doi:10.1038/nature16941

Table of Contents

1.Testing functional bias....1
2.Cyanobacterial signal in primary plastid-bearing eukaryotes....3
3.Alternative methods and datasets....5
4.Effects of database taxonomic representation and HGT....8
5.Control for other biases....10
6. Lokiarchaeota....11

1. Testing functional bias

The rate of protein evolution can be very different in different protein families. As explained in the methodology section, our normalization or the *raw stem length* by the eukaryotic branch length is precisely an attempt to correct for these differences. A necessary assumption in this correction, however, is that the evolutionary rates in the lineages preceding and post-dating LECA are correlated (i.e. not necessarily constant). Unequal shifts in rates would violate that assumption and may compromise our conclusions if they act in such a way that proteins of alpha-proteobacterial descent would have faster rates after LECA, as compared to their rates before LECA. However, several observations and controls, reject that possibility, as we explain here.

Firstly, we see no reason why the genes contributed by the mitochondrial ancestor would show slow evolutionary rates during the transformation of a free living organism to an organelle (mitochondrion) before LECA, while they would accelerate within the diversified main eukaryotic groups. It is important to note that pre-LECA branches refer to the "stem phase" a period in which those families had already been incorporated into (or inherited by) the "host", who later diversified into the current eukaryotic groups. Indeed, what we observe, is short *raw stem length* values in alpha-proteobacterial derived families compared to families of different inferred origin and average branch lengths within eukaryotes. Thus, their short stems are by no means an artefactual result of the normalization. Similarly, families with an inferred archaeal origin show a pattern of much longer stems than bacterial ones, while their branch lengths within eukaryotes appear similar. Secondly, it is commonly assumed that selection is more relaxed in operational genes as compared to informational genes. However, the protein families inferred to LECA are all widespread and mostly participate in universal pathways and

functions, information or metabolism related, which makes them more conserved than lineage specific families.

It is well established, and also supported by our results (see, for instance, Fig. 3), that, in eukaryotes, informational genes originate predominately from Archaea while operational genes largely derive from Bacteria[31]. To ensure that function is not a confounding factor driving our results, we tested for potential biases caused by different functionalities among families of different origin. Initially, we confirmed that the observed differences in *stem length* between Archaea and Bacteria were not due to different functions. For that, we compared the distribution of *stem length* between the two domains across the various functional categories, and found that differences in *stem length* remained significant, when the sample sizes to compare were sufficiently large (Extended Data Fig. 2a). In addition, looking at the two components of *stem length*, the *raw stem length* and the median of *eukaryotic branch lengths*, we also found significant differences in the former but not in the latter for the different categories. These results indicate that the observed differences in *stem length* were due to the differences in *raw stem length* values rather that the lengths within eukaryotes, used for the normalization.

We further tested this result by comparing the *raw stem lengths* among families being subject to similar selective pressures, as estimated by d*N*/d*S* values of representative sequences within the families. Because our dataset includes fairly distant species and non synonymous substitutions cannot be reliably estimated beyond a distance due to saturation, we based our analysis on d*N*/d*S* estimates provided by Ensembl (Ensembl release 81, Ensembl Fungi release 28, Ensembl Plants release 28) for representative extant species: Human, the plant *Zea mays* and the fungus *Aspergillus nidulans*. In all cases d*N*/d*S* values are computed from pairwise species comparisons: Human to macaque; *Aspergillus nidulans* to *Aspergillus niger;* and *Zea mays* to *Sorghum bicolor,* respectively. In most cases the corresponding d*N*/d*S* ratios had a value much below 1, suggesting strong purifying selective pressure, something expected given the widespread distribution and ancient origin of LECA families (see also Extended Data Fig. 2b). Then we formed matched groups of different origin but with similar d*N*/d*S* values. Briefly, we started from a randomly selected family of archaeal origin, then picking the bacterial-derived family that was closest in terms of d*N*/d*S* within a given threshold (0.01) and continued, without replacement, until exhaustion. Then, for the selected families of matched d*N*/d*S*

values, we compared their corresponding *raw stem lengths* (Extended Data Fig. 2b). The results showed that across all the different datasets used, families of archaeal origin showed significantly longer stems than those of bacterial origin for the same d*N*/d*S* values, indicating again that the differences in stems are not due to difference in selective pressures, as measured in extant organisms. The above difference was valid also when we restricted our analysis only to families involved in informational processes (P=1.1e-2 for human genes' d*N*/d*S*) or even more specifically to "Translation, ribosomal structure and biogenesis" (P=1e-3 for human genes' d*N*/d*S* values) suggesting that the differences in stems are not confounded by the strength of selection.

Finally, we evaluated the effect on the *raw stem length* of number of protein-protein interactions (i.e. degree of connectivity) and expression levels, two factors that have been previously associated to the evolutionary rate of protein families[32,33]. In the case of protein connectivity, we based the analysis on the number of known (experimental data) protein-protein interactions, as provided in the STRING database[34]. A degree of connectivity was assigned to each LECA protein family as the median of the number of interactions of each member. In the case of expression levels, gene expression data for human genes were retrieved from release 13.1 of the Bgee database[35] (GSE30611 experiment). An RPKM value was assigned to each protein family, as the median of RPKM values of the genes corresponding to the family members, across different tissues and conditions. Using the same procedure to form matched groups of different origin but similar degree of protein-protein connectivity / expression levels values (using a threshold of 10), we found as before that families of archaeal origin showed significantly longer *stems* than those of bacterial origin (P=1.4e-9 and P=5.4-12 respectively, Extended Data Fig. 2c), a result that suggests again that the observed *stem* differences are not confounded by the protein functional patterns.

2. Cyanobacterial signal in primary plastid-bearing eukaryotes

One major problem in the analysis of the phylogenetic signals mapped to LECA is the lack of events that can be used as reference points in the time during eukaryogenesis (i.e. a "positive control"). The cyanobacterial endosymbiosis is the only major event of acquisition of a set of genes with a particular phylogenetic affiliation that we can assume with certainty that occurred after LECA and the divergence of the main eukaryotic groups, in the root of the lineage that gave rise to photosynthetic eukaryotes

(common ancestor of Viridiplantae, Glaucophytes and Red algae). In fact, genomes of plants and other groups that acquired plastids through secondary or tertiary endosymbiosis carry a cyanobacterial signal, coming from the genes contributed by the cyanobacterial plastid ancestor. If our approach works as expected, we should be able to distinguish differences in time of acquisition of cyanobacterial and alpha-proteobacterial derived genes, so that cyanobacterial derived genes should be more recent. Note, in addition that mitochondria and cyanobacteria are both organelles playing key important roles in energy metabolism, and sharing relevant past evolutionary processes such as endosymbiotic gene transfer (EGT), and retargeting of protein localization. We therefore used the cyanobacterial signal in plants and red algae as a control to evaluate our methodology. For that, we used the same dataset as before, but in this case, from each NOG from the eukaryotic sequences, we selected only those from Archaeplastida (Green plants and the red alga *Cyanidioschyzon merolae*) and all prokaryotic sequences. Using the exact same standard phylogenetic pipeline as before, we reconstructed ML phylogenetic trees and inferred the nearest prokaryotic neighbour for all the Archaeplastida families that formed a monophyletic clade with three or more members and were not specific to one only specific group of Viridiplantae. First, as we expected, the distribution of different signals was similar to LECA for all prokaryotic groups other than cyanobacteria, which was the dominant one with 145 sequences (Extended Data Fig 3a). When we compared the inferred lengths of alpha-proteobacterial and cyanobacterial families (Extended Data Fig 3b), we found both the *stem length* (P=1e-3, two-sided Mann-Whitney U test) and *the raw stem length* (P=4.9e-2) being significantly shorter for cyanobacteria, pointing to its more recent acquisition. Looking at the *raw stem length* values we also found the cyanobacterial families having significantly lower values but contrary to this pattern the lengths within the eukaryotic clades being significantly higher. Indeed it has been demonstrated in previous work that plastid DNA exhibits elevated evolutionary rates[36,37], especially in primary plastid-bearing species, which may be reflected in the median of the eukaryotic branch lengths that we observe. We then compared the *stem lengths* of the families of different inferred bacterial origin and we found, as we expected, the values for the cyanobacteria derived families being smaller than the all the others (Extended Data Fig 3c). When we looked only at the plastidial proteome, as the mitochondrial proteome, it appeared to be composed overall of families with shorter *stem lengths* than the endomembrane system and the nucleus (Extended Data Fig 3d). Also, similarly to what we had observed for mitochondria, cyanobacterial derived families displayed shorter *stem lengths* as compared to plastid proteins of different phylogenetic origins (P=2.9e-4, permutation test 10^6 randomizations),

pointing to a more ancient origin of part of the plastid proteome, which eventually retargeted to the organelle. The incorporation in the organelle of proteins already available in the host of the cyanobacterium, probably explains why we do not detect a significant difference in *stem lengths* between mitochondrial and plastidial proteomes which are a chimera of proteins of endosymbiont and of other origins. The above results support the reliability of our approach in disentangling the relative order of deep ancient events by the evaluation of the evolutionary signal carried by modern protein families.

3. Alternative methods and datasets

Our analyses rely on the assessment of tree topologies and branch lengths. Both measurements depend on the methods and parameters used in the process of phylogenetic reconstruction: the selection of sequences, the alignment algorithm, the underlying evolutionary model, and the algorithm used for the phylogenetic reconstruction. We recognize the critical effect that alternative phylogenetic reconstruction approaches could have in the outcome of our analyses, and therefore we used several alternative methodological approaches to evaluate the robustness of our main result (i.e. that alpha-proteobacterial derived LECA proteins have shorter stems than other bacterial derived proteins).

3.1. Alternative LECA definitions

One of the main problems in analyses requiring a reliable inference of LECA's genetic repertoire is that the root of the eukaryotic tree of life remains highly debated[38–41]. As a result, inferences of LECA repertoires can only be approximate and usually reflect a minimal estimate. In addition, LECA's inferred proteomes vary depending on the specific rules applied to decide which protein families descend from LECA. Earlier analyses using a parsimony approach required putative LECA descendants to be present across various, diverse groups of eukaryotes. As pointed by Rochette et al.[8], the design used by Makarova et al.[42] was permissive as the selection criteria could be met for opisthokonta-specific genes, and in the analysis of Thiergart et al.[9] protists were not considered. In the Rochette et al. analysis a family was inferred to descend from LECA, if it was present in at least two groups out of Plantae, Unikonts, and Chromalveolates plus Kinetoplastids. The fact that their criterion could be met for genes shared exclusively by Plants and Chromalveolates possibly inflated the inferred cyanobacterial component of LECA, since such criteria could be met by proteins incorporated in the base of Viridiplantae through the primary plastidial endosymbiosis and in Chromalveolates through

secondary endosymbiosis[43,44]. As explained in the main text, we here used more stringent rules, and required putative LECA descendants to be present in Unikonts, plus at least one group among Viridiplantae, Chromalveolates and Excavates (Unikonts+1, Extended Data Fig. 4a). In addition we evaluated the effect of more stringent definitions in the final result. Considering only families that were present in more than one group apart from Unikonts (Unikonts+2, Unikonts+3) led to an unbalanced reduction of the number of families mapped to Bacteria, as compared to Archaea, indicating a higher plasticity of the bacterial gene families, dominated by more common lineage-specific losses throughout eukaryotic evolution (Extended Data Fig. 4b). Despite this variability, our main result was maintained. The stems of alpha-proteobacterial-derived proteins were significantly shorter than those in proteins derived from other bacteria in both Unikonts+1 and Unikonts+2 definitions with P values 4e-3 and 2.5e-2, respectively (permutation test with 10^6 randomizations, Extended Data Fig. 4c). An extremely stringent criterion (present in all 4 major eukaryotic groups, Unikonts+3), given the reduced genomes of many groups of protists, resulted in marginally insignificant differences (P=9.9e-2), likely due to lack of statistical power in the analysis due to the reduction to only 433 families in the overall set. Considering the fact that our less stringent criterion was already stricter than the ones used in most previous studies, together with our larger sampling of genomes from microbial eukaryotes, we expect our result to be robust to future addition of newly sequenced eukaryotic genomes.

3.2. Alternative datasets and branch support thresholds

In order to test the effect of using a completely different approach, we re-evaluated our results using the data provided in the analysis of Rochette et al.[8]. In the above analysis, the authors use a different sequence data source (HOGENOMv5 database[45]), as well as completely different workflows for sampling sequences in LECA families and phylogenetic reconstruction. Firstly, we used directly the phylogenetic trees computed in the Rochette et al. analysis as provided in the publication's supplementary material, and applied the same exact pipeline we applied to our data for inferring the nearest prokaryotic neighbour / sister group, assigning eukaryotic families to LECA, and measuring stem lengths (“Hogenom - original pipeline”). Secondly, we repeated the analysis with the sequence sets used by the authors but we performed the phylogenetic reconstruction as described in our main analysis (“Hogenom - main pipeline”). In both cases, the mappings to the various archaeal and bacterial taxonomic groups was similar to ours, and the *stem length*'s analysis yielded strongly significant differences between eukaryotic families of alpha-proteobacterial origin and other

non-alphaproteobacterial bacterial groups (Extended Data Fig. 5a). This result strongly supports the robustness of our observations across independent datasets and methods.

Furthermore, given the noisy nature of gene families' phylogenetic reconstruction, we evaluated the effect of more stringent branch support thresholds. Our results show that applying increasingly higher support values from 0.5 up to 0.9 does not affect our main result, and the alpha-proteobacteria stem signal remained significantly shorter compared to the whole bacterial signal's distribution (Extended Data Fig. 5b). At the 0.9 threshold, the dataset was extremely reduced to only 354 bacterial-derived families, which resulted in a lack of statistical support. Considering that the trend is conserved among a wide range of support values, we consider that our conclusions are not due to lack of sufficient phylogenetic signal. In addition, we expect that noise blurs the signal rather than creates a specific one.

3.3. Alternative phylogenetic methods and sequence sampling of EggNOG v4

One consequence of the growth of sequence databases is the high degree of redundancy (*i.e.* sequences from alternative strains of the same species), leading to orthologous group of sizes too large to be used to build accurate phylogenetic trees. As described in the Methods section, we selected 37 eukaryotic species across all eukaryotic groups and the 692 prokaryotic genera/levels present in eggNOG v4 in order to obtain a more balanced, representative sequence set that allowed us to use state of the art phylogenetic methods. Initially, for each of the orthologous groups we sampled randomly one sequence for each of the 729 taxonomic levels defined (37 eukaryotic species plus 692 prokaryotic levels, see Methods). After reconstructing phylogenetic trees using a fast ML method for each of orthologous groups of reduced sampling and detected the eukaryotic monophyletic clades, we extracted these clades together with the rest of prokaryotic sequences and used our main sophisticated and more accurate pipeline for the inference of the sister group and the computation of branch lengths (“main sub-sampling – main pipeline”, see also Methods). This last step was also repeated using faster methods to discard the possibility of biases due to algorithms or models used (“main sub-sampling – fast pipeline”, Extended Data Fig. 5a down left).

In an alternative sampling, we started by reconstructing phylogenetic trees for the whole COG/NOG datasets with a fast ML approach and selected within the large trees (including more than 20,000 sequences in some cases) monophyletic partitions of eukaryotic sequences that fulfilled the criteria for

assignment to LECA. From these partitions we extracted all sequences belonging to the 37 selected eukaryotic species. We also selected prokaryotic sequences present in neighbouring partitions upstream (i.e. towards the root) to the eukaryotic partition, which were inspected one at a time, sequentially. If a given partition contained more than one species of any of the 692 selected taxa, we kept only all sequences for one species per taxon. The selected species per taxon was chosen randomly among those present in the partition. The minimum number of neighbouring partitions was two, and we included additional partitions where necessary to achieve a minimum of 20 selected prokaryotic sequences. The selected subset of sequences was used for a more focused phylogenetic reconstruction, using again our main sophisticated and more accurate pipeline ("alternative sub-sampling – main pipeline", Extended Data Fig. 5a lower right).

Finally, recognizing the difficulty of disentangling deep evolutionary relationships using single gene phylogenies, we evaluated the effect of using the CAT profile mixture model, proposed to provide a better fit than standard matrices on saturated data[46]. Theses models, which account for across-site heterogeneities in amino acid sequences using mixture profiles, have been shown to perform better in cases of datasets with fast evolving sequences and to be less prone to phylogenetic artefacts, such as the Long Branch Attraction (LBA), mainly for big concatenated alignments in a Bayesian framework. Empirical profile mixture models have also been implemented in a Maximum Likelihood framework, providing the possibility to be applied on large scale phylogenomics analyses[47]. We used our original dataset and the same alignments calculated through our standard pipeline, to asses the effect of the C20 profile mixture model, as implemented in the IQTREE software[48]. The distribution of inferred prokaryotic origins for LECA families was very similar as before and the alpha-proteobacteria stems significantly short within bacterial families, indicating that our results are not model-dependent and the signal remains the same across the various evolutionary models used. Extended data Fig. 5a lower middle. In all cases, the trees where the eukaryotic LECA family did not form a monophyletic partition in the last phylogenetic reconstruction step, were considered unreliable and were discarded.

Consistently, in all the different datasets and methods explored, our main result remained robust (Fig. 2b, Extended Data Fig. 5). That is the *stem lengths* detected for the LECA families mapped to alpha-proteobacteria were shorter than expected by chance when compared to the stem lengths of all LECA families mapping to Bacteria. For simplicity we focused our analysis on the first reduced sampling ("main sub-sampling – main pipeline").

4. Effects of database taxonomic representation and HGT

The current sampling of sequenced genomes reflect the larger focus of studies in disease-related prokaryotic groups and model species, but also the possibility of culturing in the laboratory. As a result, some prokaryotic groups, including proteobacteria, are better sampled than others. We tested for the effect of possible sampling biases resulting from an over-representation of alpha-proteobacteria. For this we randomly removed 50% of the alpha-proteobacterial species from the final alignments of families with alpha-proteobacterial inferred origin, and recomputed the phylogenetic trees (HALF alpha sampling). When we compared the *stem lengths* for alpha-proteobacteria of the original dataset to the *stem lengths* deriving after sub-sampling alpha-proteobacteria, we found no significant differences (Extended Data Fig. 6a). We would not expect the over-representation of certain bacterial phyla to have a strong effect in the analysis, as long as other prokaryotic sources of interest are at least fairly sampled.

In addition, the diversity in putative ancestral origins detected in LECA, has been repeatedly attributed to noise caused by phylogenetic artifacts and HGT among prokaryotic species[9,10]. For HGT to result in a mapping to non-alpha-proteobacterial origin in families that were brought to LECA by the proto-mitochondrial ancestor three possible scenarios can be considered. Firstly a gene may have been transferred from alpha-proteobacteria to another lineage, and then be lost from all sampled alpha-proteobacteria except the proto-mitochondrion (eukaryotic lineage) but then we would not expect differences in the inferred *stem lengths* ("post-mito HGT from alpha", Extended Data Fig. 6b). Secondly a gene may have been acquired from another lineage into the branch that led to the proto-mitochondrion and then only survived in the proto-mitochondrial lineage ("vertical transmission / pre-mito HGT from alpha", Extended Data Fig. 6b). Both scenarios imply either a high number of gene losses or HGT being concomitant with endosymbiosis. However, given the large number of families with non-alphaproteobacterial ancestries, and the fact that they are mapped to many alternative lineages, we do not consider very plausible the possibility that all HGT events shortly predated mitochondrial endosymbiosis. To test the phylogenetic effect of massive gene loss of alpha-proteobacterial counterparts in genes descending from this lineage, we simulated the loss of all alpha-proteobacterial sequences in all the families of inferred alpha-proteobacterial origin ("NO alpha sampling"). After removing all alpha-proteobacterial sequences from all the final datasets, we repeated the phylogenetic reconstruction using the same pipeline as before. As expected, a different origin was

now inferred for all these families, mostly among related proteobacteria. However the computed *stem lengths* were significantly higher compared to the initial ones for the families mapping to these other groups (Extended Data Fig. 6c right, P=2.59e-7), suggesting that the observed signal in the first analysis cannot be explained with the loss of homologous genes in alpha-proteobacteria. Lastly a gene may have been transferred from the eukaryotic lineage, after the mitochondrial endosymbiosis, to a bacterial lineage other than alpha-proteobacteria. In the case that this would happen early before the radiation of the main eukaryotic clades (inferred to LECA), we would expect generally shorter stem lengths and certainly not longer ("post-mito HGT from protoeukaryote", Extended Data Fig. 6b).

5. Control for other biases

We tested for other possible biases in our analysis that could have an effect in its outcome. Firstly, we considered the possibility that the change in function and localization of part of the protomitochondrial proteins could contribute to the observed differences in stem lengths. We tested this by comparing the stem lengths of alpha-proteobacterial inferred LECA families annotated with mitochondrial localization and 'Energy production and conversion' function (Extended Data Fig. 6d) against those of the families without the previous annotation (54 vs 25 and 21 vs 58 respectively). In both cases there was no significant difference observed (P values 0.86 and 0.28 respectively for two-sided Mann-Whitney U-test). Secondly, we tested whether the inferred *stem lengths* of families mapped to Rickettsiales within alpha-proteobacteria could bias the result. The specific association of mitochondrial proteins to Rickettsiales has been considered by some as artifactual, with the specific alpha-proteobacterial group from which mitochondria arose remaining highly debated[49–51]. To discard that our differences were solely emerging from the inclusion of families mapped to Rickettsiales, we compared the *stem lengths* of protein families mapped to Rickettsiales (28) and other alpha-proteobacteria (51), and we found no significant difference (Extended Data Fig. 6e, P=0.29, two sided Mann-Whitney U test). Furthermore, after excluding families mapped to Rickettsiales, the alpha-proteobacterial signal remained significantly lower compared to the overall bacterial (P=1.8e-2, permutation test 10^6 randomizations) when excluding families with an inferred Rickettsiales origin. Another control consisted in excluding the groups of Excavates and Chromalveolates from the branch lengths under consideration when computing the median of branch lengths within eukaryotes (ebl). The above groups contain anaerobic species and with intracellular parasites that secondarily lost or transformed mitochondria, with a parallel loss or change of function of the proteins of alpha-proteobacterial origin. This could have

resulted in longer branches within these groups of the sequences that shifted their function, leading to longer branches and consequently higher *stem lengths*. After computing the median of the remaining sequences within the clade, the difference in stem lengths between alpha-proteobacterial and other bacterial families remained significant (P=1.7e-3), which excludes the possibility of a bias resulting from the abovementioned process. Finally, to confirm that the different functionalities do not dominate the reported differences in *stem lengths*, we compared the *stem lengths* of alpha-proteobacterial families across different functional categories and we saw no differences (Extended Data Fig. 6e), supporting that the differences are mostly driven by the phylogenetic origin, rather than the function (see also Extended Data Section 1 and Extended Data Fig. 2).

6. Lokiarchaeota

In the recent analysis of Spang et al.[11], a new archaeal phylum called Lokiarchaeota, discovered by metagenomic sequencing of samples from marine sediments was proposed as the closest known relative of eukaryotes. In phylogenetic trees constructed using gene markers, eukaryotes were placed within the Lokiarchaeota, at the base of the TACK super-phylum. Interestingly one of the main conclusions of the authors, from the analysis of the eukaryotic features with a counterpart in Lokiarchaeum's genome, was that the archaeal host cell of the mitochondrial endosymbiont was already complex and was capable of phagocytosis, thus placing the endosymbiotic event at a late stage of eukaryogenesis and favoring the phagocytic nature of the eukaryotic ancestor. As they state in the paper “emergence of cellular complexity was already underway before the acquisition of the mitochondrial endosymbiont”.

We were therefore very much interested in testing the effect of the newly sequenced genome by including sequences from the archaeon Loki in our analysis. Given the absence of this genome from EggNOG orthologous groups, we added Loki proteins in the following way. After constructing HMM profiles from the extracted, monophyletic families in the last phylogenetic reconstruction step of our workflow, we used these profiles to search for homologs within the 5,384 protein coding sequences encoded in the composite genome, as provided by the authors. The homologs detected were then added in the datasets used in this last step and the phylogenetic reconstruction pipeline was run with these sequences included. Importantly, some of the eukaryotic families previously mapped to other archaeal phyla were now instead mapped to Lokiarchaeota (30 families)(Extended Data Fig. 7a). This points to a significant signal for Lokiarchaeota in the context of gene phylogenies, supporting the conclusions of

the authors' analysis. If Lokiarchaota is the closest Archaeal relative of eukaryotes, as they proposed, we should also observe a reduction of inferred *stem lengths* for their eukaryotic descendants. Indeed we found exactly that (Extended Data Fig. 7b-c). Such a pattern was expected under a model where archaeal families are mapped to other archaeal groups in the absence of Lokiarchaeum (Extended Data Fig. 7d) with increased *raw stem lengths* and consequently *stem length* values. The fact a big part of archaeon Loki's proteome was found in the original publication to have best blast hits in bacteria rather than other archaea, pointing to an extensive bacterial gene flow to its genome, raised the possibility that part of the bacterial component detected in LECA, could have been incorporated through this lineage. In such a scenario, however, eukaryotic sequences and not bacterial would be expected as best hits. We further manually inspected the 30 phylogenetic trees where Lokiarchaeon sequences were inferred as the sister group of eukaryotic families and we found that only 4 of these eukaryotes-Loki clades were nested within bacterial sequences, while the rest where nested within archaeal sequences. Consequently we believe that most of the bacterial component of Lokiarchaeon was acquired independently, after the diversification of eukaryotes. With the current data available we cannot exclude the possibility that the signal observed in other archaeal phyla is due to the still poor sampling of the Lokiarchaeota phylum, which so far consists of a single metagenome.

References

31. Rivera, M. C., Jain, R., Moore, J. E. & Lake, J. A. Genomic evidence for two functionally distinct gene classes. *Proc. Natl. Acad. Sci. U. S. A.* **95,** 6239–6244 (1998).
32. Fraser, H. B., Hirsh, A. E., Steinmetz, L. M., Scharfe, C. & Feldman, M. W. Evolutionary Rate in the Protein Interaction Network. *Science* **296,** 750–752 (2002).
33. Zhang, J. & Yang, J.-R. Determinants of the rate of protein sequence evolution. *Nat. Rev. Genet.* **16,** 409–420 (2015).
34. Szklarczyk, D. *et al.* STRING v10: protein-protein interaction networks, integrated over the tree of life. *Nucleic Acids Res.* **43,** D447–452 (2015).
35. Bastian, F. *et al.* Bgee: Integrating and Comparing Heterogeneous Transcriptome Data Among Species. *Data Integration in the Life Sciences* **5109,** 124–131 (Springer Berlin/Heidelberg, 2008).
36. Lynch, M., Koskella, B. & Schaack, S. Mutation pressure and the evolution of organelle genomic architecture. *Science* **311,** 1727–30 (2006).
37. Zhu, A., Guo, W., Jain, K. & Mower, J. P. Unprecedented heterogeneity in the synonymous substitution rate within a plant genome. *Mol. Biol. Evol.* **31,** 1228–36 (2014).
38. Rogozin, I. B., Basu, M. K., Csürös, M. & Koonin, E. V. Analysis of rare genomic changes does not support the unikont-bikont phylogeny and suggests cyanobacterial symbiosis as the point of primary radiation of eukaryotes. *Genome Biol. Evol.* **1,** 99–113 (2009).
39. Cavalier-Smith, T. Kingdoms Protozoa and Chromista and the eozoan root of the eukaryotic tree. *Biol. Lett.* **6,** 342–5 (2010).
40. Derelle, R. *et al.* Bacterial proteins pinpoint a single eukaryotic root. *Proc. Natl. Acad. Sci. U. S. A.* **112,** 201420657 (2015).
41. Derelle, R. & Lang, B. F. Rooting the eukaryotic tree with mitochondrial and bacterial proteins. *Mol. Biol. Evol.* **29,** 1277–89 (2012).
42. Makarova, K. S., Wolf, Y. I., Mekhedov, S. L., Mirkin, B. G. & Koonin, E. V. Ancestral paralogs and pseudoparalogs and their role in the emergence of the eukaryotic cell. *Nucleic Acids Res.* **33,**

4626–38 (2005).

43. Keeling, P. J. The endosymbiotic origin, diversification and fate of plastids. *Philos. Trans. R. Soc. Lond. B. Biol. Sci.* **365,** 729–748 (2010).
44. Archibald, J. M. The puzzle of plastid evolution. *Curr. Biol. CB* **19,** R81–8 (2009).
45. Penel, S. *et al.* Databases of homologous gene families for comparative genomics. *BMC Bioinformatics* **10 Suppl 6,** S3 (2009).
46. Lartillot, N., Brinkmann, H. & Philippe, H. Suppression of long-branch attraction artefacts in the animal phylogeny using a site-heterogeneous model. *BMC Evol. Biol.* **7 Suppl 1,** S4 (2007).
47. Quang, L. S., Gascuel, O. & Lartillot, N. Empirical profile mixture models for phylogenetic reconstruction. *Bioinforma. Oxf. Engl.* **24,** 2317–2323 (2008).
48. Nguyen, L.-T., Schmidt, H. A., von Haeseler, A. & Minh, B. Q. IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. *Mol. Biol. Evol.* **32,** 268–274 (2015).
49. Rodríguez-Ezpeleta, N. & Embley, T. M. The SAR11 group of alpha-proteobacteria is not related to the origin of mitochondria. *PloS One* **7,** e30520 (2012).
50. Thrash, J. C. *et al.* Phylogenomic evidence for a common ancestor of mitochondria and the SAR11 clade. *Sci. Rep.* **1,** 13 (2011).
51. Wang, Z. & Wu, M. An integrated phylogenomic approach toward pinpointing the origin of mitochondria. *Sci. Rep.* **5,** 7949 (2015).