancestors. Among those, the order in which eukaryotic specific features originated, and in particular the timing of mitochondrial endosymbiosis relative to the emerge of cellular compartmentalization, is a matter of heated debate in the field of evolutionary biology.
In this paper, Vosseberg and colleagues try to address this question by using a phylogenomic approach to focus on the evolutionary history of duplicated genes. They argue, among others, that the genes that duplicated the most, and before the mitochondrial endosymbiosis, were inherited from Asgard archaea and are involved in cytoskeleton and membrane trafficking, suggesting that cellular complexity would have evolved prior to the acquisition of mitochondria.

It is a topic of major importance that deserves scrutiny and that I believe is of interest to the broad community of evolutionary biologists. However, I have genuine concerns about this manuscript that I outline below. Overall the methods employed lack clarity and justification; this has put me in a frustrating position where I can discuss the results themselves in any details because I was not able to grasp clearly what the authors have done and why. This kind of large-scale analysis relying entirely on the sanity of the methods, I do not feel confident about the reliability of the results. To clarify, I do not argue that the results are incorrect, I argue that the manuscript needs to be heavily edited in ordered to allow reviewers to assess those results.

I strongly encourage the authors to carefully detail the method section and provide justification for the choices made, which sometimes seem completely ad hoc.

*** Major comments ***

Core assumption: My main concern is that all the conclusions made here are based on the assumption that genes evolve somewhat clockwise and that branch length can be used as a proxy to date evolutionary events relative to one another. Although the authors do take into account the fact that some gene families evolve faster than other and try correct across gene-family rate variations by using post-LECA branch lengths, it is not clear at all that this is sufficient to make an accurate estimation of the time of duplication. In fact, the author themselves discuss this in the supplementary material: given that a duplication event gives birth to two new clades A and B, there is two branches whose (corrected) length can be used to estimate the age of this duplication – the branch leading from the duplication node to LECA in clade A, and the same in clade B. In theory, both of those should lead to a similar estimate of the duplication age. However, the authors discuss the fact that, when testing those multiple measures, “the shortest length was most consistent with the branch lengths obtained from non-duplicated acquisitions (Extended Data Fig. 5a). It should be noted that even using the shortest could not fully account for the difference in stem lengths.” and further that “This indicates that in most cases there was an accelerated evolutionary rate in at least one of the paralogues, which could not be (fully) corrected for by the post-LECA branch lengths.”. Hence, if the branch length normalization for duplicated genes is not quite comparable to the one made for “acquisitions”, this raises the a very important question which is: how can you be sure that you can compare the relative age estimates made for duplications versus acquisition events? This is an absolutely crucial point to make to be able to claim that genes of Asgard origin duplicated mostly prior to the acquisition of the mitochondrial endosymbiosis.

This should absolutely be discussed in the main text, and not vaguely mentioned in the supplementary. As currently written, the main text presents this method as if it had proven to be sound, which is not the case so far.

* Conclusion of the abstract: the conclusions of the abstract need to be toned down, particularly the sentence that says “we demonstrate that the host that engulfed the proto-mitochondrion had some eukaryote-like complexity”. "Demonstrating" is way too strong of a word in evolutionary biology in general, and furthermore when the results are extrapolated from novel methods that remain to be proven sensible.

* General comment: most of the text needs to be rewritten in a MUCH clearer way. So many sentences are overly long or imprecise, it makes it very difficult on the reader to follow what the authors are getting at. I really cannot stress this enough: please, carefully go over the text and expend and/or clarify.

* The taxon reduction procedure might lead to potential bias towards Asgards archaea inferences:

The whole section about reducing the number of sequences in trees were reduced is very opaque. First when it comes to reducing the number of prokaryotic sequences, the authors say “For each Pfam, the number of prokaryotic sequences was reduced with kClust v1.0 34 using a clustering threshold of 2.93. Asgard archaeal sequences were excluded from this reduction, because they are relatively undersampled and are already genetically diverse.”
- Is there a rationale behind choosing a threshold of 2.93?
- More importantly, I do not understand what justifies separating the Asgard sequences from the rest. If indeed they are genetically diverse, then shouldn’t they remain anyway represented after clustering? And if not, then you should better explain the criteria that are used by Kclust and how that would bias the representation of Asgard sequences. But if there is a bias after clustering, shouldn’t the authors be concerned that it will after clades other than Asgards? And that because now Asgards receive a special treatment, that this will skew the signal they later recover?

Moreover the whole section describing the eukaryotic taxon reduction is really hard to understand. Beyond explaining the step by step approach, please clarify what the goal is for each of them. In addition, please briefly explain what ScrollSaw is. Finally, please explain 1) the goal of identifying birectional best blast hits and 2) why you did this either between different supergroups OR between Diphoda and Opimoda, and how did that impact your results?

* KOG-COG clusters analyses are not understandable:
I have spent a lot of time reading this section and still cannot really grasp what the authors are doing. However, my main concern about this section is that you define a “duplication score” and a “LECA score” in order “to infer reliable duplication nodes” and as to infer gene families that were likely in LECA. What is really concerning to me is that you test out different cutoff values for those scores – including zero – and say that “it did not have a large impact on the absolute numbers and quality measures, such as the fraction of well-supported LECA and duplication nodes. This underlines the robustness of our analysis. What is the validity of such a score if even when it is zero does not change your results? Does this really prove the robustness of the analysis, or that there is a lot of randomness to it?

Other comments about this section:
- First of all, please explain what the point of this analysis before going into the griddy details.
- Here again, why do you use a different approach for Asgard archaea than for the rest of archaea? Sure, they uniquely encode some “eukaryotic” proteins, but they are not the only ones. Other archaeal lineages encode such eukaryotic proteins, some to the exclusion of Asgards.

* Tree analyses
- Here again, Asgard archaea seem to receive an unfair treatment. As much as I recognize the relevance of those taxa for this question, I think the authors need to consider that they are potentially biasing their results by not applying the same criteria to other lineages.

*** Minor comments ***

In many place the wording is just not precise. Some examples:

l. 57-58 “we attempt to reconstruct the successive stages of eukaryogenesis by systematically analysing large sets of phylogenetic trees.” Be more specific: phylogenetic trees corresponding to what? Sure, it becomes apparent later, for this should be clear already in the introduction. Same comment for “the scale of gene inventions and duplications”… in which part of the tree? Be specific.

l. 73 “To include genes having only small Pfam domains, which were excluded for computational reasons, or no domains at all, we used a linear regression analysis to obtain an estimated LECA genome containing 12,780 genes”. What does this mean? The corresponding section in the methods is equally unclear. I can guess what it means but the method description is overally vague, eg. “. The assumptions of a normal distribution of gene values at each Pfam domain value and equal variance were reasonably met after log transformation.”

- l. 434 to 440 needs to be written more clearly, it is currently jumping around and is very difficult to follow with all the mentions of euNOGs, TWOGs, ENOGs, KOGs… Please guide the reader in understanding the relevance and difference in using each of those.

- l. 470 “For the annotation of nodes in trees from the Pfam-ScrollSaw sequences the information from the eukaryotic sequences that were not in the between-supergroup or Opimoda-Diphoda BBHs were included.” What does any of this mean?

- l. 484 “the mean presence of a potential LECA family in eukaryotic species [...] should be at least 15%.” What does that mean? And why? And how did you “ weight so that each supergroup contributes equally”?

- l. 492 the other mention that “For the annotation of nodes in KOG-COG trees a slightly different approach was followed.” which made me wonder how and why there are two types of trees: the Pfam-ScollScraw and the KOG-COG trees?

- l. 497 “If there were duplication nodes in both daughters, this node had to be a duplication node as well even though its duplication consistency score was below the threshold.” How can this be a duplication node if its duplication score is below the threshold? Please clarify.

l. 538 “Eukaryotic clades with LECA nodes that were nested, i.e. they had exactly the same prokaryotic sister group, were merged.” What does that mean?

Suppl. Mat. “For both, normalisation increased the number of duplications predating mitochondrial endosymbiosis, making the mitochondrial acquisition a slightly earlier event.”
Isn’t that contradictory? If you increase the number of duplications predating the mitochondrial acquisition, then this makes the latter a later event, no?

The figure legends for Suppl Fig 1 – 10 need to be greatly expended for the reader to understand them. I have spent way too much time guessing what they represent and I’m still unsure about what to conclude from them.

Author Rebuttal to Initial comments

Reviewer #1 (Remarks to the Author):

In this manuscript, Vosseberg et al. investigate gene duplications on the eukaryote stem as a way to get a handle on the origin of eukaryotic complexity. Their analyses lead to several interesting and novel findings:

(i) duplications caused the pre-LECA genome to almost double, but most duplications were restricted to a subset of families, primarily genes involved in information processing, storage, cell processes and signalling. Consistent with this function, archaea-origin genes duplicated more frequently than bacteria-origin genes (contra the recent preprint of Tria et al).
(ii) Alphaproteobacteria and Asgards are the largest contributors to the pre-LECA stem (this has long been known for alphas, but the finding of a sister group relationship between Heimdallarchaeota and eukaryotes in many families is new, interesting, and bears on the 2D/3D tree of life debate).
(iii) Based on comparisons of normalized branch lengths, some of the archaeal origin duplications seemed to occur before the mitochondrial endosymbiosis, consistent with the "mito intermediate" view of eukaryote origins.

These findings are important, if they are correct. Findings (i) and (ii) seem well supported by the analyses, finding (iii) perhaps less so. Their validity hinges on the methods used to analyse the genomes, and I have several questions/concerns about those analyses. Before outlining those, I want to commend the authors for clearly laying out the methods used in what was clearly an elaborate analysis in sufficient detail that they could be followed (and perhaps critiqued). I also want to emphasize that this is a novel and thought-provoking study and that the purpose of the comments below is to raise issues that might conceivably be addressed.

We thank the reviewer for the excellent summary and the recognition of the implications of our findings. We appreciate the kind words about the elaborateness and detailed methods section.

(1) Pfam domains were used to trace the origins of eukaryotic gene families. But Pfams are biased towards characterized genes from model organisms. The authors use the relationship between Pfam number and gene number in a range of modern eukaryotes to extrapolate the gene content number of LECA. But I wonder, why not use "unbiased" gene families directly (e.g. using mcl clustering or a related method?) It is not clear that the evolutionary dynamics of uncharacterised gene families will be the same as those of families for which a well-annotated Pfam exists.

The reviewer addresses a valid point that is an open problem in comparative genomics, namely how to define gene families to use for phylogenetics-based detection of gene duplications. We mainly chose Pfam domains because of the sensitivity of their profile HMMs to detect more divergent homologous sequences, as it is our experience that de novo *BLAST-based methods fail to detect a substantial number of ancient homologies due to low sensitivity. Furthermore, by using this database we circumvented the need to use previously defined orthologous groups or infer homology* de novo. *A risk of using a* de novo *approach like MCL is oversplitting of greatly expanded gene families (as explicitly documented in Liebeskind* et al., *GBE, 2016), which is especially problematic when one wants to study the* *numerous duplications during eukaryogenesis.* Tria et al. *used MCL and we suspect that that is one of the main reasons they obtained very few duplications. Another complication of* de novo *clustering based methods is how it deals with protein sequences constituted by different domains; sequences that result from a fusion between FECA and LECA might end up in a single orthologous group (OG), for example, without their prokaryotic (single-domain) ancestors.*

In a different study, we applied multiple orthology inference methods, including de novo *methods, to our eukaryotic data set (Deutekom, Snel and Van Dam; preprint at biorxiv: https://www.biorxiv.org/content/10.1101/2020.05.13.092791v1). We counted the number of OGs predicted to have been present in LECA in that study that lacked a Pfam, which gives an estimate of the number of families that we have likely excluded from our analysis here by choosing Pfam. Depending on the orthology method, between 4 and 12% of the LECA OGs did not include any sequence with a Pfam hit. These OGs included much fewer sequences (mean 11-23 vs. 30-168; median 8-15 vs. 19-105), indicating that their status as LECA OG is perhaps not so certain. Thus, the vast majority of these "unbiased" LECA OGs were included in our analysis. In addition, there is no a priori reason to expect that families that are better characterized would have evolved differently during early eukaryotic evolution, as compared to less-characterized ones.*

Finally, the use of Pfam by virtue of the clan level clustering available allows even deeper homologies (paralogies) to be counted in our tally of duplications during eukaryogenesis, such as the eukaryotic tyrosine kinases that are known to be derived from eukaryotic serine/threonine protein kinases but they are in different Pfam families.

(2) The ScrollSaw method, and a criterion of presence on either side of a putative eukaryote root between Opimoda and Diphoda, was used to infer LECA families. But the position of the eukaryote root is not clear. It would be beyond the scope of this work to resolve that issue as well, but it would be worth investigating whether the conclusions change if competing root hypotheses are considered, in particular the excavate root (He et al. 2014). The "competing" Tria et al. preprint mentioned above recovers excavates at the root based on patterns of shared genes among eukaryote groups.

This concern was also raised by reviewer #2. For a more detailed response, see below. We performed ScrollSaw also with other supergroup definitions and implemented the possibility of using different root positions in our tree analysis. We have now included the analysis on the effect of the root position in the main text. When using the excavate root, we inferred 15% fewer LECA OGs compared with an Opimoda-Diphoda root but we did not observe a change in branch lengths (Extended Data Fig. 2).

(3) The use of normalized branch lengths to determine relative gene age rests on strong assumptions and has been criticized in the literature (https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5516564/). Generally, branch lengths are a product of evolutionary rate and time. Normalizing the stem length by the post-LECA lengths works if the post-LECA evolutionary rate is the same as the stem rate, but it is difficult to determine when/if this is the case. In the reference linked above, it was argued that mitochondrial genes might have shorter stems because they had to adapt less during eukaryogenesis (remaining within the same compartment). More broadly, it seems clear that many genes underwent functional divergence during eukaryotic origins: consider genes found patchily in archaea but conserved in eukaryotes. A rate slowdown post-LECA would

make stems appear longer when normalized, and faster evolutionary rates post-LECA would have the opposite effect.

The assumption behind the use of normalised branch lengths is that the evolutionary rates pre- and post-LECA are proportional, not necessarily the same. We acknowledge that the observed trends could be created by similar rate changes in proteins of the same phylogenetic origin, instead of reflecting different time points. We also refer in the main text to the concerns raised by the paper to which the reviewer refers. Pittis and Gabaldón (2016) performed matched comparisons and observed that families of similar function, selection pressure, number of protein-protein interactions or expressions levels, but with different phylogenetic origins, had different stem lengths (Extended Data Fig. 2 in Pittis and Gabaldón (2016)). Moreover, the alphaproteobacterial stem length was also shorter for families that were re-targetted outside the mitochondrion (Extended Data Fig. 6d in Pittis and Gabaldón (2016)), demonstrating that even if a gene family was subject to functional innovation during eukaryogenesis, this did not make its stem length as long as those stem lengths from families with different phylogenetic origins. Even if pre- and post-LECA shifts in evolutionary rates took place as suggested, it is difficult to explain that this creates a pattern that depends on phylogenetic origin of the group rather than on function, localisation, etc. Different time points of acquisition is the main explaining factor, and this does not preclude that rates vary as well. Additional criticisms of the referred article were addressed by the authors in more detail here: https://doi.org/10.1101/064873.

I do not know how these effects can be controlled for, but they seem to make the branch length distributions difficult to interpret. Perhaps focusing on distributions of normalized stem lengths for the subset of genes that do not show functional divergence on the eukaryote stem would help to clarify the situation.

We kindly thank the reviewer for this excellent suggestion. We implemented it by calculating branch lengths only for those families for which the functional category of the LECA family was identical to that of the prokaryotic sister group (analysing non-duplicated families only). With this, we recovered the same pattern of stem lengths (Extended Data Fig. 6b).

Finally, another issue with measuring stem lengths is that the stem is measured back to the branching point with the closest prokaryotic relative. They are therefore dependent on sampling of the prokaryotic sister groups (that is, differences in evolutionary time and differences in sampling rate of archaeal and bacterial diversity are conflated). Again, I do not see how this can be easily addressed but it should be acknowledged as another potential explanation for the branch length results. That is, better sampling of alphaproteobacterial relatives of the mitochondrion relative to Asgards would explain the differences in stem lengths.

We acknowledge this issue in the Supplementary Discussion, but we agree with the reviewer that is an important point. Generally speaking, having a better sampling of prokaryotes which would include closer ancestors will shorten the stem lengths. Having more and more closely related Asgard archaea would therefore make the difference in stem length with (alphaproteo)bacteria smaller. In this regard, including duplications in the branch length analysis provides additional information: whereas the "acquisition" represents the earliest possibility of the actual acquisition, the first duplication represents the latest possibility of the *acquisition (assuming the duplication did not predate the acquisition, which seems a fair assumption considering prokaryotic genome evolution). Because of the low numbers of these first duplications, we did not describe them separately but they might be used in further research to attenuate the acquisition time estimates.*

(4) "Removal of interspersed prokaryotes" in the analysis: when prokaryote sequences were interspersed between eukaryote clades in the trees, they were removed unless they were Asgard archaea. This would seem to straightforwardly bias the results to those observed, that Asgards are the most frequent archaeal donors of eukaryote genes.

We here would like to rectify an apparent misunderstanding, which we apparently have not explained properly in our Methods section: only if there was only one Asgard archaeal sequence present in the whole tree, it was not removed. In all other cases, interspersing Asgard archaeal sequences were removed. We made this more explicit in the methods. Our original reason for removing a single prokaryotic sequence from a tree, was that it very likely represents a eukaryote-to-prokaryote transfer or eukaryotic contamination in prokaryotic genomes, instead of a 'true' pre-LECA prokaryotic sister sequence. However, because Asgard archaea are relatively undersampled and because they encode various genes furthermore only found in eukaryotes, often in a 'patchy' manner (Zaremba-Niedzwiedzka et al.*, 2017) we considered it a likely possibility that a bona fide FECA gene would be present in only one Asgard archaeon. However, when we gave the Asgard archaea the same "normal" treatment, it did not have a large effect on the number of acquisitions and calculated duplication lengths included in the analysis: 16 Asgard archaeal acquisitions (-5%), as well as 51 Asgard archaea-derived LECA families (-8%) and 16 calculated Asgard archaea-derived duplication lengths (-6%) were lost. These small numbers did not affect the significance of comparisons and Asgard archaea remained the number one archaeal donor. Among the lost acquisitions were RPL28/MAK16, Sec23/24, UFM1 and the C-terminal tubulin domain, which are described as bona fide Asgard archaeal acquisitions in the literature (Zaremba-Niedzwiedzka* et al.*, 2017). Because of the marginal effect and the removal of bona fide Asgard archaeal acquisitions, we stayed with our decision to not prune single Asgard archaeal sequences.*

(5) The comparison with the Tria et al. preprint is interesting, and is an important paragraph because the two analyses, while in principle investigating very similar questions, come to almost diametrically opposed conclusions about the kinds of genes duplicated, and their evolutionary origin. A little more detail would be useful here to demonstrate to the reader that the Tria et al results arise from poor sequence clustering. One possibility would be to show gene trees for the families mentioned, labelling the subset of sequences obtained in the clusters of Tria et al.

We thank the reviewer for these suggestions. We adjusted the Supplementary Discussion in order to better expound our doubts about the Tria et al. *study. At the same time, we would prefer not put too much emphasis on the comparison with Tria* et al.*, as our analysis stands on its own, the data from that study - still not subjected to independent review - is not available to us, and we do not want the reader to perceive our study as a reaction to this preprint.*

Reviewer #2 (Remarks to the Author):

I very much like the conclusion that many evolutionary events occurred on the lineage from FECA to LECA, and that the emergence of eukaryotic features was a drawn-out process.

While this is a step up from Tria et al, I am concerned that the chosen approach to time estimation moved the times estimated for duplications events closer to FECA, artificially compressing the time between FECA and the duplication events. Following a gene duplication substitution rates can increase several fold. Tubulins, histones and chaperons provide examples for dramatic rate increases following duplications. The chosen estimation approach thus is likely to place these events further into the past. In addition, I suspect that the extent to which substitution rates are accelerated differs for different types of proteins – the transition from homodimer to heterodimer weakens selection pressures on the protein interface. The rate increase following duplication will be different (and smaller) for proteins that function as monomers and that were duplicated via divergence followed gene transfer (the most common way gene families expand in prokaryotes). I wonder if the different profiles for different groups of proteins might reflect differences in rate acceleration following duplications, and not differences in timing.

We agree with the reviewer that an accelerated rate after duplication in certain protein families could potentially explain the differences in duplication lengths that we see for different functions. Of note, this can only be the case if the acceleration (or deceleration) occurs in both paralogue daughter lineages, because we use only one - the shortest branch - to calculate the duplication length. We tested if such rate shifts in both daughter branches occured by calculating for each duplication in a tree the minimal stem length going through this node and compared these values with the non-duplicated families. In this way we can look at accelerated rates after duplication in both paralogue lineages within a family (instead of considering only a single post-duplication branch, as we did in our main analysis). To make a fair comparison we needed acquisitions that correspond to the same event and because we did not obtain many duplications in families with an alphaproteobacterial sister clade, we only looked at the Asgard archaea-related host lineage. It is worth noting that these values are not used in Figures 3 (only 1 sl value per acquired family) and 4 (duplication lengths are based on the distance between the duplication event and LECA). We looked for differences between the non-duplicated and duplicated stems (calculated via all duplication nodes, as described above) for informational and signalling families (not enough data points for metabolic families) (Extended Data Fig. 5c-d). The difference between non-duplicated and duplicated stem length estimates was not significant for informational families, showing that there was likely not a higher substitution rate post-duplication in both paralogue lineages for these groups of proteins. For signalling families the difference was barely significant, although this is no longer the case upon removal of the outliers (the duplicated outliers are ubiquitin duplications). Based on this analysis, we could not detect a pronounced rate shift for duplicated acquisitions.

The reviewer specifically refers to duplications resulting in the transition from a homomer to a heteromer. We specifically checked for the duplicated families from Asgard archaea if there was a difference between families that experienced a transition from homomer to heteromer during eukaryogenesis (proteasome, Snf7, TRAPP, Vps36 and OST3/OST6) and the other families. The difference was not statistically significant (Extended Data Fig. 5e).

The reviewer refers to tubulins, histones and chaperones as examples of these rate changes. Cytoskeletal duplications indeed had long duplication lengths but duplications in posttranslational modification, protein turnover, chaperones ("Protein modification" in the figures) and especially chromatin had relatively short branches. In our opinion, it is difficult to explain the observed duplication length differences between these groups of proteins by different rate changes. Rate acceleration in individual proteins can occur, but when comparing large functional groups of duplications, each comprising different types of proteins, we do not see how this could bias our results to create the patterns that we observe. For more recent duplications for which we know their age, we were able to distinguish the different time points of duplication despite functional differences (Extended Data Fig. 5g-h).

Another concern is that the analysis appears to be contingent on a particular placement of LECA relative to the genomes analyzed. Currently two placements of LECA are supported by different groups: at the split between the Amorphea/Unikonts and the Diaphoretickes, or at the base or inside the Excavates. Given that the Excavates are not thoroughly sampled, if the true placement of LECA were within the Excavates , this would imply that some of the events inferred as pre-LECA might actually have been post LECA. A more detailed discussion of the impact of LECA's phylogenetic placement on the performed analyses appears warranted. A discussion of the possible LECA nodes should make use of recent manuscripts on Eukaryotic taxonomy (e.g., Adl SM et al. Revisions to the Classification, Nomenclature, and Diversity of Eukaryotes. J Eukaryot Microbiol. 2019;66(1):4–119).

We acknowledge the current debate and discuss the effect of using a different root in the revised manuscript. The position of the eukaryotic root is used in two parts of our analysis: in the selection of eukaryotic sequences using BBHs between different eukaryotic groups (the ScrollSaw-based selection approach) and in the annotation of tree nodes as LECA nodes. We addressed the root issue by assessing the effect of using a different root position in both parts of our analysis. Note that, in this additional analysis, we restricted ourselves to Pfams that are only present in eukaryotes, for computational reasons. We performed ScrollSaw also with another supergroup definition (Amorphea, Diaphoretickes, Discoba and Metamonada) and identified BBHs between these four monophyletic groups. On the tree annotation part, we tested all seven root possibilities between the four groups. Sequences from both sides of the root had to be present to be called a LECA node. We obtained a very similar number of LECA families for non-excavate root positions (Extended Data Fig. 2a). These numbers also did not substantially differ between the three different group definitions for the BBHs. When using an excavate root, we saw a large reduction in the number of LECA families for the main Opimoda-Diphoda BBHs set. This reduction was less pronounced for the five and four group BBHs sets: 15-46% fewer LECA families compared with an Opimoda-Diphoda root for the four-group BBHs set. This likely reflects that excavate sequences, especially from Metamonada species, are rarely involved in BBHs due to high sequence divergence, unless specifically searched for, as we do with the four-group BBHs.

If after all the eukaryotic root lies on or within the excavates, this could imply that some of our LECA genes and pre-LECA duplications actually arose/occurred after LECA, after the excavate lineage diverged from the other eukaryotic lineages. However, a bona fide LECA gene could have been lost in one daughter lineage of the root (or both). Although duplications can be used to distinguish these scenarios (e.g. a well-supported node in the tree with all eukaryotes but excavates on one side, and all eukaryotes including excavates on *the other strongly suggests that the excavate lineage lost one of the paralogous genes), this information was not used when annotating the tree nodes. Given the low number and reduced and diverged nature of the excavate genomes in our data set, we consider a gene-rich LECA and losses or undetected homologies in excavates the most plausible explanation. Because of the early eukaryotic radiation, which in phylogenies often results in short and poorly supported internal branches, we do not consider a great number of acquisitions, inventions and duplications just after LECA a likely scenario.*

We also tested if a different root position had an effect on the duplication lengths and found no significant difference (Extended Data Fig. 2b).

Line 429: Given that the Excavata are possibly paraphyletic, sampling only two representatives is problematic.

We agree that in this KOG-to-COG dataset, the sampling of excavates poorly presents their phylogenetic diversity. However, we think that having only two excavates is not problematic as long as the root is not on one of the excavate lineages. Because we used the Opimoda-Diphoda root to annotate these trees, additional sampling of excavates would not have had a large effect. We would like to emphasise that the KOG-to-COG clusters analysis mainly functioned to verify our method and is not incorporated in the main figures of this paper.

Line 466ff: Presence in Opimoda and Diphoda seems a necessary but not a sufficient condition. If LECA were placed among the Excavate, some Diphoda may have the gene, without it being present in LECA – see above.

We agree with the reviewer that only the presence in Opimoda and Diphoda does not provide strong support for a gene to have been present in LECA. Therefore we additionally require a gene to be present in a sufficient proportion of species across supergroups (See Methods, 'Annotation of eukaryotic nodes'). In that way, we were strict with our LECA criteria.

Line 558: "Therefore, … " This is rather strange reasoning, resembling the streetlight effect https://en.wikipedia.org/wiki/Streetlight_effect . BBHs between the five (or three) supergroups would have been more likely to actually have been present in LECA. Declaring something a LECA node that actually is a post LECA node may lead to more and better resolved "LECA" nodes, but might not reflect reality.

We tested the effect of including BBHs between five supergroups. It is not necessarily the case that BBHs between 2 or 5 groups results in only one representative per group for one OG in the tree. Note that all eukaryotic sequences, also those not involved in BBHs, were used to annotate the tree by mapping the sequences onto their best hit in tree (See Methods, 'Annotation of eukaryotic nodes'). Given that we used the same root position for annotating two and five group BBHs trees, we did not expect a large difference in the number of LECA families between the two sets of trees. We obtained a few more LECA nodes with BBHs between five supergroups. When using a different root this changes, because excavate sequences are not regularly involved in these BBHs unless you specifically search for them (Extended Data Fig. 2a). The main difference between the two sets of trees is that there are more sequences in trees based on 5 group BBHs, which poses a challenge for resolving *heavily expanded families. Because of the higher support for LECA and duplication nodes, we chose the Opimoda-Diphoda BBHs set as our main set.*

Supplementary Materials line 75: “Such a pattern would fit into a ‘Big Bang’ hypothesis for eukaryogenesis”. This does not appear logical. If an accelerated rate following a duplication occurs, it makes it difficult to place the duplication, but why would one conclude that these happened all at the same time? In many gene families with more than one duplication the different substitution events are clearly separated from one another.

The phrasing and the quote were indeed a bit misleading. Some of the duplications might reflect larger-scale events, but our results indeed also support eukaryogenesis as a “drawn-out process”. We decided to remove this paragraph.

Line 36: Given that prokaryotes of enormous size (Thiomargarita namibiensis and Epulopiscium fishelsoni) and with endomembrane system (Gemmata obscuriglobus) are known, some qualifier should be added.

We completely agree. Although phrases like the one we used are common in eukaryogenesis papers, it does not make it justified. We made it more nuanced.

The ridgeline plots represent the data after some smoothing of the curves was applied. More details on how this was done and on how the original data look as histograms should be provided.

A ridgeline plot is basically a violin plot but then halved and overlapping. The distributions are calculated using kernel density estimation. We included the approach to create the ridgeline plots in the Methods section. Ridgeline plots allow for easy comparison of density shapes and relative heights across groups (https://serialmentor.com/dataviz/boxplots-violins.html -boxplots-violins-horizontal). With histograms this would only look good if you would plot all groups below each other in different panels, because overlapping histograms are difficult to compare.

The bandwidth parameter has a large effect on the smoothing. Histograms have a similar issue with the bin size and are as arbitrary in that respect. As expected, histograms give exactly the same shape of the distribution, especially with a small bin size. The main issue with these density plots is that they can be misleading with small sample sizes. That is why we did not include categories with very few duplications, e.g. nucleotide transport and metabolism (n=17) and cell wall/membrane/envelope biogenesis (n=14) in Fig. 4a.

The given links to additional data and code did not work.

That is strange. We checked the links and they worked (https://github.com/JulianVosseberg/feca2leca and https://doi.org/10.6084/m9.figshare.10069985.v2).

Reviewer #3 (Remarks to the Author):

Many open questions remain regarding the origin of eukaryotes from prokaryotic ancestors. Among those, the order in which eukaryotic specific features originated, and in particular the timing of mitochondrial endosymbiosis relative to the emerge of cellular compartmentalization, is a matter of heated debate in the field of evolutionary biology.
In this paper, Vosseberg and colleagues try to address this question by using a phylogenomic approach to focus on the evolutionary history of duplicated genes. They argue, among others, that the genes that duplicated the most, and before the mitochondrial endosymbiosis, were inherited from Asgard archaea and are involved in cytoskeleton and membrane trafficking, suggesting that cellular complexity would have evolved prior to the acquisition of mitochondria.

It is a topic of major importance that deserves scrutiny and that I believe is of interest to the broad community of evolutionary biologists. However, I have genuine concerns about this manuscript that I outline below. Overall the methods employed lack clarity and justification; this has put me in a frustrating position where I can discuss the results themselves in any details because I was not able to grasp clearly what the authors have done and why. This kind of large-scale analysis relying entirely on the sanity of the methods, I do not feel confident about the reliability of the results. To clarify, I do not argue that the results are incorrect, I argue that the manuscript needs to be heavily edited in ordered to allow reviewers to assess those results.

I strongly encourage the authors to carefully detail the method section and provide justification for the choices made, which sometimes seem completely ad hoc.

It is a pity to hear that the reviewer was not able to grasp what we have done and why, and we apologize for that. We thank the reviewer for pointing out paragraphs and sentences that needed clarification. We included more explanations and reasons behind the choices that we made.

*** Major comments ***

Core assumption: My main concern is that all the conclusions made here are based on the assumption that genes evolve somewhat clockwise and that branch length can be used as a proxy to date evolutionary events relative to one another. Although the authors do take into account the fact that some gene families evolve faster than other and try correct across gene-family rate variations by using post-LECA branch lengths, it is not clear at all that this is sufficient to make an accurate estimation of the time of duplication. In fact, the author themselves discuss this in the supplementary material: given that a duplication event gives birth to two new clades A and B, there is two branches whose (corrected) length can be used to estimate the age of this duplication – the branch leading from the duplication node to LECA in clade A, and the same in clade B. In theory, both of those should lead to a similar estimate of the duplication age. However, the authors discuss the fact that,
when testing those multiple measures, "the shortest length was most consistent with the branch lengths obtained from non-duplicated acquisitions (Extended Data Fig. 5a). It should be noted that even using the shortest could not fully account for the difference in stem lengths." and further that "This indicates that in most cases there was an accelerated

evolutionary rate in at least one of the paralogues, which could not be (fully) corrected for by the post-LECA branch lengths.". Hence, if the branch length normalization for duplicated genes is not quite comparable to the one made for "acquisitions", this raises the a very important question which is: how can you be sure that you can compare the relative age estimates made for duplications versus acquisition events? This is an absolutely crucial point to make to be able to claim that genes of Asgard origin duplicated mostly prior to the acquisition of the mitochondrial endosymbiosis.

We infer from the comments of the reviewer that the result in the first two-thirds of the manuscript, which detail the inferences we can make from our novel phylogenomics of eukaryogenesis, is although perhaps not clear, at least sufficiently substantiated. And the reviewer is thus mostly worried about the branch lengths. We think we understand his major comment and try to address it as follows.

The branch lengths are not intended to be interpreted as precise time estimates but they provide a sufficiently good proxy for relative time. When it comes to the trends we observe for the stem lengths, which we use as a proxy for the timing of acquisition (Fig. 3, lower plots), those can either result from a shared rate (change) in proteins of the same phylogenetic origin or can be due to different time points of acquisitions. Pittis and Gabaldón (2016) showed that the latter explanation is the most plausible one. In contrast to their paper, we here however also use stem lengths with duplications, indeed raising the possibility that those elicited an acceleration due to which we overestimate their stem lengths. We indeed confirm that duplication can result in an increased stem length: even when we only consider the shortest branch after a duplication, a clear increase in stem length can be observed in duplicated alphaproteobacterial families (Extended Data Fig. 5a). We see a difference for bacterial families in general, although this is more difficult to interpret because duplicated families could have been acquired earlier (data not shown). Similarly, we observe this difference between duplicated and non-duplicated families within the vertebrate lineage (Extended Data Fig. 5f). Contrasting the results in our first submitted manuscript, for Asgard archaeal families we no longer see an increase in stem lengths for duplications in the new set of trees in which we also reduced the number of Asgard archaeal sequences (see below) (Extended Data Fig. 5b). Likewise, when we zoom in on the duplications in genes acquired from the Asgard archaea, we do not see a clear pattern of rate acceleration for different groups of proteins (Extended Data Fig. 5c-e).

When it comes to duplication lengths, we tested if these can be compared to one another by measuring them for more recent duplications, for which we know their age (i.e. the speciation event prior to which they occurred). In this analysis, the duplication lengths of duplications of different age classes were clearly separated (Extended Data Fig. 5g). Moreover, the effect of the time point of duplication on duplication length was larger than the effect of function (Extended Data Fig. 5h). Therefore, we consider it valid to compare duplication lengths among each other, as we do in Figure 3 (upper plots) and Figure 4.

When it comes to comparing stem lengths to duplication lengths, we are indeed not sure if they can be compared directly if there is a slightly increased rate in duplicated families. However, we want to emphasise that the effect of duplications on stem lengths is largest for (alphaproteo)bacterial acquisitions and more subtle (if present) for archaeal acquisitions. To compare the timing of duplications to the timing of mitochondrial endosymbiosis, we can not only use the alphaproteobacterial acquisitions but also duplications that are intrinsically linked to the mitochondrial endosymbiosis: energy production (Fig. 4a) and the mitochondrion (Fig. 4b). Cytoskeletal and intracellular trafficking *duplications had longer branches than these (Supplementary Fig. 7 and 8), indicating that the establishment of a dynamic cytoskeleton and intracellular trafficking machinery predated the integration of the endosymbiont into the host.*

This should absolutely be discussed in the main text, and not vaguely mentioned in the supplementary. As currently written, the main text presents this method as if it had proven to be sound, which is not the case so far.

We added a detailed discussion about the use of branch lengths and duplications in the results section (see 'Using branch lengths to time acquisitions and duplications').

* Conclusion of the abstract: the conclusions of the abstract need to be toned down, particularly the sentence that says "we demonstrate that the host that engulfed the proto-mitochondrion had some eukaryote-like complexity". "Demonstrating" is way too strong of a word in evolutionary biology in general, and furthermore when the results are extrapolated from novel methods that remain to be proven sensible.

We replaced the word "demonstrate" with "infer".

* General comment: most of the text needs to be rewritten in a MUCH clearer way. So many sentences are overly long or imprecise, it makes it very difficult on the reader to follow what the authors are getting at. I really cannot stress this enough: please, carefully go over the text and expend and/or clarify.

We critically went over the text and rewrote long and/or unclear sentences.

* The taxon reduction procedure might lead to potential bias towards Asgards archaea inferences:

The whole section about reducing the number of sequences in trees were reduced is very opaque. First when it comes to reducing the number of prokaryotic sequences, the authors say "For each Pfam, the number of prokaryotic sequences was reduced with kClust v1.0 34 using a clustering threshold of 2.93. Asgard archaeal sequences were excluded from this reduction, because they are relatively undersampled and are already genetically diverse."
- Is there a rationale behind choosing a threshold of 2.93?
- More importantly, I do not understand what justifies separating the Asgard sequences from the rest. If indeed they are genetically diverse, then shouldn't they remain anyway represented after clustering? And if not, then you should better explain the criteria that are used by Kclust and how that would bias the representation of Asgard sequences. But if there is a bias after clustering, shouldn't the authors be concerned that it will after clades other than Asgards? And that because now Asgards receive a special treatment, that this will skew the signal they later recover?

We expanded this part of the methods section. We chose the threshold of 2.93, which corresponds to a sequence identity of 60%, because we expected it to retain sufficient prokaryotic diversity while removing sequences from related species to keep the analysis computationally feasible.

The exclusion of the Asgard archaeal sequences from this reduction is a fair concern. That is why we decided to combine these sequences with the other prokaryotic sequences and perform the kClust reduction on the combined set. For Pfams that had a different set of prokaryotic sequences selected, we inferred phylogenetic trees again. This new set is used for the results in the main text and other additional analyses.

We did not see drastic changes but there were some effects: (1) slightly fewer prokaryotic (8.2%) and Asgard (4.5%) acquisitions, (2) the multiplication factor for Asgard archaeal families is lower than before and not significantly different from "prokaryotic" and other archaeal families, (3) the prokaryotic stems are shorter and bacterial stems are longer (difference between bacterial sl and alphaproteobacterial sl is now statistically significant), (4) more often instead of an Asgard phylum a specific Asgard archaeon as sister group and (5) difference between duplicated and non-duplicated Asgard archaeal stem lengths is no longer significant.

Due to the inclusion of Asgard archaea in the kClust reduction, we mainly see fewer sister groups comprising both bacteria and archaea ("prokaryotic"). This probably reflects the high proportion of bacterial genes in Asgard archaeal genomes from recent inter-domain HGT events (Spang et al., *2015). Most of these acquisitions are now probably classified as bacterial acquisitions, resulting in the small shift in prokaryotic and bacterial stem lengths.*

Moreover the whole section describing the eukaryotic taxon reduction is really hard to understand. Beyond explaining the step by step approach, please clarify what the goal is for each of them. In addition, please briefly explain what ScrollSaw is. Finally, please explain 1) the goal of identifying birectional best blast hits and 2) why you did this either between different supergroups OR between Diphoda and Opimoda, and how did that impact your results?

We included an explanation and motivation of the ScrollSaw approach. We identified BBHs between different supergroups and Opimoda-Diphoda for comparative reasons. Because the LECA and duplication nodes were better supported in the Opimoda-Diphoda BBHs trees, we decided to use that set. We obtained the same higher support for BBHs between two groups in another study (Van Wijk and Snel, preprint on bioRxiv: https://www.biorxiv.org/content/10.1101/2020.01.27.920793v2).

* KOG-COG clusters analyses are not understandable:
I have spent a lot of time reading this section and still cannot really grasp what the authors are doing.

We rewrote this section.

However, my main concern about this section is that you define a "duplication score" and a "LECA score" in order "to infer reliable duplication nodes" and as to infer gene families that were likely in LECA. What is really concerning to me is that you test out different cutoff values for those scores – including zero – and say that "it did not have a large impact on the absolute numbers and quality measures, such as the fraction of well-supported LECA and duplication nodes. This underlines the robustness of our analysis. What is the validity of such a score if even when it is zero does not change your results? Does this really prove the robustness of the analysis, or that there is a lot of randomness to it?

The consistency shows that the results were not contingent on the specific set of thresholds chosen and that for most nodes the duplication consistency and LECA coverage was high. We changed the wording to make this clear. Note that, without the coverage threshold, we still required that at least one Opimoda and one Diphoda sequence should be present in the clade in order to annotate it as LECA node. We used a threshold in order to not call every node that contains both an Opimoda and Diphoda sequence a LECA node. These could also be the result of tree artefacts or, less likely, HGT events between eukaryotes.

Other comments about this section:
- First of all, please explain what the point of this analysis before going into the griddy details.

We assume this refers to using the KOG-to-COG clusters. We used this set of trees to verify our method and check if we could obtain the same patterns as observed before (Makarova et al.*, 2005). In order to clearly separate this from the main analysis, this analysis is moved to the Supplementary Methods.*

- Here again, why do you use a different approach for Asgard archaea than for the rest of archaea? Sure, they uniquely encode some "eukaryotic" proteins, but they are not the only ones. Other archaeal lineages encode such eukaryotic proteins, some to the exclusion of Asgards.

For the other prokaryotes we could simply retrieve the COG assignment directly from eggNOG. We had to do this ourselves for the Asgard archaeal proteins, because these are not included in the eggNOG database that we used. We only included prokaryotic (COG) profiles, so we did not search for eukaryote-specific proteins in Asgard proteomes in the KOG-to-COG analysis.

* Tree analyses
- Here again, Asgard archaea seem to receive an unfair treatment. As much as I recognize the relevance of those taxa for this question, I think the authors need to consider that they are potentially biasing their results by not applying the same criteria to other lineages.

This comment could refer to either the removal of interspersing prokaryotes or to the sister group identification. For the discussion of the first, see our response to reviewer #1. If one potential sister group only contains Asgard archaea and the other a diverse set of bacteria or archaea, we think it is logical, given what we know about the prokaryotic lineages involved in eukaryogenesis, to consider the Asgard archaea the most likely 'real' sister clade. The same applies to alphaproteobacteria.

*** Minor comments ***

In many place the wording is just not precise. Some examples:

l. 57-58 "we attempt to reconstruct the successive stages of eukaryogenesis by systematically analysing large sets of phylogenetic trees." Be more specific: phylogenetic trees corresponding to what? Sure, it becomes apparent later, for this should be clear already in the introduction. Same comment for "the scale of gene inventions and duplications"… in which part of the tree? Be specific.

l. 73 “To include genes having only small Pfam domains, which were excluded for computational reasons, or no domains at all, we used a linear regression analysis to obtain an estimated LECA genome containing 12,780 genes”. What does this mean? The corresponding section in the methods is equally unclear. I can guess what it means but the method description is overally vague, eg. “. The assumptions of a normal distribution of gene values at each Pfam domain value and equal variance were reasonably met after log transformation.”

- l. 434 to 440 needs to be written more clearly, it is currently jumping around and is very difficult to follow with all the mentions of euNOGs, TWOGs, ENOGs, KOGs… Please guide the reader in understanding the relevance and difference in using each of those.

- l. 470 “For the annotation of nodes in trees from the Pfam-ScrollSaw sequences the information from the eukaryotic sequences that were not in the between-supergroup or Opimoda-Diphoda BBHs were included.” What does any of this mean?

- l. 484 “the mean presence of a potential LECA family in eukaryotic species [...] should be at least 15%.” What does that mean? And why? And how did you “ weight so that each supergroup contributes equally”?

We rephrased all paragraphs and sentences pointed out by the reviewer.

- l. 492 the other mention that “For the annotation of nodes in KOG-COG trees a slightly different approach was followed.” which made me wonder how and why there are two types of trees: the Pfam-ScollScraw and the KOG-COG trees?

See our previous answer.

- l. 497 “If there were duplication nodes in both daughters, this node had to be a duplication node as well even though its duplication consistency score was below the threshold.” How can this be a duplication node if its duplication score is below the threshold? Please clarify.

A duplication node can be a duplication node if both of its daughters fulfilled the duplication criteria, but the node itself does not. In those cases the species overlap in the node was not sufficient to meet the threshold, but based on the duplications in both daughter lineages it must have been a duplication node as well. This pattern is very rare: it was only the case for two nodes in total.

l. 538 “Eukaryotic clades with LECA nodes that were nested, i.e. they had exactly the same prokaryotic sister group, were merged.” What does that mean?

They shared exactly the same prokaryotic sister group because one eukaryotic clade had in its sister clade only one prokaryotic clade and another eukaryotic clade: (acquisition, (acquisition, sister group)). In that case, the first and the second eukaryotic clades shared the same sister.

Suppl. Mat. “For both, normalisation increased the number of duplications predating mitochondrial endosymbiosis, making the mitochondrial acquisition a slightly earlier event.” Isn’t that contradictory? If you increase the number of duplications predating the mitochondrial acquisition, then this makes the latter a later event, no?

The reviewer is entirely correct; it should be “later”. We apologise for the mistake.

The figure legends for Suppl Fig 1 – 10 need to be greatly expended for the reader to understand them. I have spent way too much time guessing what they represent and I’m still unsure about what to conclude from them.

We changed the legend and removed the single-letter abbreviations. We included the significance for Figures 2 and 3 in the main figure itself.

Decision Letter, first revision:

12th August 2020

*Please ensure you delete the link to your author homepage in this e-mail if you wish to forward it to your co-authors.

Dear Toni,

Your manuscript entitled "Timing the origin of eukaryotic cellular complexity with ancient duplications" has now been seen by our reviewers, and in the light of their advice I am delighted to say that we can in principle offer to publish it. First, however, we would like you to revise your paper to address the final points made by the reviewers, and to ensure that it is as brief as possible and complies with our Guide to Authors at http://www.nature.com/natecolevol/info/final-submission.

Specifically, we agree with Reviewer #2 that you should better discuss the uncertainty of the placement of duplications in the tree.

TRANSPARENT PEER REVIEW
Nature Ecology & Evolution offers a transparent peer review option for new original research manuscripts submitted from 1st December 2019. We encourage increased transparency in peer review by publishing the reviewer comments, author rebuttal letters and editorial decision letters if the authors agree. Such peer review material is made available as a supplementary peer review file. <b>Please state in the cover letter ‘I wish to participate in transparent peer review’ if you want to opt in, or ‘I do not wish to participate in transparent peer review’ if you don’t.</b> Failure to state your preference will result in delays in accepting your manuscript for publication.
Please note: we allow redactions to authors’ rebuttal and reviewer comments in the interest of confidentiality. If you are concerned about the release of confidential data, please let us know specifically what information you would like to have removed. Please note that we cannot incorporate redactions for any other reasons. Reviewer names will be published in the peer review files if the reviewer signed the comments to authors, or if reviewers explicitly agree to release their name. For more information, please refer to our <a href="https://www.nature.com/documents/nr-transparent-peer-review.pdf" target="new">FAQ page</a>.

SPECIFIC POINTS:
In particular, while checking through the manuscript and associated files, we noticed the following specific points which we will need you to address:

1. A one-sentence editorial summary of the paper will appear on the journal homepage with the link to the paper. This is our proposed summary: ‘Combining phylogenomics with analysis of gene duplication to reconstruct the steps during eukaryogenesis the authors show that the Asgard archaea-related host already had some eukaryote-like cellular complexity, which increased further upon mitochondrial acquisition.’ Please let us know of any factual inaccuracies.

2. Please note that we have recently moved from having figures in the supplementary information to having them as Extended Data items, which are linked directly from the main text in the html version of the paper. Please see below for further details of how to submit supporting files and upload the

attached Inventory of Supporting Information.

3. In line 537 you refer to data not shown. Please note that all data must be shown either in the article or in the supplementary information.

4. Please include the name of the statistical test and sample size associated with the P values in Figure 2 and 3.

5. We recommend adding an additional citable reference to your Figshare dataset in the Data Availability section, so that it reads: “The phylogenetic trees and their annotations are available in figshare with the identifier53 doi:10.6084/m9.figshare.10069985” and where Reference 53 should be: “Vosseberg J et al. Data from: Timing the origin of eukaryotic cellular complexity with ancient duplications. Figshare fileset. https://doi:10.6084/m9.figshare.10069985 (2020)”

6. We strongly encourage that all data generated in the study is available in a public repository rather than it being available on request. Please make all data available and change the data availability statements accordingly.

7. Please complete the Editorial policy checklist and the new version of the Reporting Summary (links below) and upload them with your revised manuscript. We will publish the latter along with the paper. Please note that these forms are dynamic ‘smart pdfs’ and must therefore be downloaded and completed in Adobe Reader. Please also ensure that “Final Submission” box is checked.
Editorial policy checklist: https://www.nature.com/authors/policies/Policy.pdf
Reporting summary: https://www.nature.com/authors/policies/ReportingSummary.pdf

GENERAL POINTS:
We will also need you to check through all of the following general points when preparing the final version of your manuscript:

The main manuscript file should include the abstract, main text, methods, author contribution, data availability, code availability and competing interests statements, acknowledgements, references, and figure legends. Figures should be submitted separately as individual files. For details on other supporting material, please see below.

Figures:
Choosing the right electronic format for your figures at this stage will speed up the processing of your paper. We would like the figures to be supplied as vector files - EPS, PDF, AI or postscript (PS) file formats (not raster or bitmap files), preferably generated with vector-graphics software (Adobe Illustrator for example). Please try to ensure that all figures are non-flattened and fully editable. All images should be at least 300 dpi resolution (when figures are scaled to approximately the size that they are to be printed at) and in RGB colour format. Please do not submit Jpeg or flattened TIFF files. Please see our guidelines https://www.nature.com/documents/NRJs-guide-to-preparing-final-artwork.pdf for more details, and also our image policies http://www.nature.com/authors/editorial_policies/image.html.

We will edit your figures/tables electronically so they conform to Nature Ecology & Evolution style. If necessary, we will re-size figures to fit single or double column width. If your figures contain several parts, the parts should be labelled lower case a, b, and so on, and form a neat rectangle when assembled.

Figure legends must provide a brief description of the figure and the symbols used, within 350 words. This must include definitions of any error bars employed in the figures.

Should your Article contain any items (figures, tables, images, videos or text boxes) that are the same as (or are adaptations of) items that have previously been published elsewhere and/or are owned by a third party, please note that it is your responsibility to obtain the right to use such items and to give proper attribution to the copyright holder. This includes pictures taken by professional photographers and images downloaded from the internet. If you do not hold the copyright for any such item (in whole or part) that is included in your paper, please complete and return this <a href="http://www.nature.com/documents/thirdpartyrights-origres.doc">Third Party Rights Table</a>, and attach any grant of rights that you have collected.

Please check the PDF of the whole paper and figures (on our manuscript tracking system) VERY CAREFULLY when you submit the revised manuscript. This will be used as the 'reference copy' to make sure no details (such as Greek letters or symbols) have gone missing during file-transfer/conversion and re-drawing.

Supporting Information:
All Supporting Information must be submitted in accordance with the instructions in the attached Inventory of Supporting Information, and should fit into one of two categories:

1. EXTENDED DATA: Extended Data are an integral part of the paper and only data that directly contribute to the main message should be presented. These figures will be integrated into the full-text HTML version of your paper and will be appended to the online PDF. There is a limit of 10 Extended Data figures, and each must be referred to in the main text, cited as Extended Data 1, Extended Data 2, etc. Each Extended Data figure should be of the same quality as the main figures, and should be supplied at a size that will allow both the figure and legend to be presented on a single A4 page. Each figure should be submitted as an individual .jpg, .tif or .eps file with a maximum size of 10 MB each. All Extended Data figure legends must be provided in the attached Inventory of Supporting Information, not in the figure files themselves.

2. SUPPLEMENTARY INFORMATION: Supplementary Information is material that is essential background to the study but which is not practical to include in the printed version of the paper (for example, video files, large data sets and calculations). Each item must be detailed in the attached Inventory of Supplementary Information. Tables containing large data sets should be in Excel format, with the table number and title included within the body of the table. All textual information and any additional Supplementary Figures (which should be presented with the legends directly below each figure) should be provided as a single, combined PDF. Please note that we cannot accept resupplies of Supplementary Information after the paper has been formally accepted unless there has been a critical scientific error.

Additional Supplementary Figures and other items are not required to be referred to in your manuscript text (though they can be), but should be numbered as Supplementary Figure 1, not SI1, etc.

Methods & Notes:
Please include references for the Methods in the same list as those for the main text, following on sequentially after the main text references. Any citations in the Supplementary Information will need inclusion in a separate SI reference list.

Please include a data availability statement as a separate section after Methods but before references, under the heading "Data Availability”. This section should inform readers about the availability of the data used to support the conclusions of your study. This information includes accession codes to public repositories (data banks for protein, DNA or RNA sequences, microarray, proteomics data etc...), references to source data published alongside the paper, unique identifiers such as URLs to data repository entries, or data set DOIs, and any other statement about data availability. All data that support the findings of the study must be made available. If DOIs are provided, we also strongly encourage including these in the Reference list (authors, title, publisher (repository name), identifier, year). For more guidance on how to write this section please see:
http://www.nature.com/authors/policies/data/data-availability-statements-data-citations.pdf

Nature Research policies (https://www.nature.com/authors/policies/availability.html#data) include a strong preference for research data to be archived in public repositories and in some cases this is mandatory. If you need help complying with this policy, or need help depositing and curating your research data (including raw and processed data, text, video, audio and images) you should consider:

Contacting Springer Nature’s Research Data Helpdesk (https://www.springernature.com/gp/authors/research-data-policy/helpdesk/12327114) for advice.
Finding a suitable data repository (https://www.springernature.com/gp/authors/research-data-policy/repositories/12327124) for your data.
Uploading your data to Springer Nature’s Research Data Support service (https://springernaturedata.typeform.com/to/UeGGKT). Please note there are fees (https://www.springernature.com/gp/authors/research-data-policy/pricing/15499842) for using Springer Nature’s Research Data Support service.

Finally, we require authors to include a statement of their individual contributions to the paper, such as experimental work, project planning, data analysis, etc., immediately after the acknowledgements. The statement should be short, and refer to authors by their initials. For details please see the Authorship section of our joint Editorial policies at
http://www.nature.com/authors/editorial_policies/authorship.html

We will not send your revised paper for further review if, in the editors' judgement, the referees' comments on the present version have been addressed. If the revised paper is in Nature Ecology & Evolution format, in accessible style and of appropriate length, we shall accept it for publication immediately.

Please resubmit electronically

* the final version of the text (not including the figures) in either Word or Latex.

* publication-quality figures. For more details, please refer to our Figure Guidelines, which is available here: https://www.nature.com/documents/NRJs-guide-to-preparing-final-artwork.pdf .

* any Extended Data and Supplementary Information, as per instructed, with the associated Inventory document.

* copies of our reporting and editorial policy checklists even if they have not changed since the previous round of revision.

* a point-by-point response to any issues raised by our reviewers and to any editorial suggestions.

* any suggestions for cover illustrations, which should be provided at high resolution as electronic files. Please note that such pictures should be selected more for their aesthetic appeal than for their scientific content. I am sure you will understand that we cannot make any promise as to whether any of your suggestions might be selected for the cover of Nature Ecology & Evolution.

Please use the following link to access your home page:

[REDACTED]

*This url links to your confidential homepage and associated information about manuscripts you may have submitted or be reviewing for us. If you wish to forward this e-mail to co-authors, please delete this link to your homepage first.

Please also send the following forms as a hand-signed PDF by email to ecoevo@nature.com.

*Please sign and return the <a href="http://www.nature.com/documents/snl-ltp.docx" target="_blank">Licence to Publish form</a>

Or, if the corresponding author is a Crown government employee (including Great Britain and Northern Ireland, Canada and Australia), please sign
and return the <a href="http://www.nature.com/documents/snl-ltp-crown.docx" target="_blank"> Licence to Publish form for Crown government employees</a> , or the <a href="http://www.nature.com/documents/snl-ltp-govus.docx" target="_blank"> Licence to Publish form for US government employees</a>

For more information on our licence policy, please consult http://npg.nature.com/authors.

AUTHORSHIP

CONSORTIA -- For papers containing one or more consortia, all members of the consortium who contributed to the paper must be listed in the paper (i.e., print/online PDF). If necessary, individual authors can be listed in both the main author list and as a member of a consortium listed at the end of the paper. When submitting your revised manuscript via the online submission system, the consortium name should be entered as an author, together with the contact details of a nominated consortium representative. See https://www.nature.com/authors/policies/authorship.html for our authorship policy and https://www.nature.com/documents/nr-consortia-formatting.pdf for further consortia formatting guidelines, which should be adhered to prior to acceptance.

Nature Research journals <a href="https://www.nature.com/nature-research/editorial-policies/reporting-standards#protocols" target="new">encourage authors to share their step-by-step experimental protocols</a> on a protocol sharing platform of their choice. Nature Research's Protocol Exchange is a free-to-use and open resource for protocols; protocols deposited in Protocol Exchange are citable and can be linked from the published article. More details can found at <a href="https://www.nature.com/protocolexchange/about" target="new">www.nature.com/protocolexchange/about</a>.

We hope to hear from you within two weeks; please let us know if the revision process is likely to take longer.

[REDACTED]

****

Reviewer Comments:

Reviewer #1 (Remarks to the Author):

The authors have made a fair effort to address the points I raised about the first version of their ms. I am satisfied that this is an interesting ms. that is likely to stimulate debate and recommend publication. The results on the phylogenetic origins of the genes and gene duplications remain stronger, in my view, than the analyses of stem lengths, but the authors have spelled out the caveats, and have included additional work that explores the sensitivity of those analyses.

Reviewer #2 (Remarks to the Author):

In the revision, the authors provide a reasonable discussion of the impact of the root location on their analyses.

The discussion of rate changes following the duplication event is less convincing, in part because it was difficult to follow the pertinent Materials and Method section (addition of a schematic diagrams might make this section more readable). I still suspect that the inferred high number of duplications close to FECA is in part due to an artifact. The results in extended data Fig. 5 an and b seem to support this notion: the duplicated genes have one peak close to FECA, the non-duplicated genes have a wider distribution for the stem lengths (and this seems also to be the case for 5b, the authors statement “but not for duplicated families from Asgard archaeal origin (Extended Data Fig. 5b)” notwithstanding.
The uncertainty of the placement of duplications on the FECA to LECA could be better articulated in the discussion.

Reviewer #3 (Remarks to the Author):

In this revised version, the authors have made substantial effort to clarify the writing and justify the various approaches and interpretations that were previously unclear.

I believe it is a valuable contribution to the field at large and support its publication.

********************END********************

Final Decision Letter:

28th August 2020

Dear Toni,

We are pleased to inform you that your Article entitled "Timing the origin of eukaryotic cellular complexity with ancient duplications", has now been accepted for publication in Nature Ecology & Evolution.

Before your manuscript is typeset, we will edit the text to ensure it is intelligible to our wide readership and conforms to house style. We look particularly carefully at the titles of all papers to ensure that they are relatively brief and understandable.

The subeditor may send you the edited text for your approval. Once your manuscript is typeset you will receive a link to your electronic proof via email within 20 working days, with a request to make any corrections within 48 hours. If you have queries at any point during the production process then please contact the production team at rjsproduction@springernature.com. Once your paper has been scheduled for online publication, the Nature press office will be in touch to confirm the details.

Acceptance of your manuscript is conditional on all authors' agreement with our publication policies (see www.nature.com/authors/policies/index.html). In particular your manuscript must not be published elsewhere and there must be no announcement of the work to any media outlet until the publication date (the day on which it is uploaded onto our web site).

The Author's Accepted Manuscript (the accepted version of the manuscript as submitted by the author) may only be posted 6 months after the paper is published, consistent with our <a href="http://www.nature.com/authors/policies/license.html">self-archiving embargo</a>. Please note that the Author’s Accepted Manuscript may not be released under a Creative Commons license. For Nature Research Terms of Reuse of archived manuscripts please see: <a href="http://www.nature.com/authors/policies/license.html#terms">http://www.nature.com/authors/policies/license.html#terms</a>
If you have posted a preprint on any preprint server, please ensure that the preprint details are updated with a publication reference, including the DOI and a URL to the published version of the article on the journal website.

An online order form for reprints of your paper is available at <a href="https://www.nature.com/reprints/author-reprints.html">https://www.nature.com/reprints/author-reprints.html</a>. All co-authors, authors' institutions and authors' funding agencies can order reprints using the form appropriate to their geographical region.

We welcome the submission of potential cover material (including a short caption of around 40 words) related to your manuscript; suggestions should be sent to Nature Ecology & Evolution as electronic files (the image should be 300 dpi at 210 x 297 mm in either TIFF or JPEG format). Please note that such pictures should be selected more for their aesthetic appeal than for their scientific content, and that colour images work better than black and white or grayscale images. Please do not try to design a cover with the Nature Ecology & Evolution logo etc., and please do not submit composites of images related to your work. I am sure you will understand that we cannot make any promise as to whether any of your suggestions might be selected for the cover of the journal.

You can now use a single sign-on for all your accounts, view the status of all your manuscript submissions and reviews, access usage statistics for your published articles and download a record of your refereeing activity for the Nature journals.

To assist our authors in disseminating their research to the broader community, our SharedIt initiative provides you with a unique shareable link that will allow anyone (with or without a subscription) to read the published article. Recipients of the link with a subscription will also be able to download and print the PDF.

You can generate the link yourself when you receive your article DOI by entering it here: <a href="http://authors.springernature.com/share">http://authors.springernature.com/share<a>.

Yours sincerely,

[REDACTED]

P.S. Click on the following link if you would like to recommend Nature Ecology & Evolution to your librarian http://www.nature.com/subscriptions/recommend.html#forms

** Visit the Springer Nature Editorial and Publishing website at <a href="http://editorial-jobs.springernature.com?utm_source=ejP_NEcoE_email&utm_medium=ejP_NEcoE_email&utm_campaign=ejp_NEcoE">www.springernature.com/editorial-and-publishing-jobs</a> for more information about our career opportunities. If you have any questions please click <a href="mailto:editorial.publishing.jobs@springernature.com">here</a>.**