bioRxiv preprint doi: https://doi.org/10.1101/2021.02.09.430432; this version posted February 10, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

The asgardarchaeal-unique contribution to protein families of the eukaryotic common ancestor was 0.3%

Michael Knopp[1], Simon Stockhorst[1], Mark van der Giezen[2], Sriram G. Garg[1], Sven B. Gould[1]

[1]Institute for Molecular Evolution, Heinrich-Heine-University Düsseldorf, 40225 Düsseldorf, Germany

[2]Centre for Organelle Research, University of Stavanger, 4021 Stavanger, Norway

Keywords: eukaryote origin, eukaryogenesis, asgardarchaea, compartmentalization

Significance Statement

Ever since the first report of a new archaeal lineage, the asgardarchaea, their metagenome analyses have encouraged continued speculations on a type of cell biology ranging between that of prokaryotes and eukaryotes. While it appears a tempting notion, recent microscopic images of an asgardarchaeon suggest otherwise. We inspected the origin of eukaryotic protein families with respect to their distribution across bacteria and archaea. This reveals that the protein families shared exclusively between asgardarchaea and eukaryotes amounts to only 0.3% of the protein families conserved across all eukaryotes. Asgardarchaeal diversity is likely unrivaled across archaea, but their cell biology remains prokaryotic in nature and lends support for the importance of endosymbiosis in evolving eukaryotic traits.

bioRxiv preprint doi: https://doi.org/10.1101/2021.02.09.430432; this version posted February 10, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Summary

The difference between pro- and eukaryotic biology is evident in their genomes, cell biology, and evolution of complex and macroscopic body plans. The lack of intermediates between the two types of cells places the endosymbiotic acquisition of the mitochondrion through an archaeal host at the event horizon of eukaryote origin. The identification of eukaryote specific proteins in a new archaeal phylum, the asgardarchaea, has fueled speculations about their cellular complexity, suggesting they could be eukaryote-like. Here we analyzed the coding capacity of 150 eukaryotes, 1000 bacteria, and 226 archaea, including the only cultured member of the asgardarchaea, Candidatus *Prometheoarchaeon syntrophicum* MK-D1. Established clustering methods that recover endosymbiotic contributions to eukaryotic genomes, recover an asgardarchaeal-unique contribution of a mere 0.3% to protein families present in the last eukaryotic common ancestor, while simultaneously suggesting that asgardarchaeal diversity rivals that of all other archaea combined. Furthermore, we show that the number of homologs shared exclusively between asgardarchaea and eukaryotes is only 27 on average. Genomic and in particular cellular complexity remains a eukaryote-specific feature and, we conclude, is best understood as the archaeal host's solution to housing an endosymbiont and not as a preparation for obtaining one.

Introduction

Four billion years of prokaryotic evolution has only once resulted in the emergence of highly compartmentalized cells and eventually macroscopic body plans: following the origin of eukaryotes through endosymbiosis. The analysis of core eukaryotic features such as the nucleus, mitochondria, sex and meiosis, compartmentalization and dynamic membrane trafficking, and virtually all of the associated protein families, consistently point to their presence in the last eukaryotic common ancestor (LECA) (Fritz-Laylin et al. 2010, Koumandou et al. 2013, Koonin et al. 2013, Garg and Martin 2016). We possess a reasonable understanding of the basic cellular features and coding capacity of LECA, owing to the growing number of genome sequences spanning all of eukaryotic diversity. All eukaryotes stem from a single ancestor that in terms of cellular and genomic complexity rivaled those of extant eukaryotic supergroups (Fritz-Laylin et al. 2010, Koumandou et al. 2013, Koonin et al. 2013). It is certain that LECA was a product of the integration of an alphaproteobacterium into an archaeal host following endosymbiosis (Lane 2011, Blackstone 2013, Martin et al. 2015, Dacks et al. 2016, Spang et al. 2019).

bioRxiv preprint doi: https://doi.org/10.1101/2021.02.09.430432; this version posted February 10, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Through the description of the asgardarchaea, current debates once again concern the cellular complexity of the host that came to house the endosymbiont and what contribution the mitochondrion could have played in establishing the eukaryotic cell (Dacks et al. 2016, Martin et al. 2017). Asgardarchaea, a novel phylum assembled from metagenome data, are viewed as bridging the gap between pro- and eukaryotic cells, because they encode proteins homologous to eukaryotic ones that are e.g. involved in intracellular vesicle trafficking and the regulation of actin cytoskeleton dynamics (Spang et al. 2015, Zaremba-Niedzwiedzka et al. 2017, Neveu et al. 2020). The cellular complexity of the host cell that acquired the alphaproteobacterial endosymbiont has been a matter of speculation ever since the realization that endosymbiosis was pivotal in the transition to eukaryotic life. Modern models of eukaryogenesis differ regarding the timing of mitochondrial acquisition, the extent of the cellular complexity of the host, and the selective reasons provided for explaining the presence, function, and emergence of eukaryotic traits prior or ensuing endosymbiosis (O'Malley 2010, Martin et al. 2015, Gould et al. 2016, Tria et al. 2019, Vosseberg et al. 2020).

Understanding the steps of eukaryogenesis is a demanding intellectual challenge that explores the past of life and one of its most radical transitions. It holds the key to understanding the steps towards cellular complexity, the timing of mitochondrial entry, and what limits prokaryotes to frequently evolve eukaryote-like complexity. Was eukaryogenesis really a matter of luck (Booth and Doolittle 2015) and how important was the energetic superiority provided by the mitochondrion to the host cell (Lane and Martin 2010, Lynch and Marinov 2015, Lane and Martin 2016)? Any model that views endosymbiosis as some kind of terminal coincidence on the evolutionary roadmap to the eukaryotic domain of life needs to explain the singularity that is eukaryogenesis and the lack of comparable complexity among prokaryotes.

A consistent motivation for speculating on the archaeal host cell's grade of complexity is trying to understand whether the host cell was phagocytotic or not (Cavalier-Smith 1987, Yutin et al. 2009, Martijn and Ettema 2013, Martin et al. 2017). This is complicated by the description of a phagocytosis-like process in a planctomycete (Shiratori et al. 2019) and the conflicting evidence for intracellular prokaryotic endosymbionts in the absence of phagocytosis (Fenchel and Bernard 1993, Emblay and Finlay 1993, Schmid 2003, Zientz et al. 2004, Duplessis et al. 2004, Thacker 2005, Woyke et al. 2006, Husnik et al. 2013). It has

bioRxiv preprint doi: https://doi.org/10.1101/2021.02.09.430432; this version posted February 10, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

also been concluded that asgardarchaea are not phagocytotic (Burns et al. 2018), although they encode actin-regulating profilins (Akıl and Robinson 2018), small Rab-like GTPases (Surkont and Pereira-Leal 2016), and prototypic SNARE proteins (Neveu et al. 2020). Phagocytosis might have evolved multiple times independently (Yutin et al. 2009, Mills 2020) and is a mode of feeding, which is incompatible with the syntrophic foundation that underpins eukaryogenesis (Martin et al. 2015, Spang et al. 2019, Martin and Müller 1998, Vellai et al. 1998, Imachi et al. 2020). A sole focus on this single eukaryotic trait might distract and furthermore discounts the complexity of the transition that was involved. What is certain is that images of an asgardarchaeon, Candidatus *Prometheoarchaeum syntrophicum* MK-D1, reveal cells with typical archaeal morphology, half a micron in diameter, with obligate syntrophy, and devoid of intracellular complexity (Imachi et al. 2020).

Here we clustered the available genomes of 150 eukaryotes, 1000 bacteria and 226 archaea (including asgardarchaea metagenomic assemblies, and for comparison the complete genome of the cultured Candidatus *P. syntrophicum* strain MK-D1) in order to evaluate the asgardarchaeal-unique contribution to eukaryogenesis that is understood as support for early cellular complexity in asgardarchaea.

Results

In order to evaluate to what degree asgardarchaea bridge the prokaryotic and eukaryotic protein families, we performed a global comparison of clustered gene families across eleven asgardarchaeal metagenome-assembled genomes (MAGs), the closed MAG of the cultured asgardarchaeon Candidatus *P. syntrophicum* MK-D1, 214 other archaea, 1,000 bacteria, and 150 eukaryotes. Protein families for 150 eukaryotes were taken from Brueckner and Martin (Brueckner and Martin 2020), which included 239,012 clusters. We further clustered proteins from the prokaryotes, resulting in 352,384 bacterial clusters and 49,855 archaeal clusters. Subsequently, the eukaryote and prokaryote clusters were merged in a reciprocal best cluster approach previously described in Ku et al. 2015 (Ku et al. 2015), yielding Eukaryote-Prokaryote-Clusters (EPCs). These EPCs contained proteins from eukaryotes and proteins from either archaeal (Eukaryote-Archaea clusters, EA) or bacterial (Eukaryote-Bacteria clusters, EB), or both (Eukaryote-Archaea-Bacteria clusters, EAB). This approach yielded 2,590 EPCs, of which 867 or 33.5% (330 EA clusters + 537 EAB clusters; Suppl. Table 1)

bioRxiv preprint doi: https://doi.org/10.1101/2021.02.09.430432; this version posted February 10, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

contained an archaeal component. Among these 867 EPCs, asgardarchaeal protein sequences, including those of Candidatus *P. syntrophicum*, were present in about 75% of the protein families (75% of EAB clusters and 75.2% of EA clusters).

A presence-absence pattern (PAP) of all 867 protein families with archaeal contribution to the EPCs, including the 537 protein families present in all domains is shown in Fig. 1 (and Suppl. Table 1). While gene distributions among eury- and crenarchaeota is highly similar, those of the asgardarchaea are patchier and more diverse. Among all of our 239,012 eukaryotic clusters, we could identify only six EA clusters with asgardarchaeal-unique contributions to eukaryotes (Suppl. Table 2), representing 0.0025% of all extant eukaryotic diversity. To calculate the asgardarchaeal-unique contribution to LECA, we filtered the eukaryotic clusters for those that include at least one representative of each of the six eukaryotic supergroups, resulting in 1880 LECA clusters and consequently an asgardarchaeal-unique contribution of 0.3191%.

FIG 1

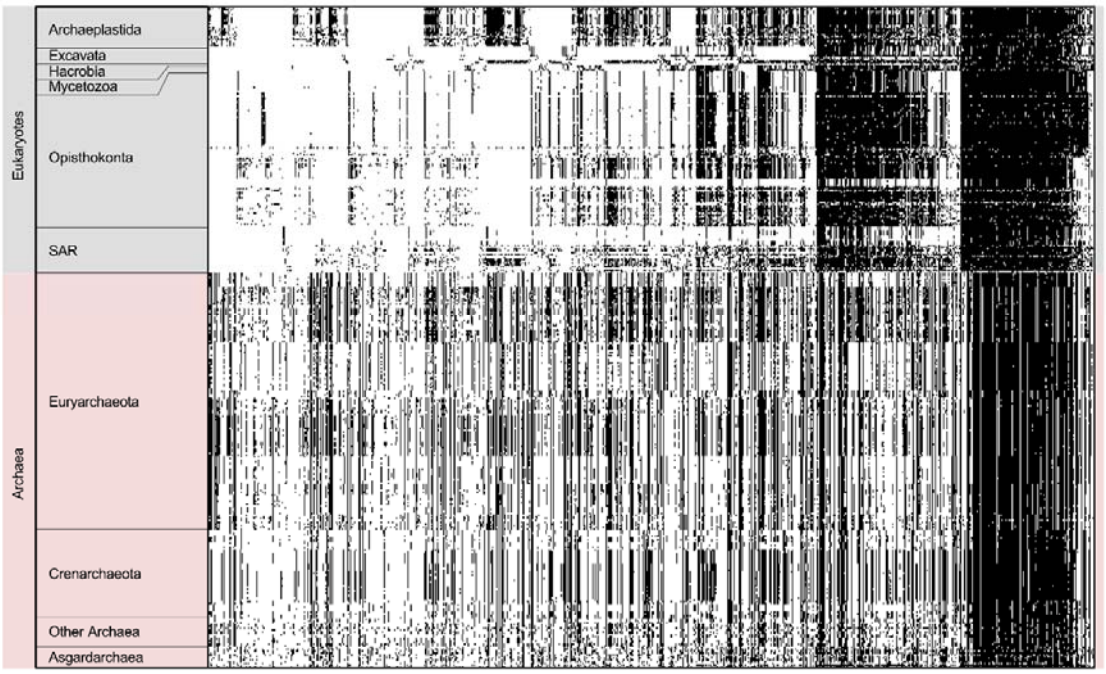


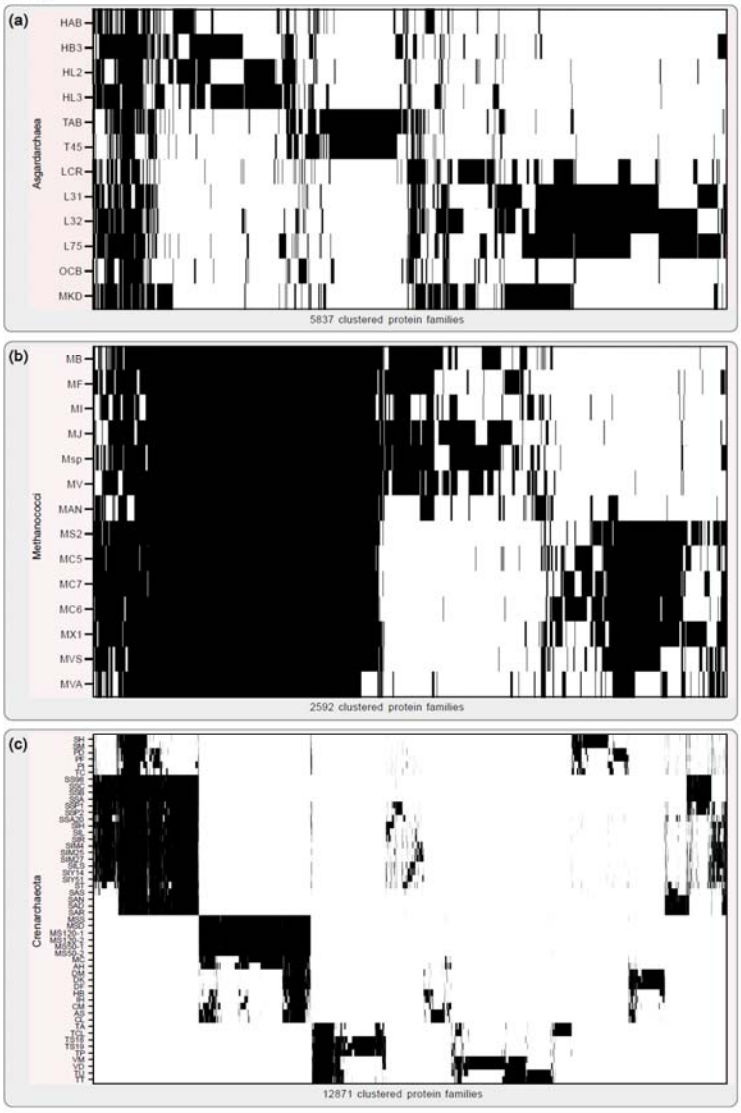
Fig. 1: Presence-absence pattern (PAP) of all 867 eukaryote-prokaryote clusters (EPCs) with archaeal contribution among the investigated 150 eukaryotes, 212 archaea and twelve asgardarchaea. *The protein families were sorted by their distribution among six eukaryotic supergroups (SAR; Stramenopila, Alveolata, Rhizaria), their presence is indicated in black along the X-axis. The group of "Other Archaea" is comprised of all archaea within our*

bioRxiv preprint doi: https://doi.org/10.1101/2021.02.09.430432; this version posted February 10, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

database that have less than 15 members. The distribution of these EPCs among the asgardarchaea does not reveal a particular pattern such as that seen for the other archaeal groups, demonstrating their large genetic differences.

But how closely are the asgardarchaea related to each other with respect to other archaea? To investigate their kinship, we constructed protein families of only the asgardarchaeal protein sequences, which generated a set of 5,837 protein families. The distribution of these protein families among the asgardarchaea reveals a pronounced diversity, with only a small proportion of the protein families being shared across all of them, which suggests one is dealing with a kind of superphylum (Fig. 2a). This is evident from a comparison wherein, we clustered all protein sequences of 14 members of the genus Methanococci and 52 crenarchaeote members from the TACK superphylum (Fig. 2b and Fig. 2c), revealing 2,592 and 12,871 protein families, respectively.

FIG 2



bioRxiv preprint doi: https://doi.org/10.1101/2021.02.09.430432; this version posted February 10, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Fig. 2: Archaeal protein family distributions. (a) *Distribution of all 5,837 calculated asgardarchaeal protein families among the investigated asgardarchaea. The protein families were obtained by globally comparing all asgardarchaeal protein sequences in a pairwise all-vs-all Diamond BLASTp approach including subsequent clustering via MCL. The result was sorted via hierarchical clustering along the X-axis. The vast majority of protein families is not shared among all asgardarchaea but they are rather specific to the individual taxon. Candidatus Prometheoarchaeon syntrophicum MK-D1 shares the highest number of its protein families with members of the Lokiarchaeota.* ***(b)*** *For comparison, we calculated protein families for 14 members of the class Methanococci in the same manner, generating a total of 2,592 protein families. Hierarchical clustering revealed a striking difference in protein families shared between members within the two clusterings, underlining the investigated asgardarchaea's diversity among each other.* ***(c)*** *On the contrary, clustering of 52 members of the crenarchaeota, a highly diverse taxonomic group, results in 12,871 protein families. The matrix was sorted along the X- and Y-axis via hierarchical clustering. All Abbreviations used are listed in Suppl. Table 6.*

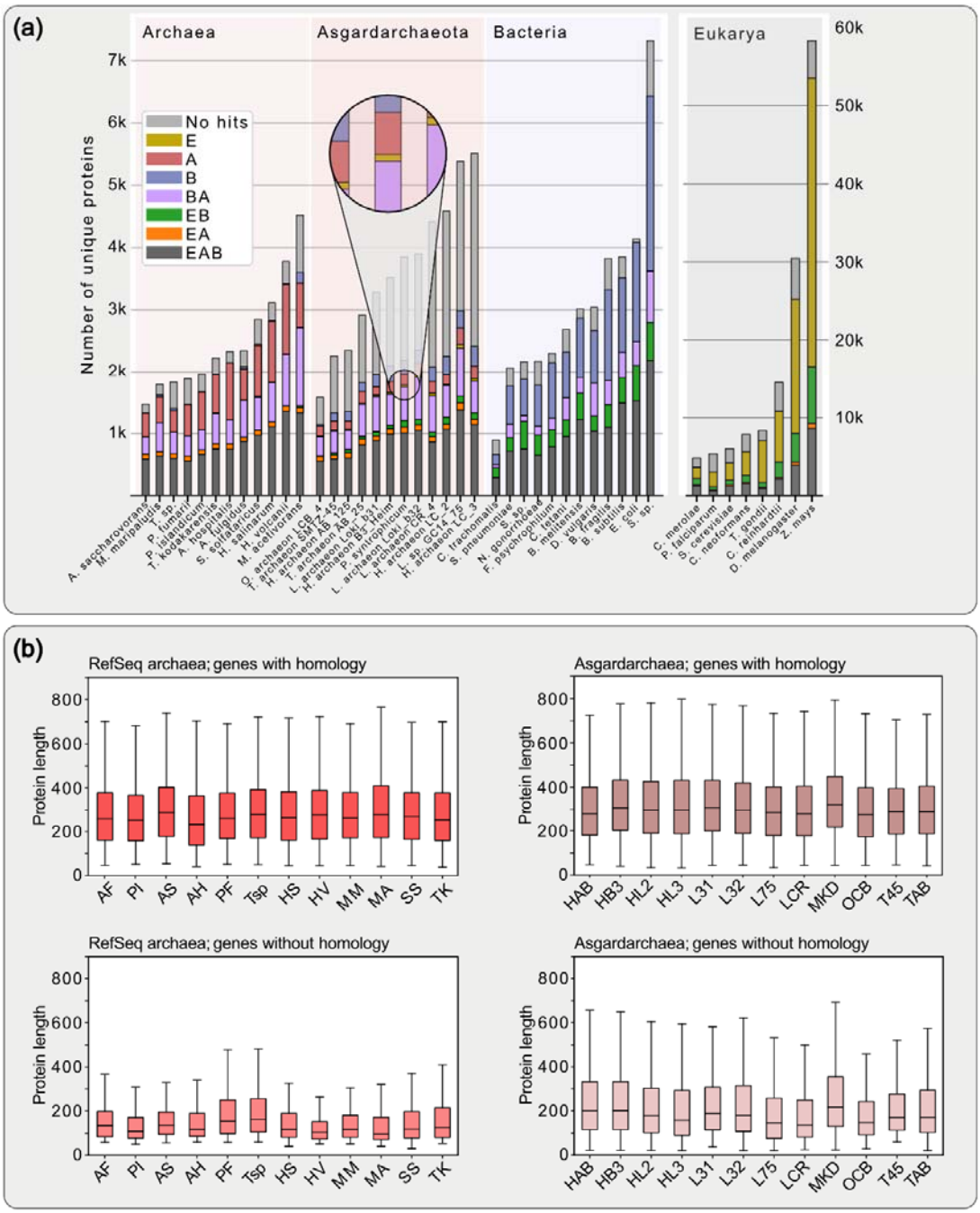
A key difference between archaea and eukaryotes is the difference in the number of protein families. In a previous study, 239,012 protein families were generated on the basis of all pairwise protein sequence comparisons of 150 eukaryotic proteomes (Brueckner and Martin 2020). The same approach yields a total of 49,855 protein families on the basis of all pairwise protein sequence comparisons between 226 archaea, including 11 asgardarchaea and Candidatus *P. syntrophicum* MK-D1. The patchy distribution of the asgardarchaeal EPCs (Fig. 1) and the high diversity between each other (Fig. 2a) prompts a closer look, but the vast difference in the number of protein families reflects the sudden inflation of protein family emergence at the origin of eukaryotes.

Since many asgardarchaeal protein sequences could not be clustered due to a lack of homology, we conducted an analysis of BLASTp hits against a bacterial database consisting of 5,443 proteomes, an archaeal database consisting of 212 proteomes (excluding asgardarchaea) and a eukaryotic database consisting of 150 proteomes. For all investigated asgardarchaea and a selection of twelve well known and diverse archaea and bacteria each, as well as eight eukaryotes, we quantified the amount of sequences with homologs in one of these databases, any combination of these, or no significant homologs at all (Fig. 3a), while ignoring hits from the same genus to counter database composition biases. For asgardarchaea, only 27 protein sequences on average have homologs unique to eukaryotes (Fig 3a, magnified area), supporting the initial EPC analysis which only uncovered six protein families that were unique to eukaryotes and asgardarchaea. Furthermore, this test reveals that high proportions of the asgardarchaeal proteomes do not retrieve significant hits in our three tested databases

bioRxiv preprint doi: https://doi.org/10.1101/2021.02.09.430432; this version posted February 10, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

(Fig. 3a, grey bars). In a few cases, such as for example for Candidatus *Heimdallarchaeota LC_3* or Candidatus *Lokiarchaeota archaeon CR_4*, the number of proteins for which no homology was detected in any other species, exceeds half of the respective genome's coding capacity (Fig. 3a).

FIG 3



bioRxiv preprint doi: https://doi.org/10.1101/2021.02.09.430432; this version posted February 10, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Fig. 3: Distribution of protein homologs and length comparison of archaeal genes with and without homology. (a) *BLAST hit analysis comparing the proteomes of known archaea and bacteria as well as the twelve investigated asgardarchaea. Sequences were blasted against a prokaryotic database comprising 5,655 prokaryotic genomes (including the 1000 bacteria and 212 archaea which were used for protein family calculation) and a eukaryotic database comprised of 150 genomes. For each organism we quantified the amount of sequences with subjects among eukaryotes, bacteria or archaea only, or any combination of these three. To counter possible biases in database composition, hits from the same genus were excluded. E, Eukaryotes; A, Archaea; B, Bacteria; AB, Archaea-Bacteria; et cetera.* ***(b)*** *For proteins with- or without database hits, protein length distributions are shown as box and whiskers plots. This reveals a noticeable difference in protein lengths of sequences without database hits from only the asgardarchaea compared to their sequences with database hits, and in comparison, to those of Refseq archaea. p-values of FDR-corrected, pairwise double-sided Kolmogorov-Smirnov tests in Suppl. Table 7. All used abbreviations are listed in Suppl. Table 6.*

We plotted the protein sequence length distributions for each proteome, separating hits with and hits without significant database hits (Fig. 3b). While the length distributions of sequences with significant database hits were comparable between the asgard- and the RefSeq archaea, the distributions of hits without significant database hits show a major difference. (Fig. 3b). Furthermore, simulating missing data by ignoring not only hits from the same genus but also the same phylum produced a more similar result for the reference organisms and asgardarchaea (Suppl. Fig.1), recapitulating the extensive diversity within the asgardarchaea.

Discussion

The identification of asgardarchaea from deep-sediment metagenome data (Seitz et al. 2016; Zaremba-Niedzwiedzka et al. 2017) provides valuable new information from which to re-evaluate key issues surrounding the tree of life and the emergence of its eukaryotic branch. To begin with, the iconic tree that introduced three aboriginal lineages (Woese and Fox 1977) might require a revision. Phylogenomic analysis of asgardarchaea provides evidence for a two-domains tree and the emergence of the host cell lineage of eukaryogenesis from within the archaeal domain (Cox et al. 2008, McInerney et al. 2014, Hug et al. 2016, Williams et al. 2020) with some skepticism, however, remaining (Forterre 2015, Da Cunha et al. 2018, Liu et al. 2020). Parallel to the discovery of the asgardarchaea were immediate speculations regarding their cellular complexity (Dacks et al. 2016, Pittis and Gabaldón 2016, Rout and Field 2017, Akıl and Robinson 2018, Zachar et al. 2018, Neveu et al. 2020) and a faith of having identified the missing link between pro- and eukaryotic biology based on the identification of a few eukaryote signature proteins (ESPs), that we find to be only 27 on

bioRxiv preprint doi: https://doi.org/10.1101/2021.02.09.430432; this version posted February 10, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

average. In light of these numbers, the potential of these archaea to display eukaryote-like cell complexity is hard to maintain.

Our analysis confirms a patchy distribution of ESPs among asgardarchaea (Dacks et al. 2016, Zaremba-Niedzwiedzka et al. 2017, Imachi et al. 2020, Liu et al. 2020, Klinger et al. 2016, Inoue et al. 2020, Bulzu et al. 2019) (Fig. 1). While of course the archaeal host brought in 1000s of genes, the unique contribution of this new archaeal lineage to the eukaryotic protein families is substantially less than what one might infer from the original metagenome reports and subsequent interpretations (Dacks et al. 2016, Pittis and Gabaldón 2016, Rout and Field 2017, Zachar et al. 2018, Akıl and Robinson 2018, Neveu et al. 2020, Vosseberg et al. 2020). The irregular gene distribution (Fig. 2a) might reflect differential gene loss upon the segregation of the common ancestor of asgardarchaea and the archaeal host cell lineage (Eme et al. 2017). Considering the role of pangenomes in the transformation of prokaryotic lineages and the conquering of ecological niches (McInerney et al. 2017), pangenomes offer a complementary explanation to the differential loss of genes. The archaeal ancestor of eukaryotes might have tapped a shared gene pool more extensively than the sister lineages leading to extant asgardarchaea.

The notion that asgardarchaeal contributions to eukaryotes were higher only to be eventually replaced by bacterial (endosymbiotic) contributions might be brought up (Pittis and Gabaldon 2016; Eme et al. 2017; Vosseberg et al. 2020). The absence of extant archaeal relatives with similarly higher ESPs, however, indicates that archaea neither have the necessity nor the selective pressure for maintaining ESPs in the absence of an endosymbiont. Any theory that hinges upon a larger presence of ESPs in archaea or bacteria prior to mitochondrial endosymbiosis ignores the complete lack of “accumulated ESPs” in extant prokaryotes to a degree that even remotely matches that of any given eukaryotic lineage. Extinction and absence of geological conditions that promoted eukaryote origins considered for ESPs, must also be considered for any genes that are currently thought to be recent independent gene transfer events from prokaryotes to eukaryotes and not of endosymbiotic (mitochondrial) origin.

Our protein family clustering method, which readily detects the mitochondrial contribution (and the cyanobacterial contributions in the case of the Archaeplastida; Suppl. Fig. 2) failed to detect a comparable asgardarchaeal-unique contribution. A stacked bar diagram puts gene

bioRxiv preprint doi: https://doi.org/10.1101/2021.02.09.430432; this version posted February 10, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

family cluster contribution in each lineage into a global perspective (Fig. 3a; Supp. Fig 1). There is a small proportion of eukaryotic homologs (E, mustard yellow) visible, e.g. in the Candidatus *P. syntrophicum* MK-D1, but it is substantially smaller in comparison to the eukaryote-bacteria (EB, green) specific homologs evident in eukaryotes (and bacteria *vice versa*) that reflects the mitochondrial contribution to eukaryogenesis (Brueckner and Martin 2020). Neglecting the surprisingly low number of asgardarchaeal-unique homologs to eukaryotic genomes, our analysis demonstrates that asgardarchaea are among the most genetically diverse group of archaea when comparing it to the genus Methanococci and the phylum Crenarchaeota (Fig. 2). The odd length distributions of the asgardarchaeal proteins with no homology (Fig. 3b) are strange as well, since protein length across pro- and eukaryotes is usually well conserved (Xu et al. 2006). This could hint at an assembly and/or binning issue, which was also observed regarding the anomalous phylogenetic behavior of their ribosomal proteins and concatenated gene trees (Da Cunha et al. 2018; Garg et al. 2021). If not, it is a biological phenomenon absent in other sequenced prokaryotes.

Considering the amount of data gathered in just the few years (Imachi et al. 2020, Liu et al. 2020, Klinger et al. 2016, Inoue et al. 2020, Bulzu et al. 2019, Villanueva et al. 2016), it is surprising asgardarchaea have escaped identification for so long. Their habitats had been sampled before, so it is likely that the method used and maybe an obligate dependency on syntrophy, hindered culturing except for one imposing exception (Imachi et al. 2020). Dedicated phylogenomic efforts are necessary to resolve their taxonomic classification, while only culturing can picture their cell morphology.

Analyses of asgardarchaeal ESPs shows they have the potential to function similar to their eukaryotic homologs in a eukaryotic system (Neveu et al. 2020, Akıl and Robinson 2018, Rout and Field 2017, Klinger et al. 2016), but cross-kingdom inferences have their limits (Dey et al. 2016). The analysis of archaeal small GTPases (Surkont and Pereira-Leal 2016) and homologs of ESCRT proteins, the CDVs (Lindås et al. 2008, Lindås and Bernander 2013, Caspi and Dekker 2018), serve as examples. One needs to interpret asgardarchaeal ESPs in their prokaryotic context and in cells lacking an endosymbiont. The first images of an asgardarchaeon, those of Candidatus *P. syntrophicum* MK-D1, and its dependency on a bacterial partner (Imachi et al .2020) define the current standard from which to plot eukaryogenesis and the steps leading to eukaryotic cell- and genome complexity.

bioRxiv preprint doi: https://doi.org/10.1101/2021.02.09.430432; this version posted February 10, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

The identification of the asgardarchaea and the culturing of one representative represent an important milestone in micro- and evolutionary biology. Their phylogenetic analysis echoes two previously predicted outcomes: (i) eukaryotes to branch from within archaea, solidifying the two-domains tree of life, and (ii) that the closer we zoom in on the two prokaryotic partners from which eukaryotes evolved, the higher the number of otherwise eukaryote-typical genes we identify in prokaryotes. The description of Candidatus *P. syntrophicum* MK-D1 (Imachi et al. 2020) reminds us to not conflate genotypic with phenotypic complexity. This predicts that future asgardarchaea we see cultured will lack eukaryotic traits, too, and most, if not all, will depend on syntrophy. Our analysis also predicts that much of asgardarchaeal diversity remains to be described, but that the gap between the pro- and eukaryotic protein families will remain decisive and to change little. Placing the endosymbiotic event and the energetic benefit offered by mitochondria to fuel the transition early in eukaryogenesis, explains the lack of physical evidence for eukaryote-like complexity in asgardarchaea despite them encoding ESPs. It offers a comprehensive full-service theory for the singularity that is the origin of the eukaryotic cell that mitochondria-late models fail to provide.

Acknowledgements: SBG would like to thank the German research council (267205415 – SFB 1208) and the VolkswagenStiftung (Life) for funding. SGG was furthermore supported by the Moore–Simons Project on the Origin of the Eukaryotic Cell GBMF9743. Computational infrastructure and support were provided by the Centre for Information and Media Technology at HHU Düsseldorf.

Author Contributions: SGG, MK and SBG conceived the analysis, which was carried out predominantly by MK but also SS. SGG, MK, MvdG and SBG wrote the manuscript, whose final version was approved all authors.

bioRxiv preprint doi: https://doi.org/10.1101/2021.02.09.430432; this version posted February 10, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Methods

Calculation of protein families

Prokaryotic gene families were calculated from complete genomes of 1000 bacteria and 226 archaea of the Refseq database (Version September 2016) including eleven representatives of the asgardarchaea[13] and Candidatus *P. syntrophicum* MK-D1 (Imachi et al. 2020), separately. Bacterial and archaeal protein families were calculated via MCL (van Dongen 2000, Enright 2002) (--abc -scheme 7) from all reciprocal best BLAST hits with pairwise global identities (Rice et al. 2000) of at least 25% identity and a maximum e-value of $1x10^{-10}$ among all investigated bacteria and archaea, respectively. Eukaryotic gene families on the basis of 150 eukaryotic genomes were calculated as part of a previous study (Brueckner et al. 2020). Prokaryotic and eukaryotic cluster were combined into EPCs if at least 50% of all sequences of a eukaryotic cluster had their best hit in a prokaryotic cluster and vice versa according to the 'reciprocal best cluster approach' described in Ku et al. 2015 (Ku et al. 2015). All proteomes used in this study are listed in Supplementary Table 5. Protein families of Methanococci and Crenarchaeota were calculated from all pairwise global identities of all protein sequences, using the above-mentioned identity and e-value cutoffs, via MCL (van Dongen 2000, Enright 2002).

BLAST hit analysis

We performed a Diamond BLASTp (Buchfink et al. 2015) hit analysis on all protein sequences of the eleven asgardarchaeal proteomes including MK-D1. We compared the results to twelve bacterial and archaeal proteomes each, plus eight eukaryotic ones. We blasted all protein sequences of the chosen proteomes against our database of 5655 prokaryotic and 150 eukaryotic proteomes. For each proteome, we quantified the number of sequences that showed significant hits (at least 25% identity and a minimal e-value of $1x10^{-5}$) within bacteria, archaea, eukaryota or any combination of these three. To counter overrepresentation of some genera within our database, we excluded hits of the same genus for all tested protein sequences.

InterProScan ESP analysis

InterProScan version 5.39-77.0 (Quevillon et al. 2005) with standard parameters was used to annotate all asgardarchaeal proteomes, the MK-D1 proteome and all 212 archaeal proteomes within our prokaryotic database. As in Zaremba et al. 2017 (Zaremba-Niedzwiedzka et al. 2017), we searched for InterPro-Identifiers that correspond with Eukaryote-specific-proteins

bioRxiv preprint doi: https://doi.org/10.1101/2021.02.09.430432; this version posted February 10, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

and plotted the results of all investigated asgard archaea together with the results of 14 model Refseq archaea for comparison. (see Suppl. Figure 3)

LECA cluster filtering

Since eukaryotic inheritance is strictly vertical the eukaryotic protein families were filtered for families that contained at least one protein sequence from one member of each of the six supergroups included in our dataset of 150 eukaryotes, being *Archaeplastida*, *Opisthokonta*, *SAR*, *Hacrobia*, *Excavata* and *Mycetozoa* resulting in 1880 protein families passing this criterion.

Suppl. material titles and legends

Suppl. Fig. 1: **BLAST hit analysis.** The BLAST hit analysis underlying Fig. 3a was redone, this time ignoring all hits from the same phylum. This was done in order to test for the asgardarchaeas' taxonomic rank, since the amount of protein sequences without any significant database hits, could either stem from poor sequence quality or just from the fact that asgardarchaea are inherently different from all archaea we know so far. As expected, the amount of Refseq archaeal protein sequences without significant database hits did increase but the effect was not sufficient to recover the same proportion of "unknown" protein sequences as we see in asgardarchaea.

Suppl. Fig. 2: PAPs of sets of Eukaryote-Prokaryote-Clusters. (a) Presence-Absence pattern of all 537 EPCs containing members from eukaryotes, bacteria and archaea. **(b)** Presence-Absence pattern of all 1723 EPCs containing only proteins from eukaryotes and bacteria. The cyanobacterial contribution to the Archaeplastida is marked with red boxes. In both cases, protein families were sorted by their distribution among the six eukaryotic supergroups Archaeplastida, Excavata, Hacrobia, Mycetozoa, Opisthokonta and the SAR group.

Suppl. Fig. 3: InterProScan results for all investigated asgardarchaea as well as 14 reference archaea. InterProScan was used to annotate the proteomes of all 12 investigated asgardarchaea and 14 reference archaea. InterProScan results were used to search for ESP candidate proteins within all analyzed proteomes and their presence indicated by a black dot. InterPro identifiers pointing towards the presence of ESPs were obtained from Zaremba et al.

bioRxiv preprint doi: https://doi.org/10.1101/2021.02.09.430432; this version posted February 10, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

2017 and in some cases revised to be more strict. For most ESPs results show a clear divide between asgardarchaea and the chosen reference archaea, while the DNA-directed RNA-Polymerases I/III and II, Cyclins or Cyclin-like proteins and the STT3 subunit of the OST complex could also be detected in the majority of the analyzed reference archaea.

Suppl. Table 1. EPC overview including protein family counts, protein sequence listing for each protein family and relative frequency of occurrence within protein families of all taxonomic groups present in the clustering. One protein family per line, one member each cell.

Suppl. Table 2. Six identified Eukaryote-Asgardarchaea unique protein families annotated via InterProScan using standard parameters and showing results from all subject databases.

Suppl. Table 3. Archaeal and bacterial clusterings. One protein family per line, one member each cell.

Suppl. Table 4. InterProScan results for all investigated Asgardarchaea, underlying Fig. 3. in addition to the list of InterPro accessions used to identify ESPs within the investigated asgardarchaeal MAGs.

Suppl. Table 5. Listing of all proteome assemblies used in cluster creation and BLAST hit analysis.

Suppl. Table 6. Species abbreviations used in Figure 2 and Figure 3.

Suppl. Table 7. FDR-corrected (Benjamini-Hochberg) p-values of pairwise double-sided KS tests of gene length distributions.

References

Akıl C, & Robinson RC (2018). Genomes of Asgard archaea encode profilins that regulate actin. *Nature*, **562**, 439–443.

Blackstone NW (2013). Why did eukaryotes evolve only once? Genetic and energetic aspects of conflict and conflict mediation. *Phil trans R Soc B*, **368**, 20120266.

bioRxiv preprint doi: https://doi.org/10.1101/2021.02.09.430432; this version posted February 10, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Booth A & Doolittle WF (2015). Eukaryogenesis, how special really? *Proc Natl Acad Sci USA*, **112**, 10278–10285.

Brueckner J, & Martin WF (2020). Bacterial genes outnumber archaeal genes in eukaryotic genomes *Genome Biol Evol*, **12**, 282-292.

Buchfink B, Xie C, & Huson DH (2015). Fast and sensitive protein alignment using DIAMOND. *Nature Methods*, **12**, 59–60.

Bulzu P-A, Andrei A-S, Salcher MM, Mehrshad M, Inoue K, Kandori H, Beja O, Ghai R, & Banciu HL (2019). Casting light on Asgardarchaeota metabolism in a sunlit microoxic niche. *Nat Microbiol*, **4**, 1129–1137.

Burns JA, Pittis AA & Kim E (2018). Gene-based predictive models of trophic modes suggest Asgard archaea are not phagocytotic. *Nat Ecol Evol*, **2**, 697–704.

Caspi Y & Dekker C (2018). Dividing the archaeal way: the ancient cdv cell-division machinery. *Front Microbiol*, **9**, 174.

Cavalier-Smith T (1987) The origin of eukaryote and archaebacterial cells. *Ann N Y Acad Sci*, **503**, 17–54.

Cox CJ, Foster PG, Hirt RP, Harris SR &, Embley TM (2008). The archaebacterial origin of eukaryotes. *Proc Natl ional Acad Sci USA*, **105**, 20356–20361.

Da Cunha V, Gaia M, Nasir A &, Forterre P (2018). Asgard archaea do not close the debate about the universal tree of life topology. *PLoS Genet*, **14**, e1007215.

Dacks JB, Field MC, Buick R, Eme L, Gribaldo S, Roger AJ, Brochier-Armanet C &, Devos DP. (2016). The changing view of eukaryogenesis – fossils, cells, lineages and how they all come together. *Journal of Cell Science Sci*, **129**, 3695-3703.

Dey G, Thattai M & Baum B (2016). On the archaeal origins of eukaryotes and the challenges of inferring phenotype from genotype. *Trends Cell Biol*,. **26**, 476–485.

Duplessis MR, Ziebis W, Gros O, Caro A, Robidart J &, Felbeck H. 2004. Respiration strategies utilized by the gill endosymbiont from the host lucinid Codakia orbicularis (Bivalvia: Lucinidae). *Appl Environ Microbiol*, **70**, 4144–4150.

Embley TM & Finlay BJ. 1993. Systematic and morphological diversity of endosymbiotic methanogens in anaerobic ciliates. *Antonie Van Leeuwenhoek*, **64**, 261–271.

Eme L, Spang A, Lombard J, Stairs CW & Ettema TJG. 2017. Archaea and the origin of eukaryotes. *Nat Rev Microbiol*, **15**, 711–723.

Enright, AJ. 2002. An efficient algorithm for large-scale detection of protein families. Nucleic Acids Res, **30**, 1575–1584.

bioRxiv preprint doi: https://doi.org/10.1101/2021.02.09.430432; this version posted February 10, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Fenchel T & Bernard C. 1993. A purple protist. *Nature*, **362**, 300–300.

Forterre P. 2015. The universal tree of life: an update. *Front Microbiol*, **6**, 717.

Fritz-Laylin LK, Prochnik SE, Ginger ML, Dacks JB, Carpenter ML, et al…, Dawson SC. (2010). The genome of *Naegleria gruberi* illuminates early eukaryotic versatility. *Cell*. **140**, 631–642.

Garg SG, Kapust N, Lin W, Knopp M, Tria F, Nelson-Sathi S, Gould SB, Fan L, Zhu R, Zhang C, Martin WF. 2021. Anomalous phylogenetic behavior of ribosomal proteins in metagenome assembled genomes. *Genome Biol Evol* , **13,** evaa238.

Garg SG &, Martin WF. 2016. Mitochondria, the cell cycle, and the origin of sex via a syncytial eukaryote common ancestor. *Genome Biol Evol*,. **8**, 1950-1970.

Gould SB, Garg SG. & Martin WF. 2016. Bacterial vesicle secretion and the evolutionary origin of the eukaryotic endomembrane system. *Trends Microbiol*,. **24**, 525–534.

Hug LA, Baker BJ, Anantharaman K, Brown CT, Probst AJ, …, Banfield JF. 2016. A new view of the tree of life. *Nat. Microbiol*,. **1**, 16048.

Husnik F, Nikoh N, Koga R, Ross L, Duncan RP, …, McCutcheon JP. 2013. Horizontal gene transfer from diverse bacteria to an insect genome enables a tripartite nested mealybug symbiosis. *Cell*, **153**, 1567.

Imachi H, Nobu MK, Nakahara N, Morono Y, Ogawara M, Takaki Y, Takano Y, Uematsu K, Ikuta T, Ito M, Matsui Y, Miyazaki M, Murata K, Saito Y, Sakai S, Song C, Tasumi E, Yamanaka Y, Yamaguchi T, …, Takai K. 2020. Isolation of an archaeon at the prokaryote–eukaryote interface. *Nature*, **577**, 519–525.

Inoue K, Tsunoda SP, Singh M, Tomida S, Hososhima S, …, Kandori H. 2020. Schizorhodopsins: A family of rhodopsins from Asgard archaea that function as light-driven inward H+ pumps. *Sci Adv*, **6**, eaaz2441.

Klinger CM, Spang A, Dacks JB & Ettema TJG. 2016. Tracing the archaeal origins of eukaryotic membrane-trafficking system building blocks. *Mol Biol Evol*, Mol. Biol. Evol., **33**, 1528–1541.

Koonin EV, Csuros M & Rogozin IB. 2013. Whence genes in pieces: reconstruction of the exon-intron gene structures of the last eukaryotic common ancestor and other ancestral eukaryotes. *Wiley Interdiscip Rev RNA*, **4**, 93–105.

Koumandou VL, Wickstead B, Ginger ML, van der Giezen M, Dacks JB &, Field MC. 2013. Molecular paleontology and complexity in the last eukaryotic common ancestor. *Crit Rev Biochem Mol*, , **48**, 373–396.

bioRxiv preprint doi: https://doi.org/10.1101/2021.02.09.430432; this version posted February 10, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Ku C, Nelson-Sathi S, Roettger M, Sousa FL, Lockhart PJ, Bryant D, Hazkani-covo E, Mcinerney JO, Landan G, & Martin WF. 2015. Endosymbiotic origin and differential loss of eukaryotic genes. *Nature*, **524**, 427–437.

Lane N & Martin WF. 2010. The energetics of genome complexity. *Nature*, **467**, 929–934.

Lane N & Martin WF. 2016. Mitochondria, complexity, and evolutionary deficit spending. *Proc National Acad Sci*, **113**, E666–E666.

Lane N. 2011. Energetics and genetics across the prokaryote-eukaryote divide. *Biol. Direct*, **6**, 35.

Lindås A-C & Bernander R. 2013. The cell cycle of archaea. *Nat Rev Microbiol*, **11**, 627–38.

Lindås A-C, Karlsson EA, Lindgren MT, Ettema TJG & Bernander R. 2008. A unique cell division machinery in the Archaea. *Proc National Acad Sci* **105**, 18942–18946.

Liu Y, Makarova KS, Huang WC, Wolf YI, Nikolskaya A, Zhang X, Cai M, Zhang CJ, Xu W, Luo Z, Cheng L, Koonin EV &, Li M. 2020. Expanding diversity of Asgard archaea and the elusive ancestry of eukaryotes. bioRxiv

Lynch M & Marinov GK. 2015. The bioenergetic costs of a gene. *Proc Natl Acad Sci USA*, **112**, 15690–15695.

Martijn J &, Ettema TJ. 2013. From archaeon to eukaryote: the evolutionary dark ages of the eukaryotic cell. *Biochem Soc Trans*, **41**, 451-457.

Martin W & Müller M. 1998. The hydrogen hypothesis for the first eukaryote. *Nature*, **392**, 37–41.

Martin WF, Garg SG & Zimorski V. 2015. Endosymbiotic theories for eukaryote origin. *Phil Trans R Soc B*, **370**, 20140330.

Martin WF, Tielens AGM, Mentel M, Garg SG & Gould SB. 2017. The physiology of phagocytosis in the context of mitochondrial origin. *Microbiol Mol Biol Rev*, **81**, 1–36.

McInerney JO, O'Connell MJ & Pisani D. 2014. The hybrid nature of the Eukaryota and a consilient view of life on Earth. *Nat Rev Microbiol*, **12**, 449–455.

McInerney J, McNally A & O'Connell M. 2017. Why prokaryotes have pangenomes. *Nat Microbiol*, **2,** 17040.

Mills DB. 2020. The origin of phagocytosis in earth history. *Interface Focus* **10**, 20200019.

Neveu E, Khalifeh D, Salamin N, & Fasshauer D. 2020. Prototypic SNARE proteins are encoded in the genomes of Heimdallarchaeota, potentially bridging the gap between the prokaryotes and eukaryotes. *Curr Biol*, **30**, 2468-2480.

bioRxiv preprint doi: https://doi.org/10.1101/2021.02.09.430432; this version posted February 10, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

O'Malley MA. 2010. The first eukaryote cell: an unfinished history of contestation. *Stud Hist Philos Biol Biomed Sci*, **41**, 212-224.

Pittis AA & Gabaldón T. 2016. Late acquisition of mitochondria by a host with chimaeric prokaryotic ancestry. *Nature*, **531**, 101–104.

Quevillon E, Silventoinen V, Pillai S, Harte N, Mulder N, Apweiler R, & Lopez R. 2005. InterProScan: protein domains identifier. *Nucleic Acids Research*, **33**(Web Server), W116–W120.

Rice P, Longden L, & Bleasby A. 2000. EMBOSS: The European Molecular Biology Open Software Suite. *Trends Genet*, **16**, 276–277.

Rout MP & Field MC. 2017. The evolution of organellar coat complexes and organization of the eukaryotic cell. *Annu Rev Biochem*, **86**, 637–657.

Schmid AMM. 2003. Endobacteria in the diatom Pinnularia (Bacillariophyceae). I. 'Scattered ct-nucleoids' explained: DAPI-DNA complexes stem from exoplastidial bacteria boring into the chloroplasts. *J Phycol*, **39**, 122–138.

Seitz KW, Lazar CS, Hinrichs K-U, Teske AP & Baker BJ. 2016. Genomic reconstruction of a novel, deeply branched sediment archaeal phylum with pathways for acetogenesis and sulfur reduction. *ISME J* **10**, 1696–1705.

Shiratori T, Suzuki S, Kakizawa Y & Ishida K-I. 2019. Phagocytosis-like cell engulfment by a planctomycete bacterium. *Nat Commun*, **10**, 5529.

Spang A, Saw JH, Jørgensen SL, Zaremba- Niedzwiedzka K, Martijn J, …, Ettema TJG. 2015. Complex archaea that bridge the gap between prokaryotes and eukaryotes. *Nature*, **521**, 173–179.

Spang A, Stairs CW, Dombrowski N, Eme L, Lombard J, …, Ettema TJG. 2019. Proposal of the reverse flow model for the origin of the eukaryotic cell based on comparative analyses of Asgard archaeal metabolism. *Nat Microbiol*, **4**, 1138–1148.

Surkont J & Pereira-Leal JB. 2016. Are there Rab GTPases in archaea? *Mol Biol Evol*, **33**, 1833–1842.

Thacker RW. 2005. Impacts of shading on sponge-cyanobacteria symbioses: A comparison between host-specific and generalist associations. *Integr Comp Biol*, **45**, 369–376.

Tria FDK, Brückner J, Skejo J, Xavier JC, Zimorski V, Gould SB., Garg SG, & Martin WF. 2019. Gene duplications trace mitochondria to the onset of eukaryote complexity. *BioRxiv*, **781211**. https://doi.org/10.1101/781211

van Dongen S. 2000. A cluster algorithm for graphs. *Inf Syst*, **R0010**, 1–40.

bioRxiv preprint doi: https://doi.org/10.1101/2021.02.09.430432; this version posted February 10, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Vellai T, Takács K &, Vida G. 1998. A new aspect to the origin and evolution of eukaryotes. *J Mol Evol*, **46**, 499-507.

Villanueva L, Schouten S & Damsté JSS. 2016. Phylogenomic analysis of lipid biosynthetic genes of Archaea shed light on the "lipid divide". *Environ Microbiol*, **19**, 54–69.

Vosseberg J, van Hooff JJE, Marcet-Houben M, van Vlimmeren A, van Wijk LM, Gabaldón T & Snel B. 2020. Timing the origin of eukaryotic cellular complexity with ancient duplications. *Nat Ecol Evol*, **5,** 92–100.

Williams TA, Cox CJ, Foster PG, Szöllősi GJ & Embley TM. 2020. Phylogenomics provides robust support for a two-domains tree of life. *Nat Ecol Evol*, **4**, 138–147.

Woese CR. & Fox GE. 1977. Phylogenetic structure of the prokaryotic domain: The primary kingdoms. *Proc Natl Acad Sci USA*, **74**, 5088–5090.

Woyke T, Teeling H, Ivanova NN, Huntemann M, Richter M, …, Dubilier N. 2006. Symbiosis insights through metagenomic analysis of a microbial consortium. *Nature*, **443**, 950–955.

Xu L, Chen H, Hu X, Zhang R, Zhang Z, & Luo ZW. 2006. Average gene length is highly conserved in prokaryotes and eukaryotes and diverges only between the two kingdoms. *Mol Biol Evol*, **23**, 1107–1108.

Yutin N, Wolf MY, Wolf YI & Koonin EV. 2009. The origins of phagocytosis and eukaryogenesis. *Biol Direct*, **4**, 9.

Zachar I, Szilágyi A, Számadó S & Szathmáry E. 2018. Farming the mitochondrial ancestor as a model of endosymbiotic establishment by natural selection. *Proc Nat Acad Sci USA*, **115**, 201718707.

Zaremba-Niedzwiedzka K, Caceres EF, Saw JH, Bäckström D, Juzokaite L, …, & Ettema TJG. 2017. Asgard archaea illuminate the origin of eukaryotic cellular complexity. *Nature*, **541**, 353–358.

Zientz E, Dandekar T & Gross R. 2004. Metabolic Interdependence of Obligate Intracellular Bacteria and Their Insect Hosts. *Microbiol Mol Biol Rev*, **68**(4), 745–770.